Forward modelling the large-scale structure:
perfectly parallel simulations and simulation-based inference

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

Wolfgang Enzi, Baptiste Faure, Alan Heavens,
Andrew Jaffe, Jens Jasche, Guilhem Lavaux,
Will Percival, Benjamin Wandelt, Camille Noûs

and the Aquila Consortium
www.aquila-consortium.org

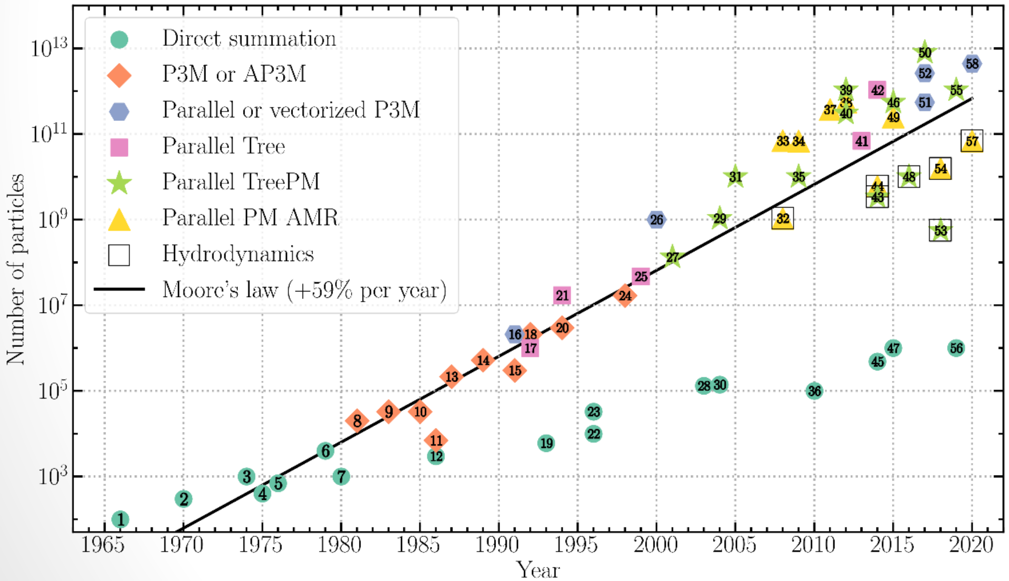
13 October 2020

ICIC
Imperial Centre
for Inference & Cosmology

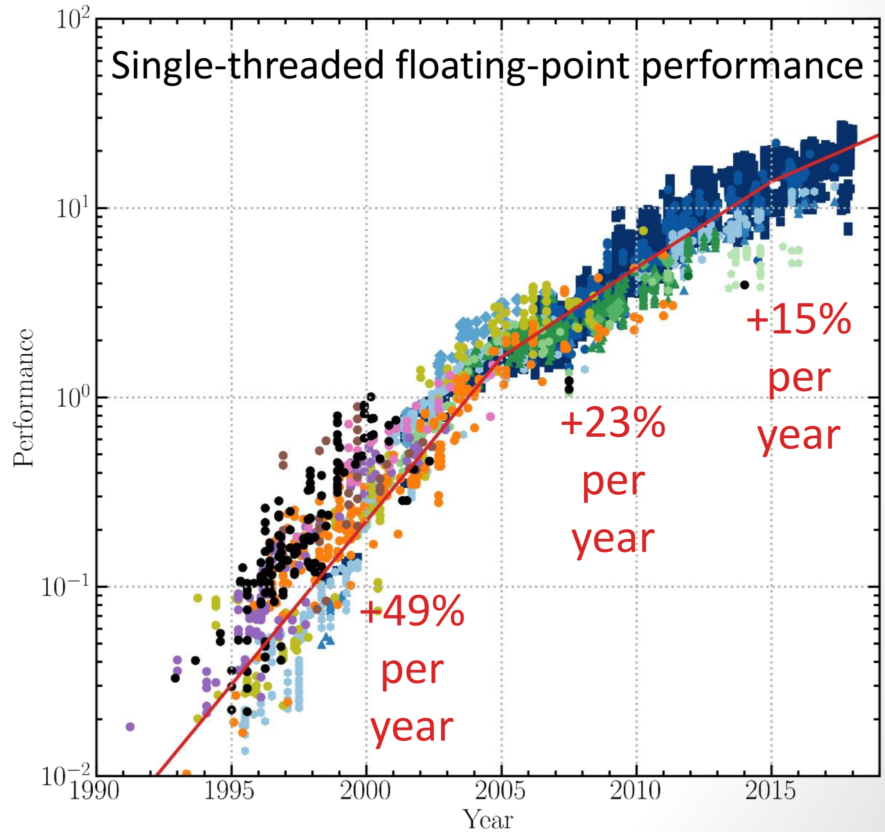
Imperial College
London

Parallelisation of *N*-body codes: the challenge

- Most of the work on numerical cosmology so far has focused on algorithms (such as tree, multipole, and mesh methods) that reduce the need for communications across the full computational volume
- But per-core compute performance is slowing down



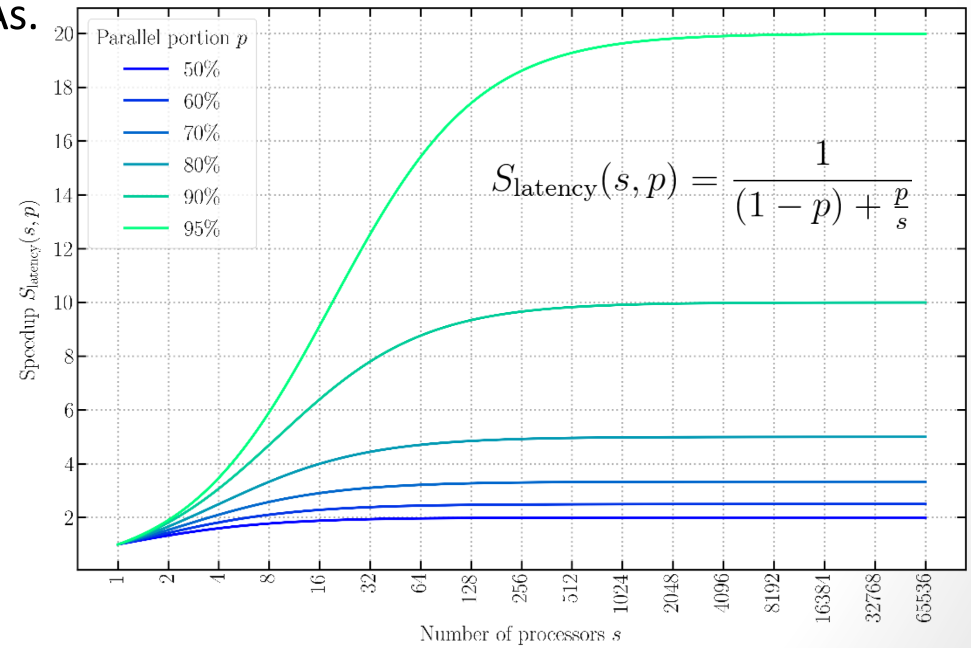
References for all points available at http://florent-leclercq.eu/blog.php?page=2



Based on adjusted SPECfp® results, http://spec.org

Cosmological simulations in the exascale world

- Traditional hardware architectures are reaching their physical limit.
- Current hardware development focuses on:
  - Packing a larger number of cores into each CPU: currently $\mathcal{O}(10^5)$, soon $\mathcal{O}(10^{6-7})$ in systems that are currently being built.
  - Developing hybrid architectures with cores + accelerators: GPUs and reconfigurable chips such as FPGAs.



- Compute cycles are no longer the scarce resource. The cost is driven by interconnections.
- Amdahl's law: latency kills the gains of parallelisation

Amdahl 1967, doi:10.1145/1465482.1465560

➡ Cosmological simulations cannot merely rely on computers becoming faster to reduce the computational time.

tCOLA: *Comoving Lagrangian Acceleration* (temporal domain)

- Write the displacement vector as: $\boldsymbol{\Psi} = \boldsymbol{\Psi}_{\mathrm{LPT}} + \boldsymbol{\Psi}_{\mathrm{res}} \quad (\mathbf{x} = \mathbf{q} + \boldsymbol{\Psi})$

Tassev & Zaldarriaga 2012, 1203.5785

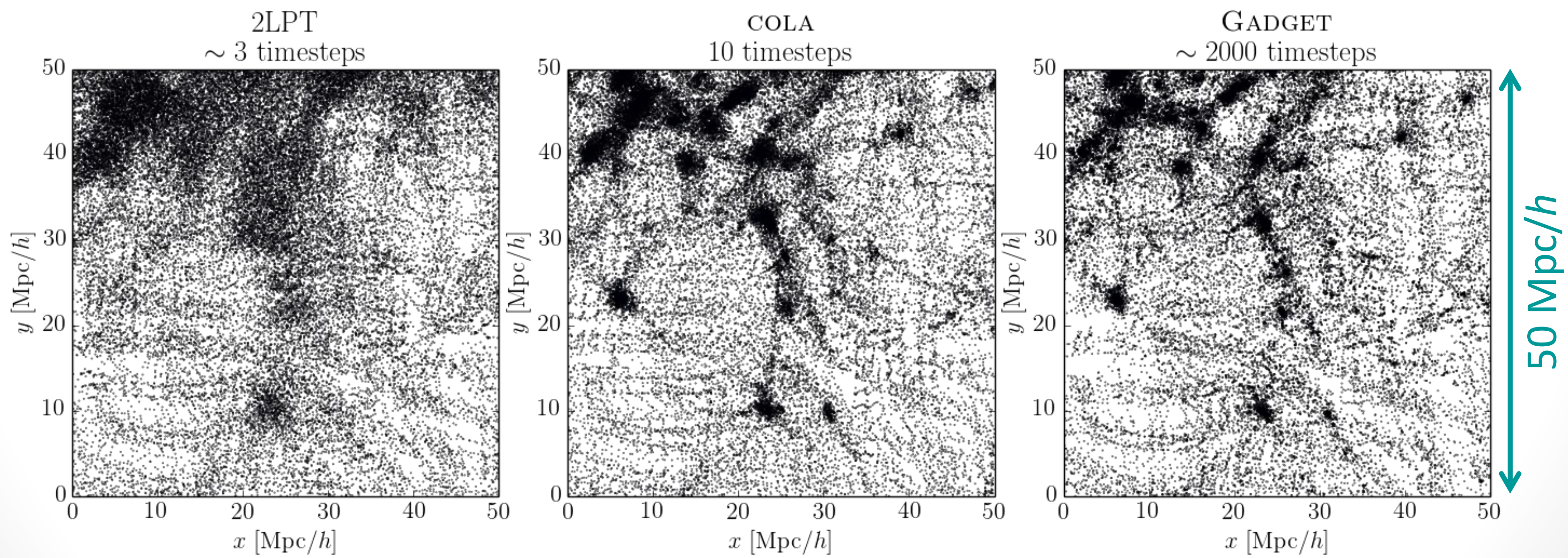
Analytical solutions!

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

Standard: $\partial_a^2 \boldsymbol{\Psi} = -\boldsymbol{\nabla}_{\mathbf{x}} \Phi$

Modified: $\partial_a^2 \boldsymbol{\Psi}_{\mathrm{res}} = \partial_a^2 (\boldsymbol{\Psi} - \boldsymbol{\Psi}_{\mathrm{LPT}}) = -\boldsymbol{\nabla}_{\mathbf{x}} \Phi - \partial_a^2 \boldsymbol{\Psi}_{\mathrm{LPT}}$

Tassev, Zaldarriaga & Einsenstein 2013, 1301.0322



Beneficial gain of efficiency… but the real problem is not CPU-hours, but the inability to run on a very large number of cores due to latencies/parallelisation overhead.

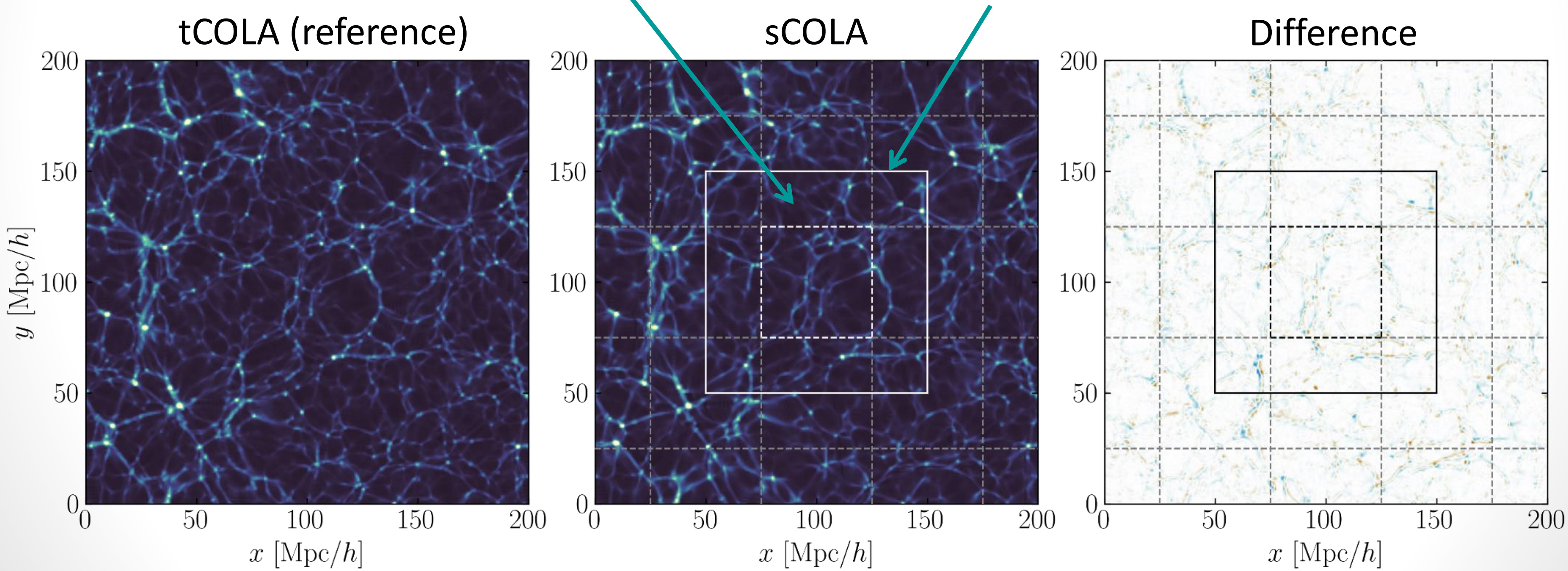
sCOLA: Extension to the spatial domain

- Computing the LPT reference frame suggests a new strategy:

Can we decouple sub-volumes by using the large-scale analytical solution?

Proof of concept using one sub-box: Tassev, Eisenstein, Wandelt & Zaldarriaga 2015, 1502.07751

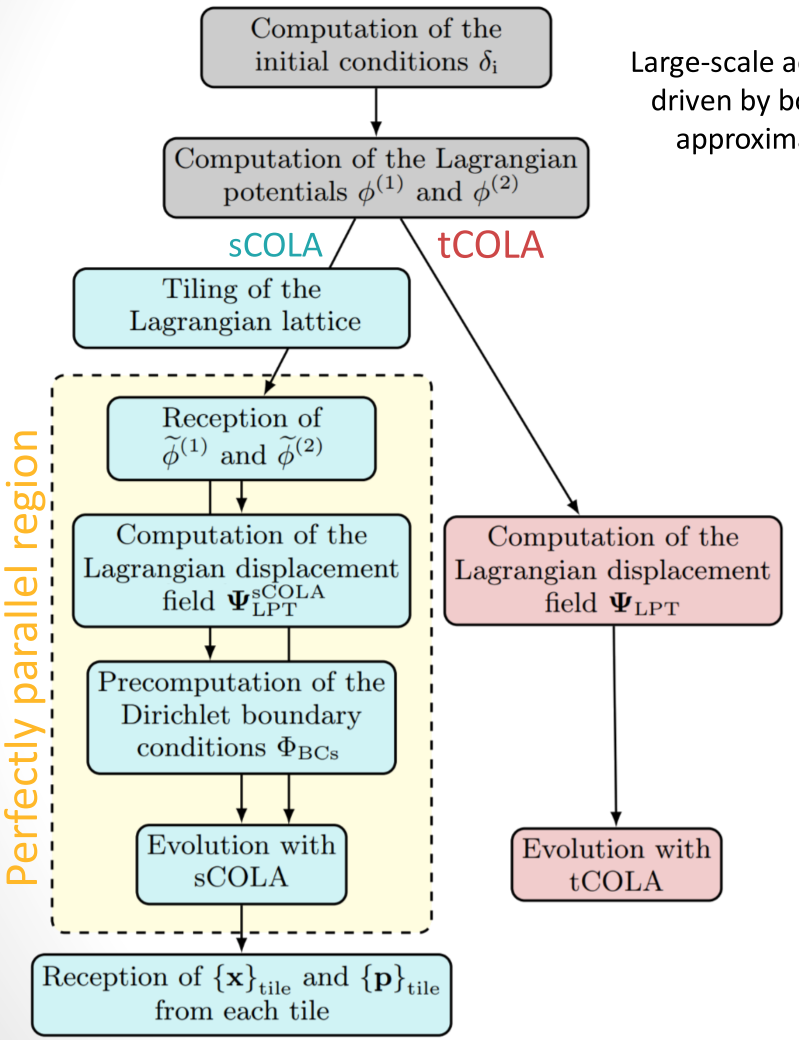
1. A buffer region around each tile
2. Appropriate Dirichlet boundary conditions for the potential



The Poisson solver uses discrete sine transforms (DSTs) instead of FFTs.

FL, Faure, Lavaux, Wandelt, Jaffe, Heavens, Percival & Noûs 2020, 2003.04925

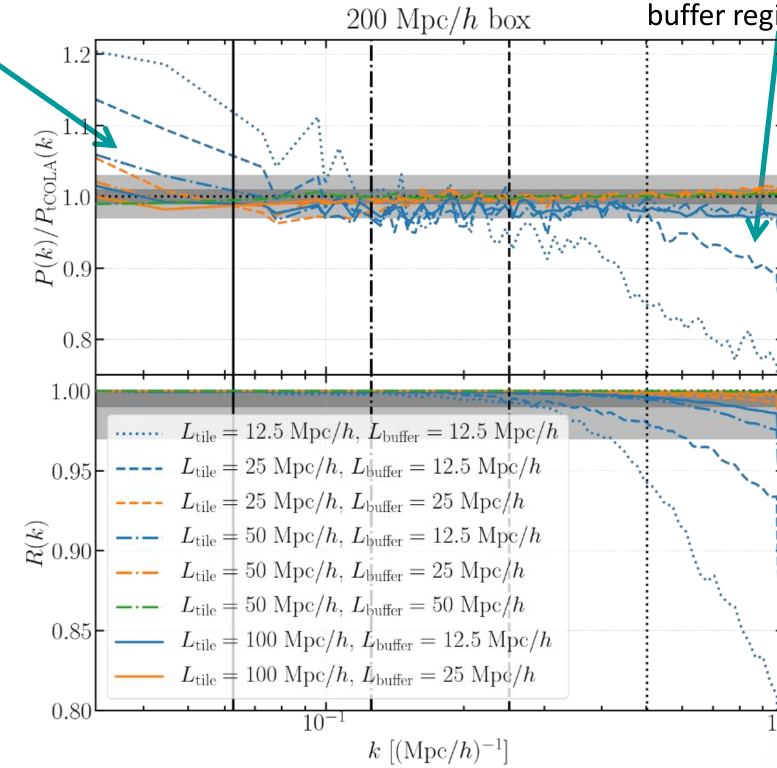
The perfectly parallel algorithm and its accuracy



- Parameter investigation (size of tiles and buffer regions):

Large-scale accuracy is driven by boundary approximations

Small-scale accuracy is driven by the size of buffer regions



- Some setups reach 3 to 1% accuracy at all scales, as required for the next-generation of surveys

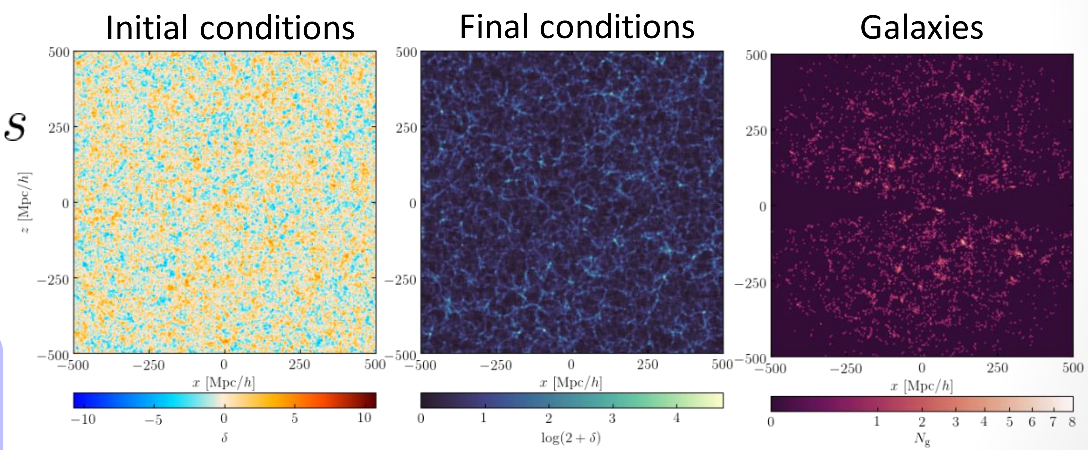
FL, Faure, Lavaux, Wandelt, Jaffe, Heavens, Percival & Noûs 2020, 2003.04925

Memory requirements, parallelisation potential & speed

- Buffer regions require to oversimulate the volume by a factor r
- But small *N*-body simulations can be run in the L3 cache of CPUs, on GPUs or FGPAs: hardware speed-up factor of s
- Parallelisation potential factor:

$$p = s\,\frac{N_{\mathrm{tiles}}}{r} = s\left(\frac{L}{L_{\mathrm{sCOLA}}}\right)^3$$

sCOLA is implemented in the publicly available Simbelmynë code (v. ≥ 0.4):
https://bitbucket.org/florent-leclercq/simbelmyne/

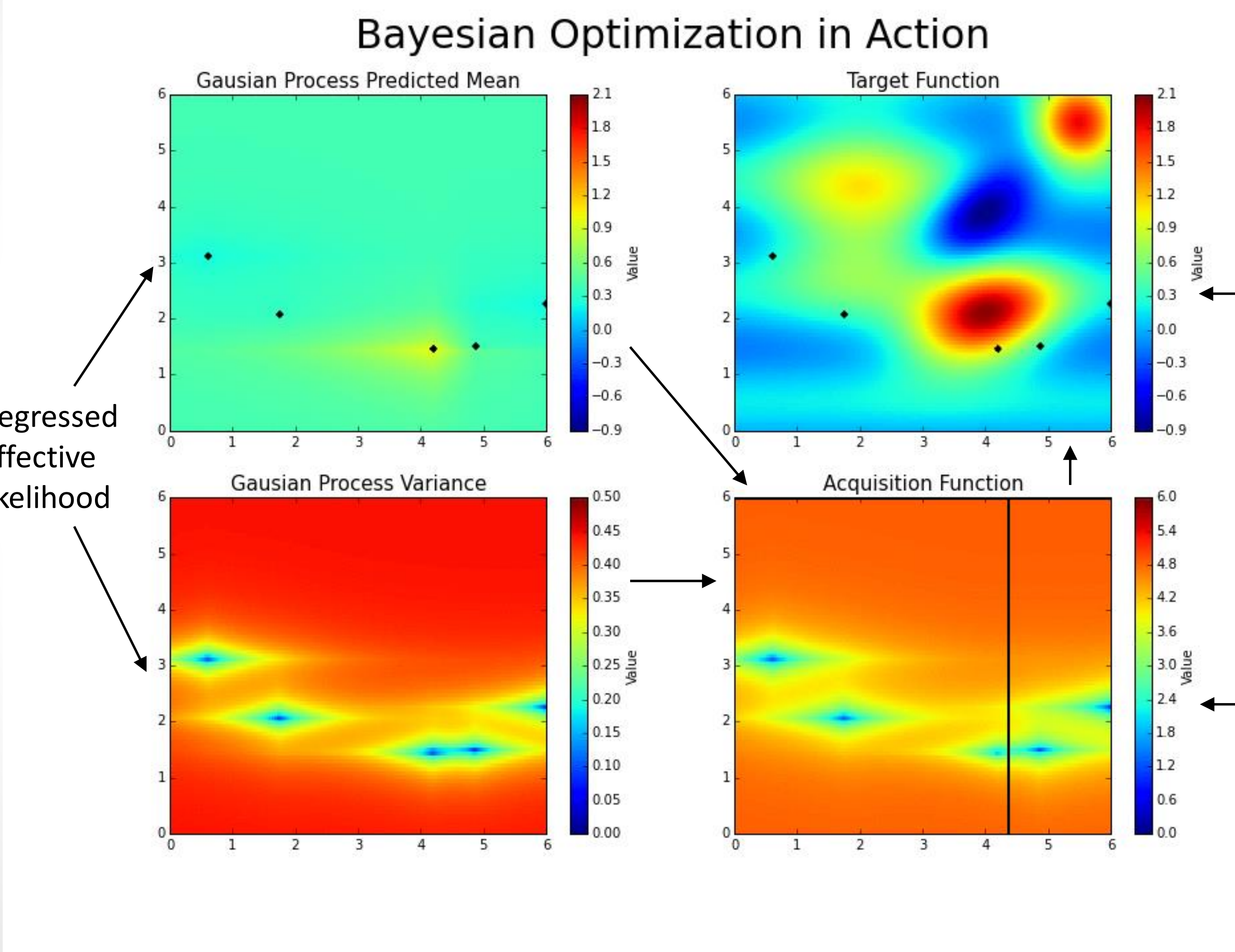


Voir aussi : Leclercq & Lavaux, *Vers une simulation de l'Univers sur un téléphone portable* (The Conversation France, Mai 2020)

FL, Faure, Lavaux, Wandelt, Jaffe, Heavens, Percival & Noûs 2020, 2003.04925

BOLFI: Data acquisition

Simulations are obtained from sampling an adaptively-constructed proposal distribution, using the regressed effective likelihood



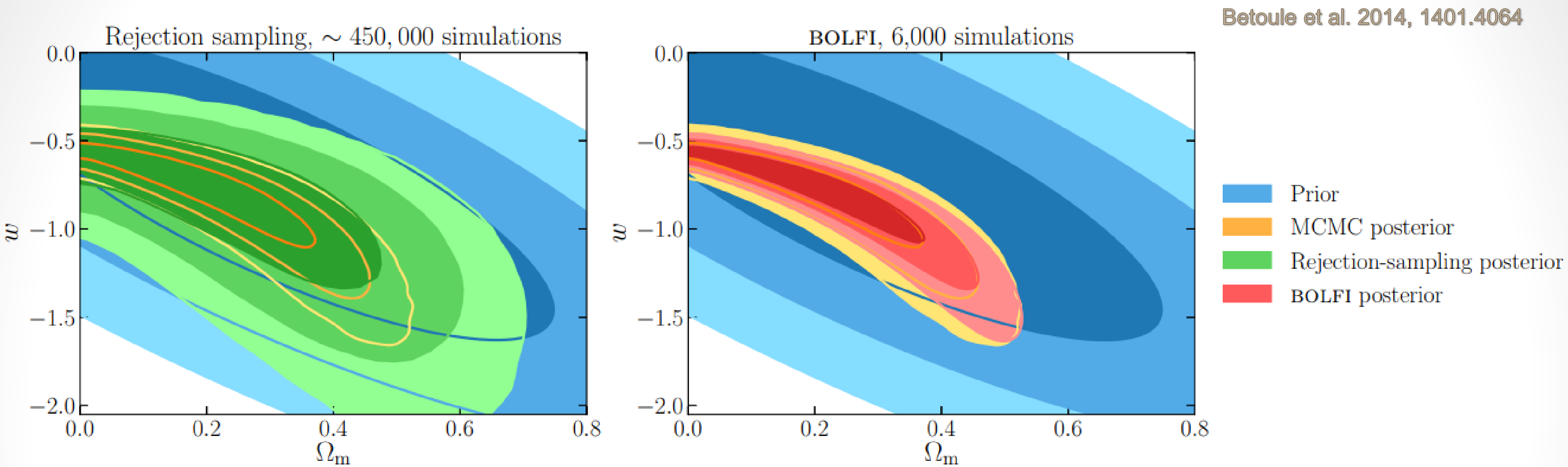
Regressed effective likelihood

True (unknown) effective posterior

The acquisition function finds a compromise between exploration & exploitation

F. Nogueira, https://github.com/fmfn/BayesianOptimization

BOLFI: Re-analysis of the JLA supernova sample

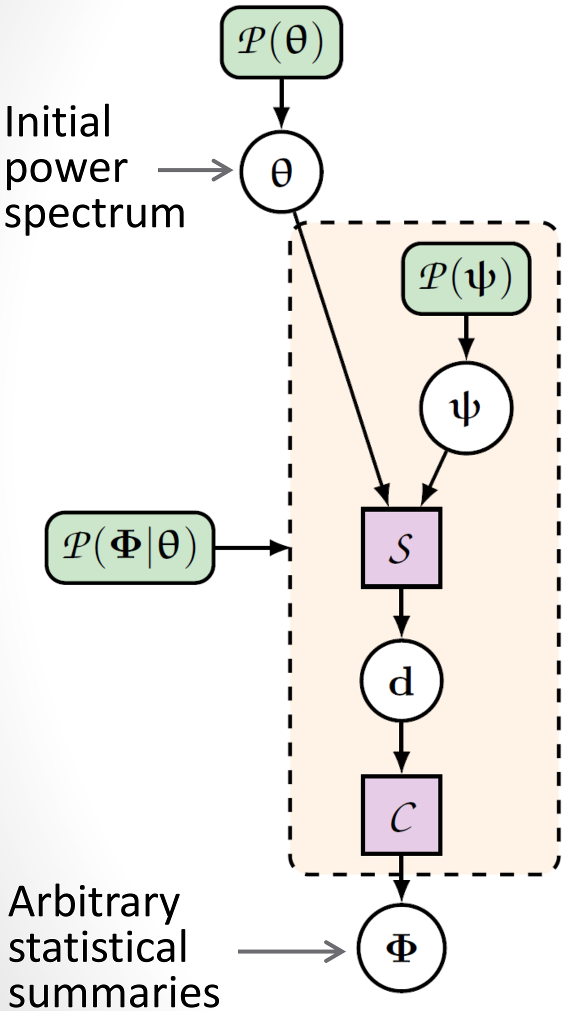


- The number of required simulations is reduced by:
  - 2 orders of magnitude with respect to likelihood-free rejection sampling (for a much better approximation of the posterior)
  - 3 orders of magnitude with respect to exact Markov Chain Monte Carlo sampling

FL 2018, 1805.07152

- Bayesian optimisation can also be applied to the “true” likelihood (if known) or to iteratively build an emulator of the data model

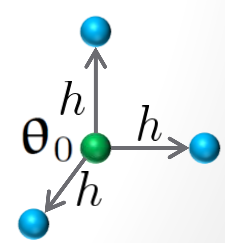
SELFI: Expansion of black-box data models



- We aim at inferring the initial 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:

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

- Gradients of the black-box can be evaluated via finite differences in parameter space

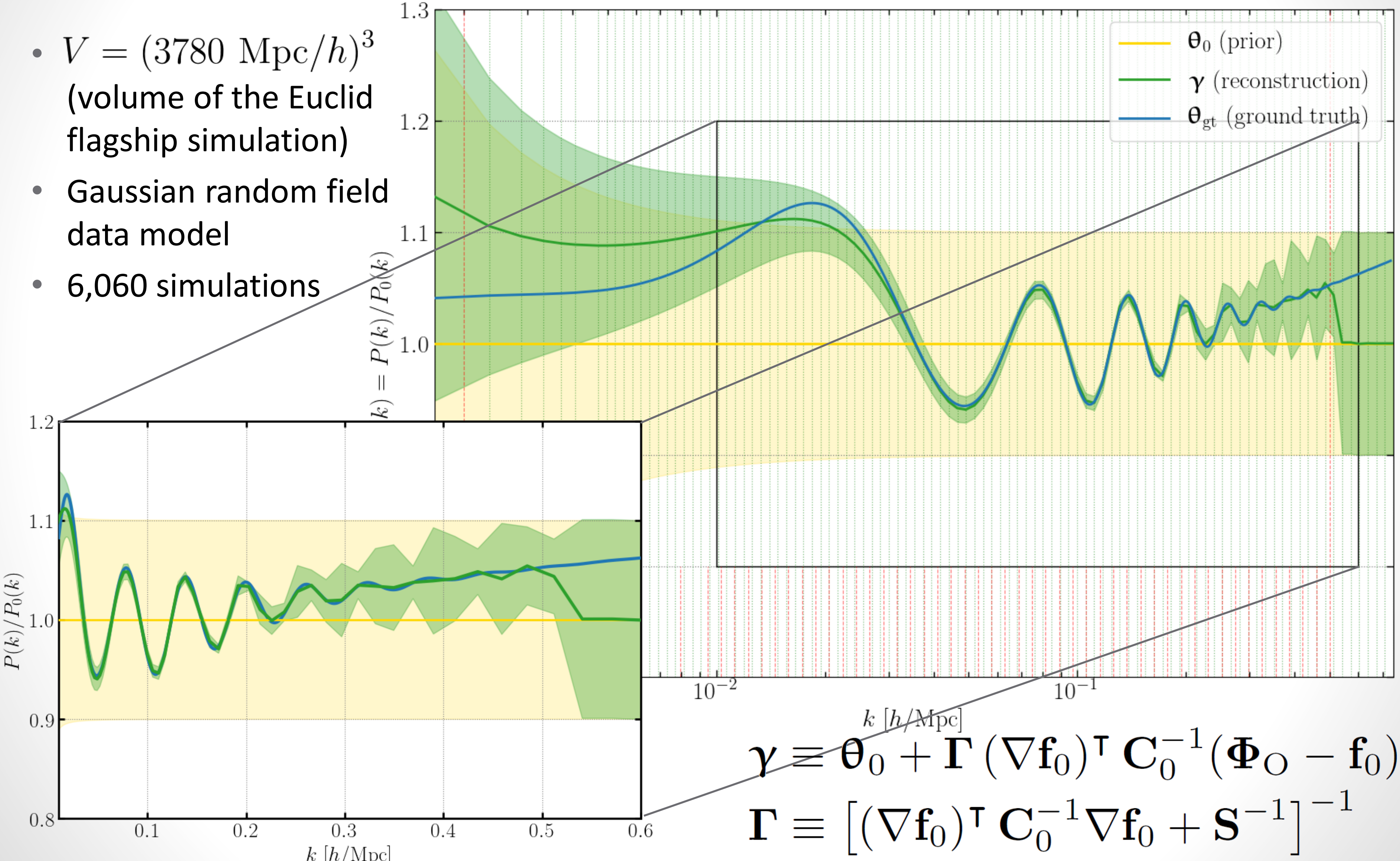


FL, Enzi, Jasche & Heavens 2019, 1902.10149

SELFI Euclid forecast (cosmic variance limit)

"GCLFI" project

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



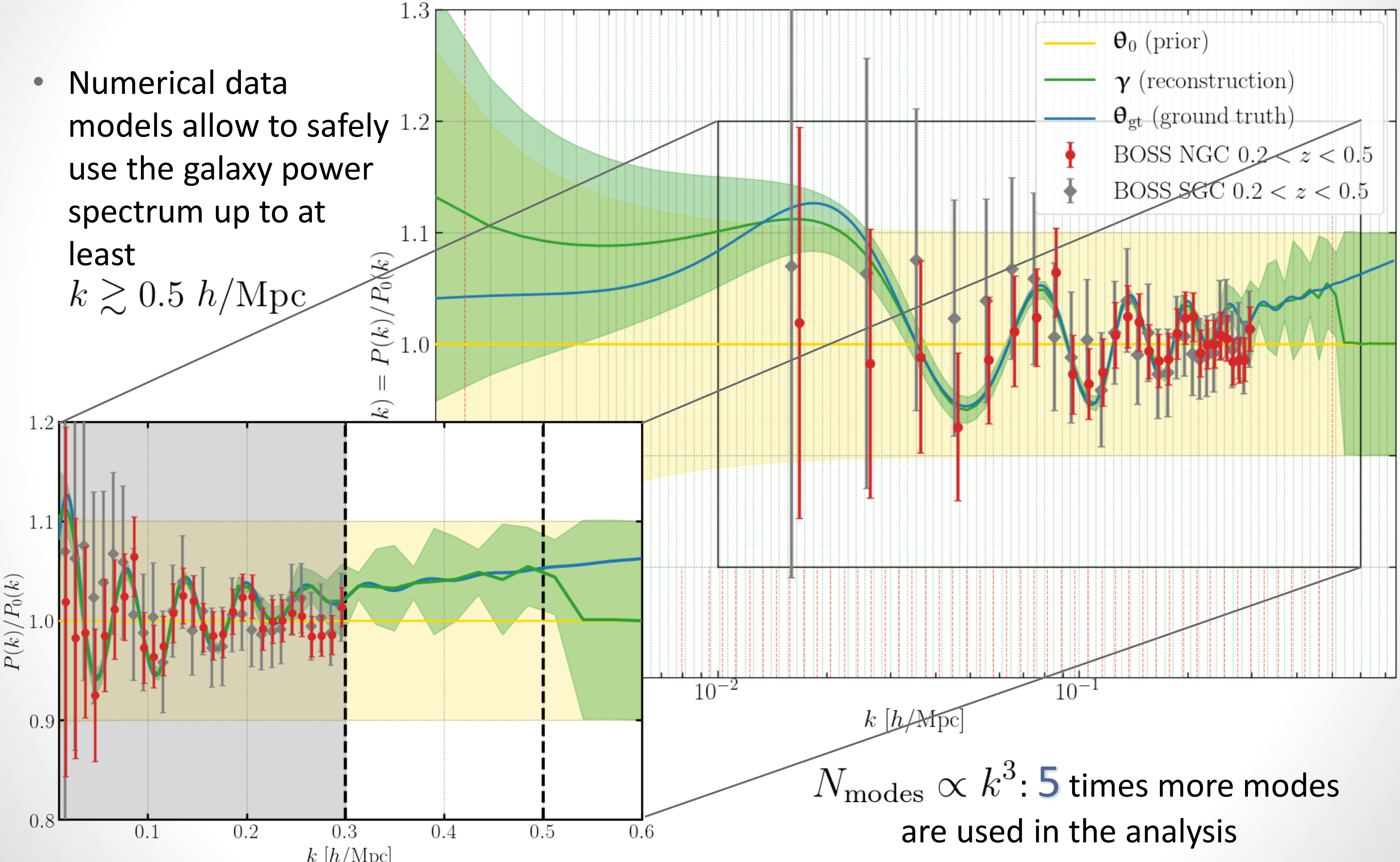
$$\gamma \equiv \theta_0 + \Gamma\,(\nabla \mathbf{f}_0)^\intercal\,\mathbf{C}_0^{-1}(\Phi_\mathrm{O} - \mathbf{f}_0)$$

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

SELFI 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
  $k \gtrsim 0.5\ h/\mathrm{Mpc}$

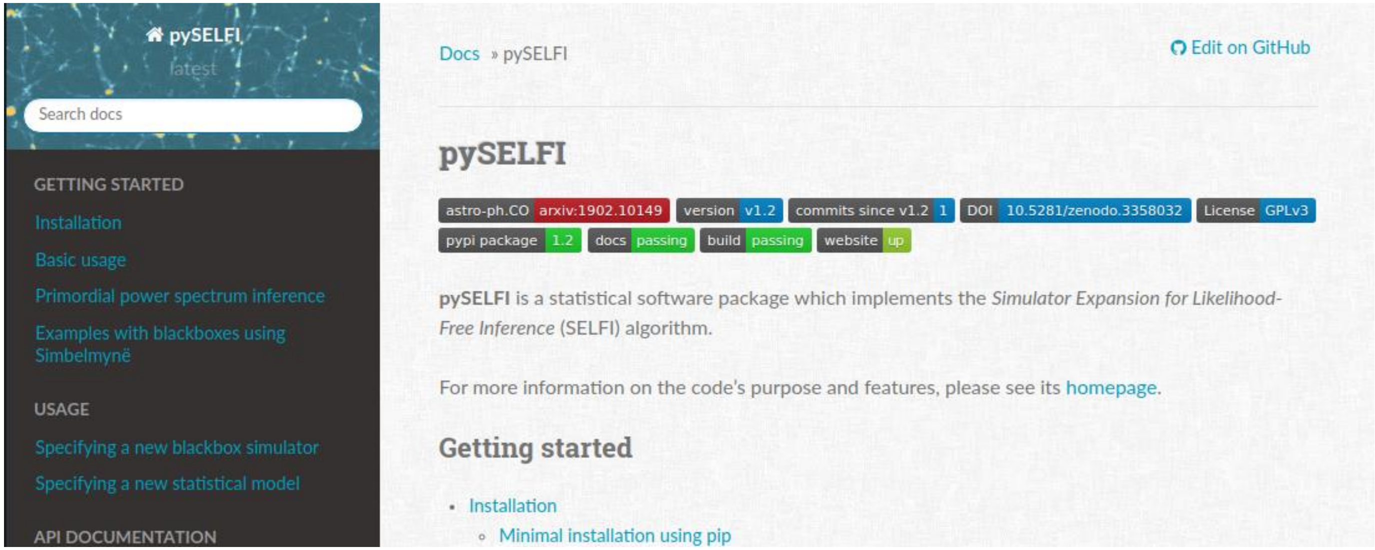


$N_{\mathrm{modes}} \propto k^3$: 5 times more modes are used in the analysis

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 own black-box)




```
pip install pyselfi
```

Concluding thoughts

- In the age of peta-/exa-scale computing, we introduced a perfectly parallel and easily applicable algorithm for cosmological simulations using sCOLA, a hybrid analytical/numerical technique.
- Bayesian analyses of galaxy surveys with fully non-linear numerical black-box models is not an impossible task!
- BOLFI allows inference within specific cosmological models with a very limited simulation budget.
- SELFI allows inference of the initial power spectrum and cosmological parameters.