Lightening black-box models in field-based implicit-likelihood cosmological inference

Tristan Hoellinger

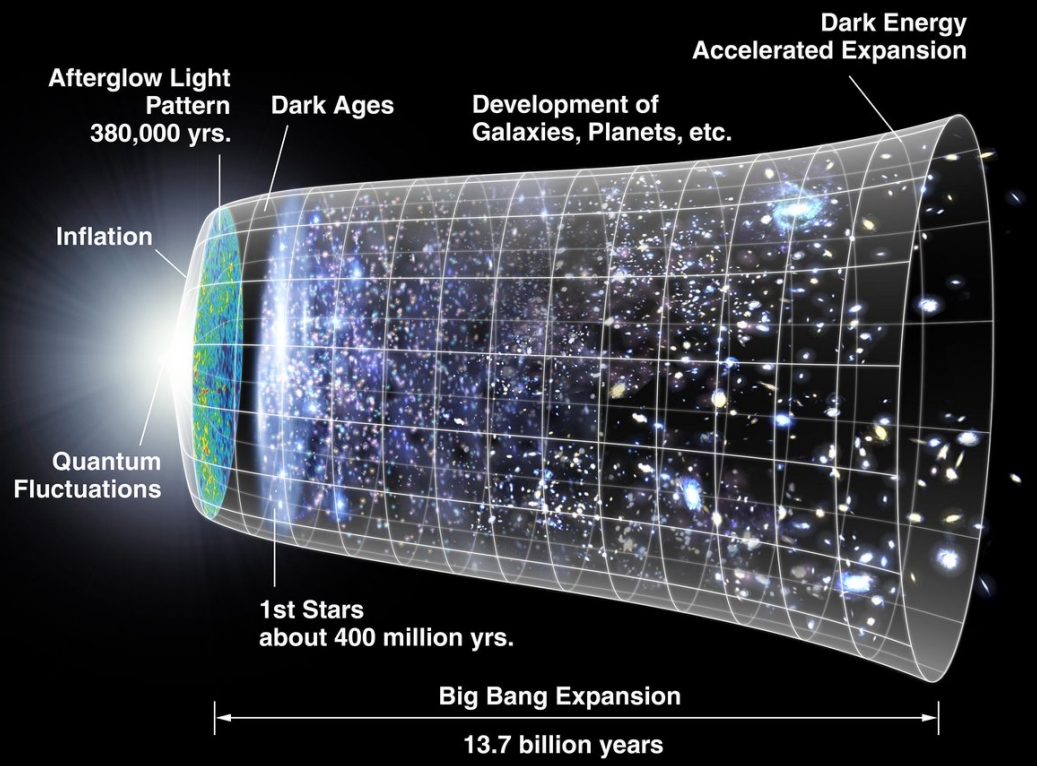
PhD candidate at the Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

In collaboration with:
Florent Leclercq (IAP), Guilhem Lavaux (IAP)
and the Aquila Consortium

aquila-consortium.org

October 18th, 2024

The big picture
Cosmological parameters



Parameters in Λ-CDM

Λ Ω_m Ω_b n_S A_S h τ

Basic extensions of Λ-CDM

w $\sum m_\nu$

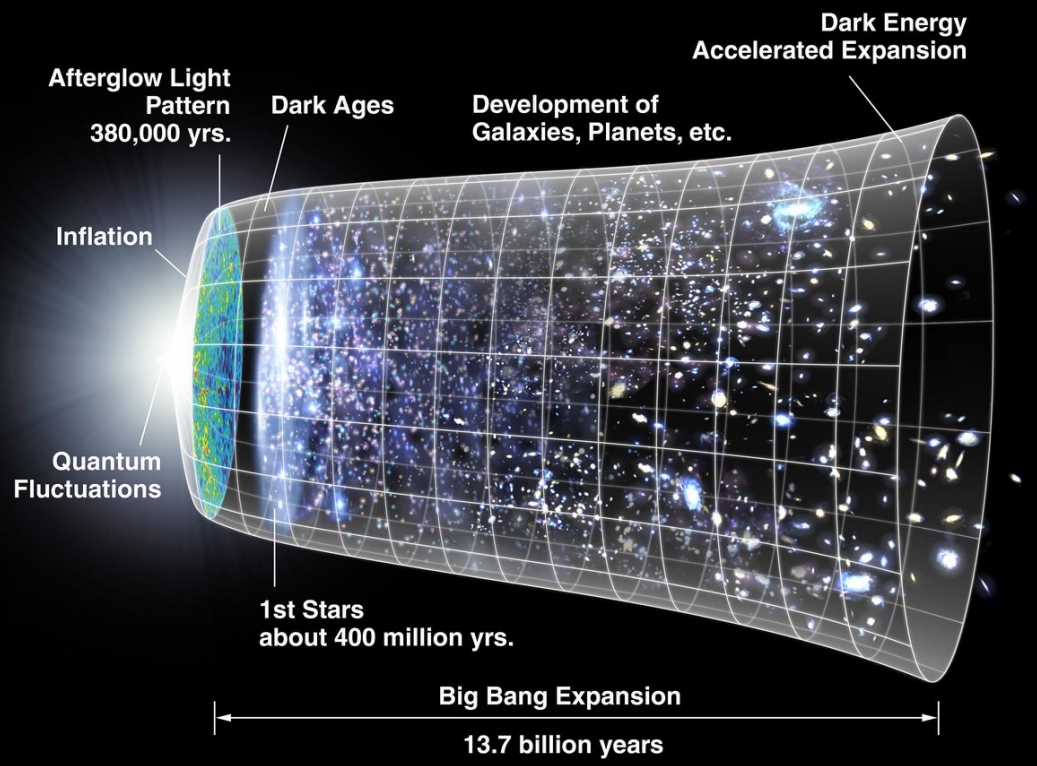
Signatures of inflation

f_{NL} r

Cosmic beginning
Cosmic content
Cosmic fate

The big picture

Emphasis on Dark energy



- Λ Dark energy

Basic extensions of Λ-CDM

- w Quintessence

- *Dynamical component?*
- *Relation to fundamental physics?*

euclid

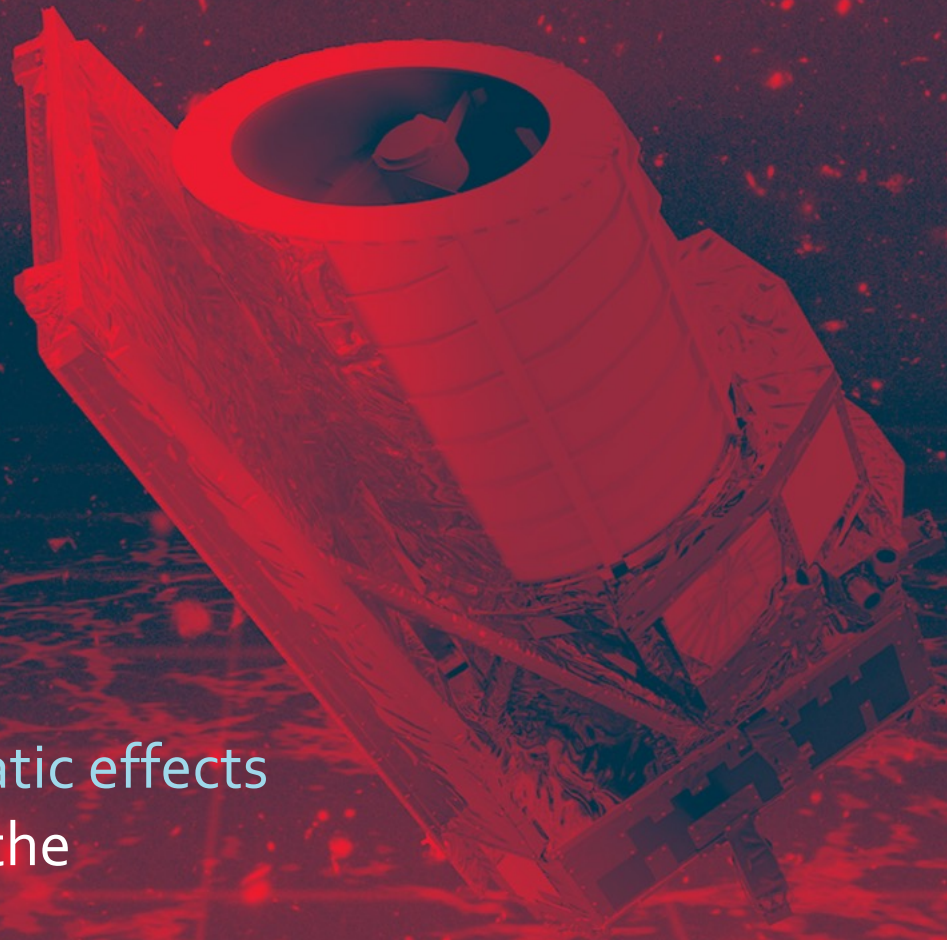
Exploring the dark Universe

Exponential growth of the size of the Universe,

but also of the amount of data to probe dark energy

euclid

Exploring the dark Universe



Many instrumental systematic effects may bias the posteriors on the cosmological parameters.

Issue of model misspecification

How to attain robust Implicit Likelihood Inference (ILI) from Stage-IV galaxy surveys?

Model misspecification

Causes biased posteriors.
Several solutions for *explicit* likelihood inference.

So far, no solution within ILI frameworks in cosmology.

Issue of model misspecification

How to attain robust Implicit Likelihood Inference (ILI) from Stage-IV galaxy surveys?

- ✓ Diagnose systematic effects using the inferred initial matter power spectrum after recombination.

We propose a generic solution within ILI frameworks in cosmology.

Field-based Implicit likelihood cosmological inference

Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ $\xrightarrow{?}$ Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

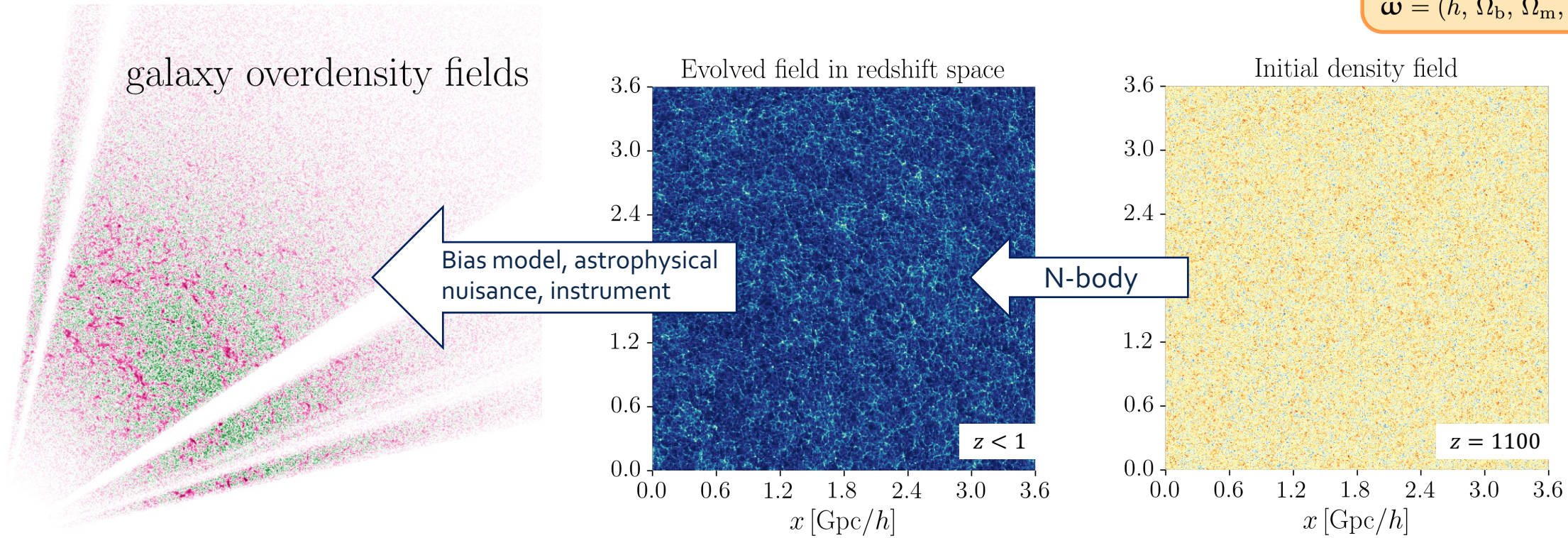
Cosmological parameters

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

Field-based Implicit likelihood cosmological inference

Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ —?→ Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

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



Field-based Implicit likelihood cosmological inference

Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ $\xrightarrow{?}$ Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

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

Our prior knowledge of what Bayes looks like

P(Bayes)

The likelihood is represented implicitly through simulations.

$$\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}}) \propto \mathcal{L}(\boldsymbol{\omega})\mathcal{P}(\boldsymbol{\omega})$$

Field-based Implicit likelihood cosmological inference

Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ $\xrightarrow{?}$ Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

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

Our prior knowledge of what Bayes looks like

P(Bayes)

The likelihood is represented implicitly through simulations.

$$\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}}) \propto \mathcal{L}(\boldsymbol{\omega})\mathcal{P}(\boldsymbol{\omega})$$

- Non-linear gravity, selection effects
- Modelling assumptions, redshift uncertainties, galaxy biases +
  - ✓ In Bayesian analysis, whatever is uncertain gets a pdf.

Field-based Implicit likelihood cosmological inference

Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ $\xrightarrow{?}$ Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

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

What Bayes is likely to look like when panicking

P(Bayes | panicking)

The likelihood is represented implicitly through simulations.

$$\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}}) \propto \mathcal{L}(\boldsymbol{\omega})\mathcal{P}(\boldsymbol{\omega})$$

- Non-linear gravity, selection effects
- Modelling assumptions, redshift uncertainties, galaxy biases +
  - ✓ In Bayesian analysis, whatever is uncertain gets a pdf.

Main challenge: **model misspecification**.

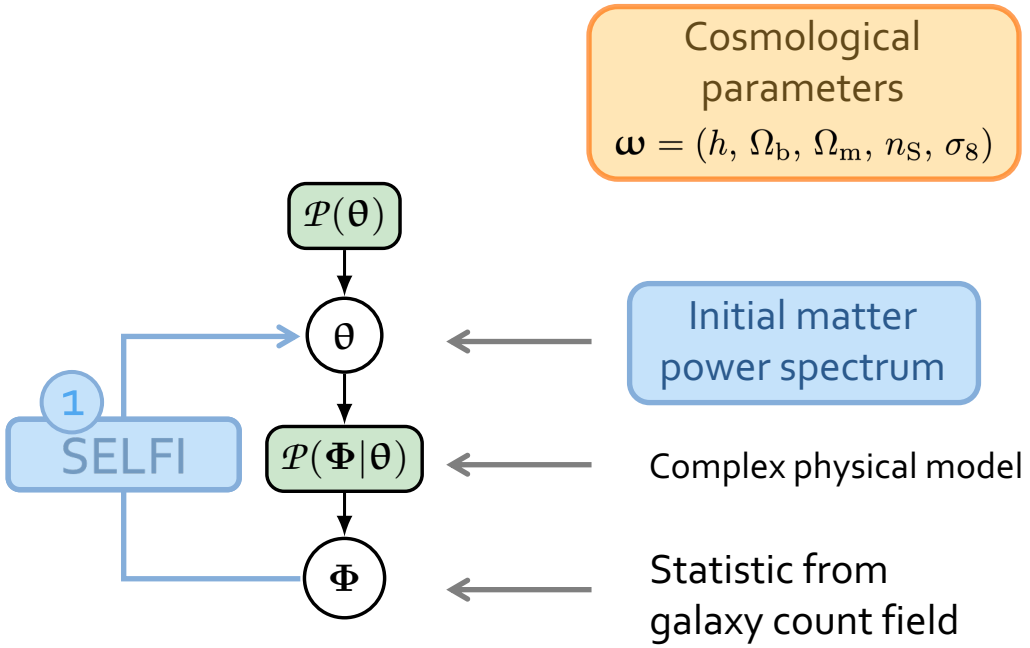
Robustify field-based inference with SELFI

Obervations $\mathbf{\Phi}_\mathrm{O}$ → Posterior $\mathcal{P}(\boldsymbol{\omega}|\mathbf{\Phi}_\mathrm{O})$

SELFI (Simulator Expansion for Likelihood-free Inference)

Leclercq *et al.* 2019, 1902.10149

1. Infer a latent function, the initial matter power spectrum $\boldsymbol{\theta}$



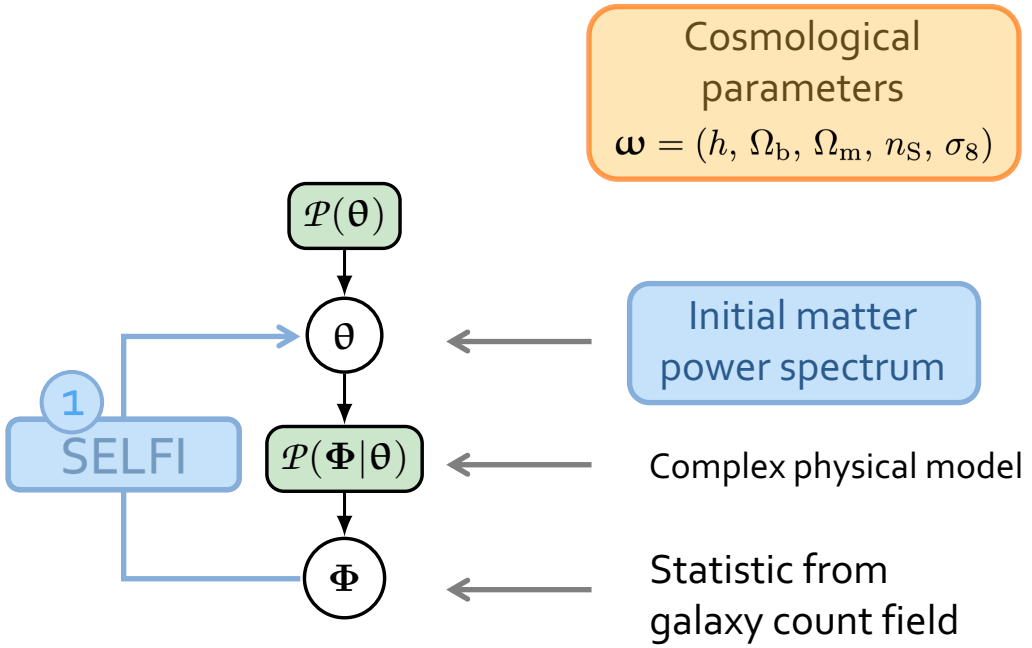
Robustify field-based inference with SELFI

Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ → Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

SELFI (Simulator Expansion for Likelihood-free Inference)

Leclercq *et al.* 2019, 1902.10149

1. Infer a latent function, the initial matter power spectrum $\boldsymbol{\theta}$
   - ✓ Utilise $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$ to thoroughly diagnose systematic effects



Robustify field-based inference with SELFI

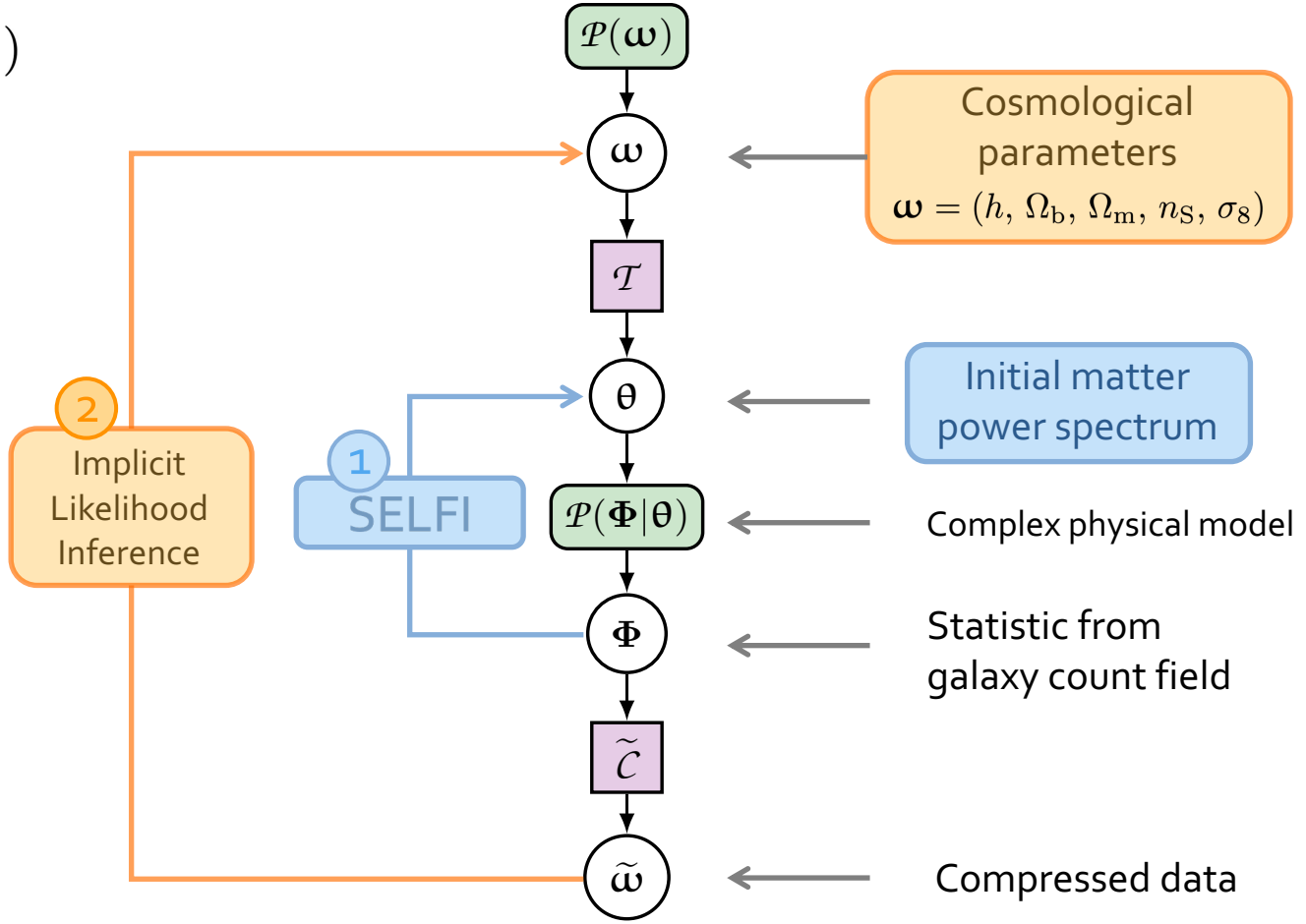
Obervations $\boldsymbol{\Phi}_{\mathrm{O}}$ → Posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$

SELFI (Simulator Expansion for Likelihood-free Inference)

Leclercq *et al.* 2019, 1902.10149

1. Infer a latent function, the initial matter power spectrum $\boldsymbol{\theta}$
   - ✓ Utilise $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi}_{\mathrm{O}})$ to thoroughly diagnose systematic effects
2. Infer the top-level cosmology $\boldsymbol{\omega}$

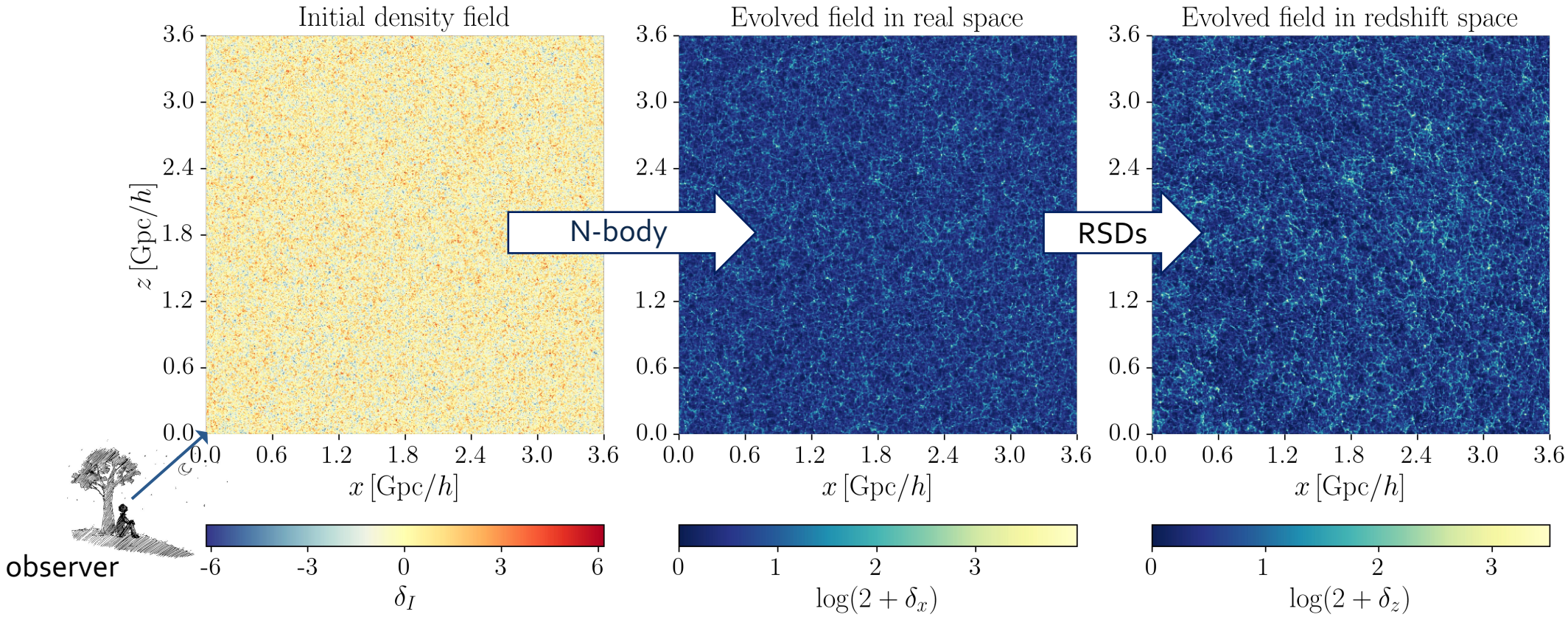
   With any Implicit Likelihood Inference technique



Forward model of spectroscopic galaxy surveys

Leclercq, Jasche & Wandelt 2015, 1502.02690

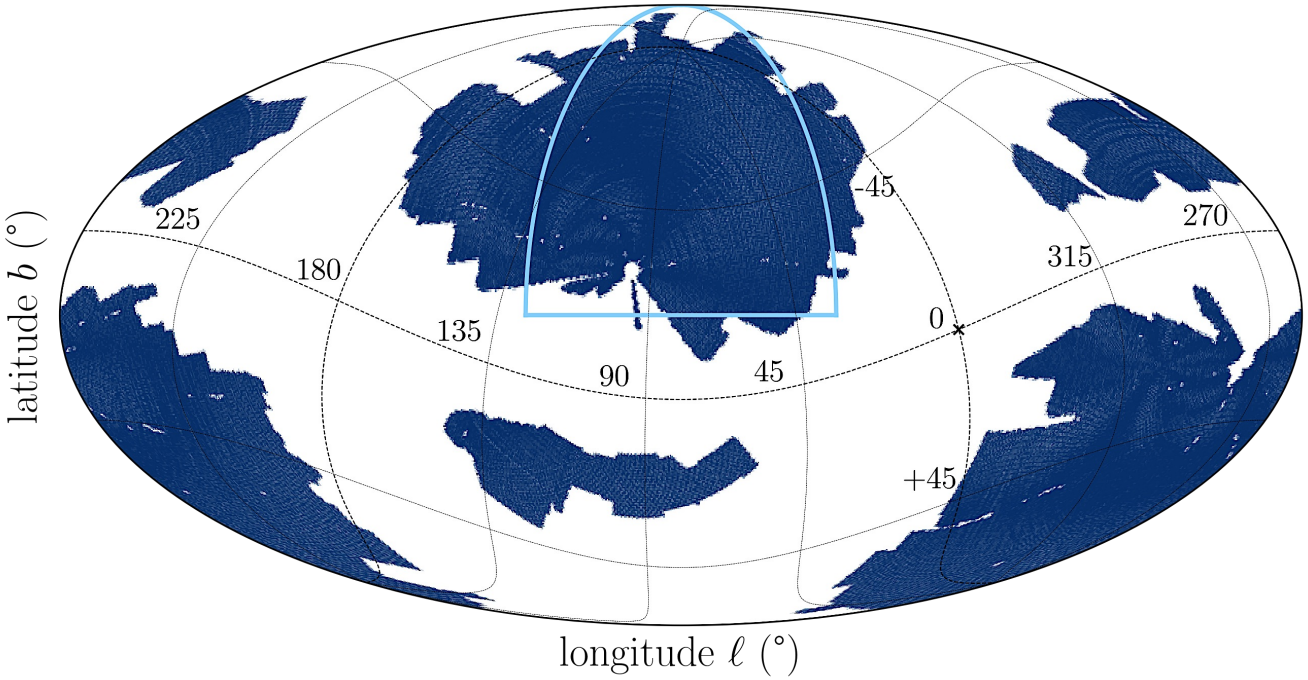
- θ defined on S = 64 support wavenumbers
- Flat Λ-CDM
- Gravitational evolution (N-body) using Simbelmynë
  - 1024^3 dark matter particles on a 1024^3 grid
    Tassev, Zaldarriaga & Eisenstein 2013, 1301.0322



Hoellinger & Leclercq, in prep.

Forward model of spectroscopic galaxy surveys

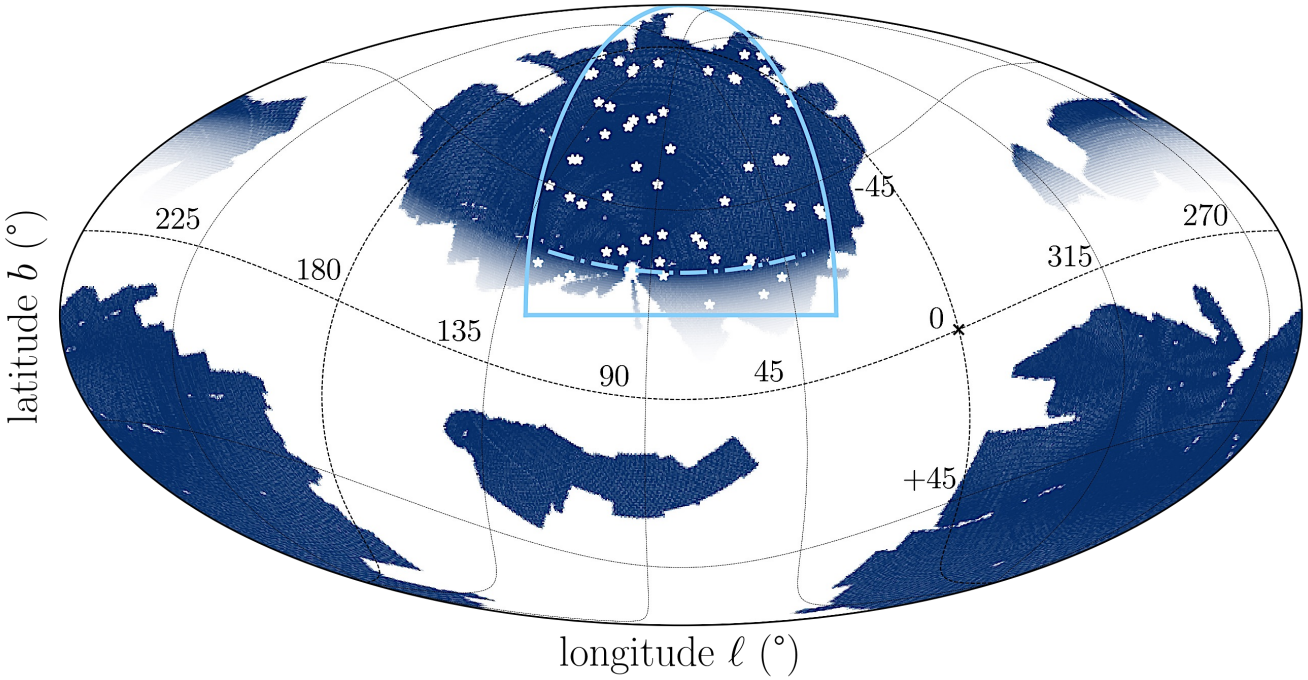
Observer at the corner of a cubic box
covering 1 octant of the sky.



Hoellinger & Leclercq, in prep.

Forward model of spectroscopic galaxy surveys

Observer at the corner of a cubic box
covering 1 octant of the sky.



Model A

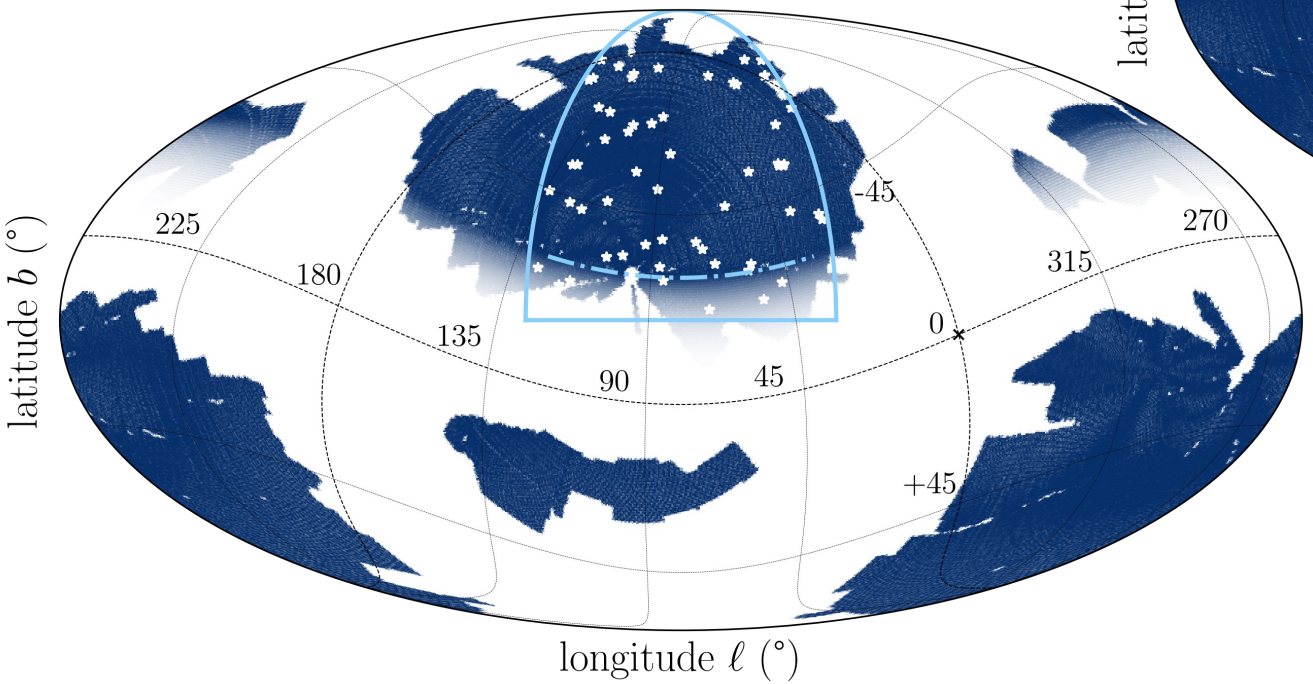
additional holes

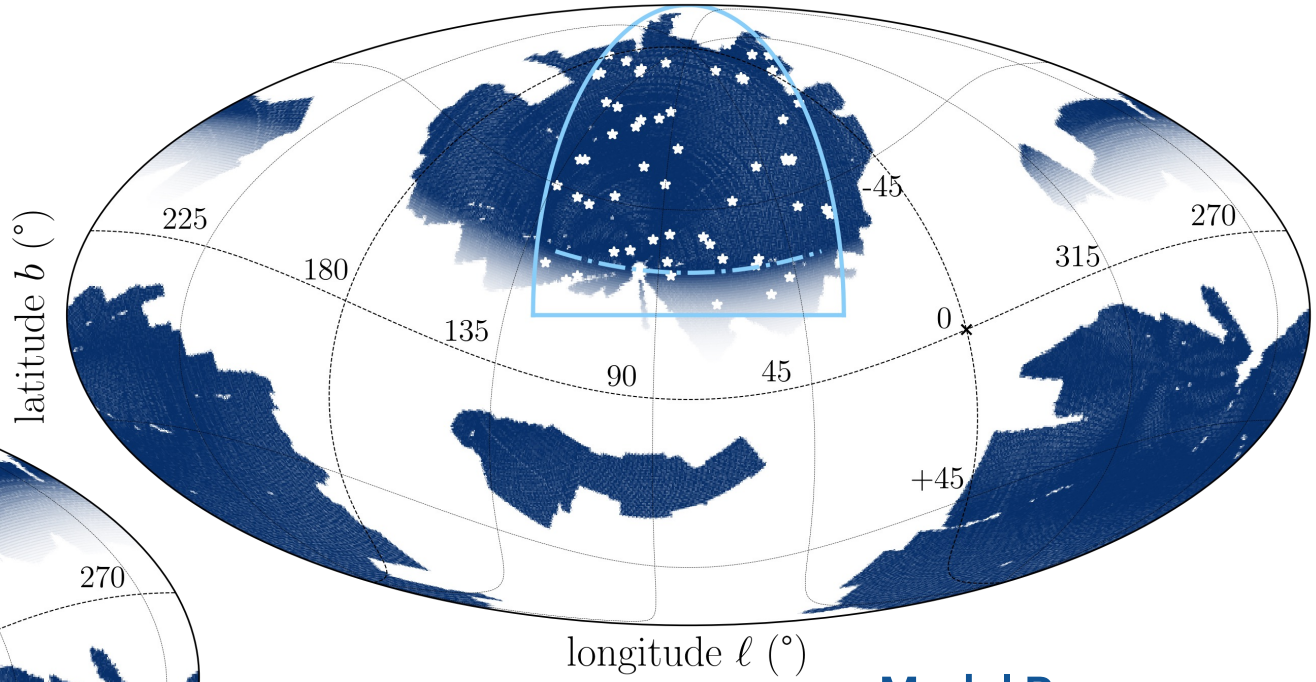
extinction from -60° to 0° latitude (galactic)

Hoellinger & Leclercq, in prep.

Forward model of spectroscopic galaxy surveys

Observer at the corner of a cubic box covering 1 octant of the sky.





Model A

additional holes

extinction from -60° to 0° latitude (galactic)

Model B

additional holes

extinction from -59° to 0°

Hoellinger & Leclercq, in prep.

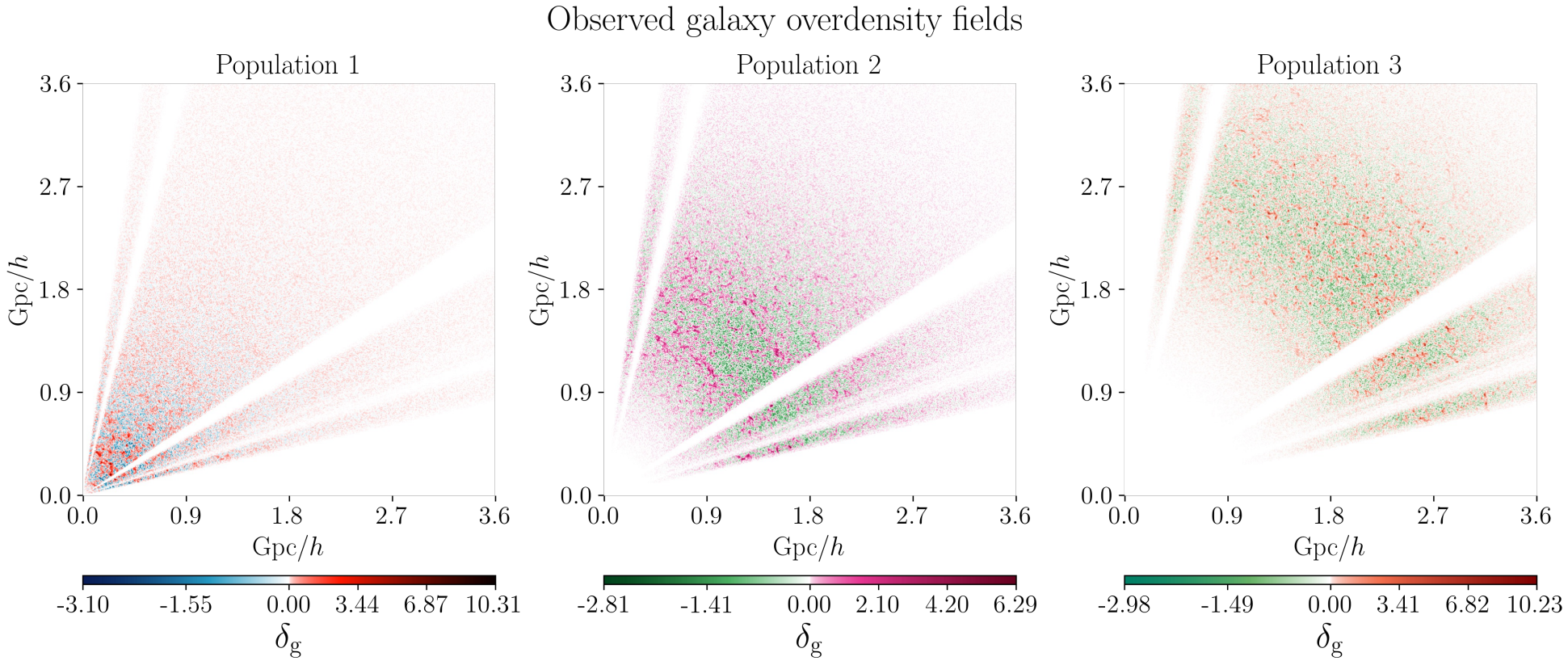
Forward model of spectroscopic galaxy surveys

Model A: correct

- Log-normal selection functions
- First order linear biases

Model B: misspecified

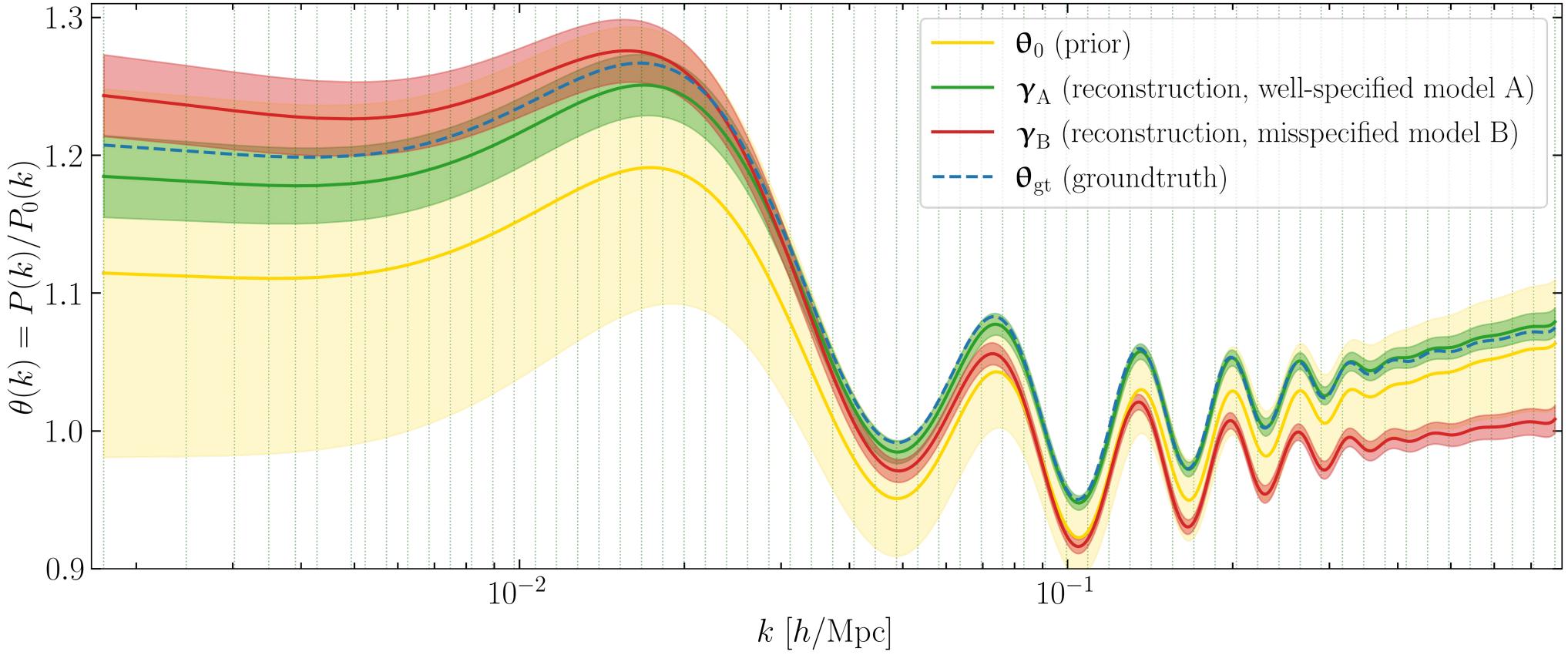
- Misspecified selections functions & biases
- Effect sizes $\mathcal{O}(1\%)$



Hoellinger & Leclercq, in prep.

SELFI posterior for the initial power spectrum

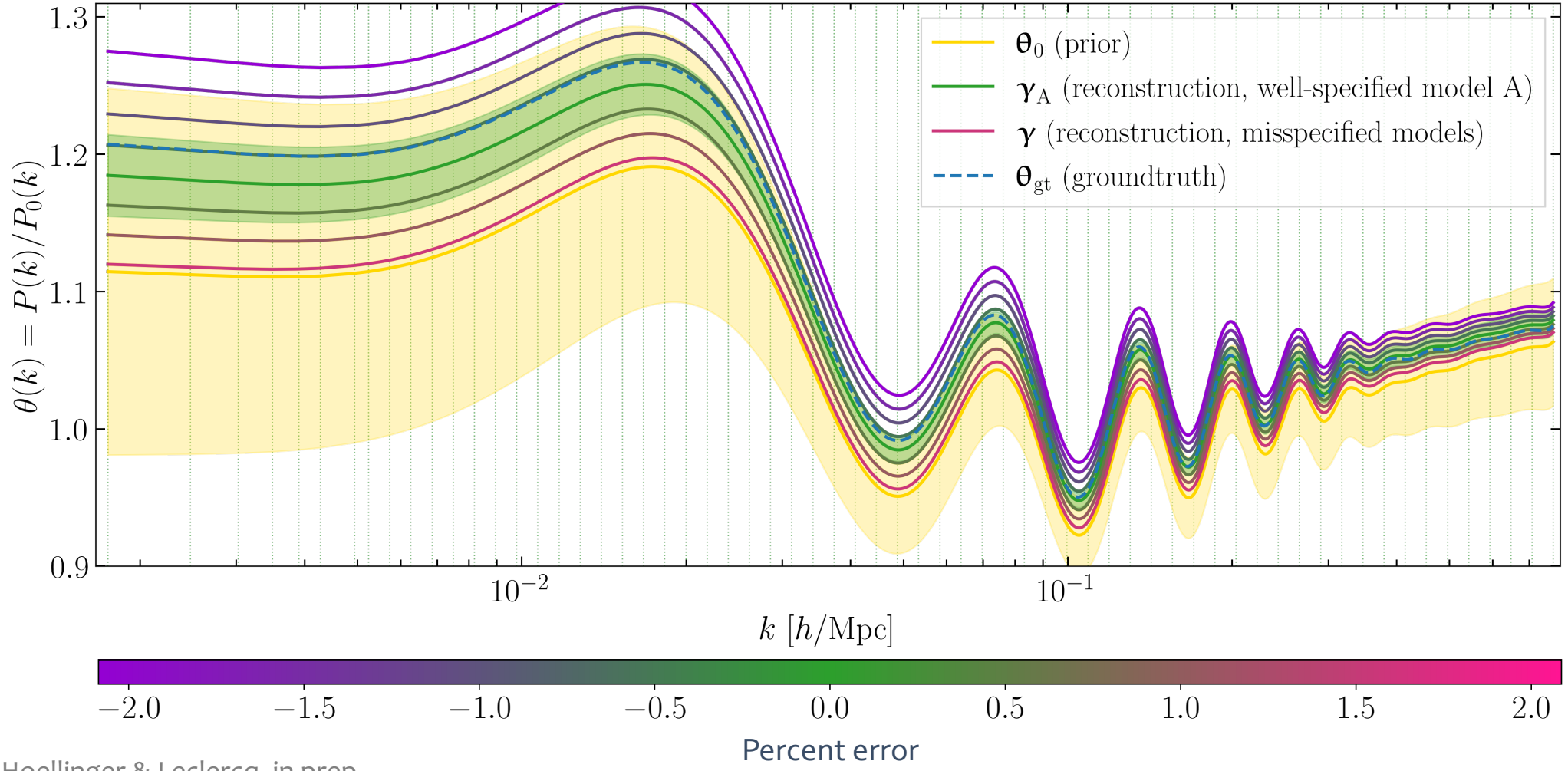
Prior and posterior on the initial matter power spectrum after recombination



Hoellinger & Leclercq, in prep.

SELFI posterior for diagnosing systematic effects

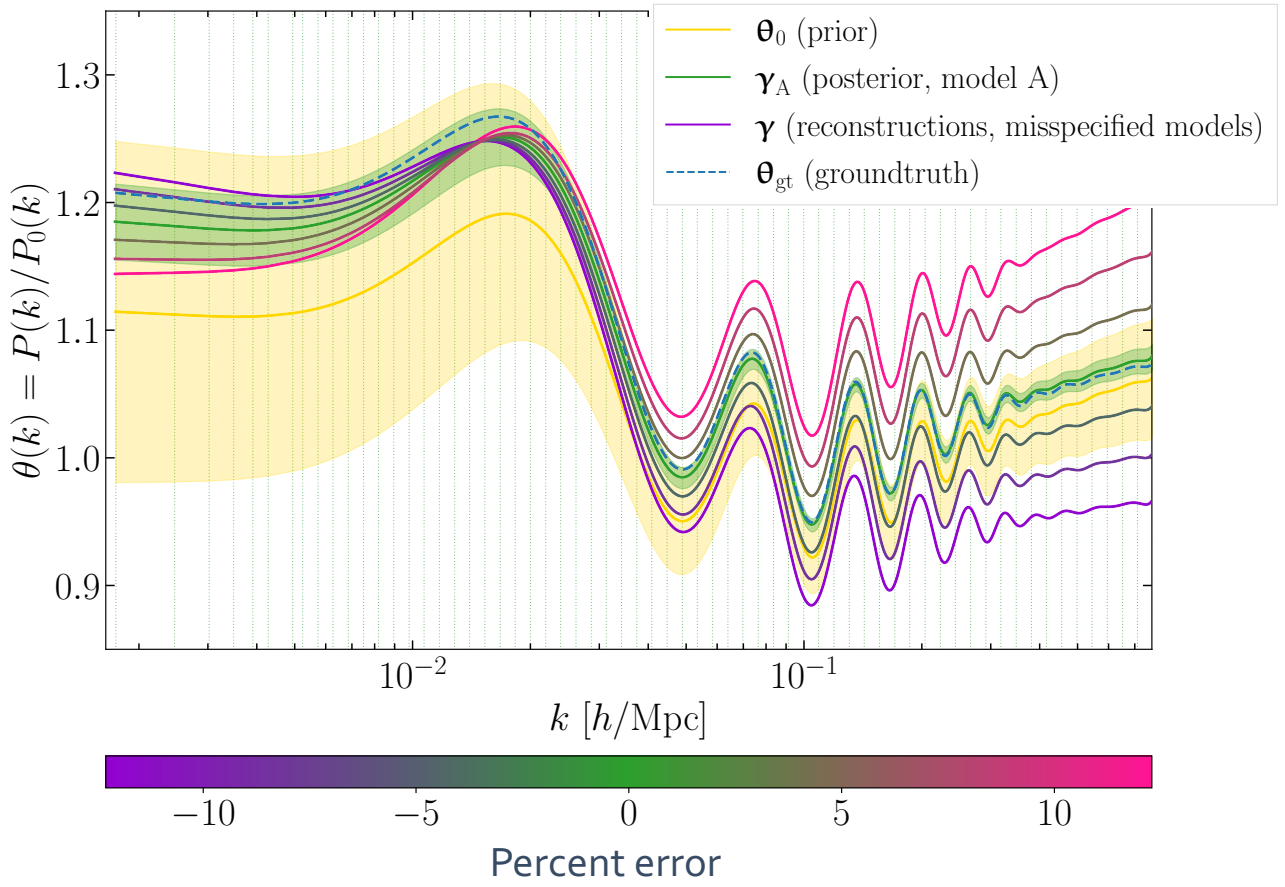
Impact of misspecified galaxy biases on the posterior



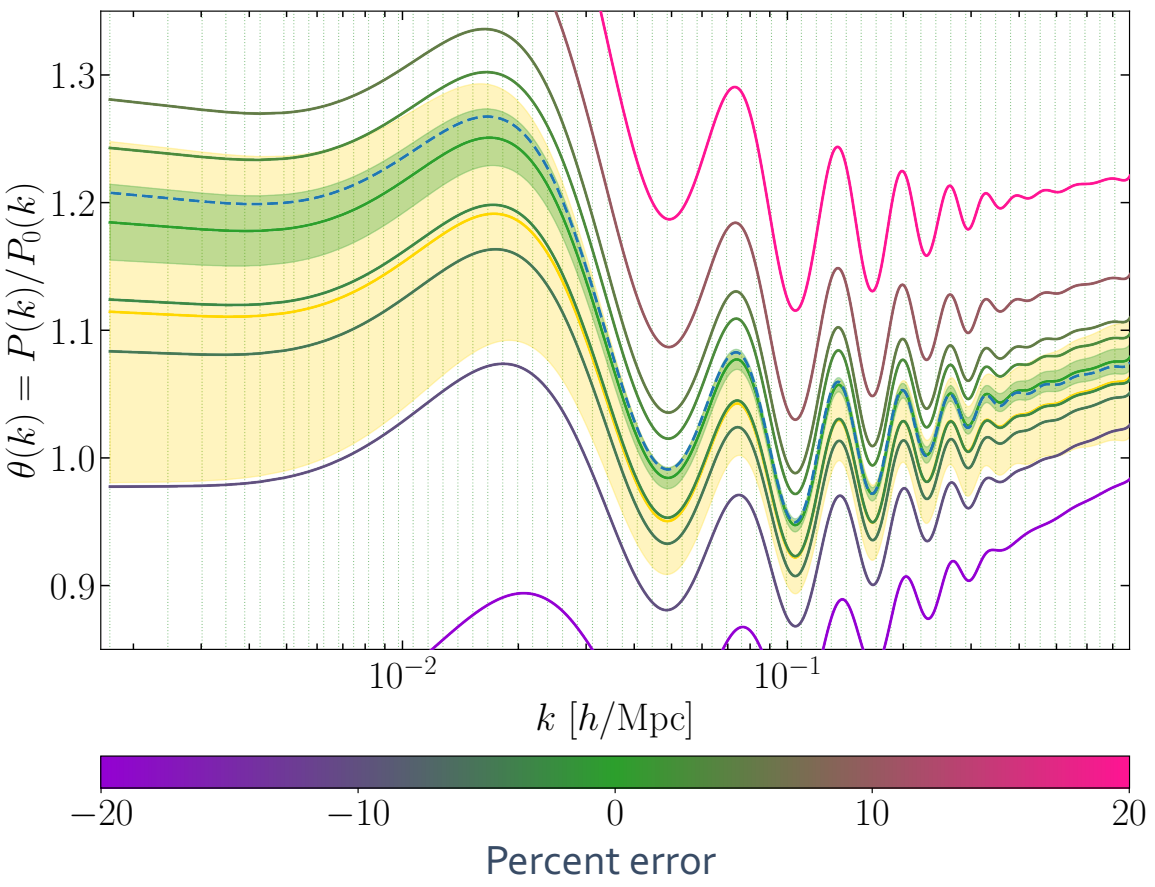
Hoellinger & Leclercq, in prep.

SELFI posterior for diagnosing systematic effects

Impact of misspecified extinction



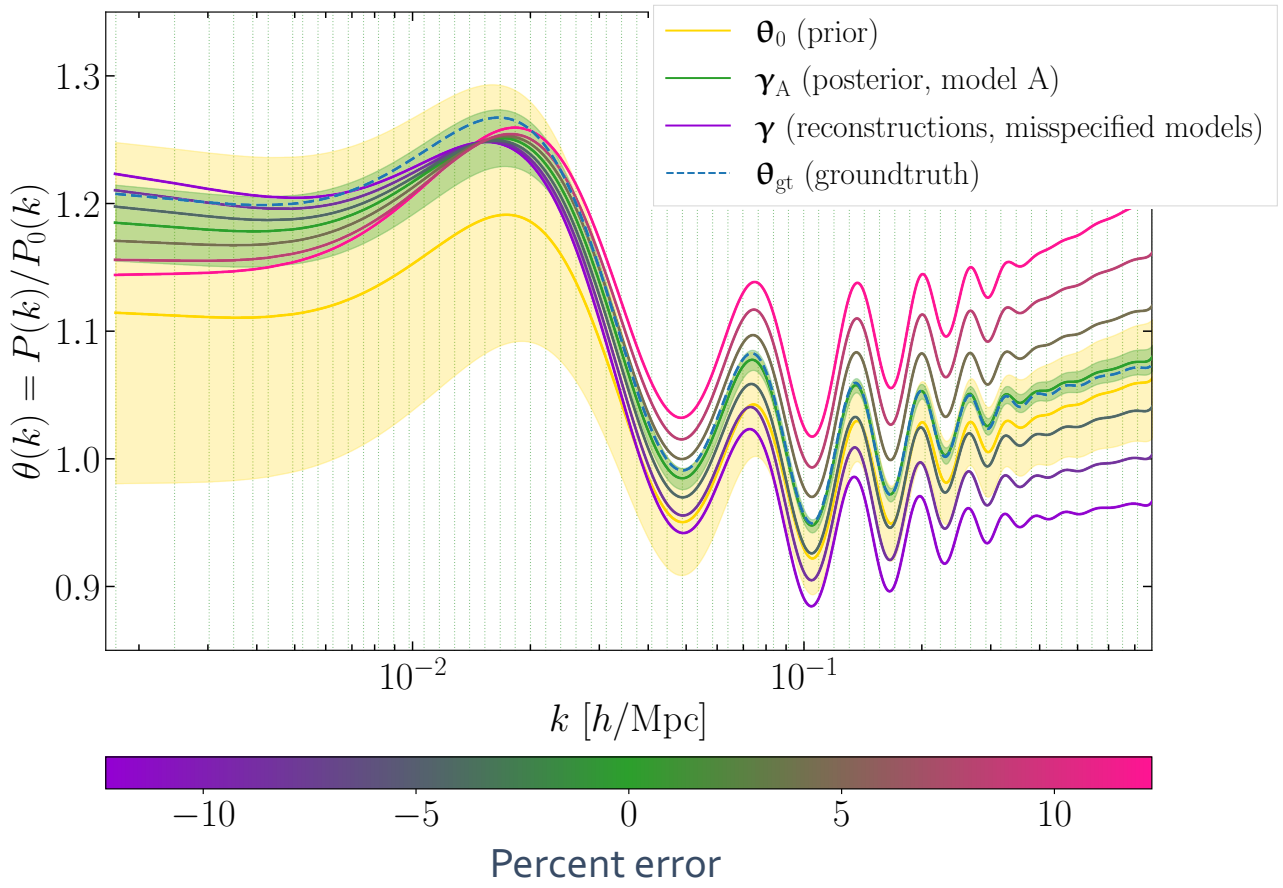
Impact of misspecified selection function variance



Hoellinger & Leclercq, in prep.

SELFI posterior for diagnosing systematic effects

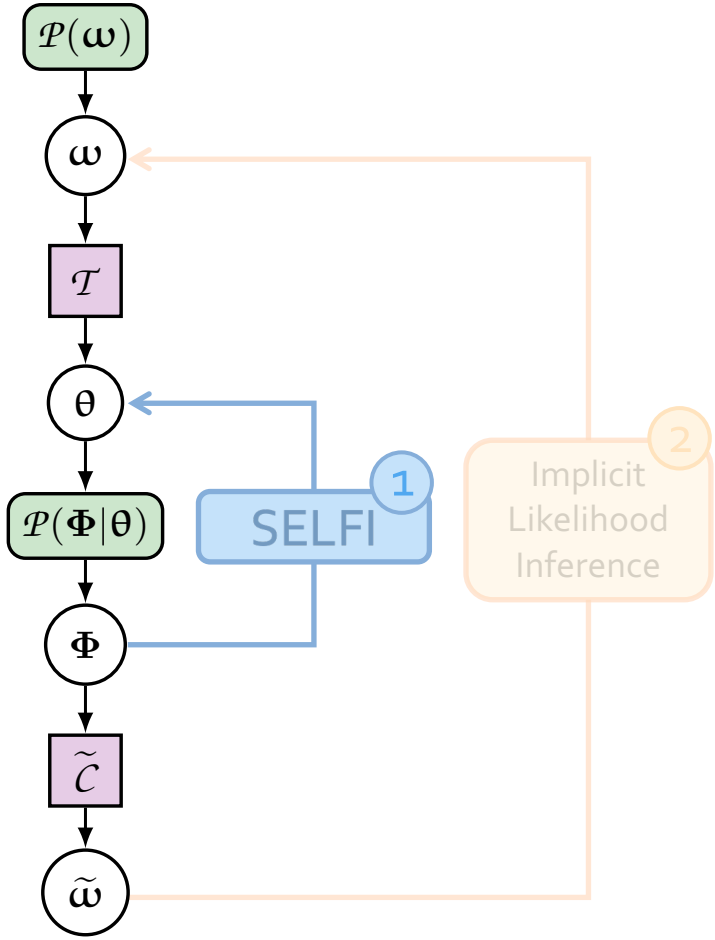
Impact of misspecified extinction



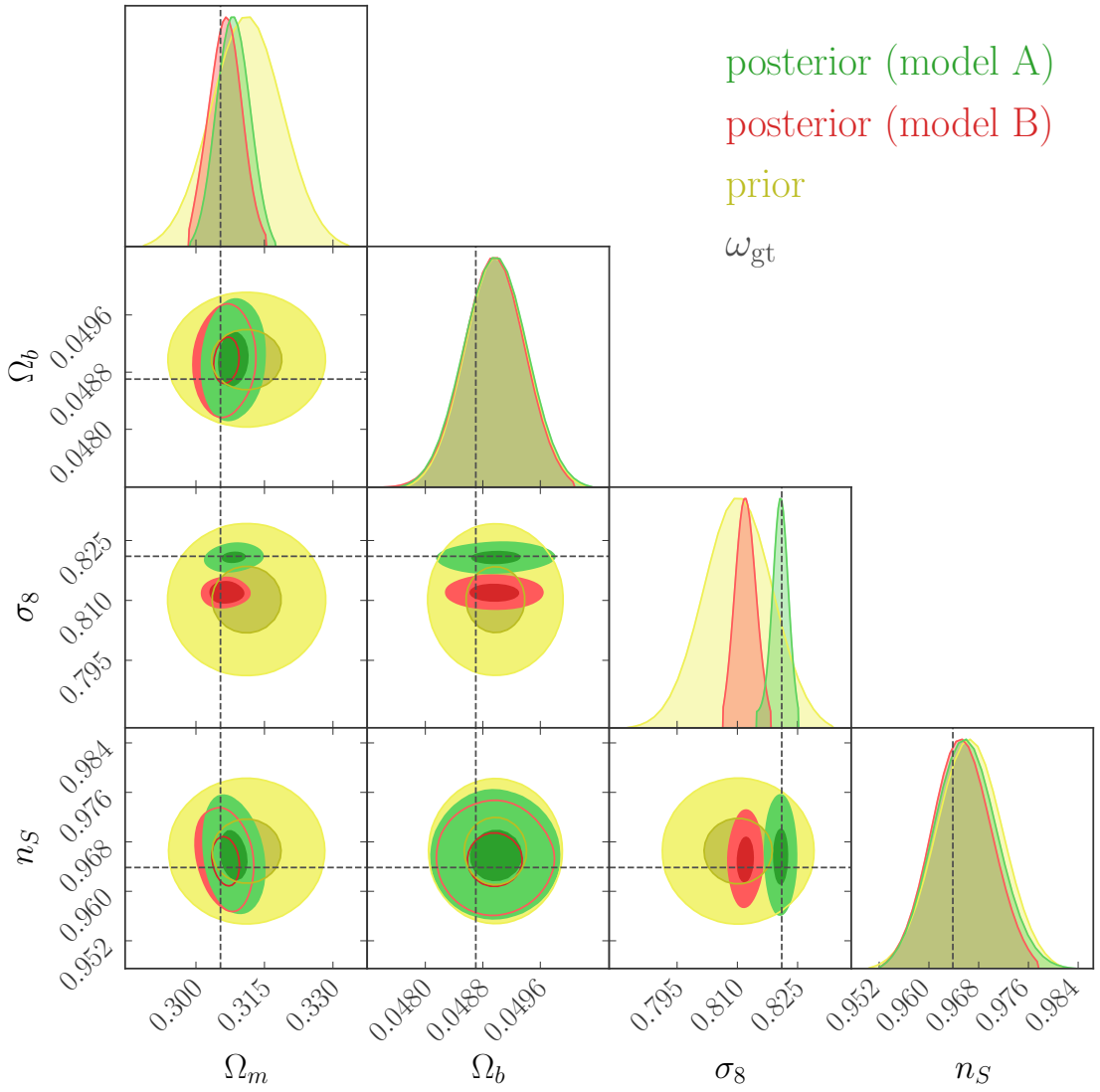
Hoellinger & Leclercq, in prep.

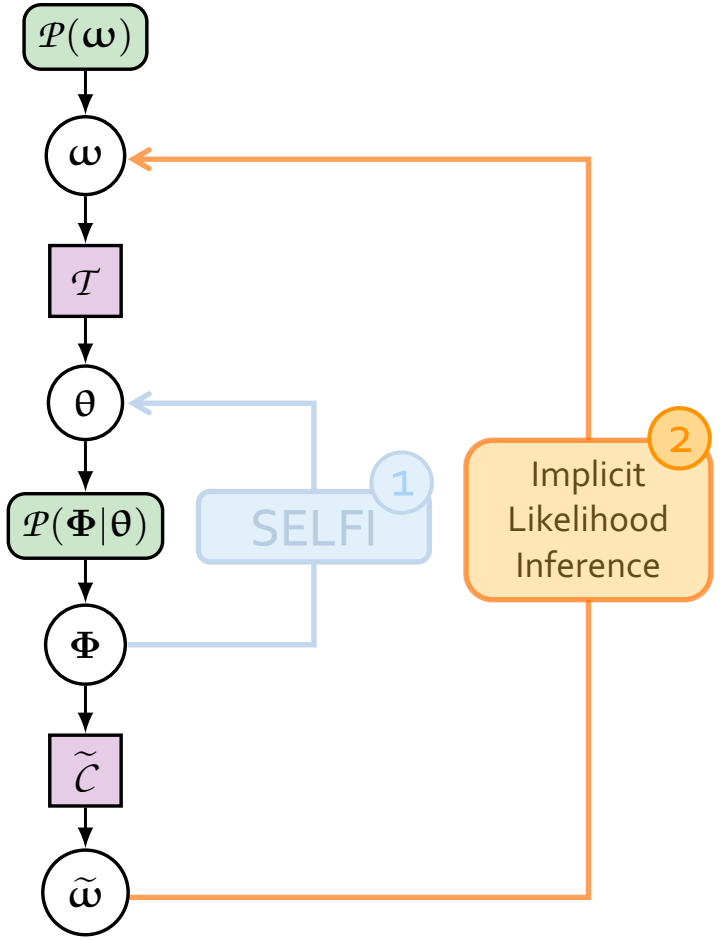
Step 1 diagnose sytematic effects

1,140 N-body simulations



Implicit Likelihood Inference of cosmological parameters





Step 2
ABC-PMC

18,052 N-body simulations for Model A

Hoellinger & Leclercq, in prep.

Conclusion

A two-step framework to address model misspecification in field-based, implicit likelihood cosmological inference:

- utilise our prior knowledge & theoretical insight of the initial matter power spectrum,
- arbitrarily complex physics and systematic effects can be included in the forward model,
- additionally, the N-body simulations are recycled for optimal data compression.

Thanks!

Main references:

Alsing & Wandelt 2018, 1712.00012

Leclercq *et al.* 2019, 1902.10149

Leclercq 2022, 2209.11057

Leclercq, Jasche & Wandelt 2015, 1502.02690

SELFI: github.com/florent-leclercq/pyselfi

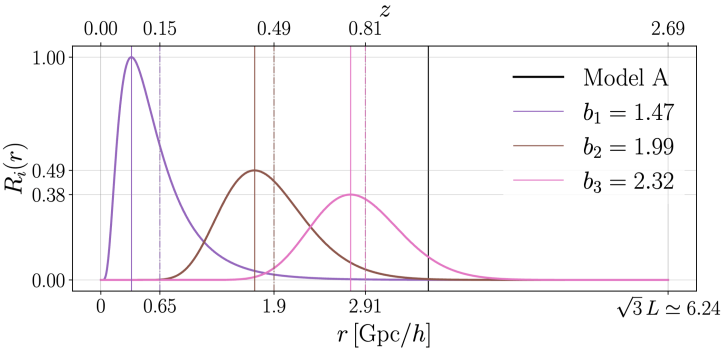
Code & paper: in prep.

Backup slides

The Bayesian Hierarchical Model in details

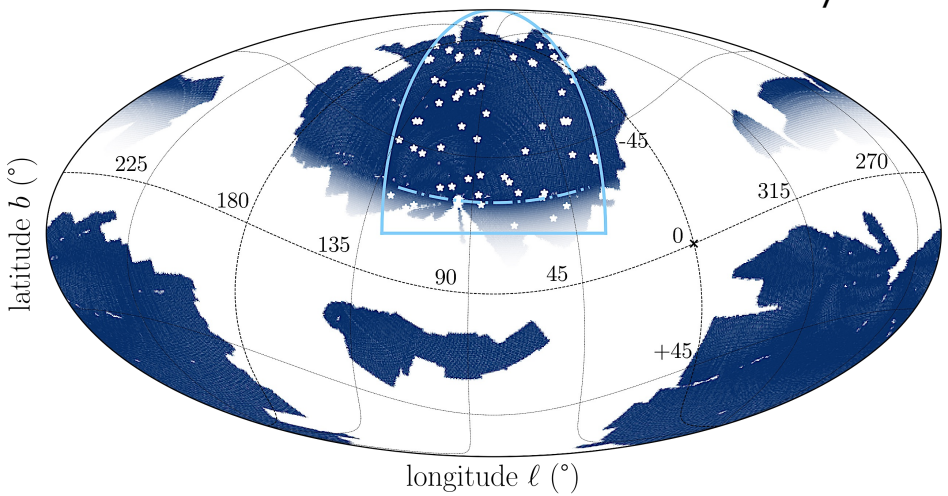
Complex probabilistic observational process

2-LPT, N-body (e.g. COLA, PM), galaxy formation

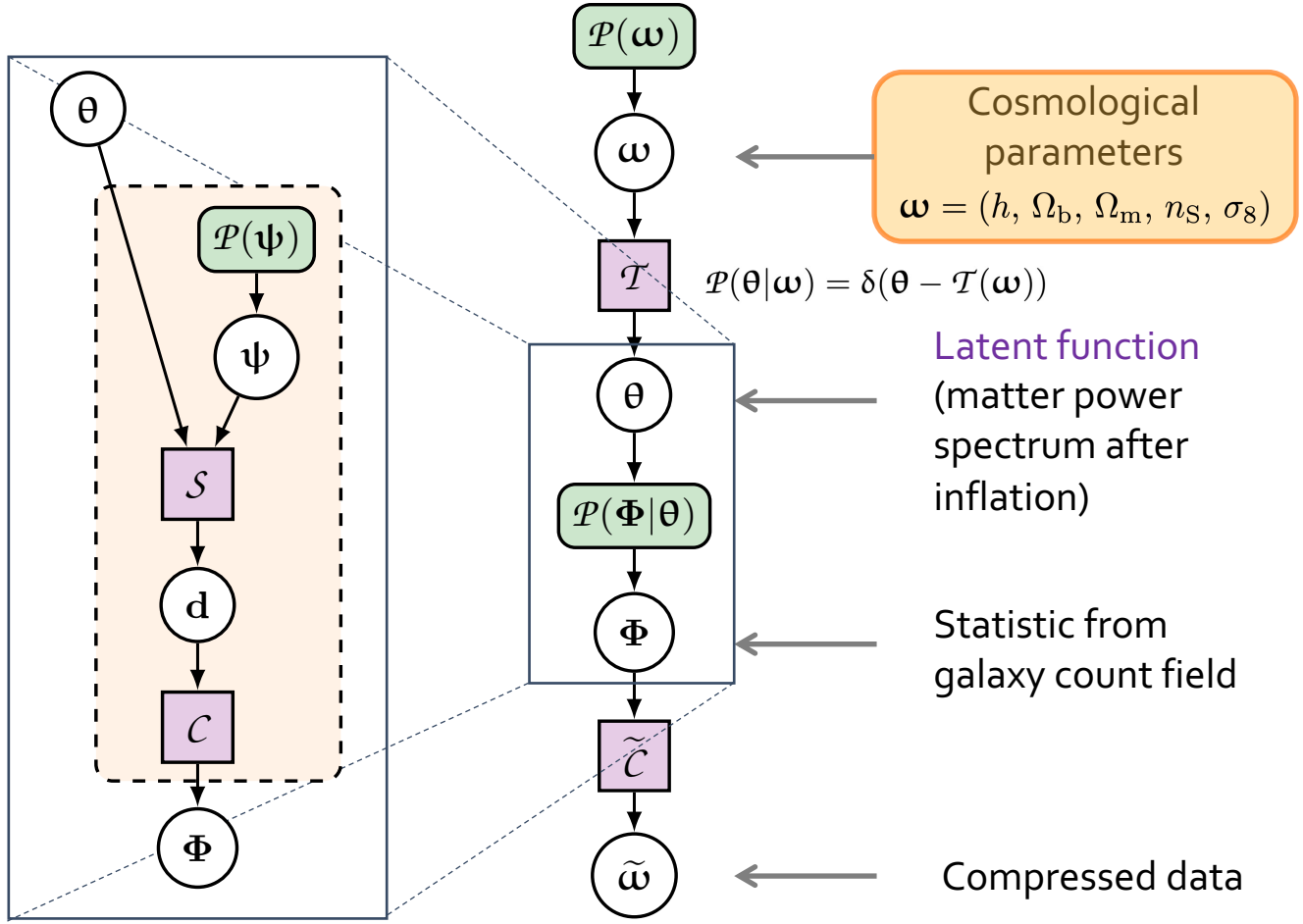


Nuisance parameters $\boldsymbol{\psi}$
- redshift uncertainties
- galaxy biases

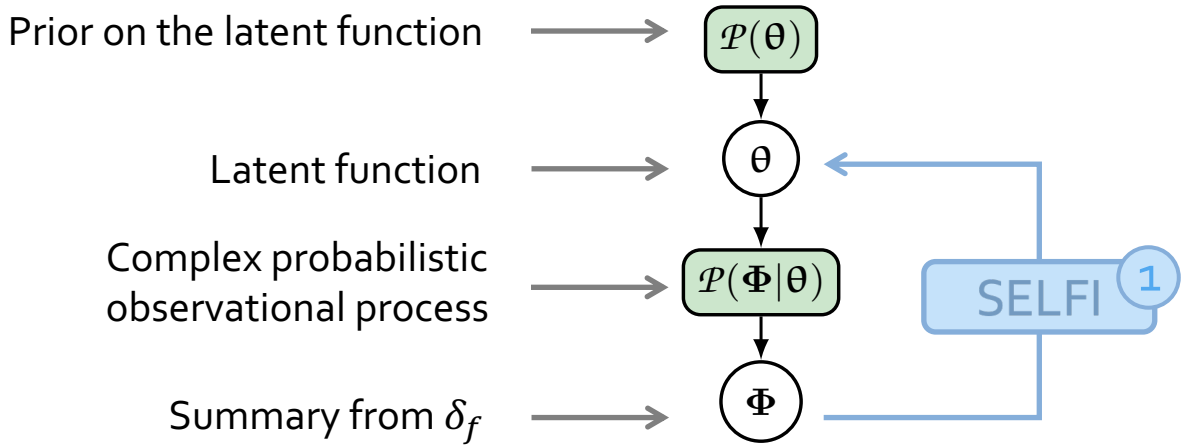
Instrumental effects
- radial selection
- survey mask



Hoellinger & Leclercq, in prep.



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

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

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

➢ Effective posterior (when θ_0 = prior mean)

expansion point — observed summaries

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

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

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

- $\mathbf{f}_0, \mathbf{C}_0$ and $\nabla\mathbf{f}_0$ evaluated through simulations
- The number of simulations is fixed *a priori* (contrary to MCMC)
- These 2 assumptions are not even mandatory if one is ready to use MCMC to get $\mathcal{P}(\theta|\Phi_O)$

Optimal data compression

2 Infer the top-level cosmology $\boldsymbol{\omega}$

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

The compression is optimal in the sense that it preserves the Fisher content of the data.

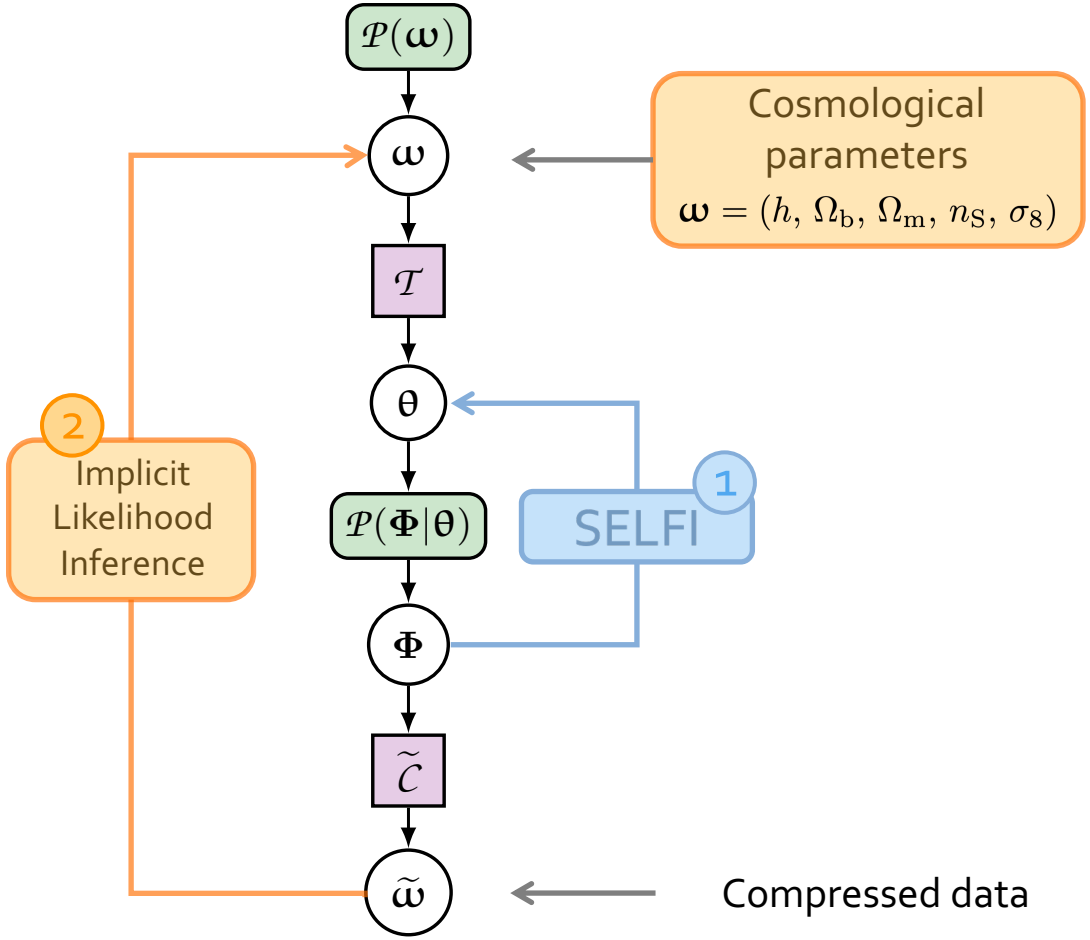
$$\mathcal{C}(\boldsymbol{\Phi}) = \tilde{\boldsymbol{\omega}} \equiv \boldsymbol{\omega}_0 + \mathbf{F}_0^{-1}\left[(\nabla_{\boldsymbol{\omega}}\mathbf{f}_0)^\intercal \mathbf{C}_0^{-1}(\boldsymbol{\Phi} - \mathbf{f}_0)\right]$$

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



The compression is optimal in the sense that it preserves the Fisher content of the data. Hypothesis:

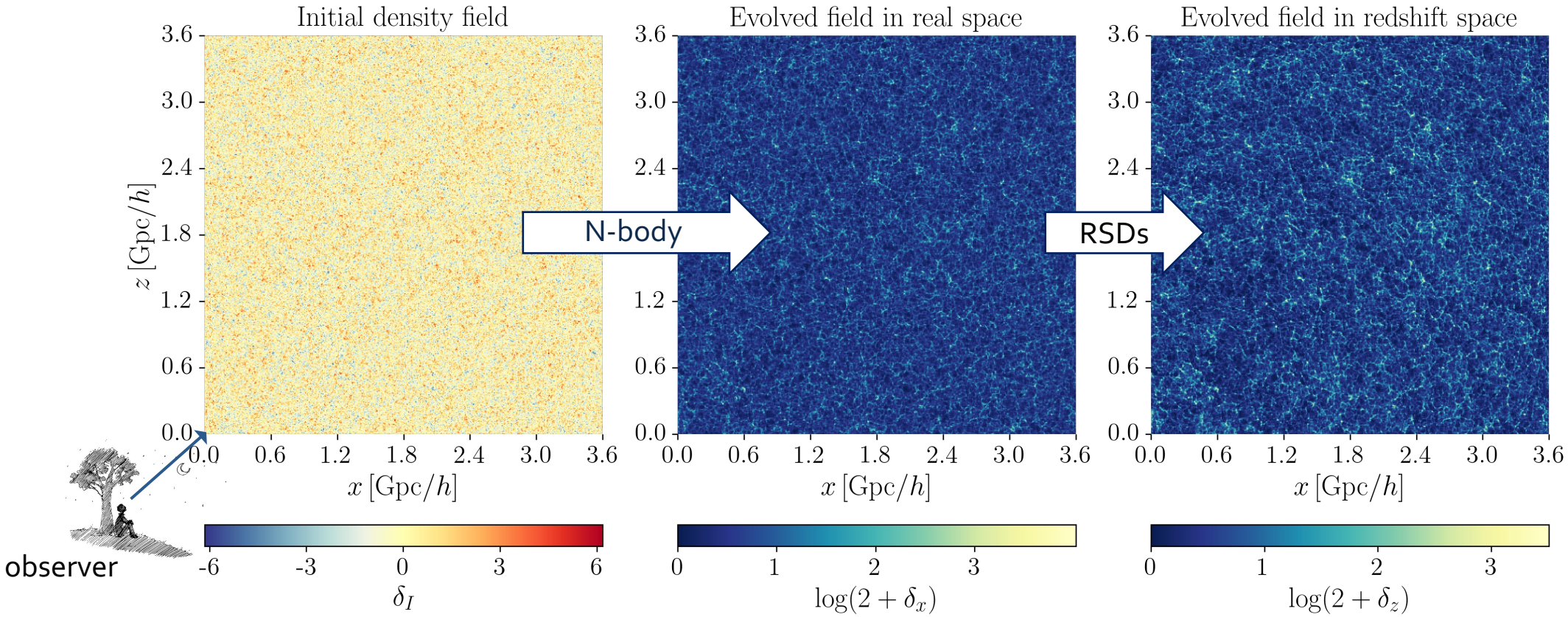
- Locally Gaussian likelihood or: $\nabla\,\mathbb{E}_\theta\left[\nabla^T\mathcal{L}\right] = \mathbb{E}_\theta\left[\nabla\nabla^T\mathcal{L}\right]$
- covariance matrix ~constant close to the expansion point $\nabla_{\boldsymbol{\omega}}\mathbf{C} = 0$

Alsing & Wandelt 2018, 1712.00012 & Leclercq 2022, 2209.11057

Simulator-based data model of galaxy surveys

Leclercq, Jasche & Wandelt 2015, 1502.02690

- θ defined on S = 64 support wavenumbers
- Flat Λ-CDM
- Gravitational evolution (N-body) using Simbelmynë
  - COLA with 1024^3 dark matter particles on a 1024^3 grid
  - Tassev, Zaldarriaga & Eisenstein 2013, 1301.0322

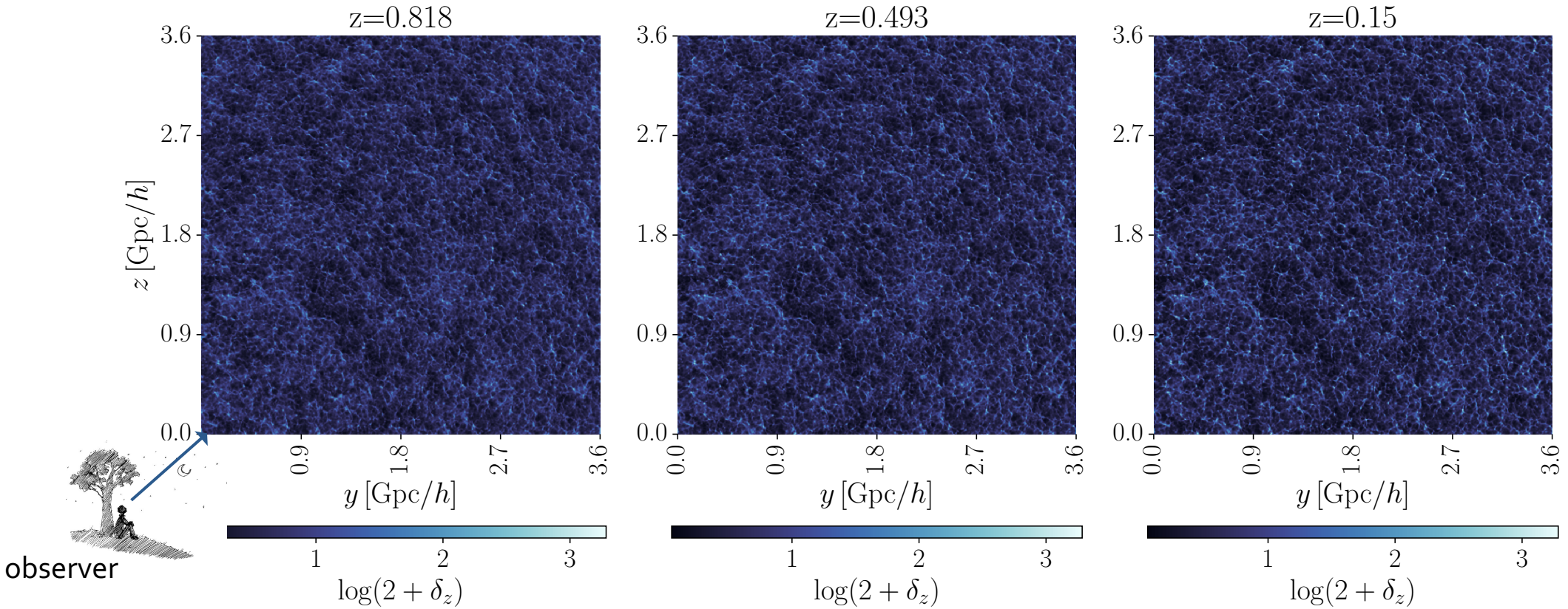


Hoellinger & Leclercq, in prep.

Simulator-based data model of galaxy surveys

Leclercq, Jasche & Wandelt 2015, 1502.02690

- θ defined on S = 64 support wavenumbers
- Flat Λ-CDM
- Gravitational evolution (N-body) using Simbelmynë
  - COLA with 1024^3 dark matter particles on a 1024^3 grid
    Tassev, Zaldarriaga & Eisenstein 2013, 1301.0322



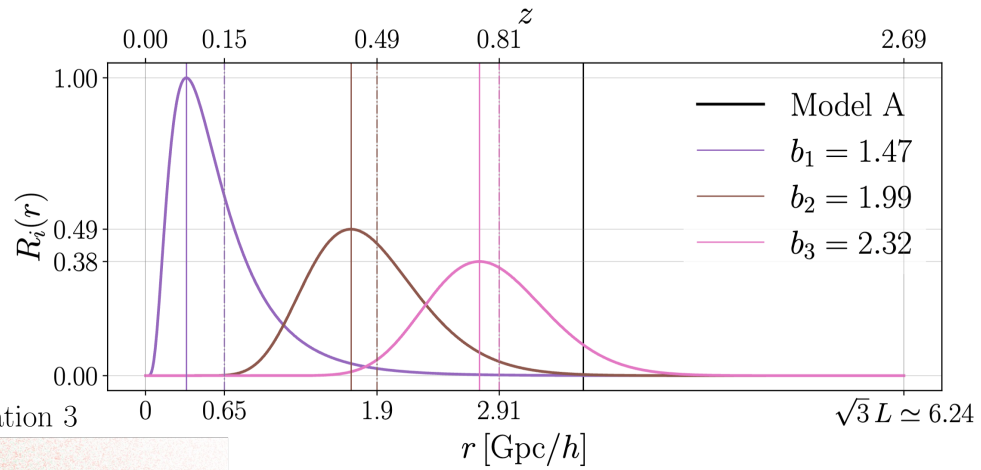
Hoellinger & Leclercq, in prep.

Forward model of spectroscopic galaxy surveys

Model A: correct

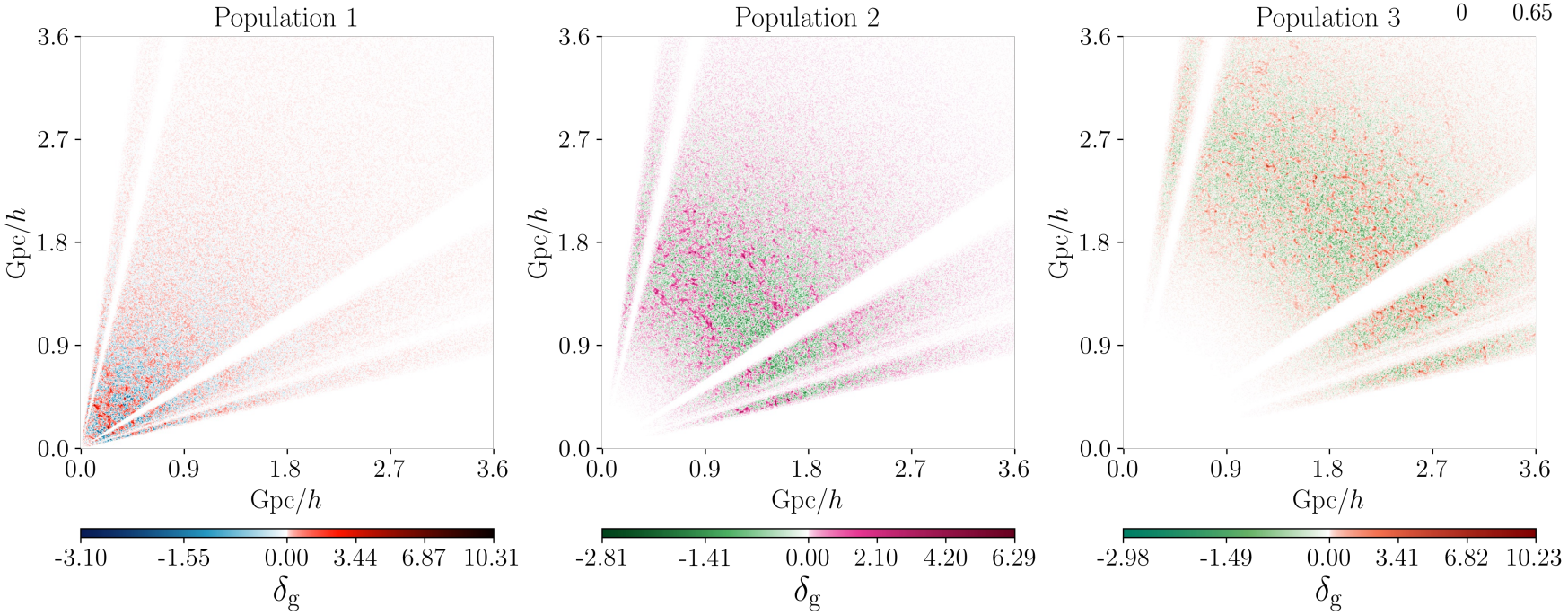
3 mock galaxy populations (1 nearby, 2 LRGs)

- Log-normal selection functions
- First order linear galaxy biases



Biases based on:
Howlett *et al.* 2015, 1409.3238
Gil-Marín *et al.* 2015, 1407.5668

Observed galaxy overdensity fields



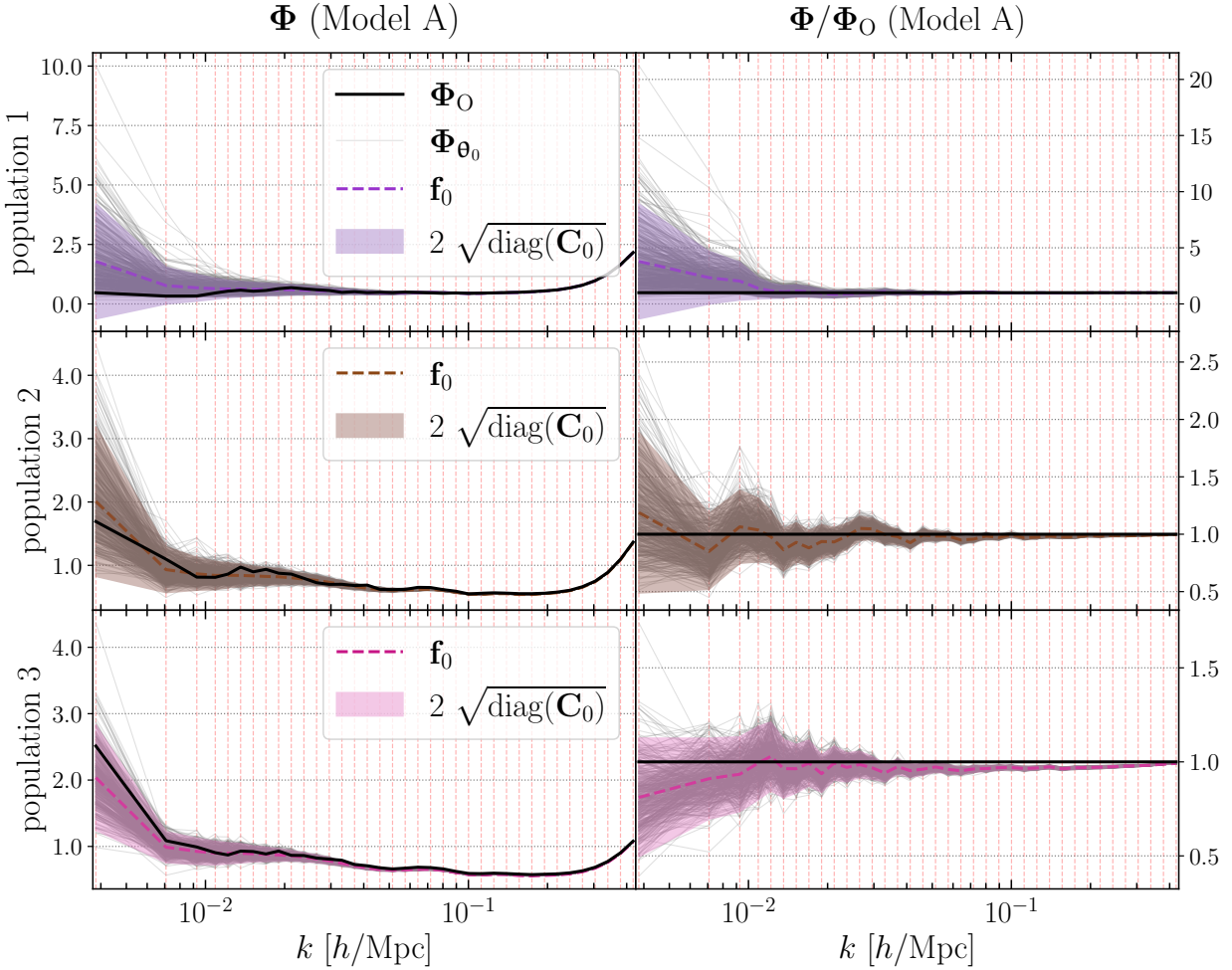
Model B: misspecified

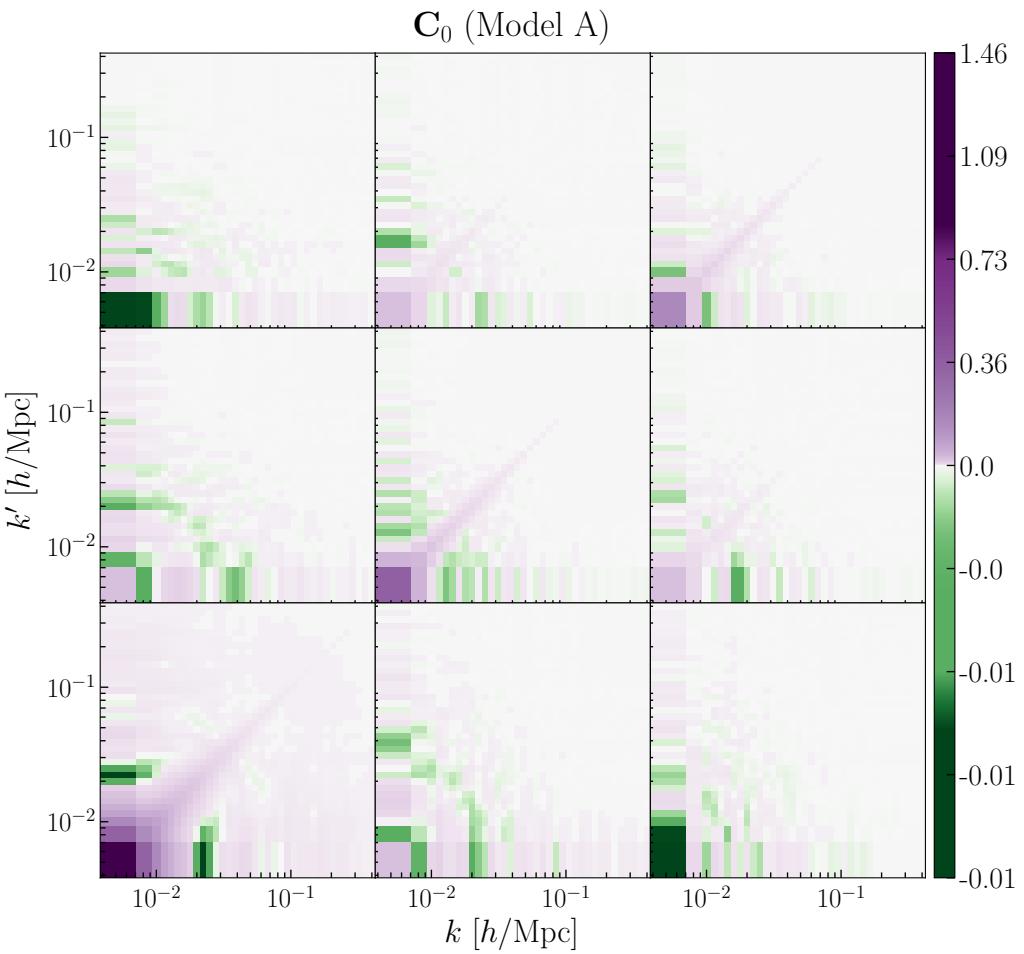
- Misspecified selection functions
- Misspecified biases
- Effect sizes $\mathcal{O}(1\%)$

Hoellinger & Leclercq, in prep.

Generating the simulations for SELFI

Observations, mock data and their covariance. Hard to distinguish well- from mis- specified model.

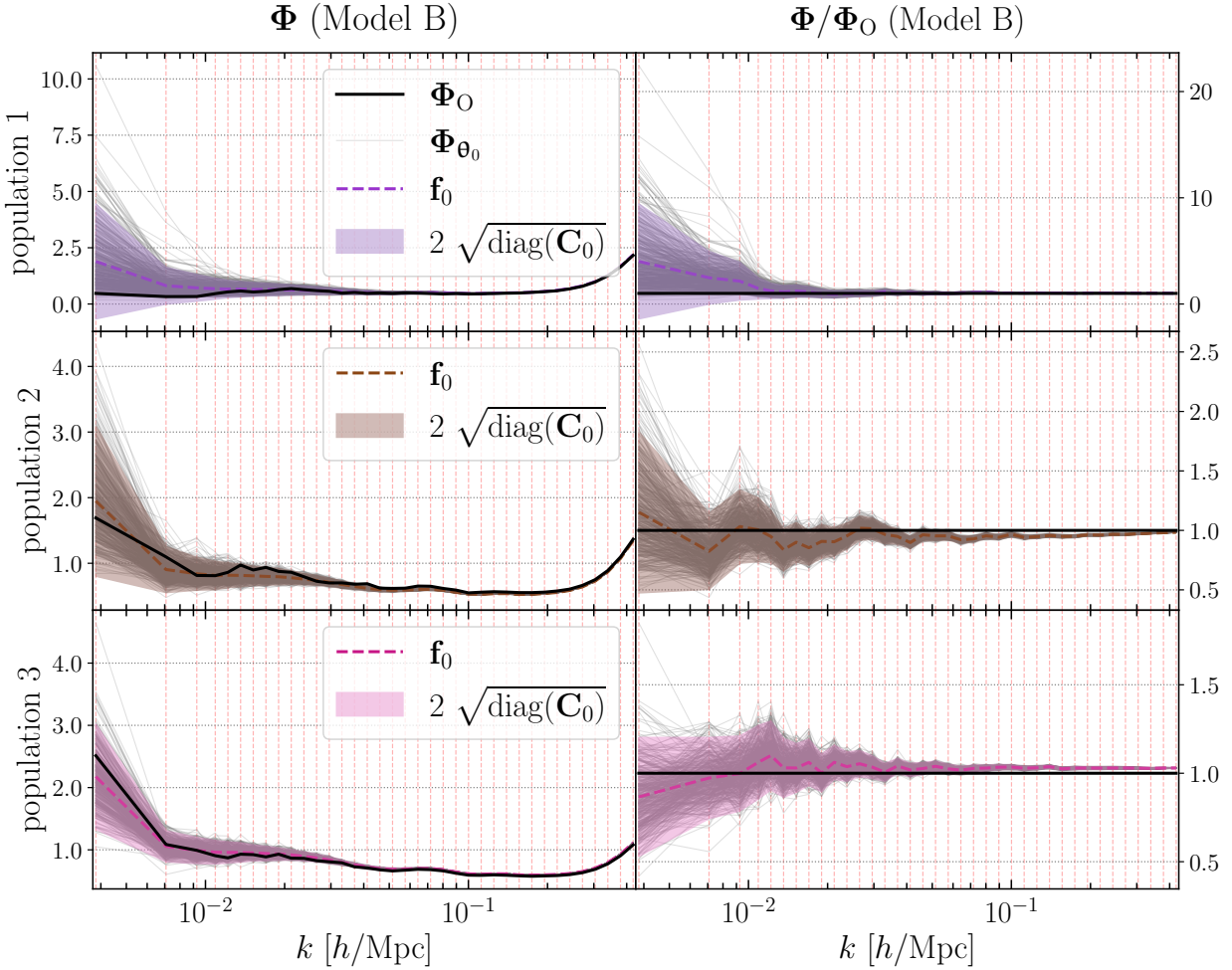


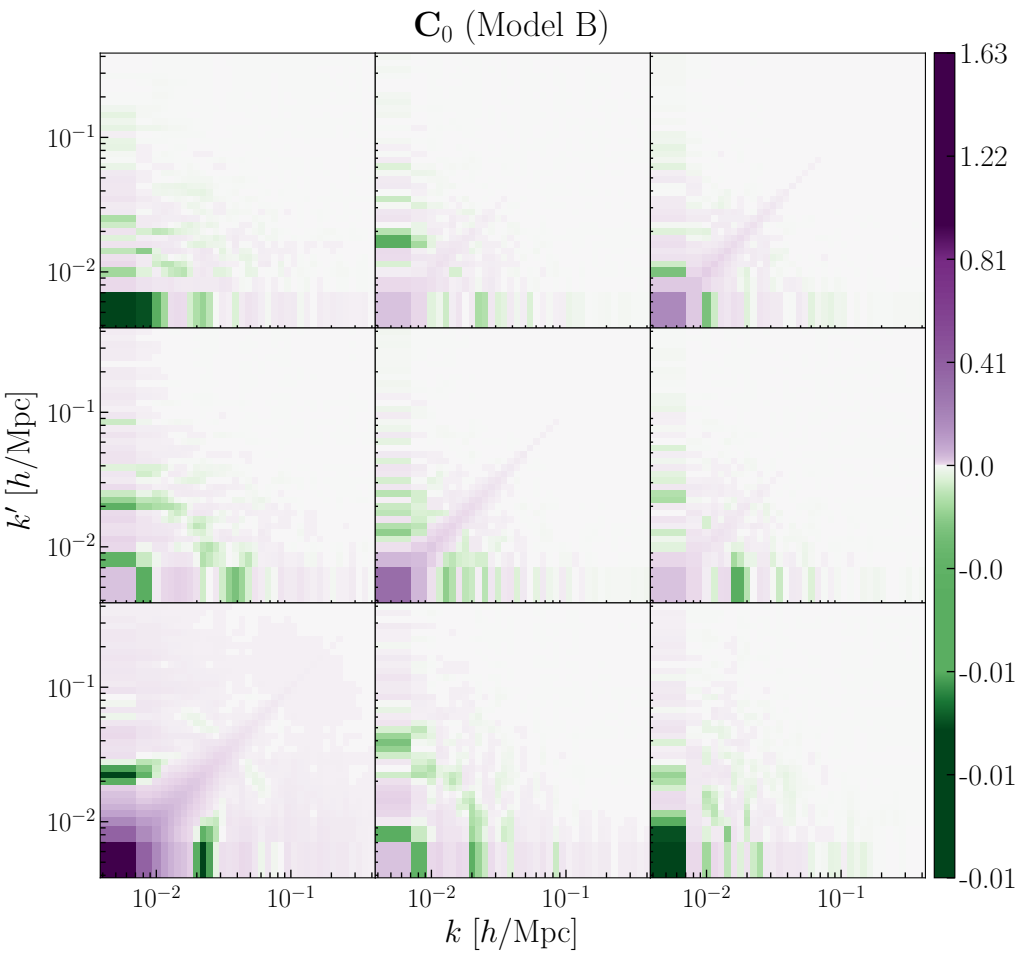


Hoellinger & Leclercq, in prep.

Generating the simulations for SELFI

Observations, mock data and their covariance. Hard to distinguish well- from mis- specified model.

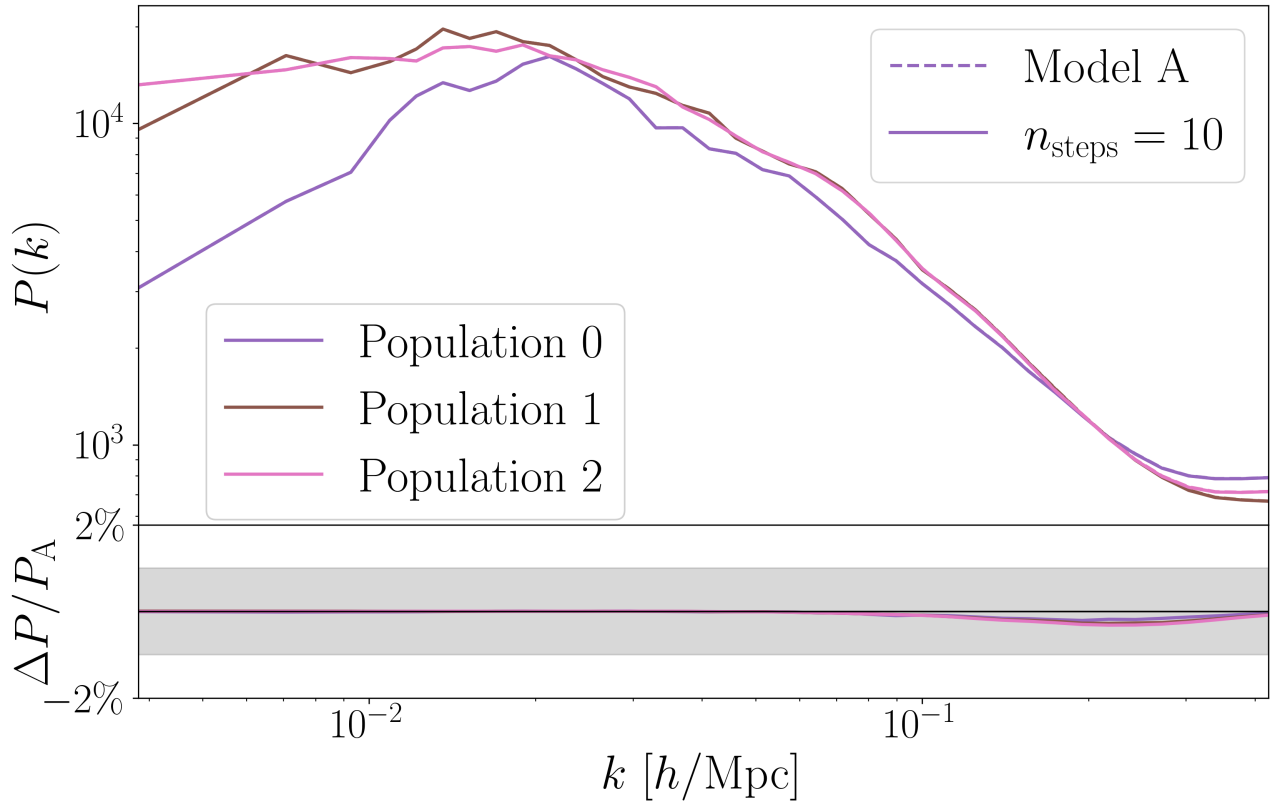




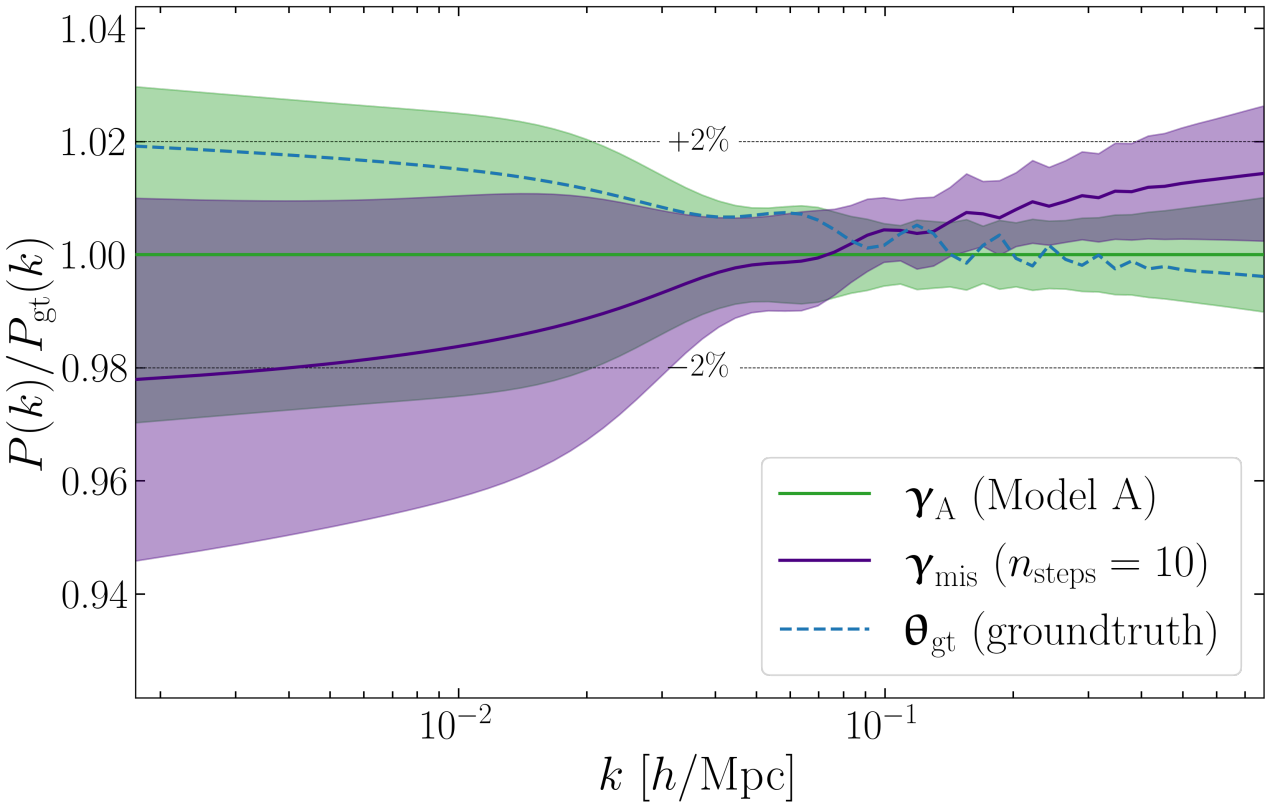
Hoellinger & Leclercq, in prep.

SELFI posterior to compute the inverse error of the gravity solver

Direct error on the measured galaxy spectra with 10 vs 20 time steps for the gravitational evolution



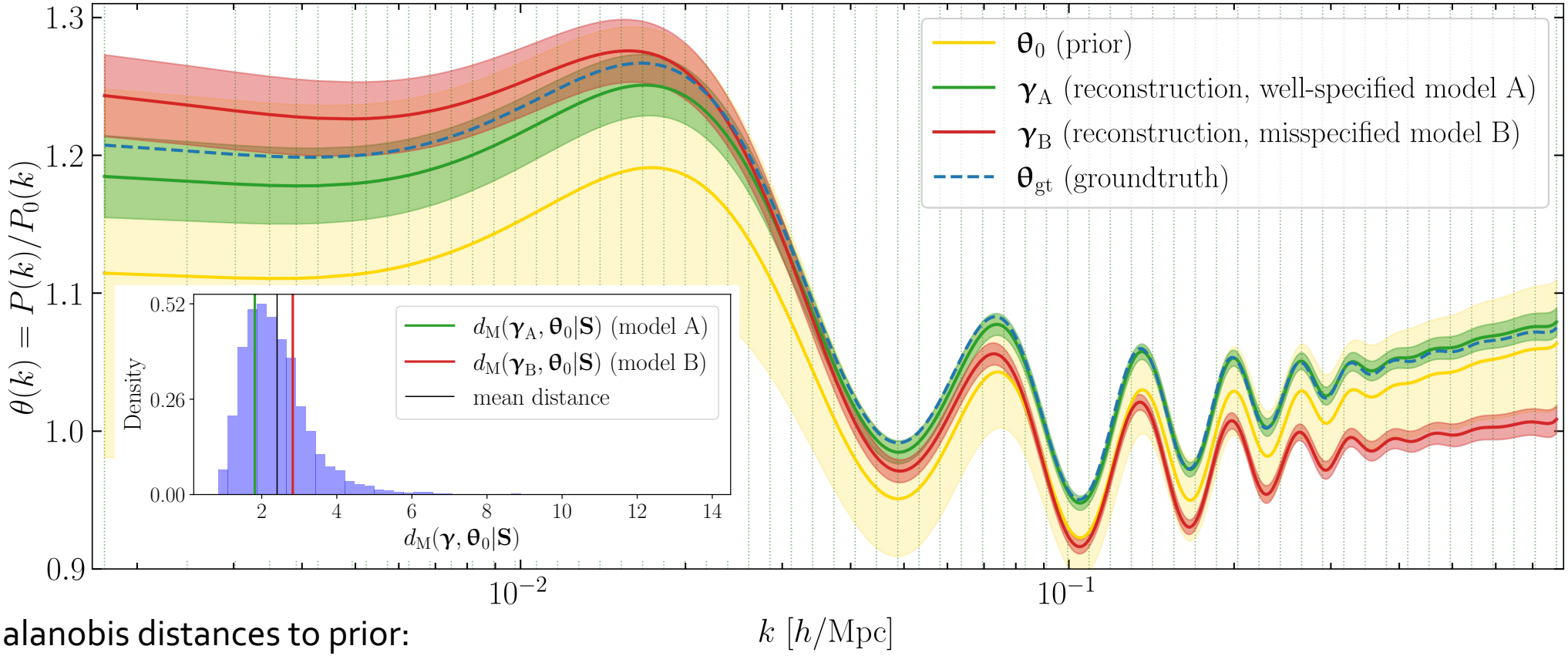
Corresponding inverse error on the SELFI posterior



Hoellinger & Leclercq, in prep.

SELFI posterior for the initial power spectrum

Prior and posterior on the initial matter power spectrum after inflation



Mahalanobis distances to prior:

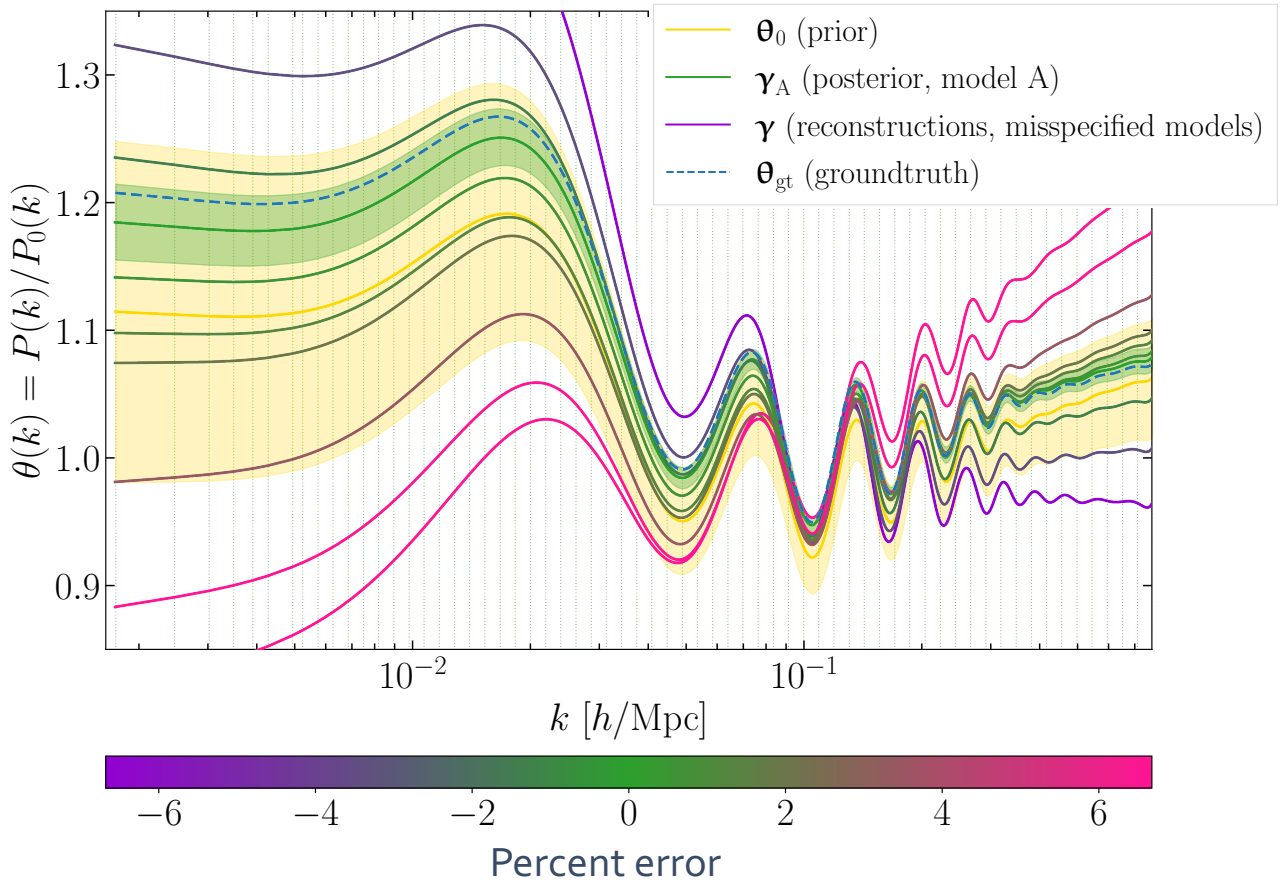
$$d_{\rm M}(\boldsymbol{\gamma}, \boldsymbol{\theta}_0|\mathbf{S}) \equiv \|(\boldsymbol{\gamma} - \boldsymbol{\theta}_0)\|_{\mathbf{S}^{-1}}$$

Model A: 1.81 **Model B:** 2.82

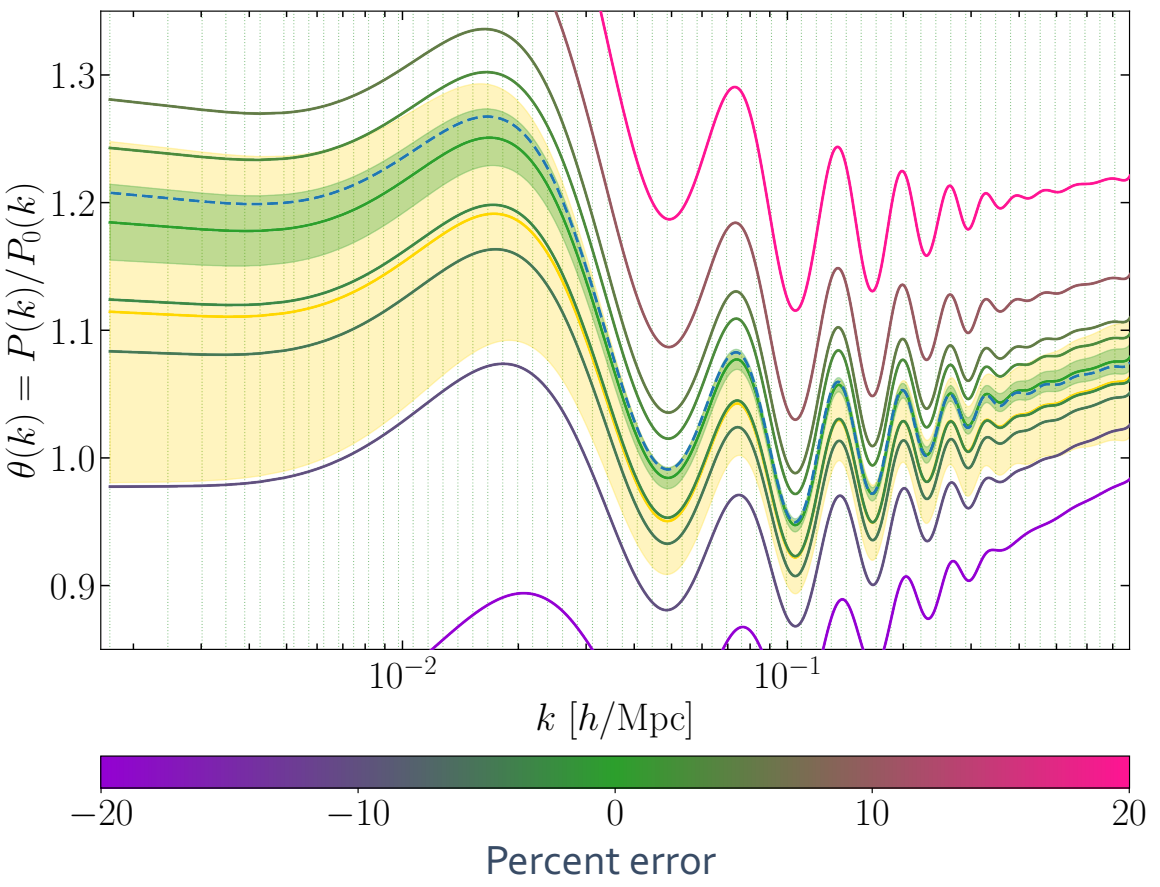
Hoellinger & Leclercq, in prep.

SELFI posterior for diagnosing systematic effects

Impact of misspecified redshifts



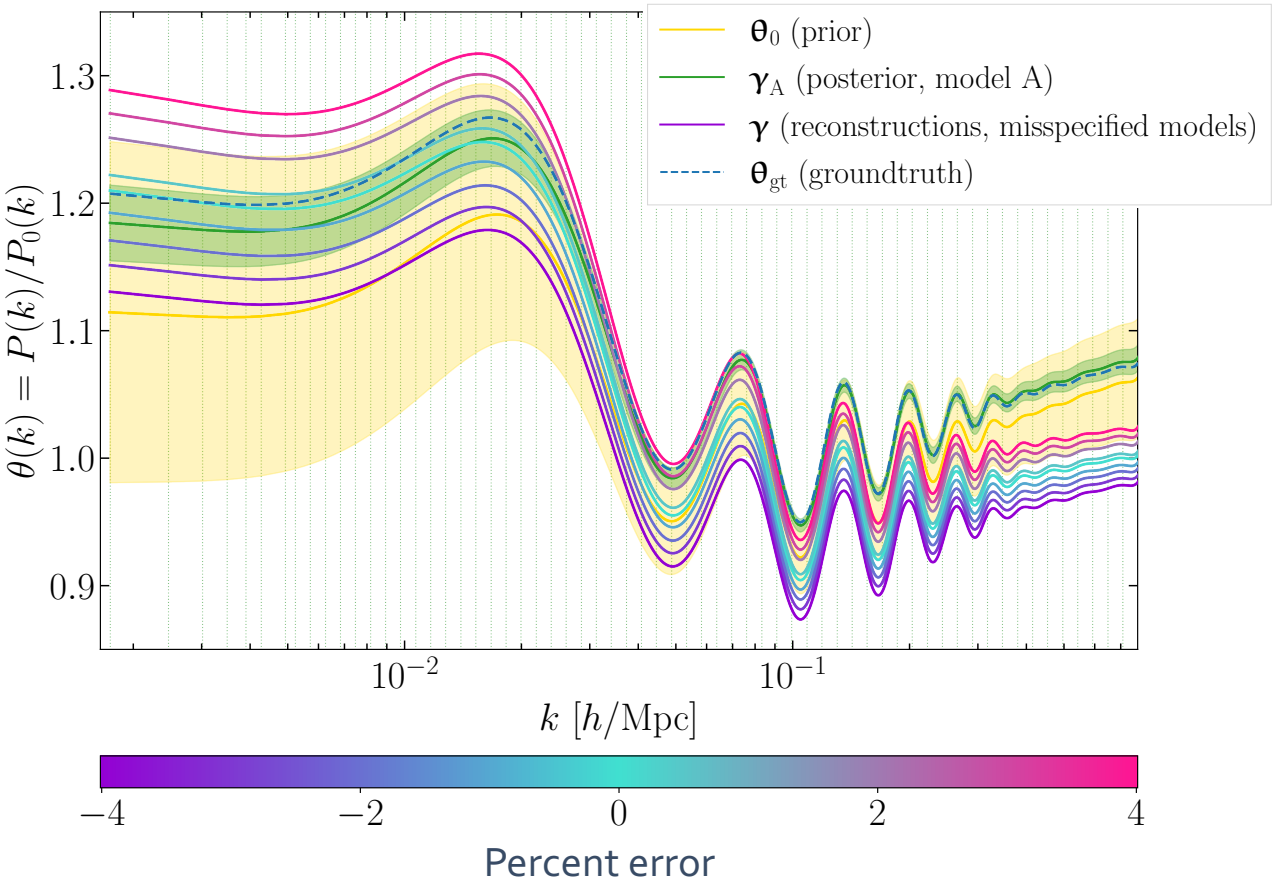
Impact of misspecified selection function variance



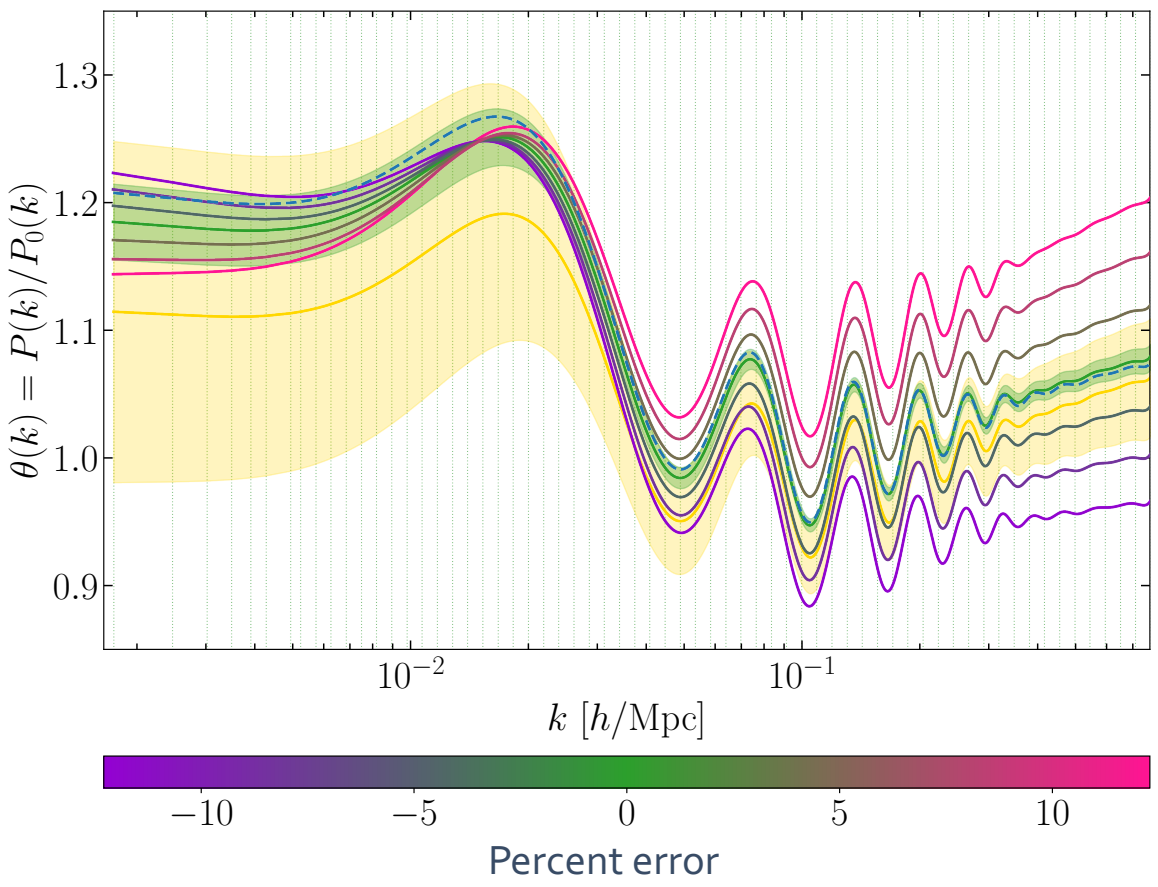
Hoellinger & Leclercq, in prep.

SELFI posterior for diagnosing systematic effects

Impact of misspecified selection function variances*



Impact of misspecified linear extinction

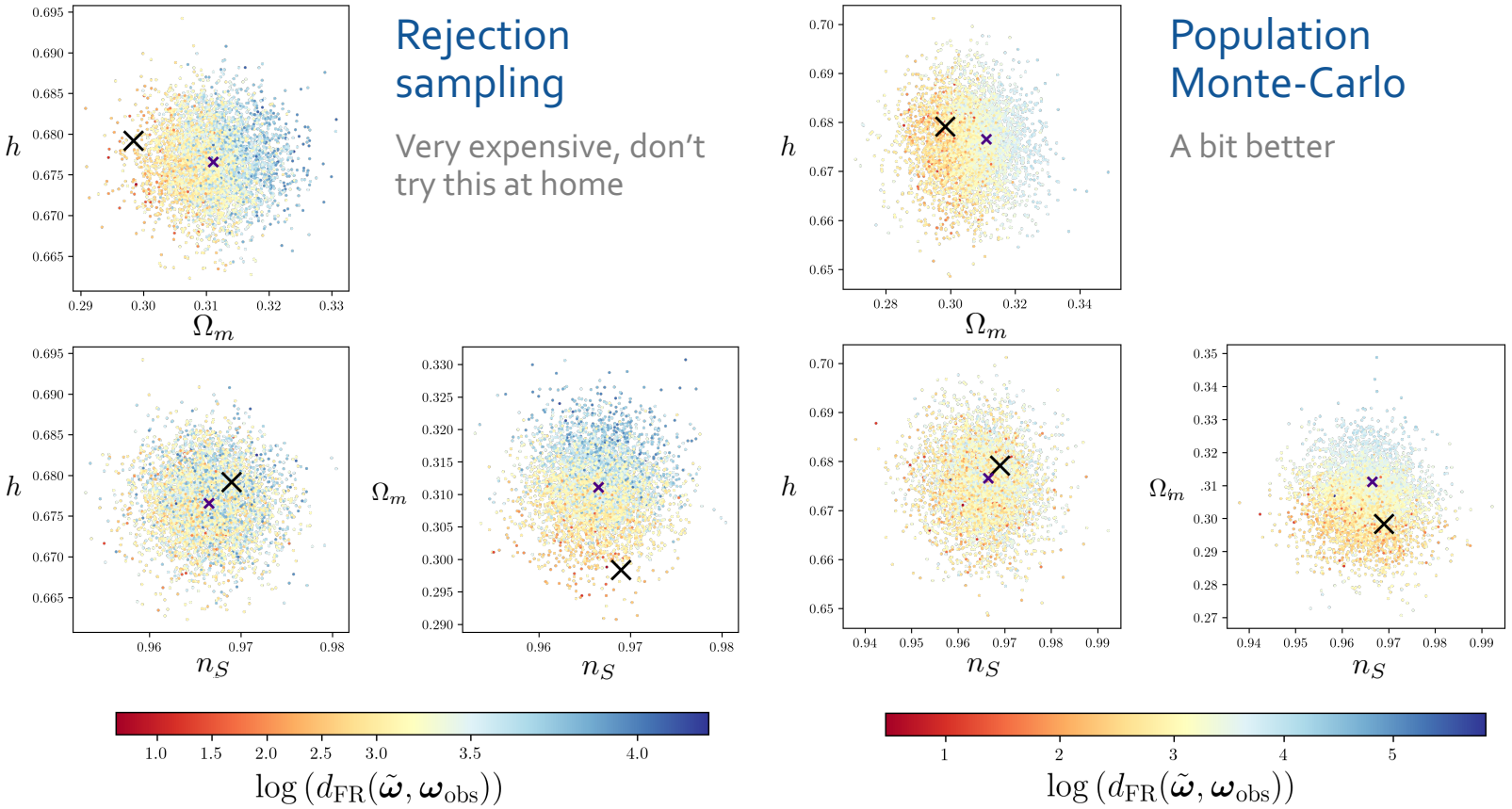


*under misspecified linear extinction

Hoellinger & Leclercq, in prep.

2 Infer the top-level cosmology $\boldsymbol{\omega}$

Any ILI method such as Approximate Bayesian Computation with any sampler

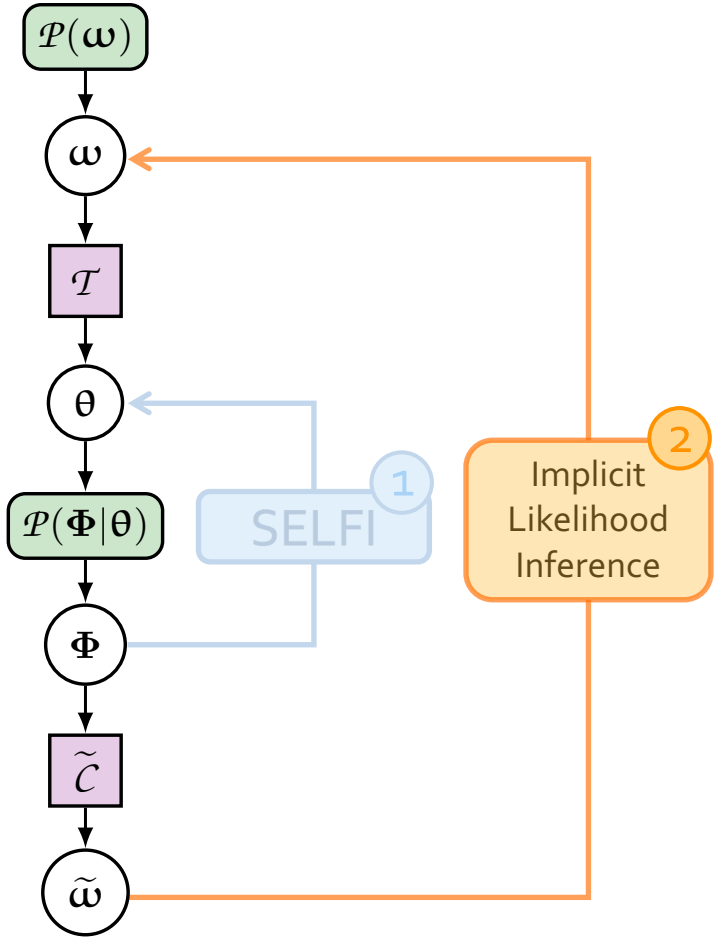


Rejection sampling

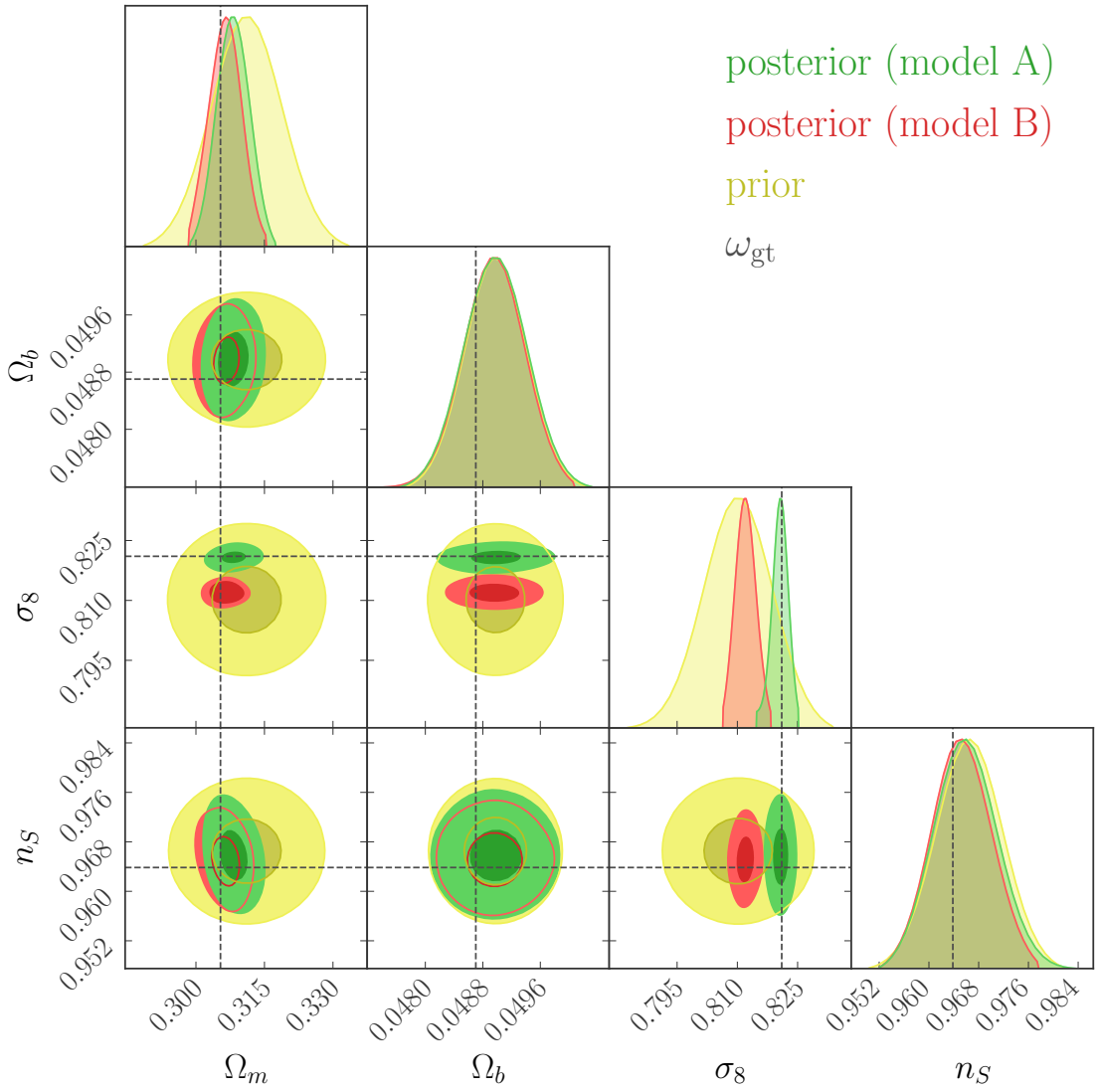
Very expensive, don't try this at home

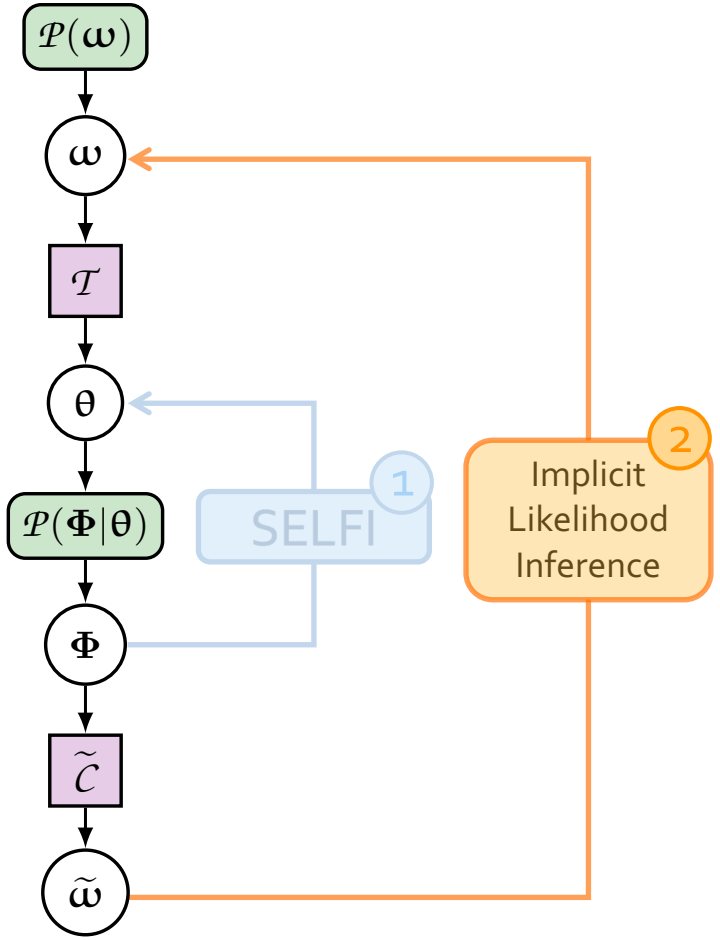
Population Monte-Carlo

A bit better



Implicit Likelihood Inference of cosmological parameters



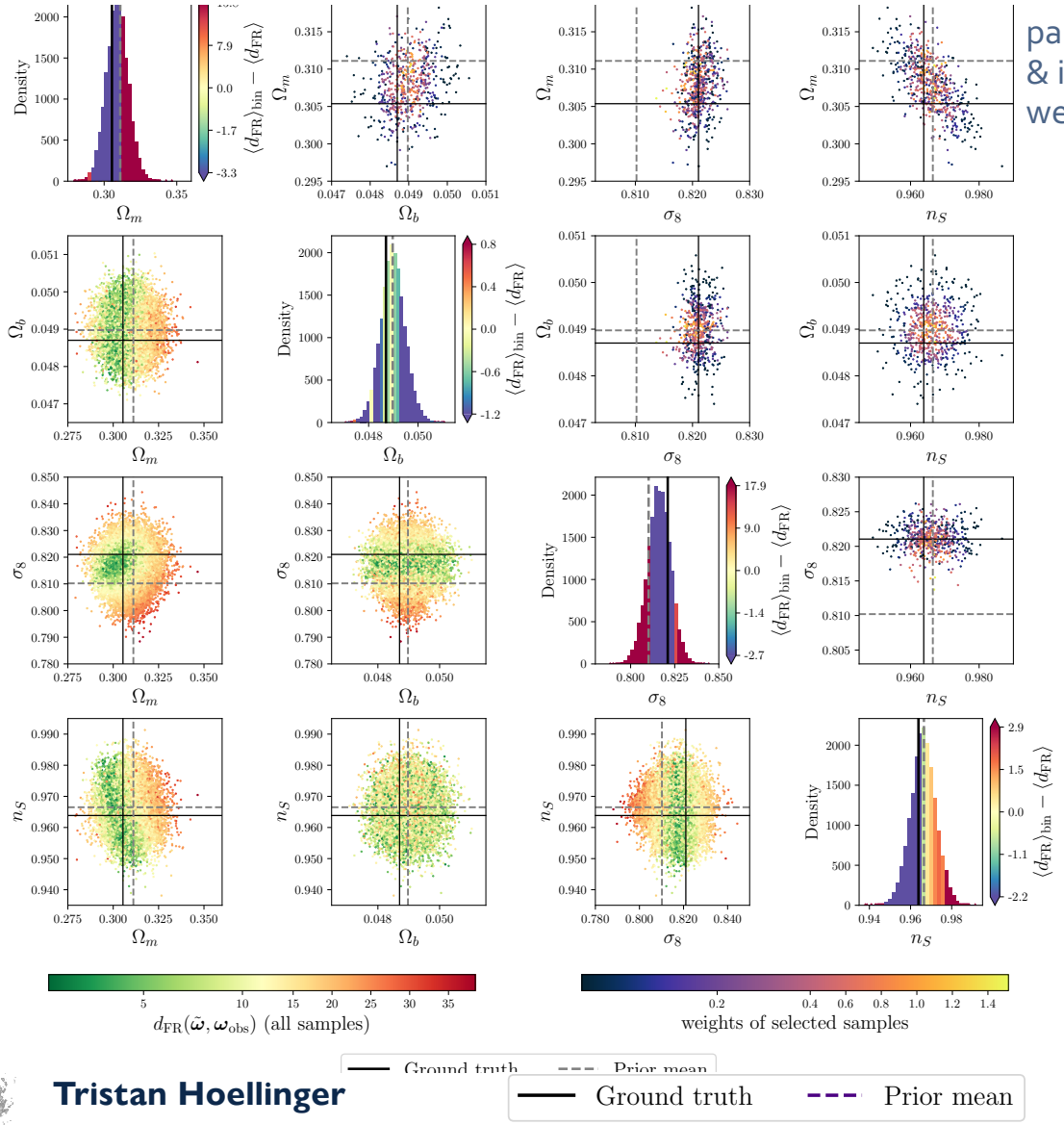


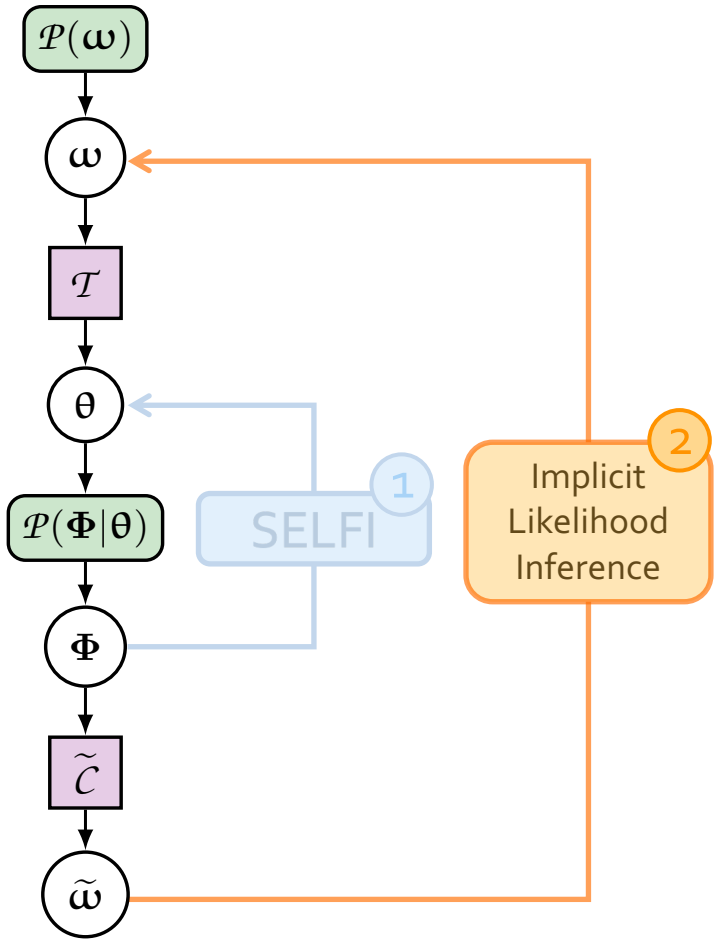
Step 2
ABC-PMC

18,052 N-body simulations for Model A
512 particles

Hoellinger & Leclercq, in prep.

ILI of cosmological parameters: Model A, well specified





Step 2
ABC-PMC

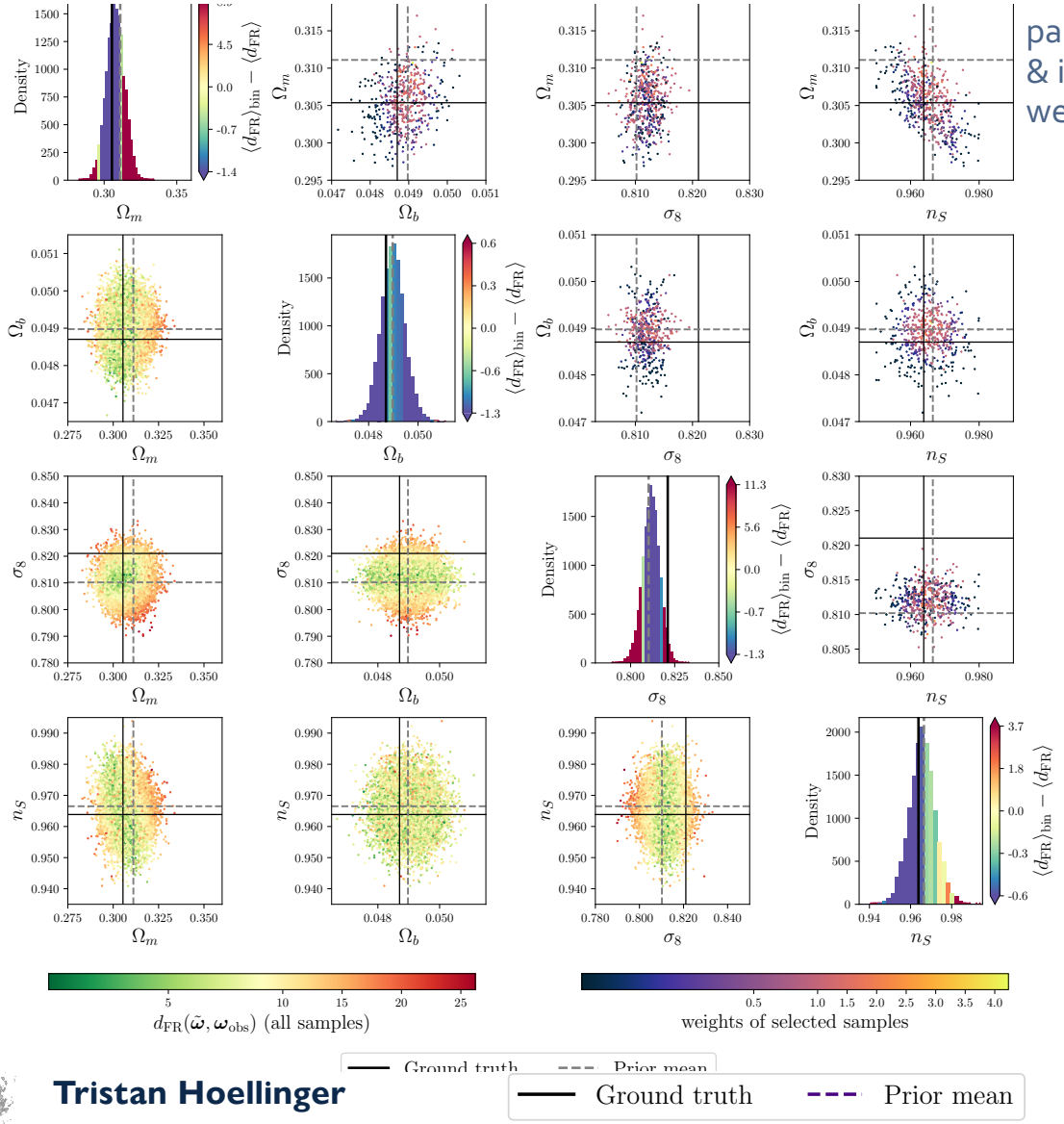
18,052 N-body
simulations for Model A
512 particles

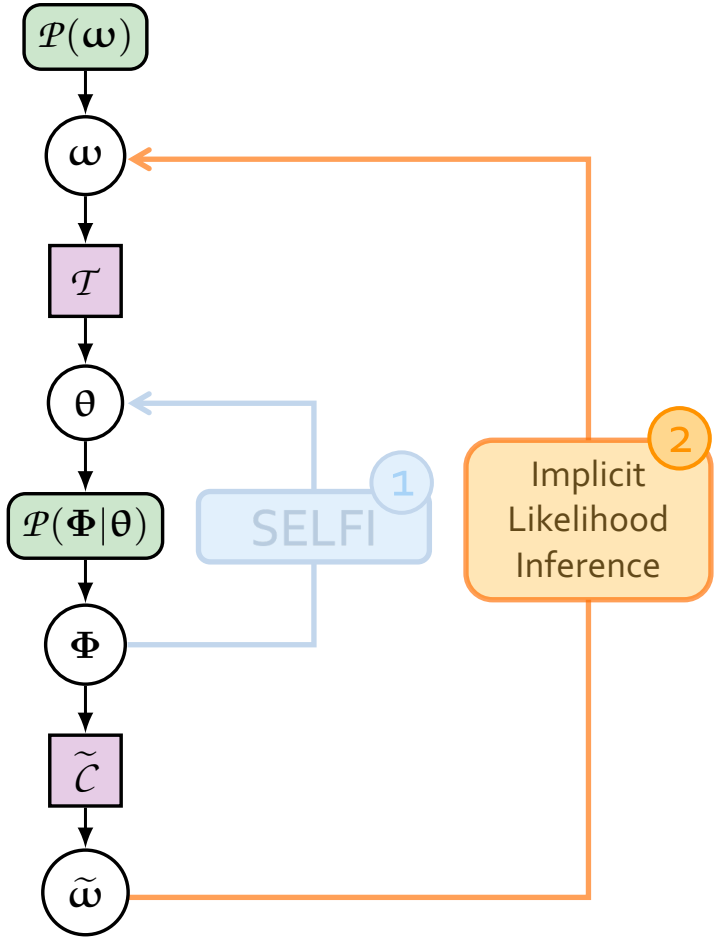
Hoellinger & Leclercq, in prep.

ILI of cosmological parameters: Model B, misspecified

particles
& importance
weights

samples
& Fisher





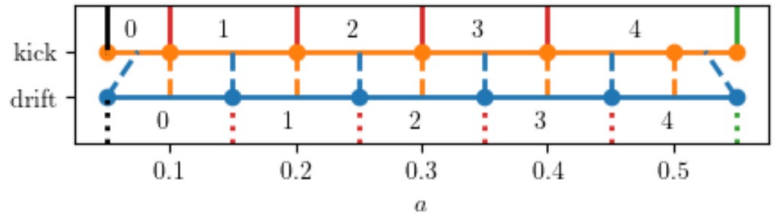
Step 2
ABC-PMC

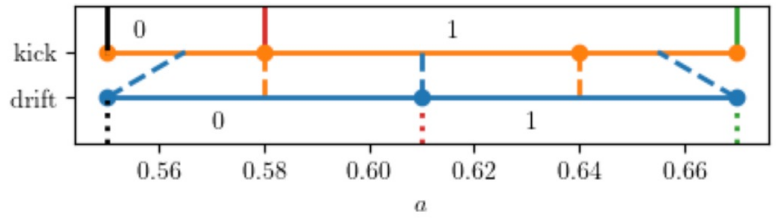
14,668 N-body
simulations for Model B
512 particles

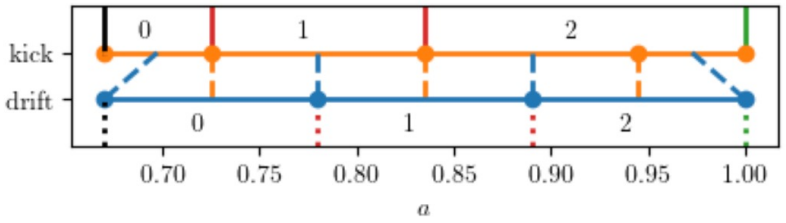
Hoellinger & Leclercq, in prep.

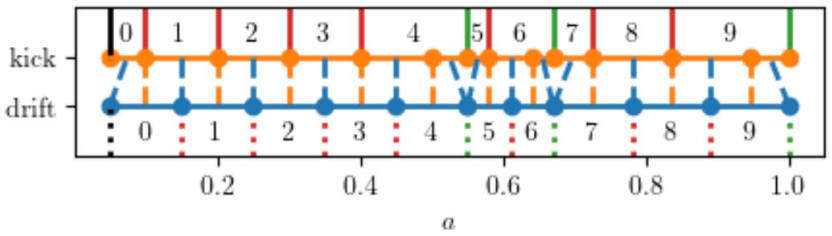
Gravitational evolution

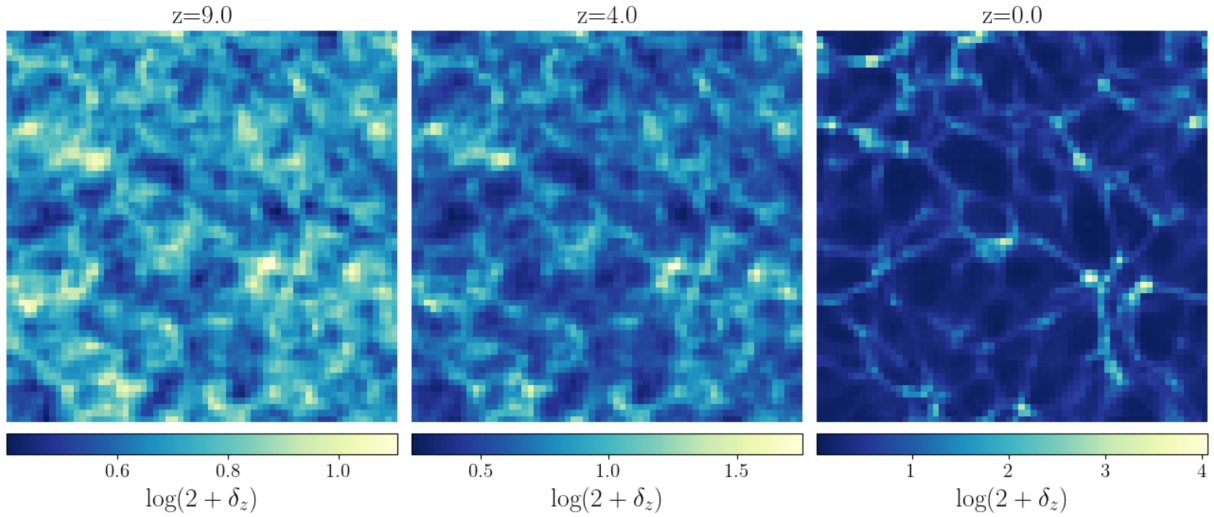
To approximate light cone effects we use 3 distinct shapshots at different redshifts, for the 3 galaxy populations (obtained with COLA using the Simbelmynë code). Leclercq, Jasche & Wandelt 2015, 1502.02690



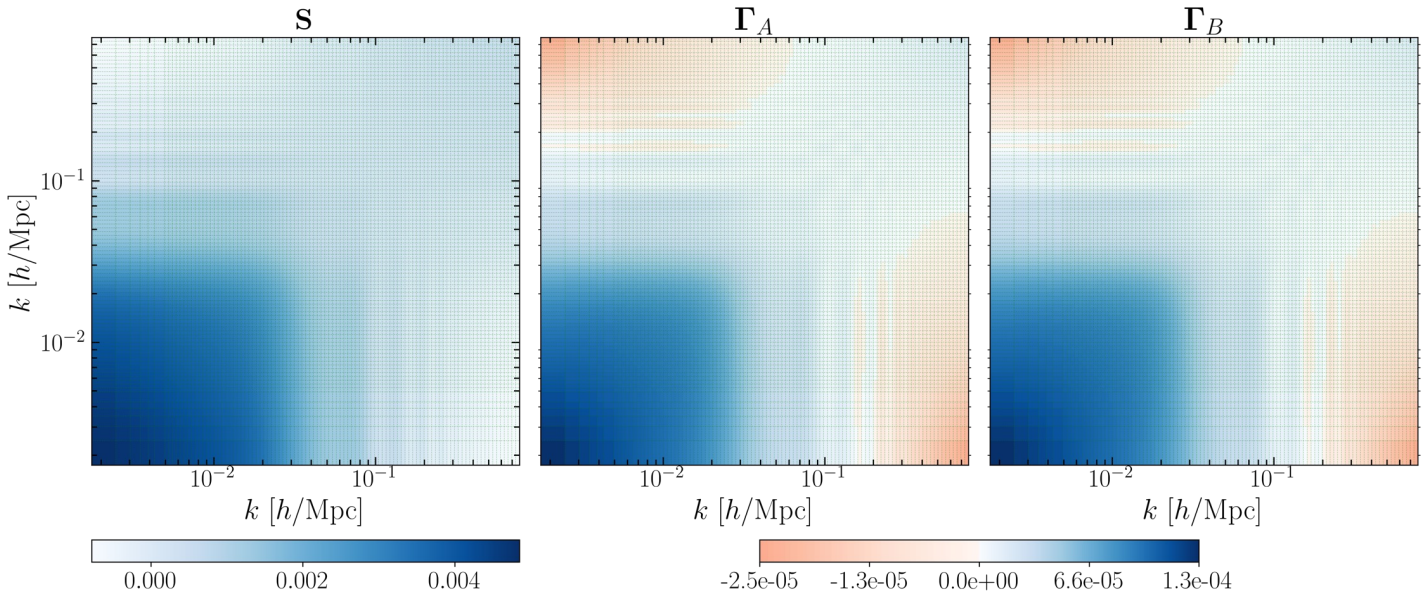


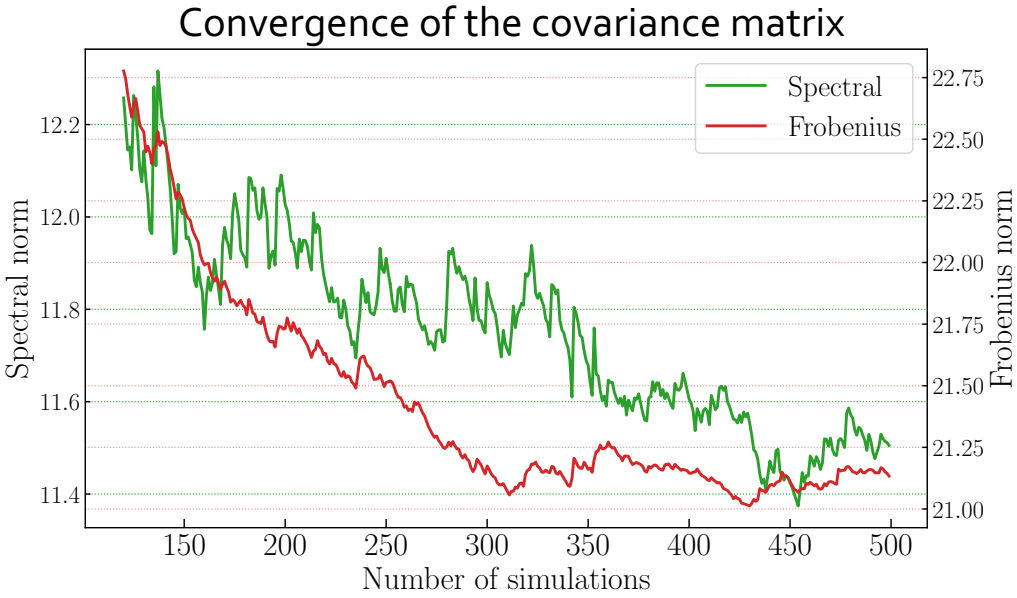




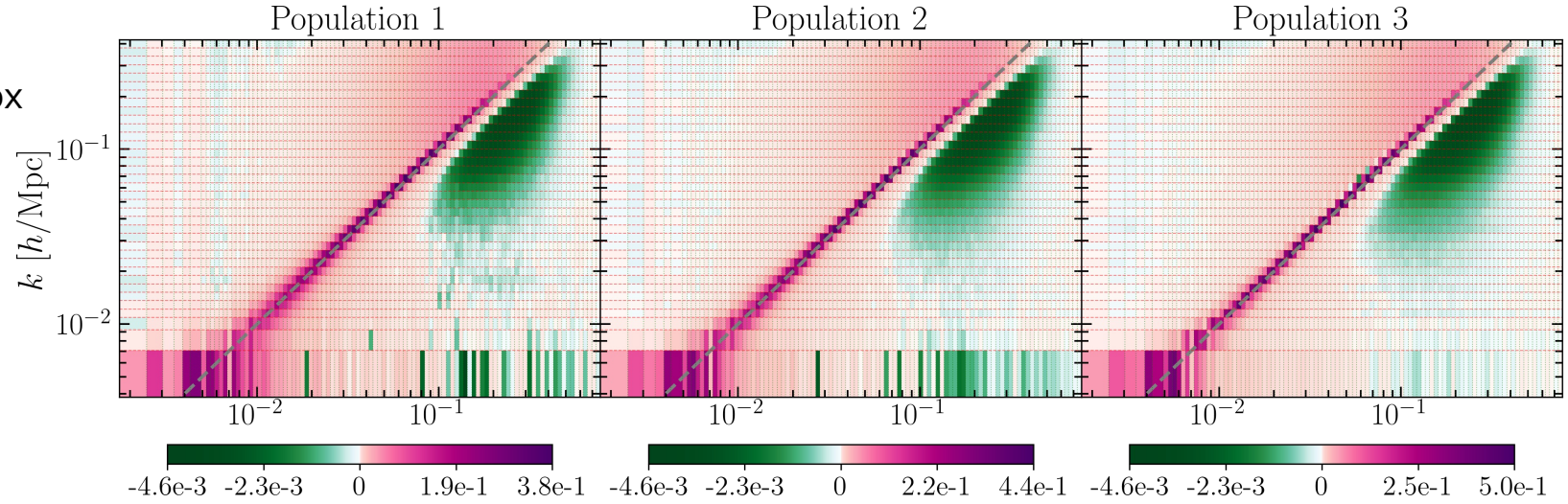


Diagnostic of SELFI





Gradients $\nabla \mathbf{f}_0$ of the blackbox at the expansion point



Prior agnostic to the BAOs and corresponding SELFI posterior