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

Baptiste Faure, Alan Heavens,
Andrew Jaffe, Guilhem Lavaux,
Will Percival, Benjamin Wandelt, Camille Noûs

and the Aquila Consortium
www.aquila-consortium.org

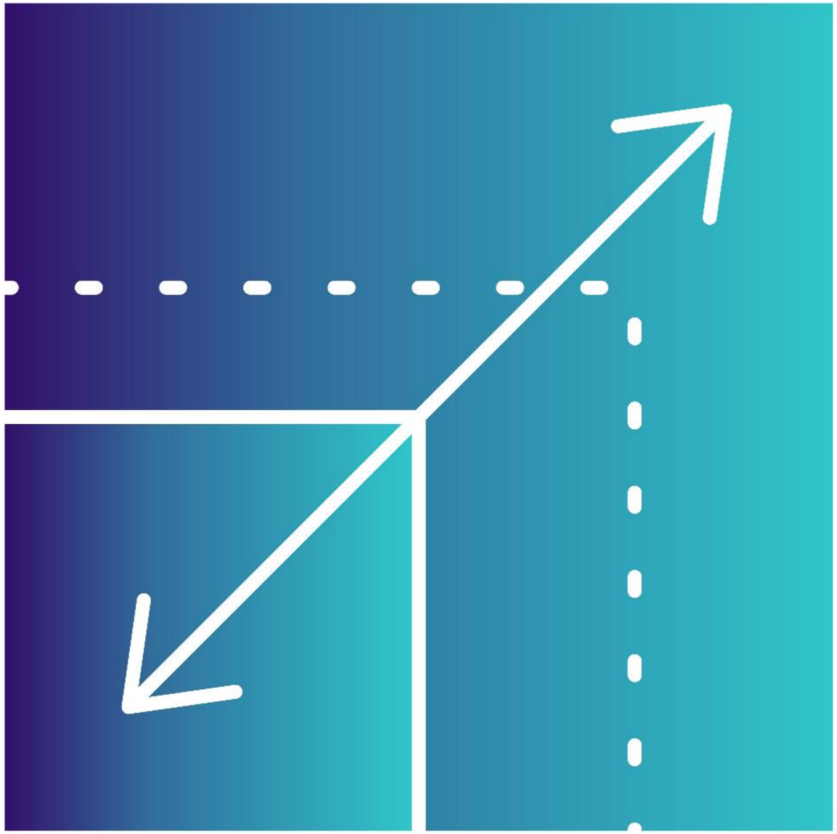
20 January 2022

ICIC
Imperial Centre
for Inference & Cosmology

Imperial College
London

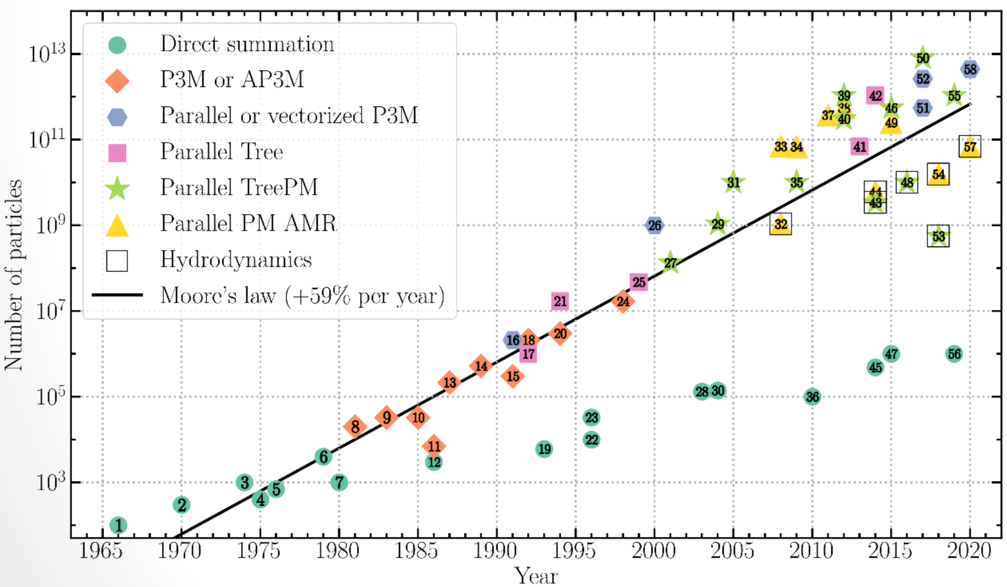
The challenge addressed in this talk: scalability

- Scalability: the property of analysis techniques to handle a growing amount of data under computational resource constraints.
- The challenge is twofold:
  - in the data models: how can we best use modern computers and their architecture?
  - in the inference techniques: how can we perform rigorous Bayesian reasoning given a limited computational budget?

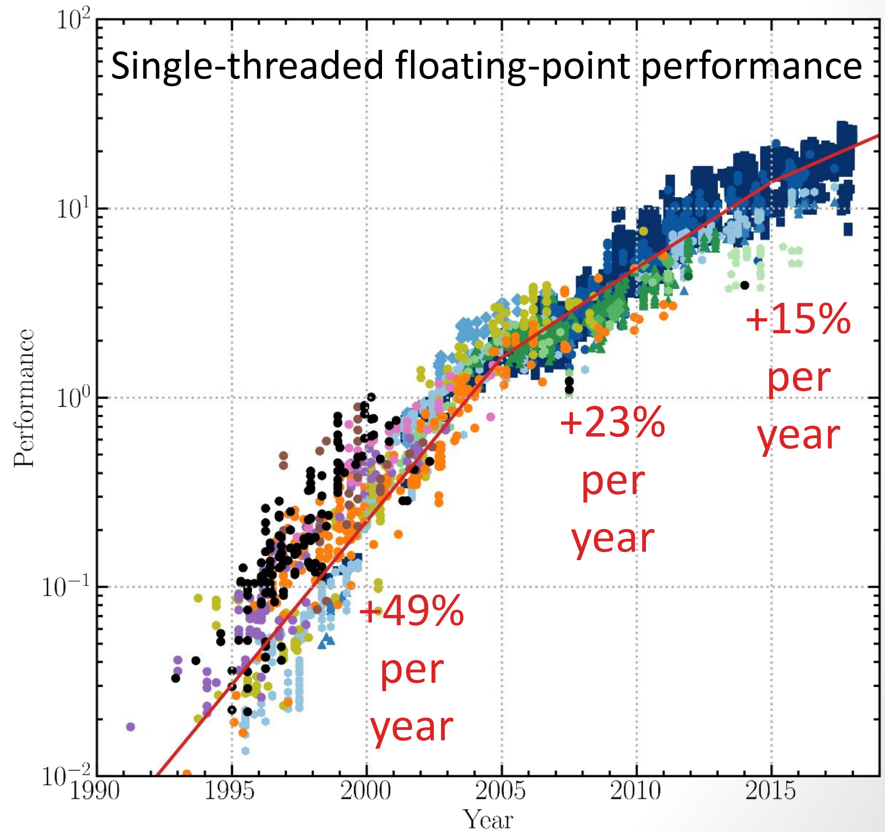


Parallelisation of *N*-body codes: the challenge

- 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

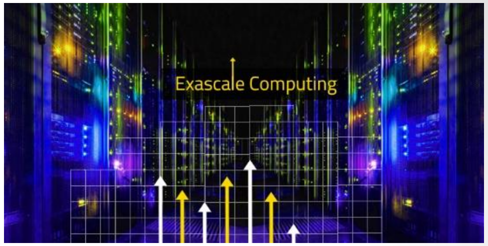


References for all points available at http://florent-leclercq.eu/blog.php?page=2



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

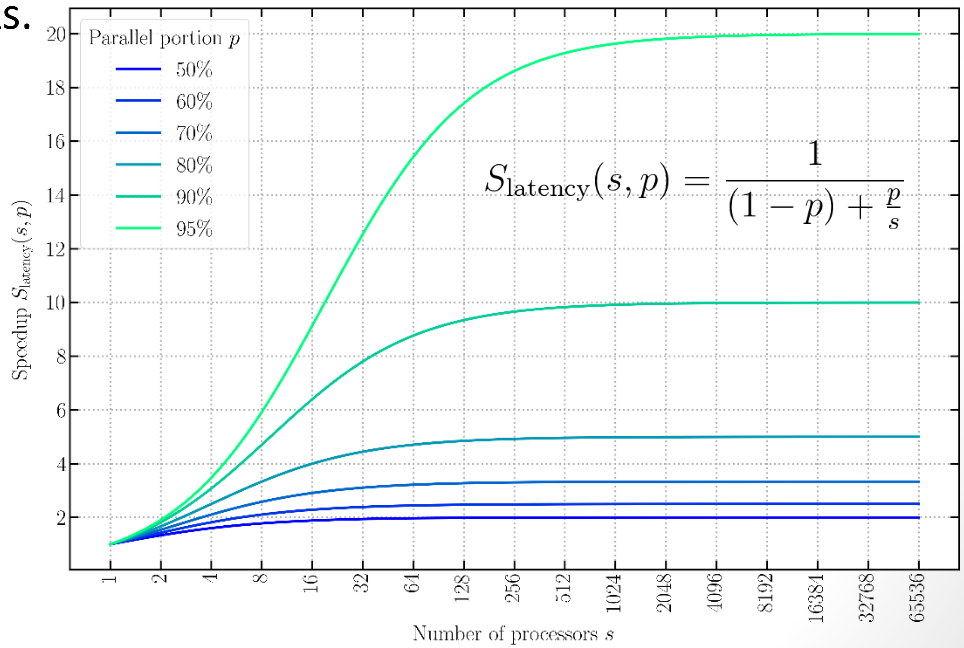
Numerical data models in the exascale world



- 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: $\mathbf{\Psi} = \mathbf{\Psi}_{\mathrm{LPT}} + \mathbf{\Psi}_{\mathrm{res}} \quad (\mathbf{x} = \mathbf{q} + \mathbf{\Psi})$

Tassev & Zaldarriaga 2012, 1203.5785

Analytical solutions!

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

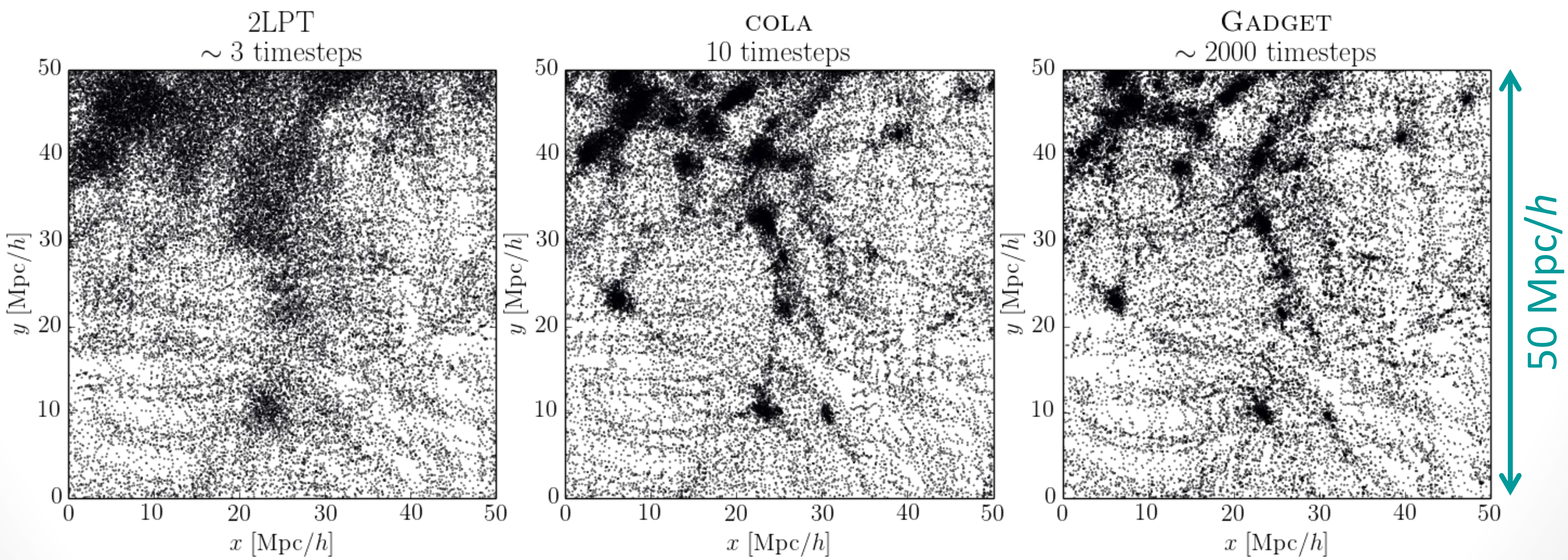
Standard:

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

Modified:

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

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.

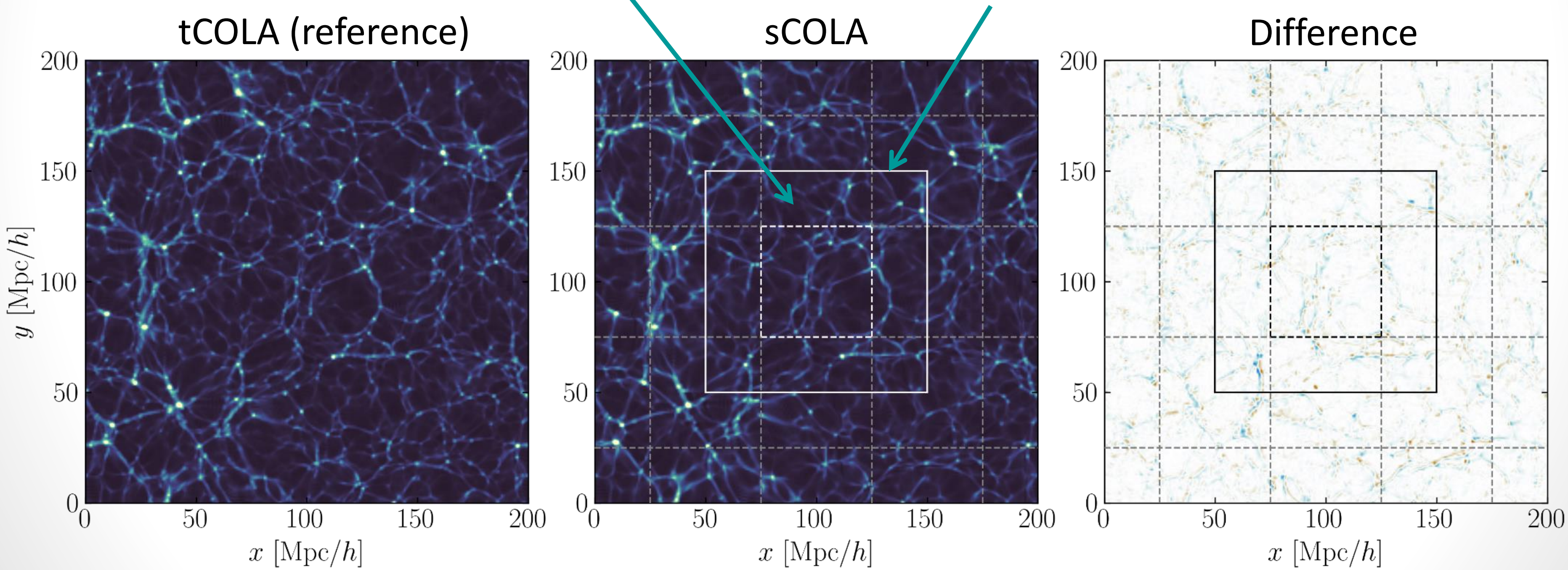
sCOLA: Extension to the spatial domain

- 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.

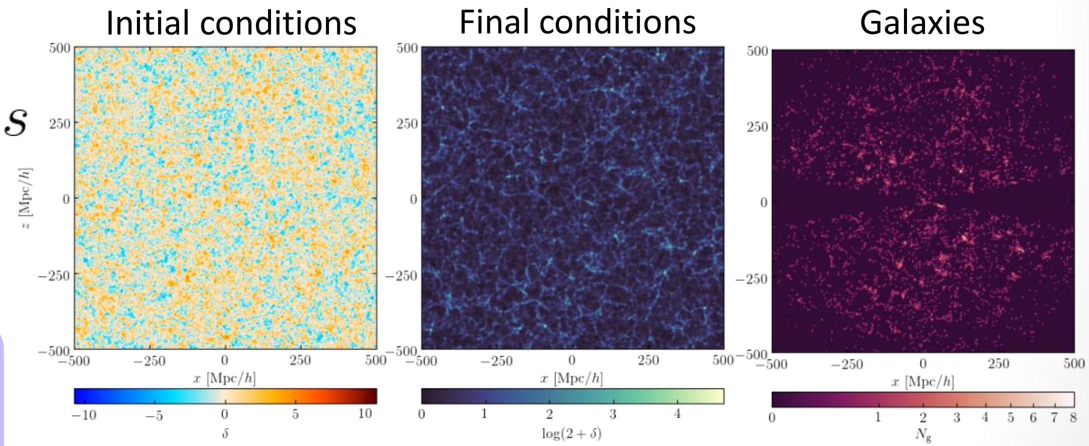
FL, Faure, Lavaux, Wandelt, Jaffe, Heavens, Percival & Noûs 2020, 2003.04925

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:

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

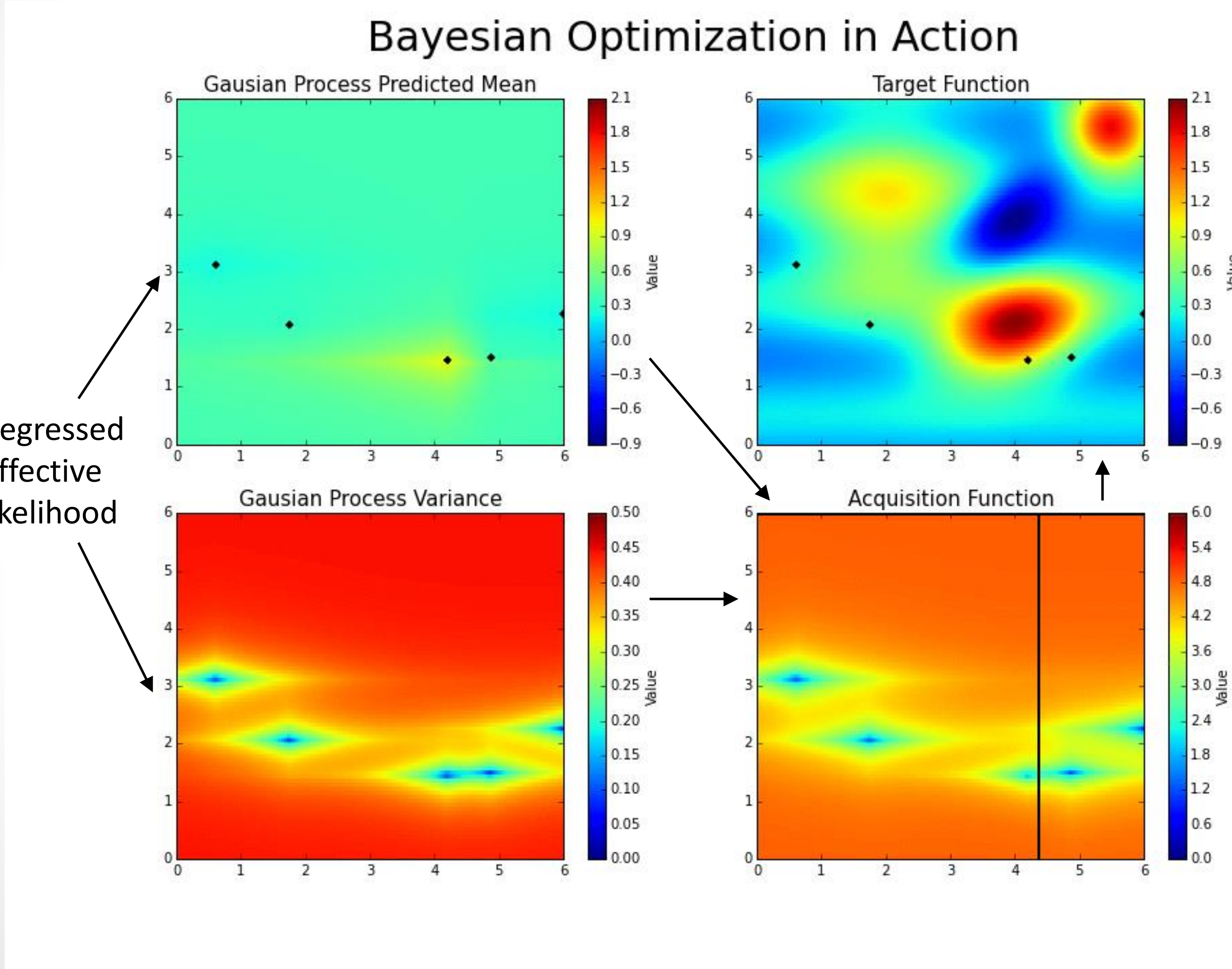
sCOLA is implemented in the publicly available Simbelmynë code (v. ≥ 0.4):
https://bitbucket.org/florent-leclercq/simbelmyne/



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



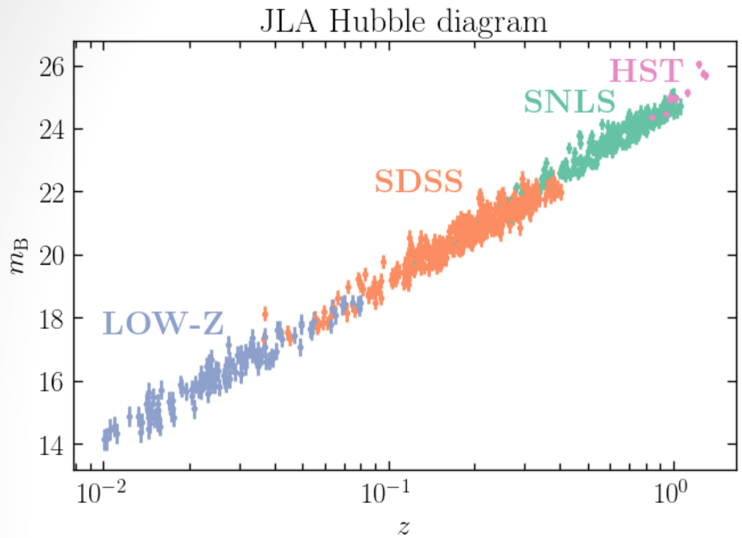
Regressed effective likelihood

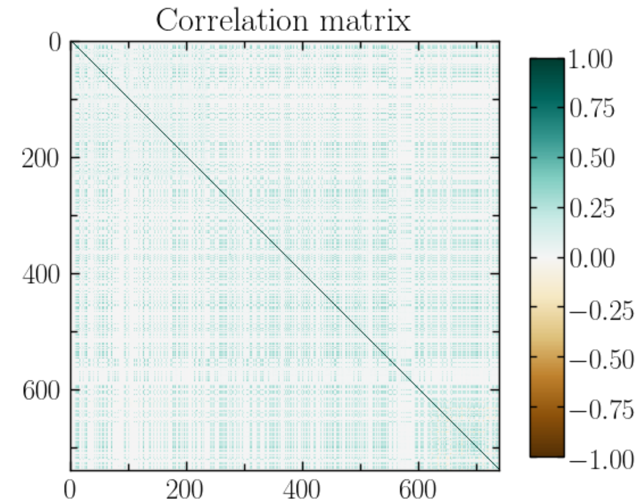
True (unknown) effective posterior

The acquisition function finds a compromise between exploration & exploitation

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

BOLFI: Re-analysis of the JLA supernova sample





Betoule et al. 2014, 1401.4064

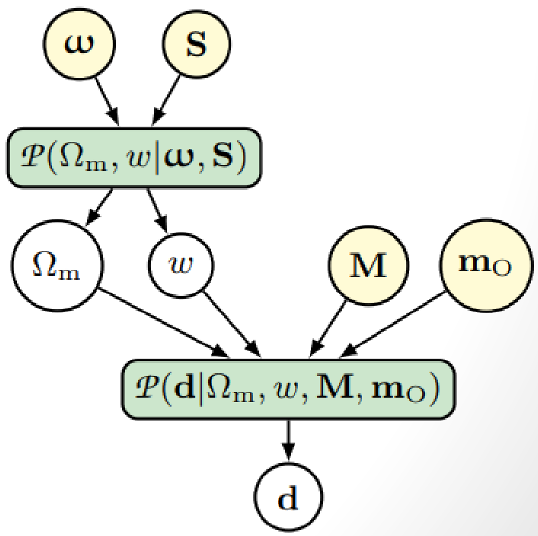
- 6-parameter model:
  2 cosmological parameters + 4 nuisance parameters

$$m_B = 5\log_{10}\left[\frac{D_L(z)}{10\ \mathrm{pc}}\right] + \widetilde{M}_B(M_{\mathrm{stellar}}, M_B, \delta M) - \boxed{\alpha} X_1 + \boxed{\beta} C$$

$$\widetilde{M}_B(M_{\mathrm{stellar}}, M_B, \delta M) = \boxed{M_B} + \boxed{\delta M}\,\Theta\left(M_{\mathrm{stellar}} - 10^{10} M_\odot\right)$$

$$D_L(z) = \frac{(1+z)\,c}{H_0}\int_0^z \frac{dz'}{E(z')}$$

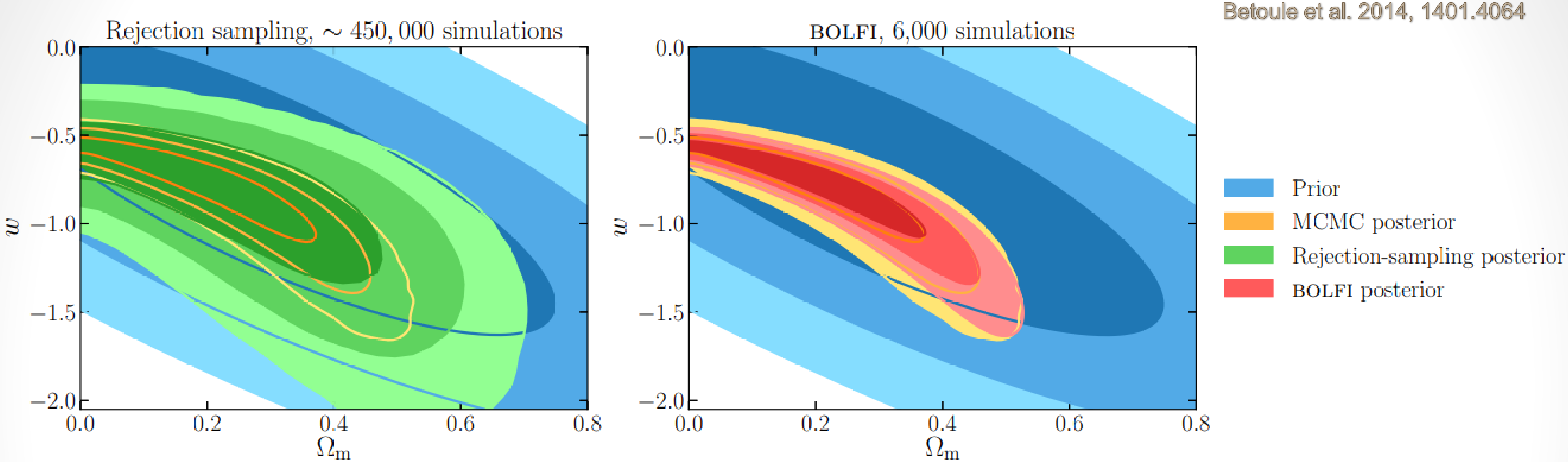
$$E(z) \equiv \sqrt{\boxed{\Omega_m}(1+z)^3 + (1-\boxed{\Omega_m})(1+z)^{3(\boxed{w}+1)}}$$



FL 2018, 1805.07152

BOLFI: Re-analysis of the JLA supernova sample

Betoule et al. 2014, 1401.4064



- 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

Concluding thoughts

- Scalability (of data models and of inference techniques) is the highest priority we have for the analysis of next-generation data.
- 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.
- BOLFI is a general algorithm for Gaussian Process surrogate Bayesian inference. It allows inference within specific cosmological models with a very limited simulation budget.