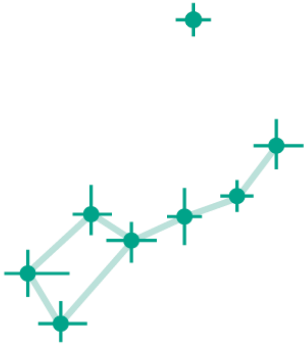
Primordial power spectrum and cosmology from black-box galaxy surveys

Prospects for Euclid

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London



Wolfgang Enzi, Alan Heavens,
Jens Jasche, Guilhem Lavaux,

and the Aquila Consortium
www.aquila-consortium.org

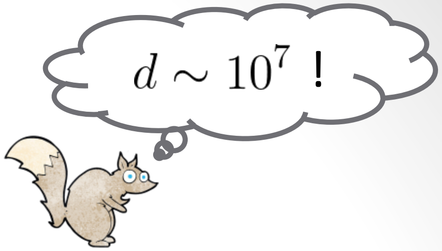
December 11th, 2019

ICIC
Imperial Centre
for Inference & Cosmology

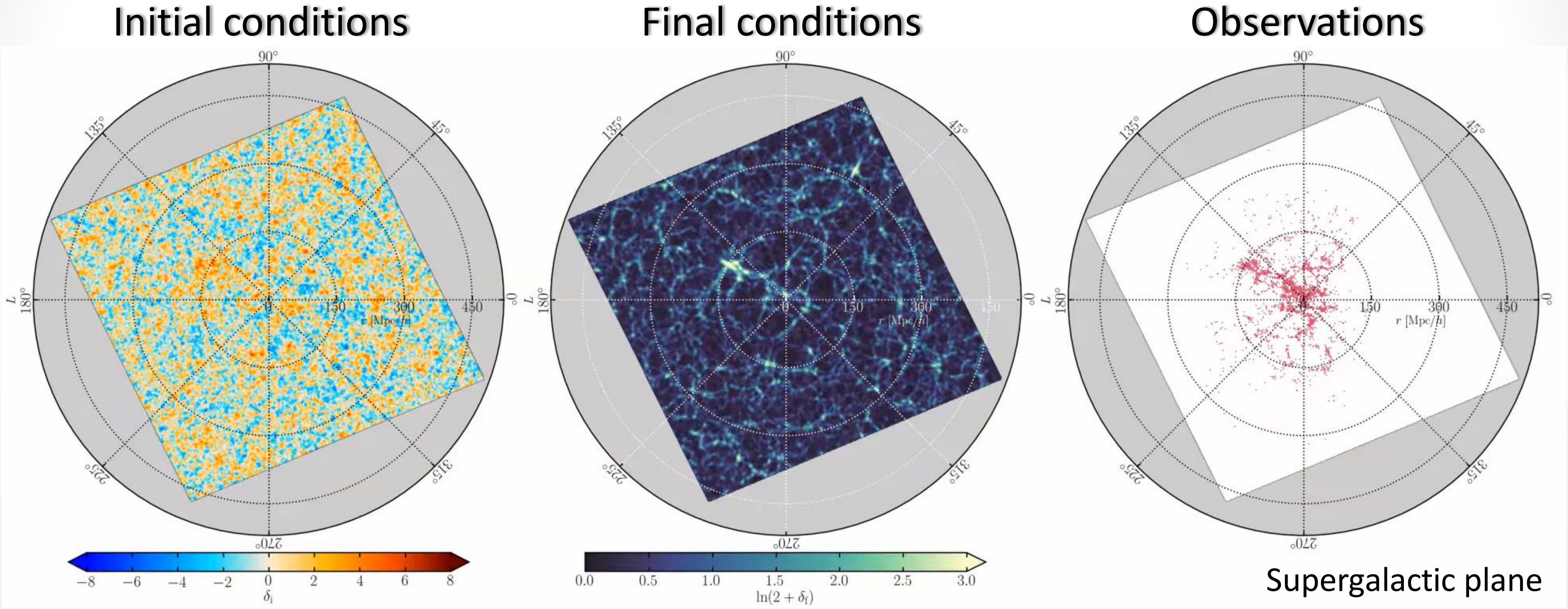
Imperial College
London

Vocabulary considerations I:

What is the likelihood?



In cosmology, the (true?) likelihood should live at the level of the map of the CMB or LSS. e.g. Wiener filtering for the CMB, BORG for the LSS (a 256^3-dimensional Poisson likelihood):


Initial conditions

Final conditions

Observations

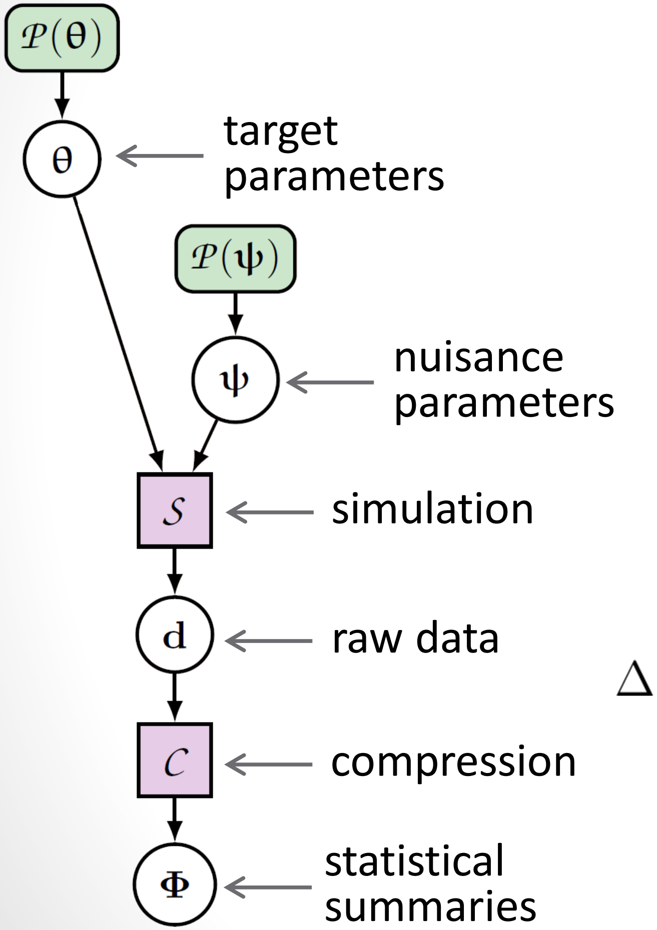
Supergalactic plane

Jasche & Lavaux 2019, 1806.11117 – FL, Lavaux & Jasche, in prep.

Expert knowledge of the likelihood is needed to beat the curse of dimensionality: conditionals/gradients of the likelihood are required by the samplers (Gibbs/Hamiltonian).

Vocabulary considerations II:

You may already be an LFI specialist!



$\mathcal{P}(\theta)$

θ target parameters

$\mathcal{P}(\psi)$

ψ nuisance parameters

$\mathcal{S}$ simulation

$\mathbf{d}$ raw data

$\mathcal{C}$ compression

$\mathbf{\Phi}$ statistical summaries

- Likelihood-free inference (LFI) techniques bypass the need for a map-level likelihood, by relying instead only on a “black-box”.
- The likelihood is replaced by a measure of the distance/discrepancy Δ between simulated and observed statistical summaries of the data.

e.g. $\mathbf{d} =$ full galaxy survey data
$\mathbf{\Phi} = \{\widehat{P}(k)\}$ estimated power spectrum
$\Delta =$ Mahalanobis distance with covariance matrix Σ

$$\Delta(\mathbf{\Phi}_\theta, \mathbf{\Phi}_\mathrm{O}) = \sqrt{\sum_{k,k'} \left[\widehat{P}_\theta(k) - \widehat{P}_\mathrm{O}(k)\right]^\intercal \Sigma^{-1}_{k,k'} \left[\widehat{P}_\theta(k') - \widehat{P}_\mathrm{O}(k')\right]}$$

Note that this is what many people would call…
(square root of -2 times) the log-likelihood!

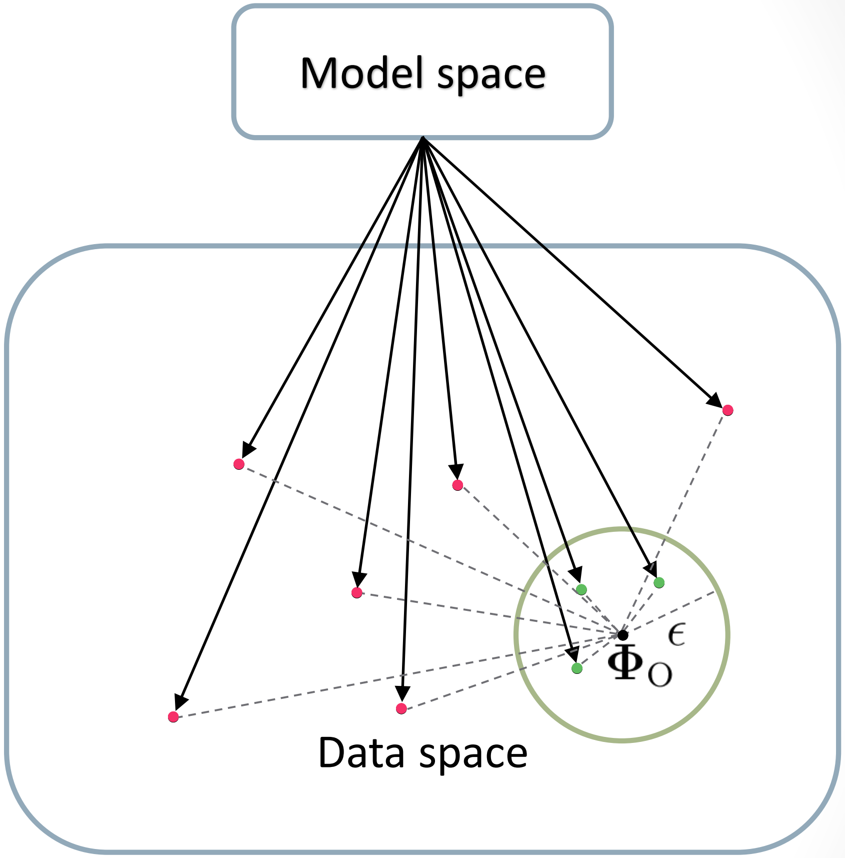
- What is “primordial” depends mostly on your ambition…

Cosmological synergies with LFI

- Advantages of likelihood-free inference:
  - No expert knowledge (conditionals/gradients of the likelihood) is required
  - Summary statistics need not be physically modelled and can be chosen robustly to model misspecification, e.g.:
    - Microwave sky: cross-power spectra between different frequency maps
    - Imaging surveys: cross-correlation between different bands
  - Joint and self-consistent analyses of correlated data sets is straightforward
- Drawbacks of likelihood-free inference:
  - No inference of the map
  - Relies on (lossy) data compression and statistical approximations

Likelihood-free rejection sampling (LFRS)

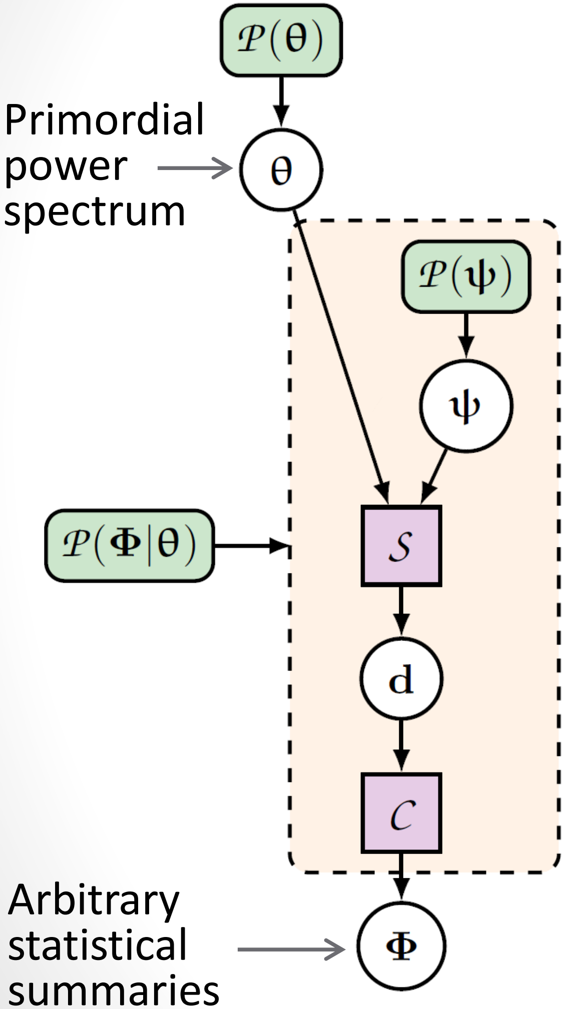
- Iterate many times:
  - Sample θ from a proposal distribution $q(\theta)$
  - Simulate Φ_θ using the black-box
  - Compute the distance $\Delta(\Phi_\theta, \Phi_\mathrm{O})$ between simulated and observed data
  - Retain θ if $\Delta(\Phi_\theta, \Phi_\mathrm{O}) \leq \epsilon$, otherwise reject



ϵ can be adaptively reduced (Population Monte Carlo)

Beyond LFRS: the SELFI approach

Simulator Expansion for Likelihood-Free Inference

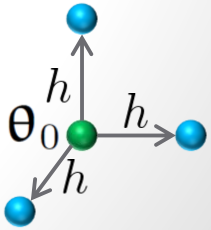


- We aim at inferring the primordial power spectrum, which contains (almost?) all of the information
- This requires doing LFI in $d = \mathcal{O}(100) - \mathcal{O}(1,000)$
- If we trust the results of earlier experiments, we can Taylor-expand the black-box around an expansion point θ_0:

$$\hat{\boldsymbol{\Phi}}_\theta \approx \mathbf{f}_0 + \nabla\mathbf{f}_0 \cdot (\theta - \theta_0) \boxed{+ \frac{1}{2}(\theta - \theta_0)^\intercal \cdot \mathbf{H} \cdot (\theta - \theta_0)} + \dots$$

SELFI-2 (second-order): coming soon!

- Gradients, Hessian matrix, etc. of the black-box can be evaluated via finite differences in parameter space



SELFI-1: linearization of the black-box

- Linearization of the black-box:

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

- Gaussian prior + Gaussian effective likelihood

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

expansion point

observed summaries

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

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

prior covariance

covariance of summaries

gradient of the black-box

$\mathbf{f}_0$, $\mathbf{C}_0$ and $\nabla \mathbf{f}_0$ can be evaluated through simulations only.

The number of required simulations is fixed *a priori* (contrary to MCMC).

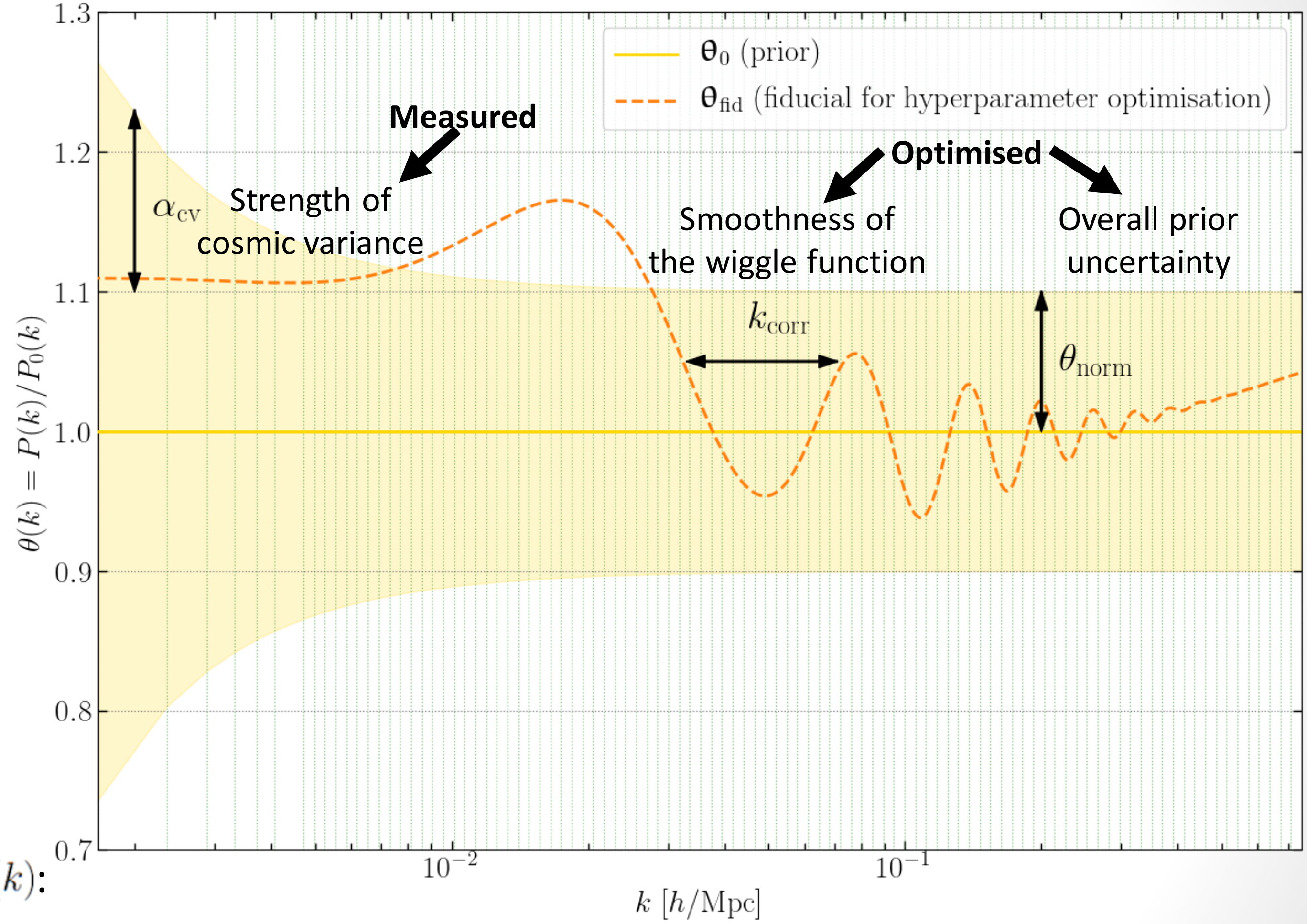
The workload is perfectly parallel.

FL, Enzi, Jasche & Heavens 2019, 1902.10149

A prior for the primordial power spectrum

Assumptions:

1. the power spectrum is Gaussian-distributed
2. it is strongly constrained to live close to P_0,
3. it is a smooth function of wavenumber,
4. and the power spectrum P_0 is subject to cosmic variance



Measured

Strength of cosmic variance

Optimised

Smoothness of the wiggle function

Overall prior uncertainty

➡ Prior for $\theta(k) = P(k)/P_0(k)$:

Mean: $\boldsymbol{\theta}_0 = \mathbf{1}_{\mathbb{R}^S}$ (without baryon acoustic oscillations wiggles)

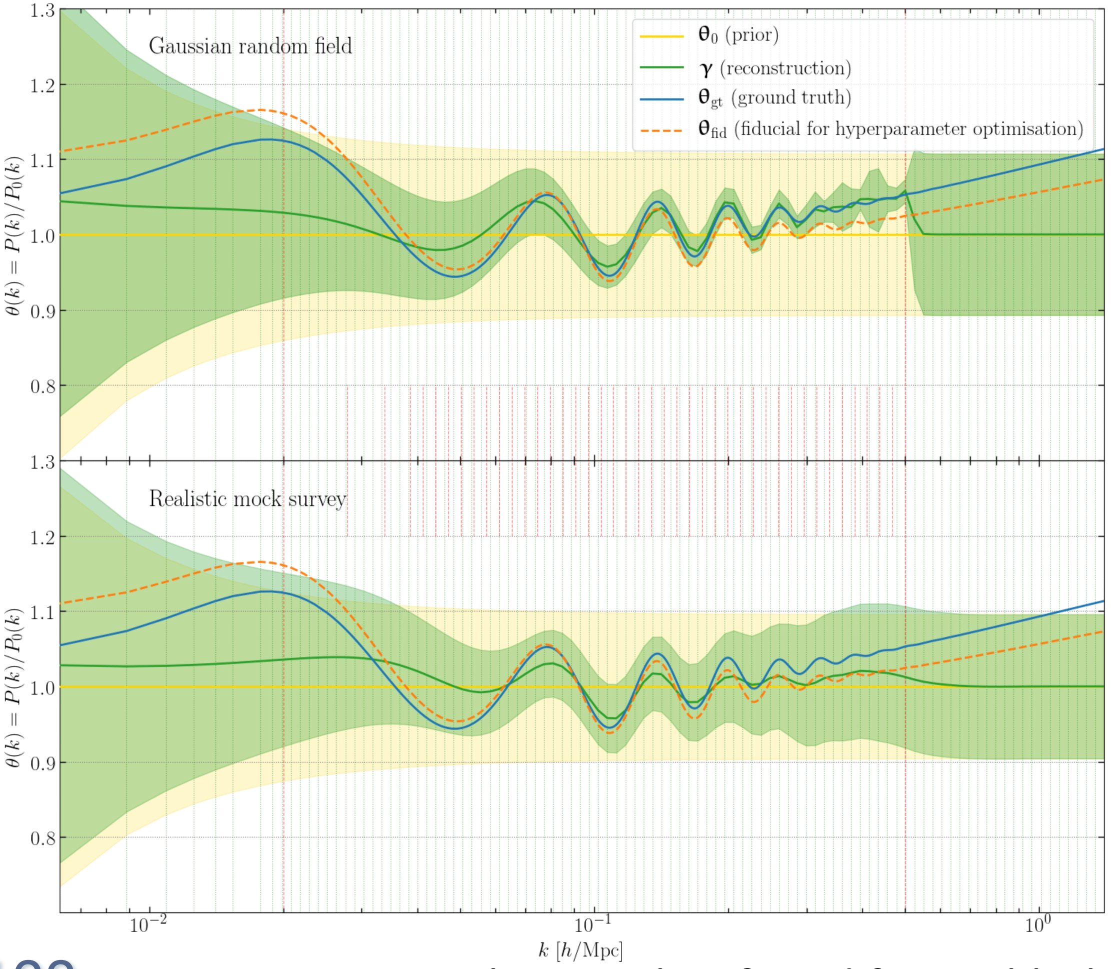
Covariance: $\mathbf{S} \equiv \theta^2_\mathrm{norm}\, \mathbf{u}\mathbf{u}^\intercal \circ \mathbf{K}$ $\quad (\mathbf{K})_{ss'} \equiv \exp\left[-\frac{1}{2}\left(\frac{k_s - k_{s'}}{k_\mathrm{corr}}\right)^2\right]$ $\quad (\mathbf{u})_s \equiv 1 + \sigma_s = 1 + \frac{\alpha_\mathrm{cv}}{k_s^{3/2}}$

FL, Enzi, Jasche & Heavens 2019, 1902.10149

SELFI + numerical model: Proof-of-concept

Gaussian random field

Realistic mock survey with Simbelmynë



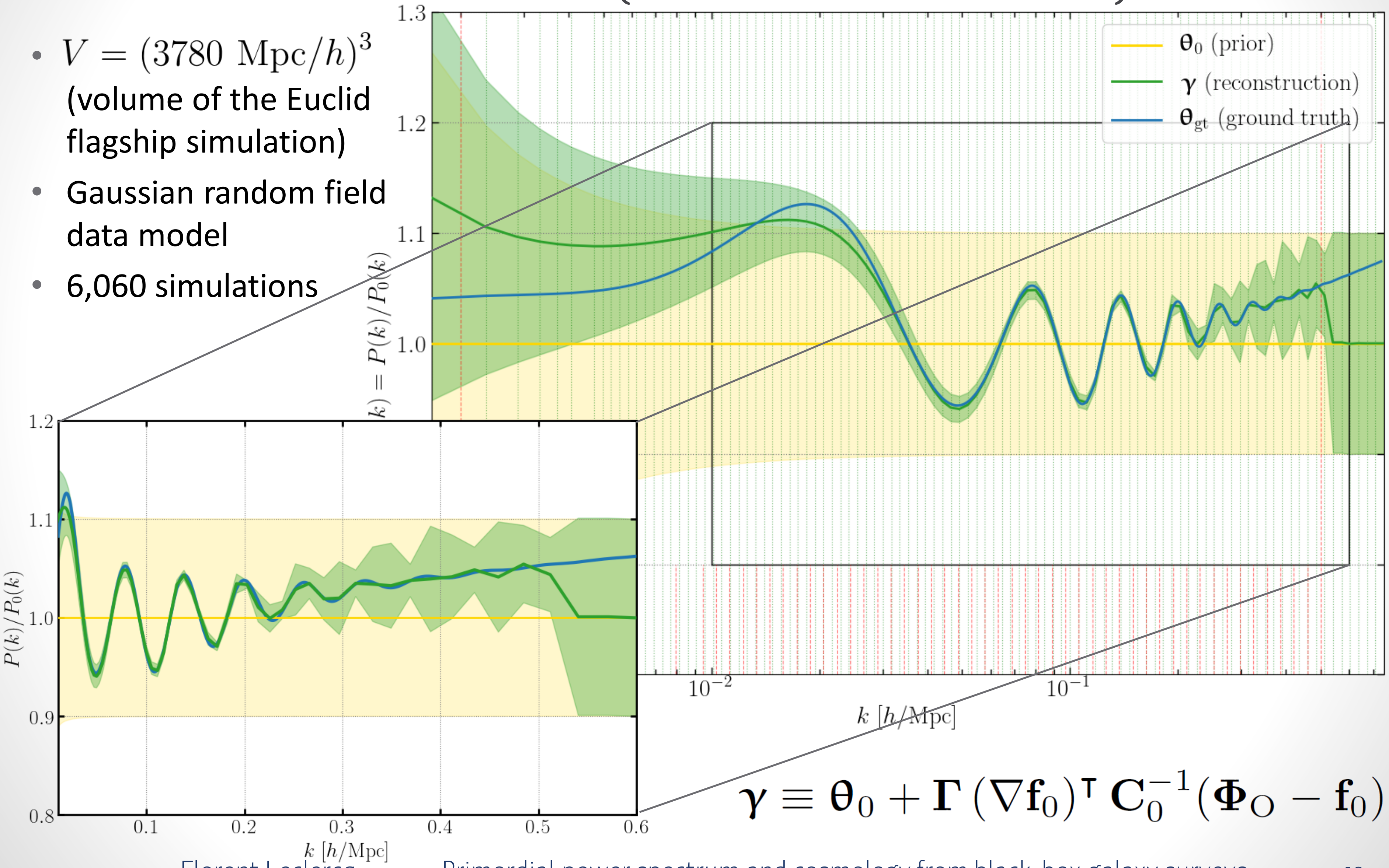
100 parameters are simultaneously inferred from a black-box data model

1 (Gpc/h)^3 only! Much more potential for upcoming data…

FL, Enzi, Jasche & Heavens 2019, 1902.10149

SELFI-1 Euclid forecast (cosmic variance limit)

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

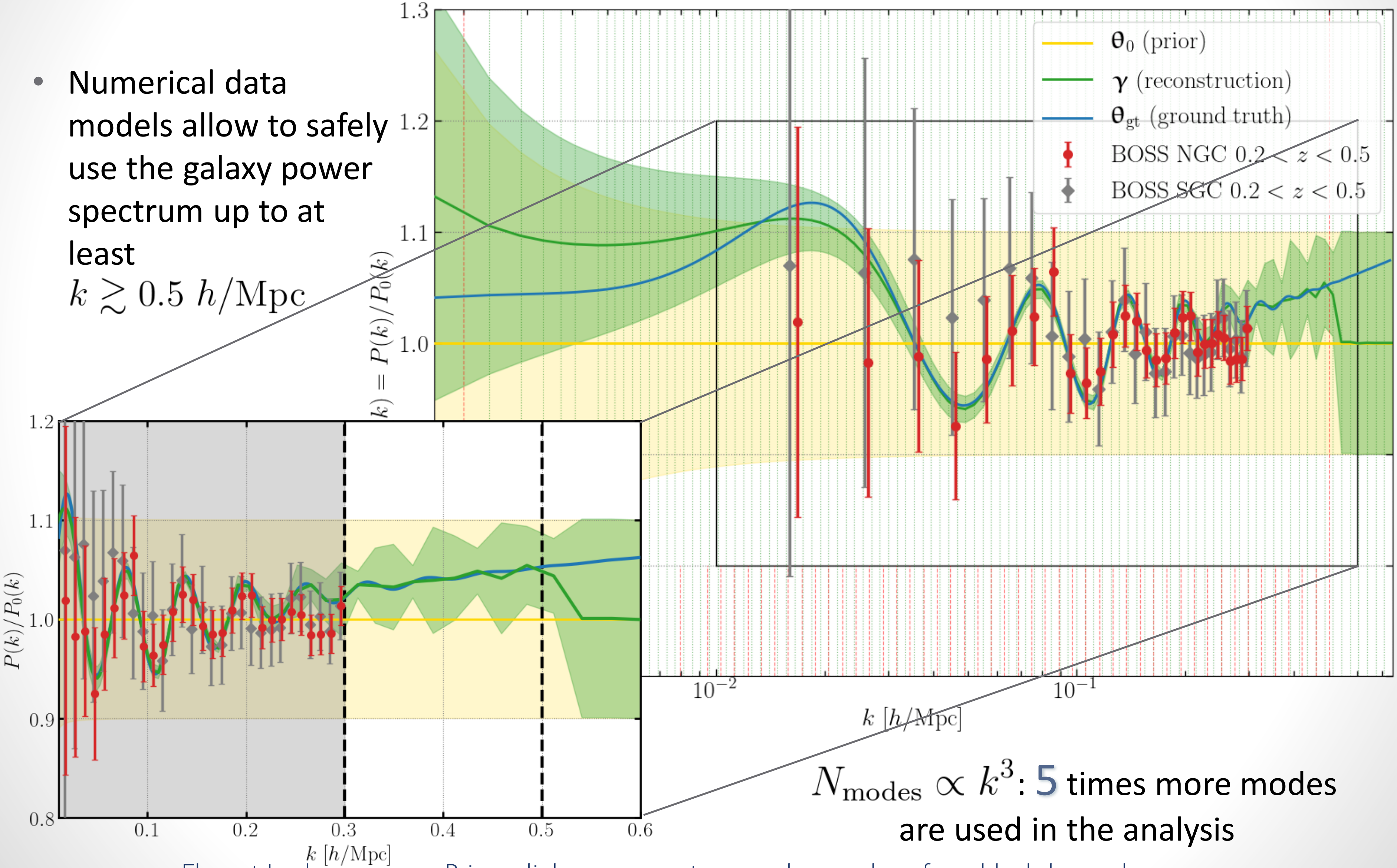


$$\boldsymbol{\gamma} \equiv \boldsymbol{\theta}_0 + \boldsymbol{\Gamma} \left(\nabla \mathbf{f}_0\right)^\intercal \mathbf{C}_0^{-1} (\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0)$$

SELFI-1 Euclid versus BOSS

Data points from
Beutler *et al.* 2016, 1607.03149

- Numerical data models allow to safely use the galaxy power spectrum up to at least
  $k \gtrsim 0.5\ h/\mathrm{Mpc}$



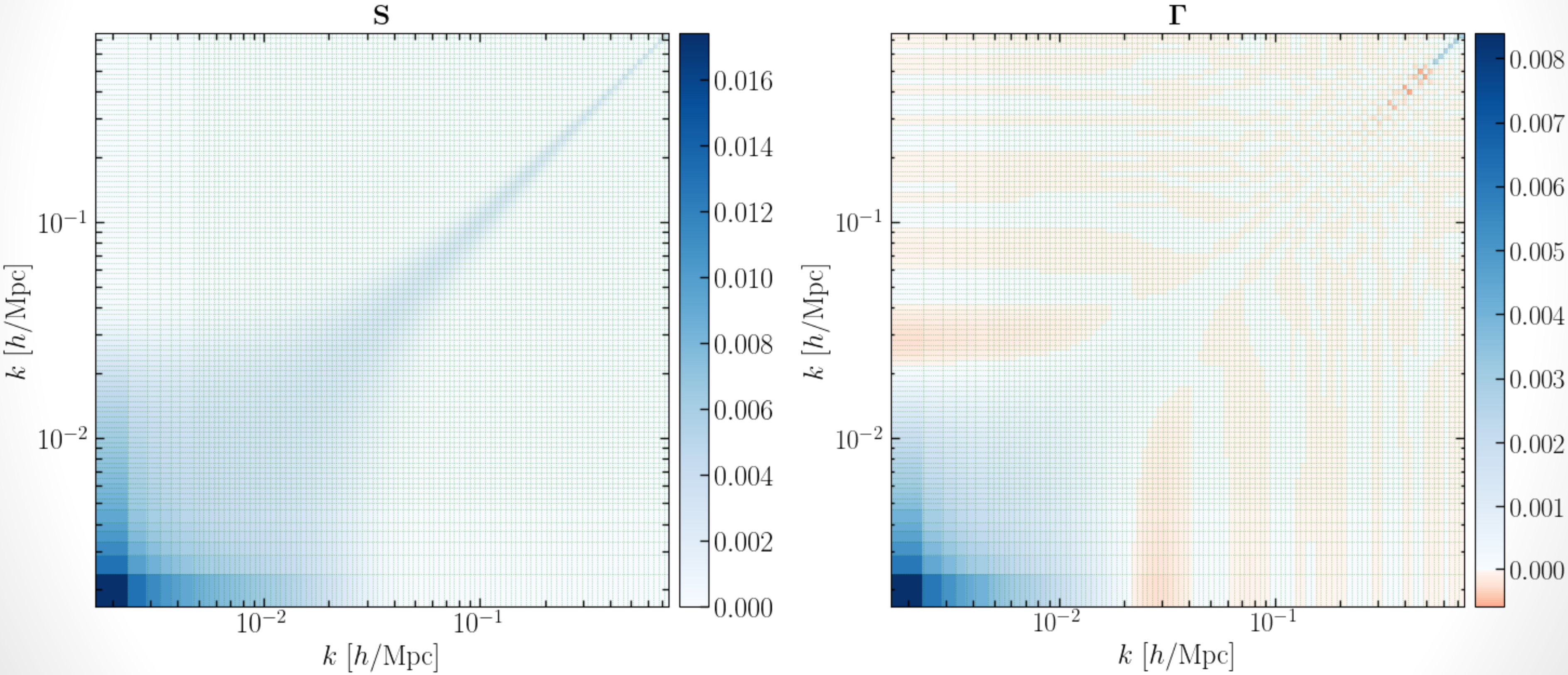
$N_{\mathrm{modes}} \propto k^3$: 5 times more modes are used in the analysis

Uncertainty quantification

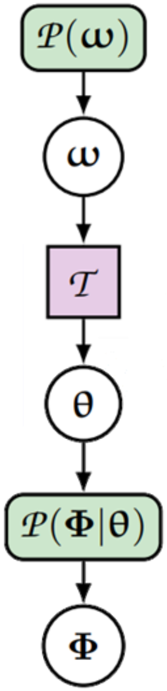
$$\mathbf{\Gamma} \equiv \left[(\nabla \mathbf{f}_0)^\intercal \, \mathbf{C}_0^{-1} \nabla \mathbf{f}_0 + \mathbf{S}^{-1} \right]^{-1}$$

Prior covariance matrix

Posterior covariance matrix



From primordial power spectrum to cosmology



Cosmological parameters → ω

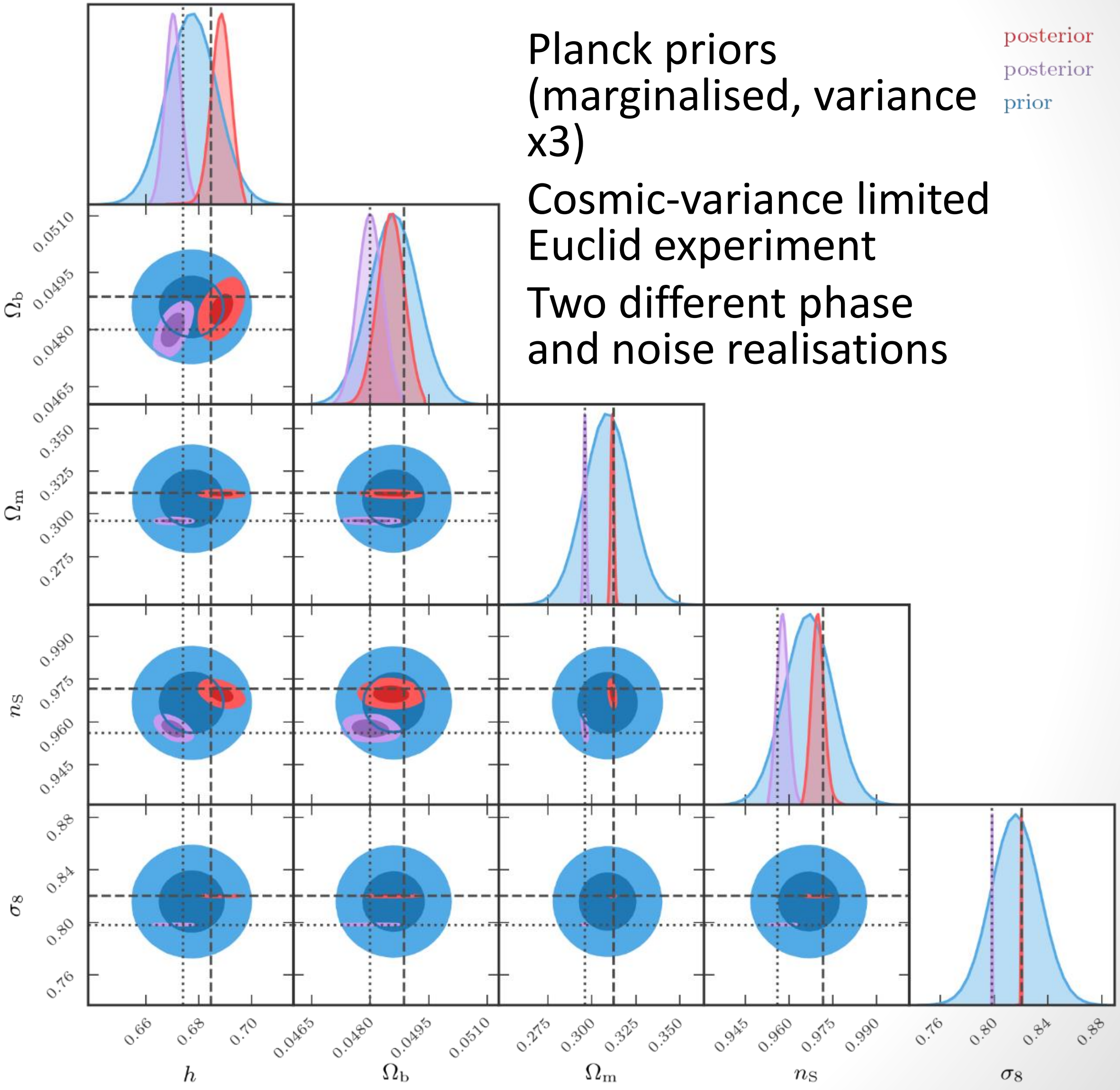
Primordial power spectrum → θ

Taylor-expanded black-box → $\mathcal{P}(\Phi|\theta)$

Arbitrary statistical summaries → Φ

- Robust inference of cosmological parameters can be easily performed *a posteriori* once the linearized data model is learnt

FL, Enzi, Jasche & Heavens 2019, 1902.10149



Planck priors (marginalised, variance x3)

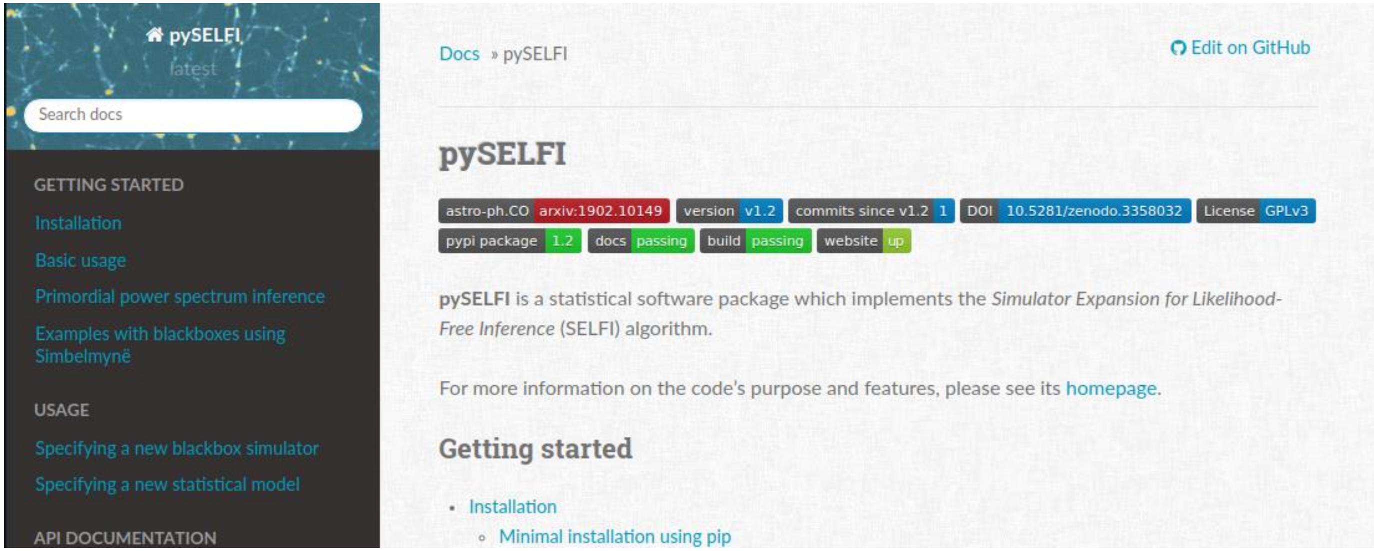
Cosmic-variance limited Euclid experiment

Two different phase and noise realisations

pySELFI is publicly available

- Code homepage: http://pyselfi.florent-leclercq.eu/
- Source on GitHub: https://github.com/florent-leclercq/pyselfi/
- Documentation on ReadtheDocs: https://pyselfi.readthedocs.io/en/latest/

(with templates to use your on black-box)



GETTING STARTED

- Installation
- Basic usage
- Primordial power spectrum inference
- Examples with blackboxes using Simbelmynë

USAGE

- Specifying a new blackbox simulator
- Specifying a new statistical model

API DOCUMENTATION

pySELFI

pySELFI is a statistical software package which implements the *Simulator Expansion for Likelihood-Free Inference* (SELFI) algorithm.

For more information on the code's purpose and features, please see its homepage.

Getting started

- Installation
  - Minimal installation using pip

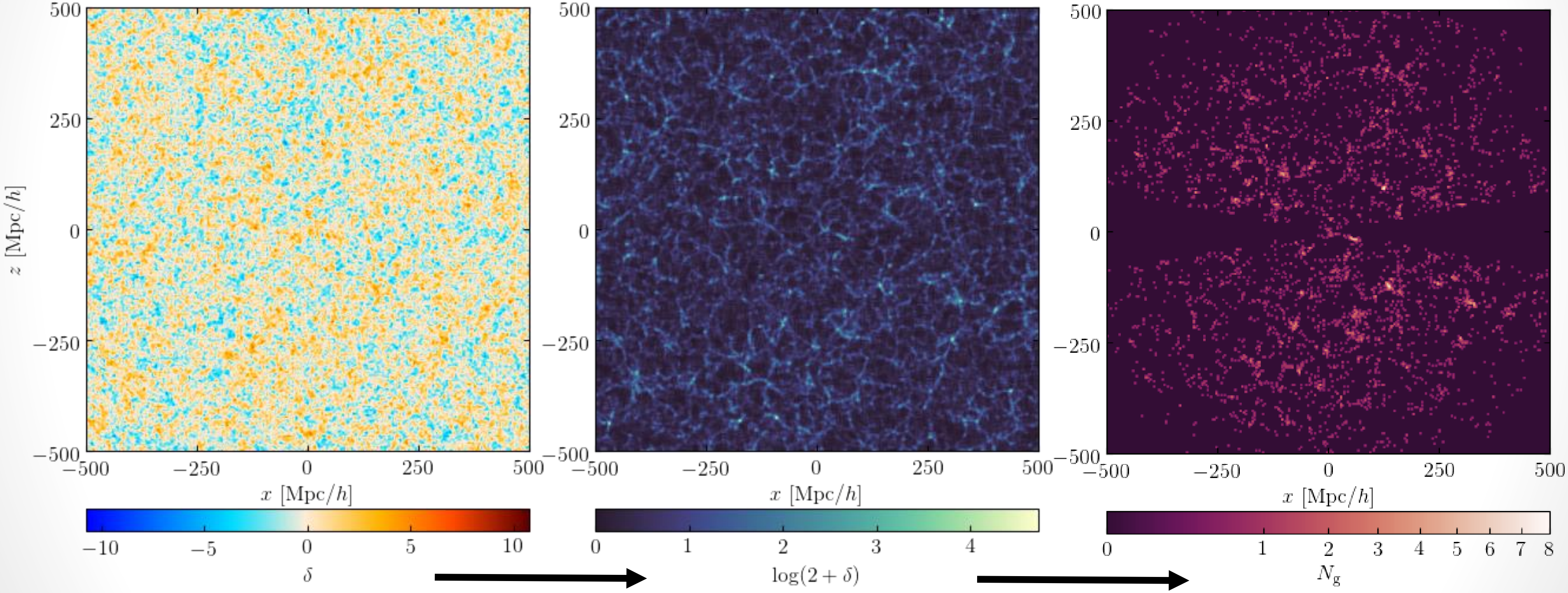
```
pip install pyselfi
```

A black-box: Simbelmynë

I'm happy to explain the name during the coffee break...

Publicly available code:

https://bitbucket.org/florent-leclercq/simbelmyne/



Initial conditions

Final conditions (dark matter)

Galaxies

Dark matter simulation with COLA

Tassev, Zaldarriaga & Eisenstein 2013, 1301.0322

Survey simulation: Redshift-space distortions, galaxy bias, selection effects, survey geometry, instrumental noise

tCOLA: Comoving Lagrangian Acceleration (temporal domain)

- Write the displacement vector as: $\mathbf{s} = \mathbf{s}_{\mathrm{LPT}} + \mathbf{s}_{\mathrm{MC}}$

Tassev & Zaldarriaga 2012, arXiv:1203.5785

- Time-stepping (omitted constants and Hubble expansion):

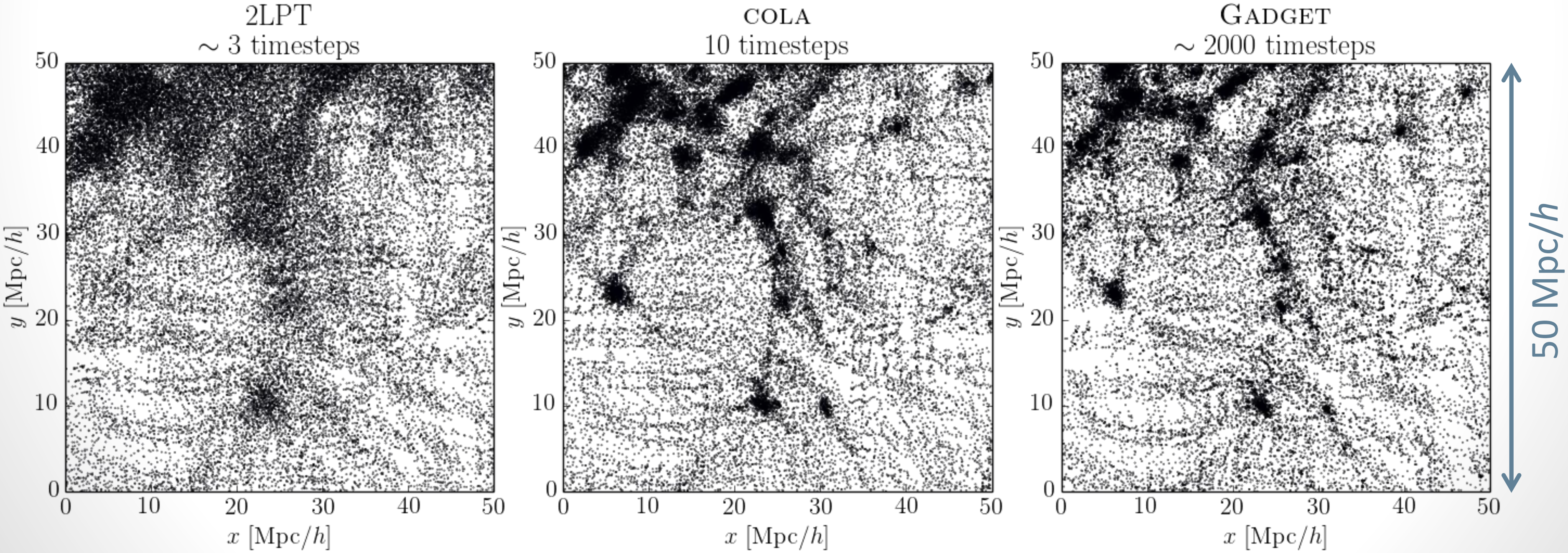
Standard:

$$\partial_\tau^2 \mathbf{s} = -\nabla\Phi$$

⇒

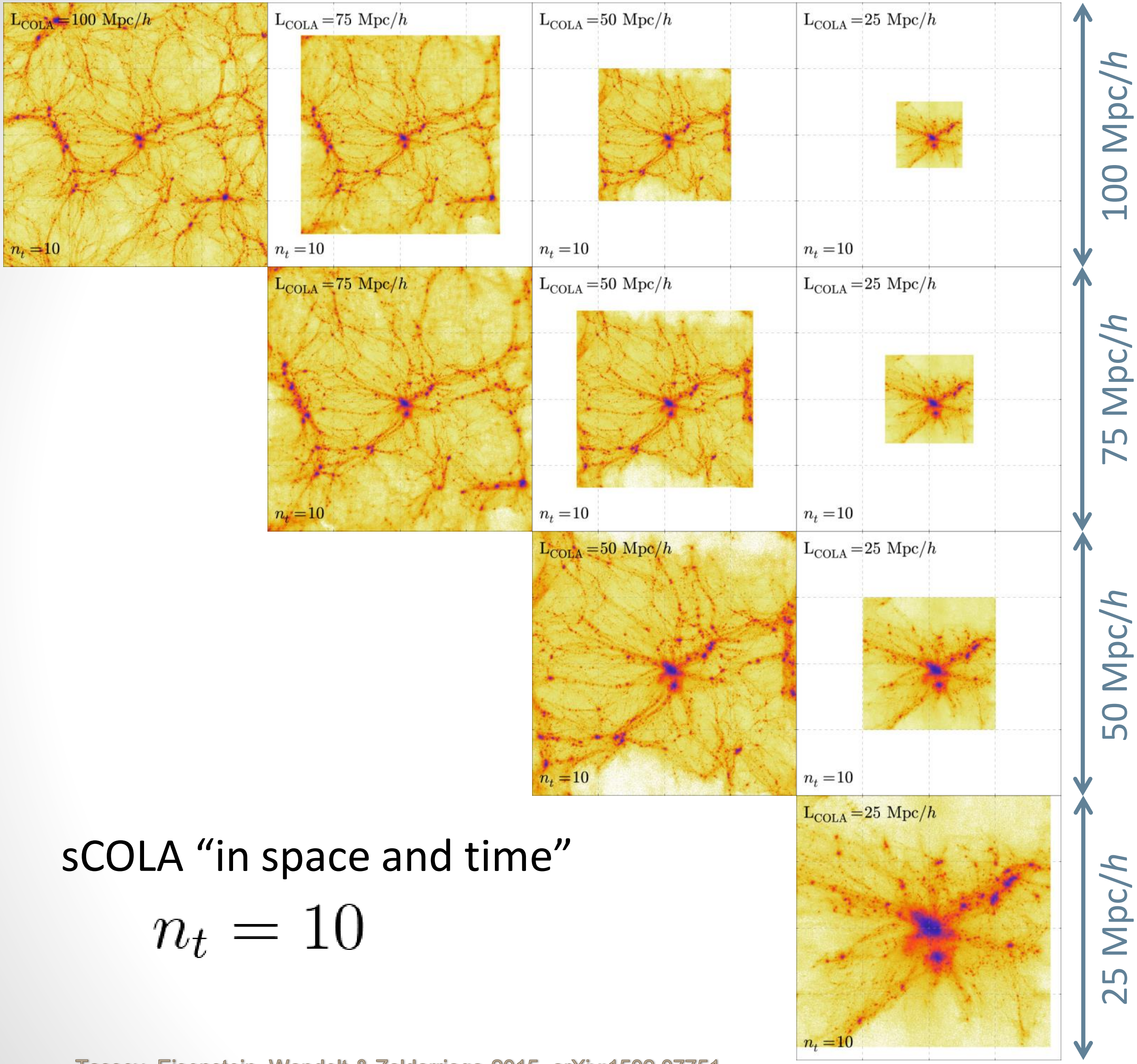
Modified:

$$\partial_\tau^2 \mathbf{s}_{\mathrm{MC}} = \partial_\tau^2(\mathbf{s} - \mathbf{s}_{\mathrm{LPT}}) = -\nabla\Phi - \partial_\tau^2 \mathbf{s}_{\mathrm{LPT}}$$



Tassev, Zaldarriaga & Einsenstein 2013, arXiv:1301.0322

sCOLA: Extension to the spatial domain

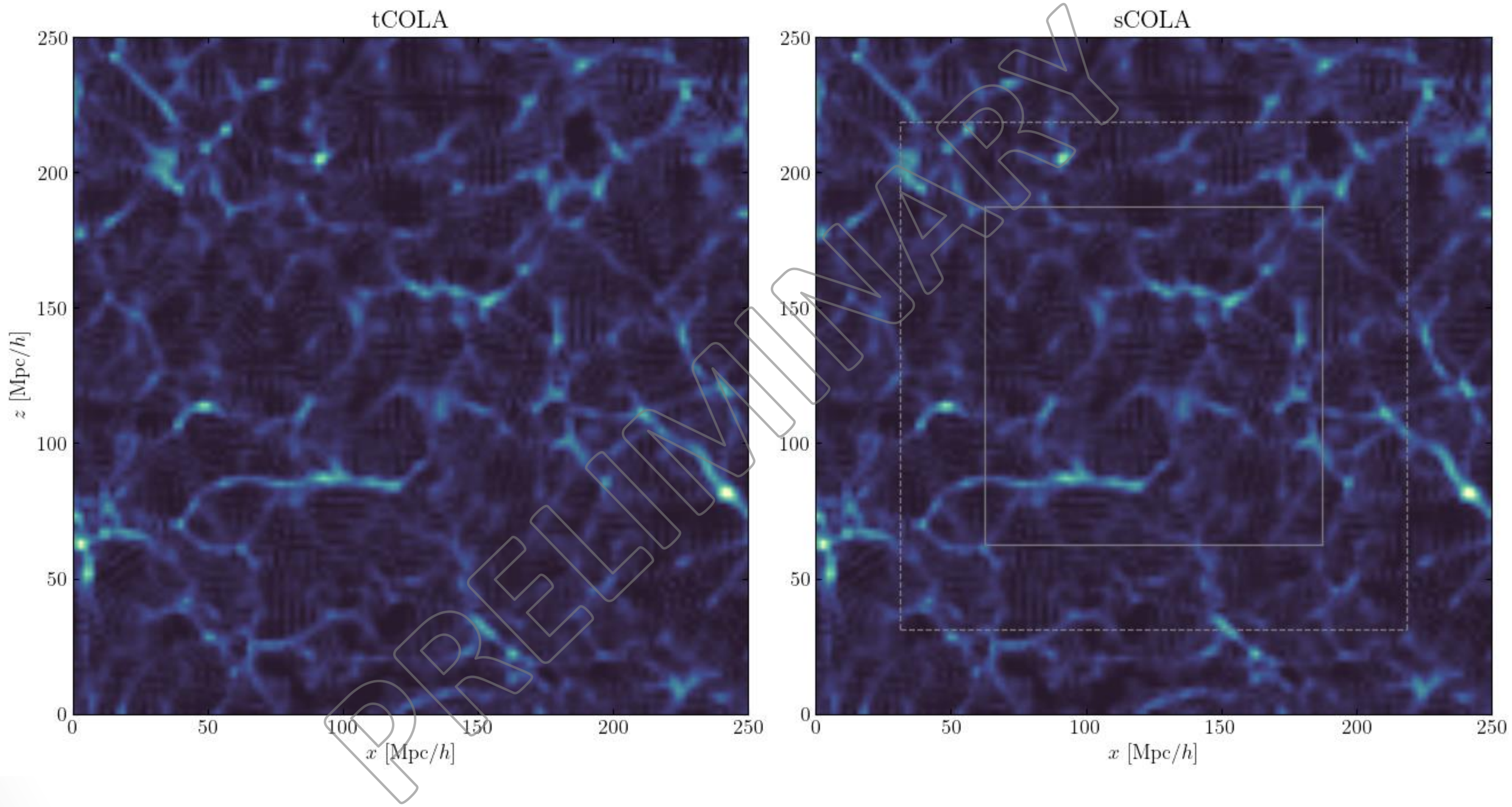


sCOLA “in space and time”

$$n_t = 10$$

Tassev, Eisenstein, Wandelt & Zaldarriaga 2015, arXiv:1502.07751

Perfectly parallel simulations with sCOLA tiling



FL *et al.*, in prep. (with B. Faure, G. Lavaux, A. Heavens, A. Jaffe, W. Percival, B. Wandelt, M. Zaldarriaga)

Concluding thoughts

- Goal: developing an algorithm for targeted questions, allowing the use of simulators including all relevant physical and observational effects.
- Bayesian analyses of galaxy surveys with fully non-linear numerical black-box models is not an impossible task!
- Likelihood-free inference is an easy way to account for cosmological synergies.
- The "number of parameters route" beyond likelihood-free rejection sampling (SELFI):
  - High-dimensional likelihood-free problems can be addressed.
  - The computational workload is fixed *a priori* and perfectly parallel.
- SELFI allows inference of the primordial power spectrum and cosmological parameters.