Implicit likelihood inference from galaxy survey data with robustness to model misspecification

Simulation based inference in Astrophysics,
RAS Specialist Discussion Meeting

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

In collaboration with:
Wolfgang Enzi (MPA), Alan Heavens (Imperial College), Tristan Hoellinger (IAP), Jens Jasche (Stockholm University), Guilhem Lavaux (IAP)

and the Aquila Consortium

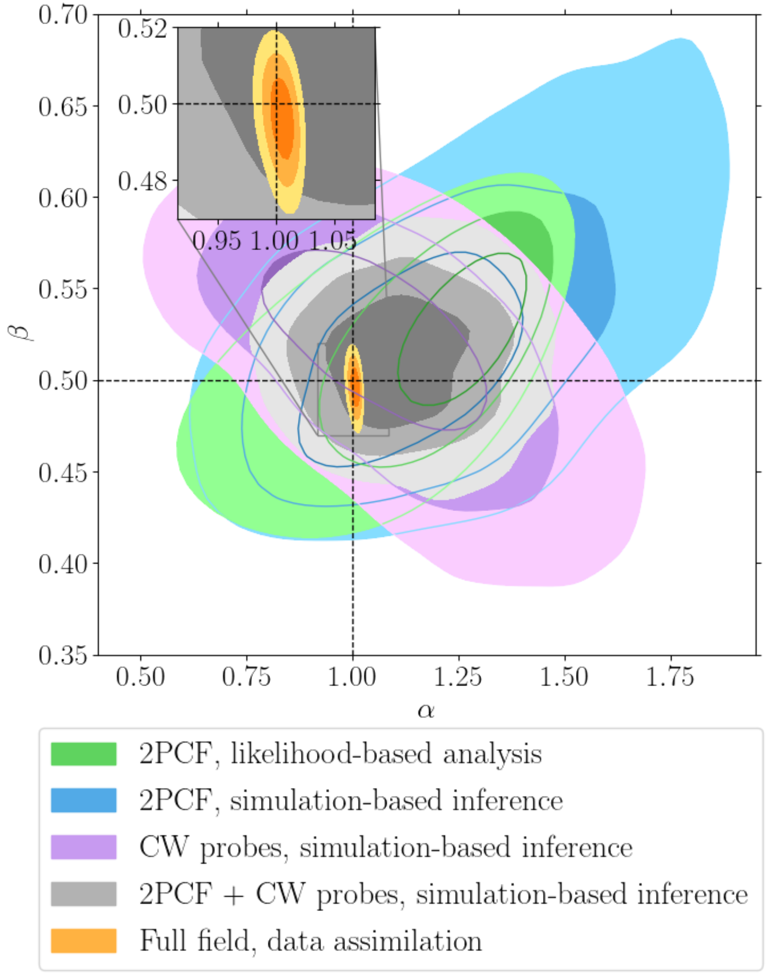
www.aquila-consortium.org

12 January 2024

Why I decided to go "implicit" for galaxy clustering additional probes

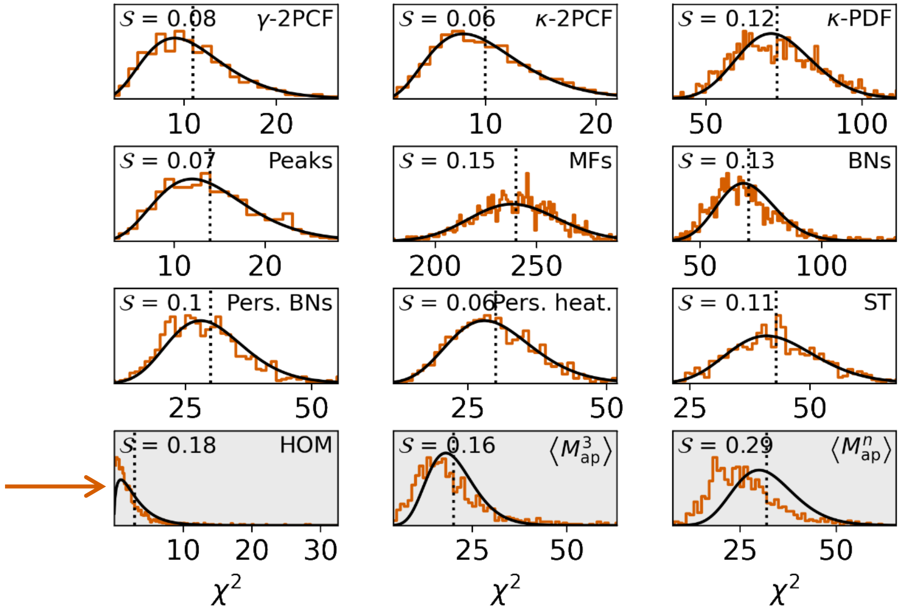
Note: likelihood-free inference (LFI) ≈ simulation-based inference (SBI) ≈ implicit likelihood inference (ILI)

- A question of accuracy: first, avoid biases.



FL & Heavens, 2103.04158

- Some weak lensing additional probes also have a non-Gaussian distribution.



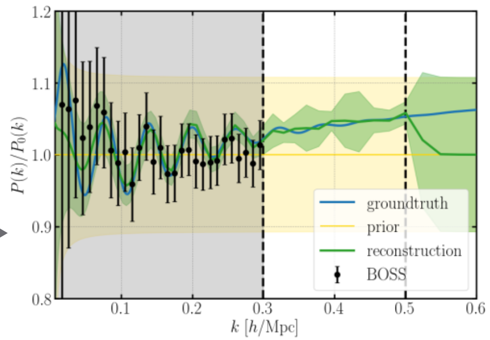
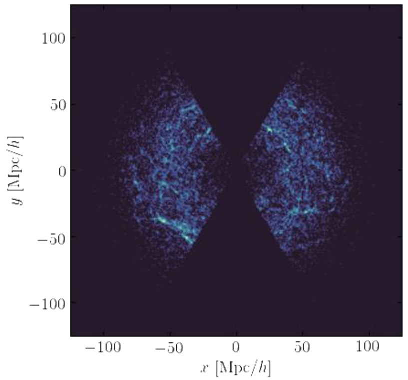
- A question of precision: can numerical forward models be used to push further than $k \gtrsim 0.15\ h/\mathrm{Mpc}$? The full field contains much more information.

Euclid HOWLS-KP paper 1, Ajani *et al.*, 2301.12890

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

$\mathcal{P}(\omega)$

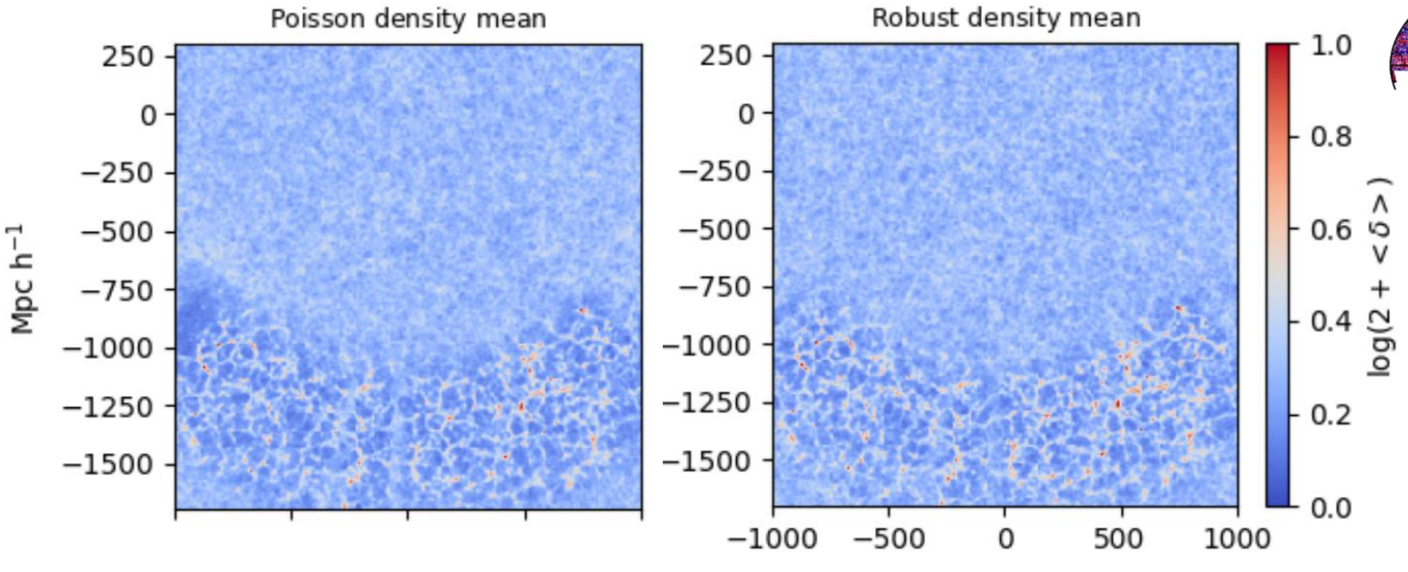
Example from cosmology:

Target parameters	ω	Cosmological parameters	$\Omega_\mathrm{m}, \Omega_\mathrm{b}, H_0, n_\mathrm{S}, \sigma_8, \tau$
Cheap (deterministic) simulator	$\mathcal{T}$	Solver of the Boltzmann equation (CAMB, CLASS)	
Latent function	θ	Matter power spectrum	
Complex probabilistic observational process	$\mathcal{P}(\Phi\|\theta)$	*N*-body simulation, galaxy formation, survey modelling...	
Raw data	Φ	Galaxy survey data	
Compressor	$\mathcal{C}$	Score compression	
Compressed summary statistics	$\tilde{\omega}$	Quasi maximum-likelihood estimators	$\widetilde{\Omega_\mathrm{m}}, \widetilde{\Omega_\mathrm{b}}, \widetilde{H_0}, \widetilde{n_\mathrm{S}}, \widetilde{\sigma_8}, \widetilde{\tau}$

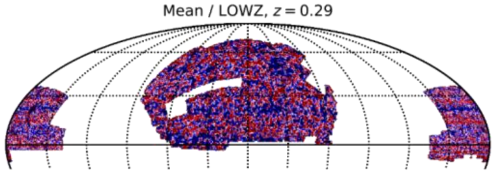
Model misspecification and unknown systematics with an explicit field-level likelihood

- 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.

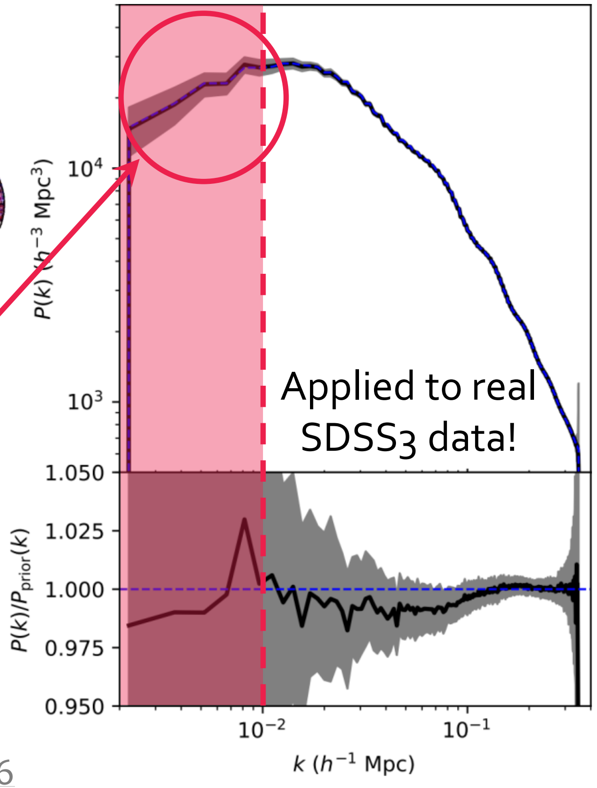
- In cosmology, we are sometimes unable to formulate ***any*** model that fits the data in some regimes.
- Machine-aided report of unknown systematic effects is possible with an explicit field-level likelihood (BORG):



Porqueres, Ramanah, Jasche & Lavaux, 1812.05113

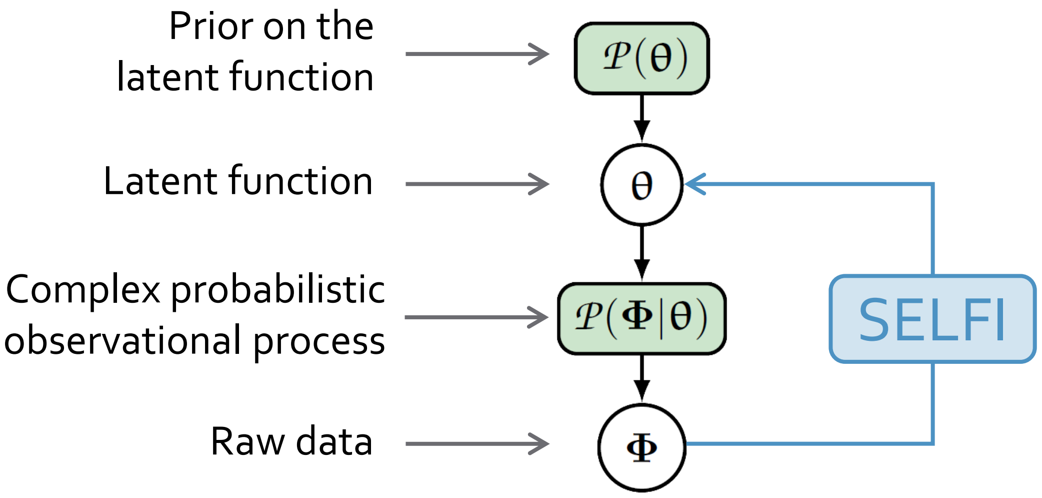


No apparent contamination, even well beyond the turn-over



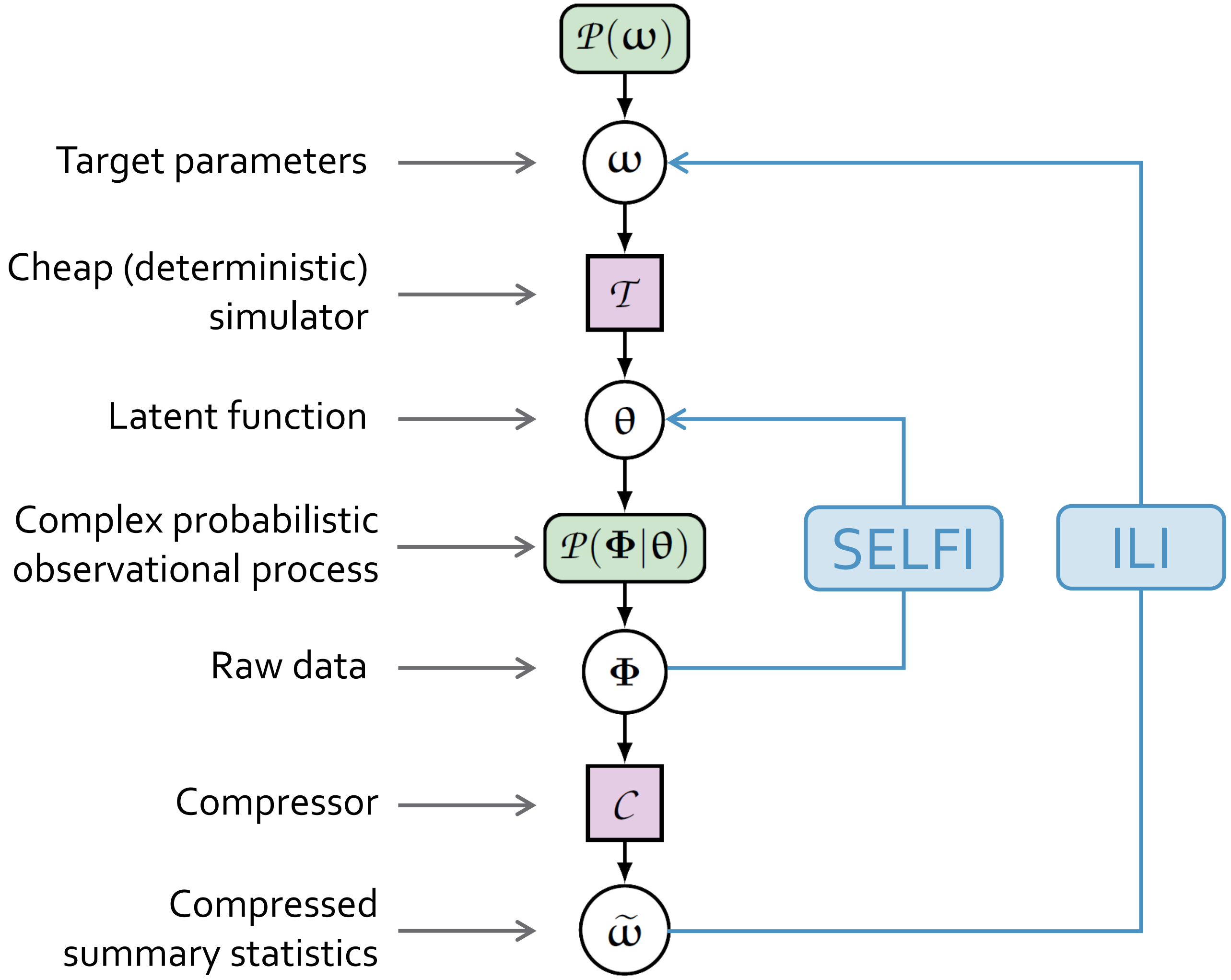
Lavaux, Jasche & FL, 1909.06396

Key idea: a two-step ILI process that recycles simulations



1. Inference of the latent function θ, to check for model misspecification:
   - SELFI algorithm

Key idea: a two-step ILI process that recycles simulations

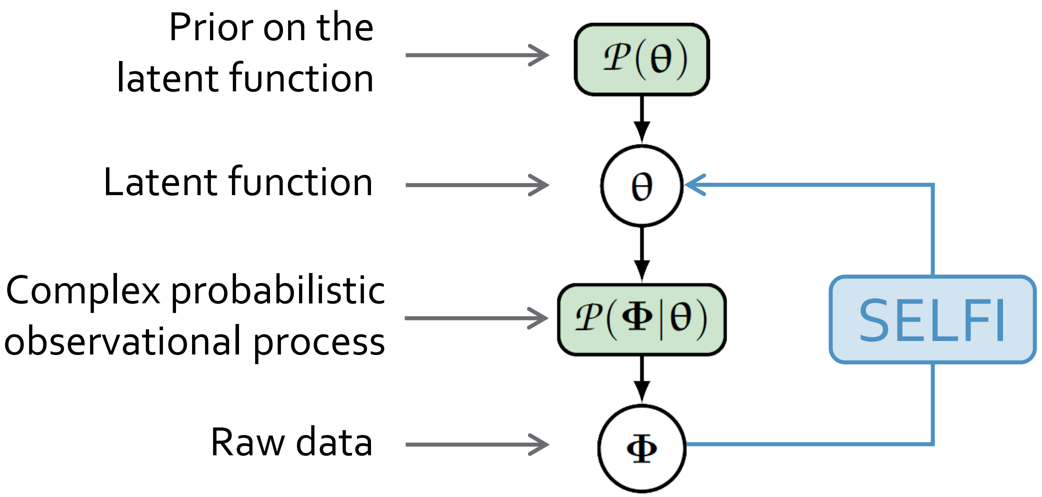


1. Inference of the latent function θ, to check for model misspecification:
   - SELFI algorithm

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.

Step 1: latent function inference:
The SELFI approach (*Simulator Expansion for Likelihood-Free Inference*)



- Linearisation of the black-box data model:

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

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

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

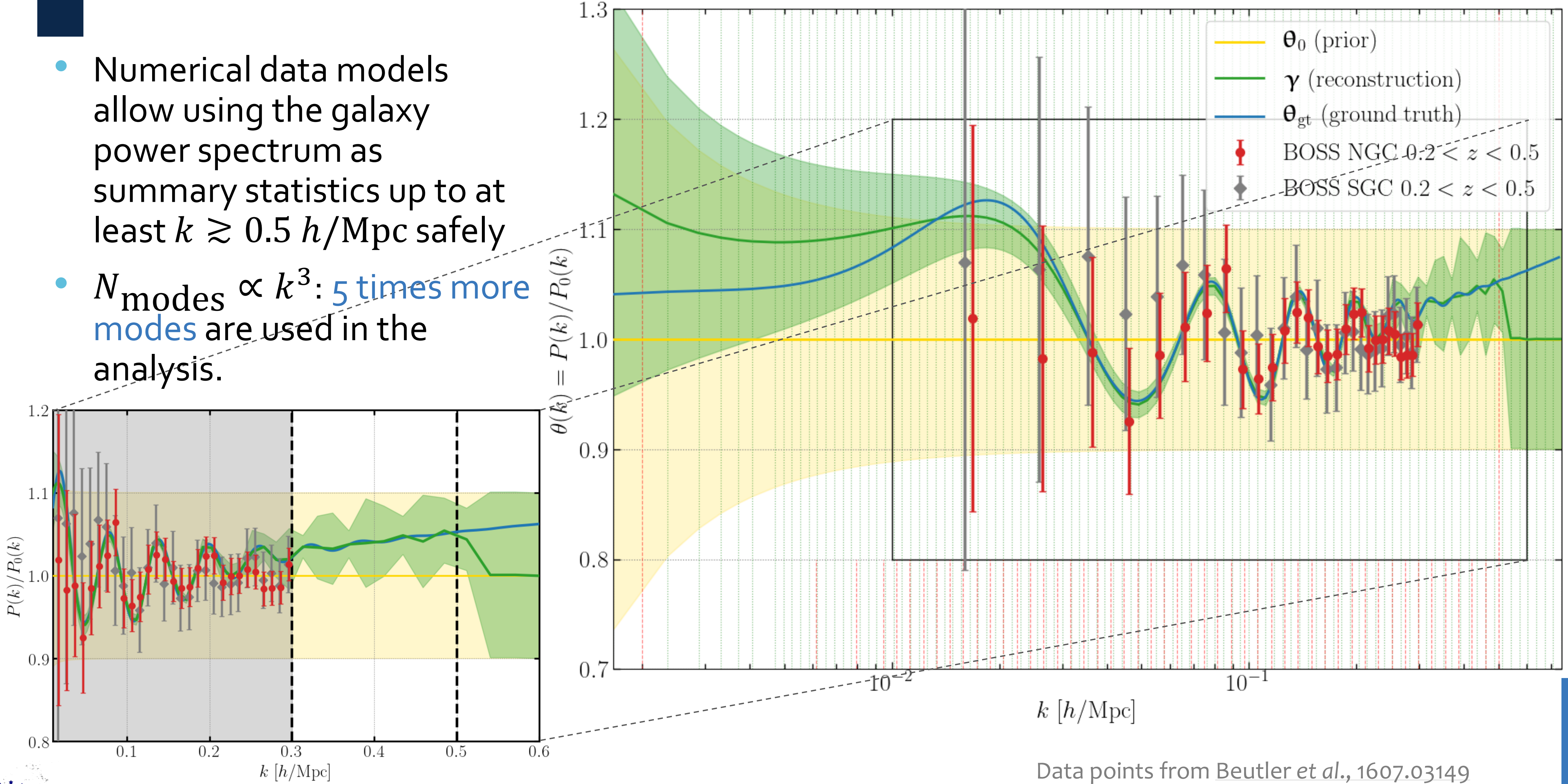
covariance of summaries — gradient of the black-box — prior covariance

- $\mathrm{f}_0, \mathbf{C}_0$ and $\nabla\mathrm{f}_0$ 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 Euclid forecast (cosmic variance limit) vs BOSS

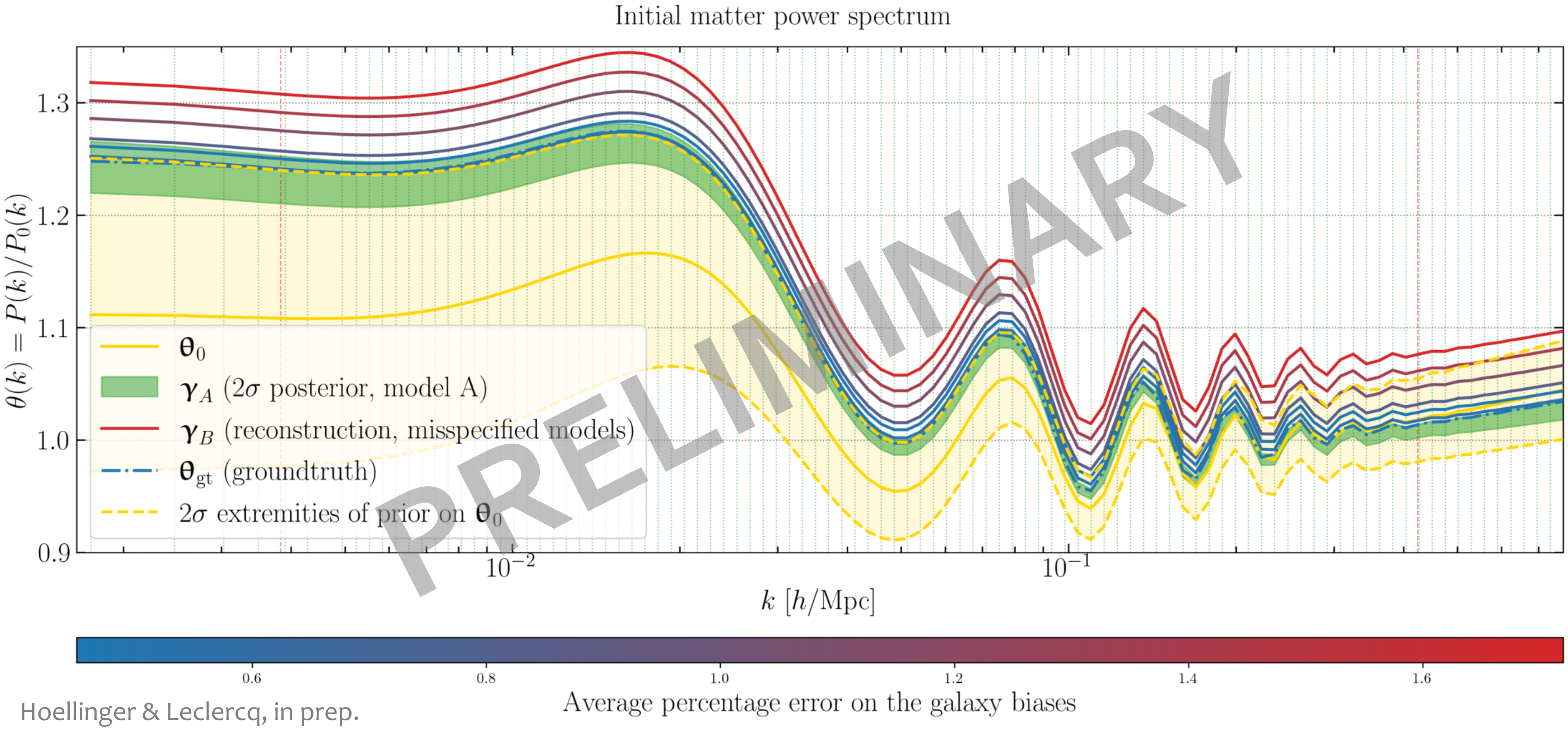
- Numerical data models allow using the galaxy power spectrum as summary statistics up to at least $k \gtrsim 0.5\ h/\mathrm{Mpc}$ safely
- $N_{\mathrm{modes}} \propto k^3$: 5 times more modes are used in the analysis.



Data points from Beutler *et al.*, 1607.03149

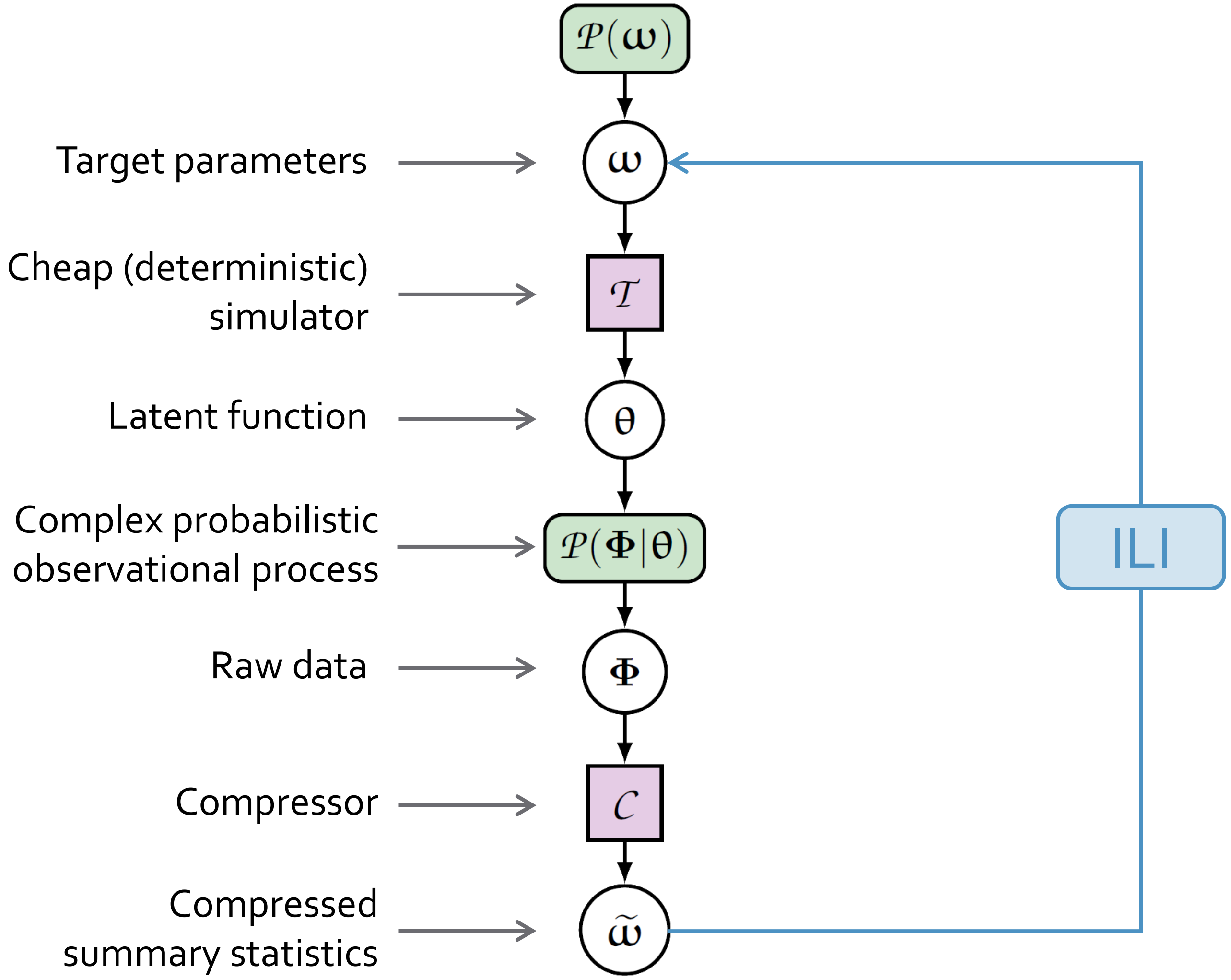
Checking for systematics in ILI problems with SELFI as a first step

- One can utilise the initial matter power spectrum to check for systematics.



Hoellinger & Leclercq, in prep.

Step 2: implicit likelihood inference of top-level target parameters



- The simulations used for step 1 can be recycled to write a free score compressor for step 2.
- 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 $\tilde{\omega}_\mathrm{O}$:

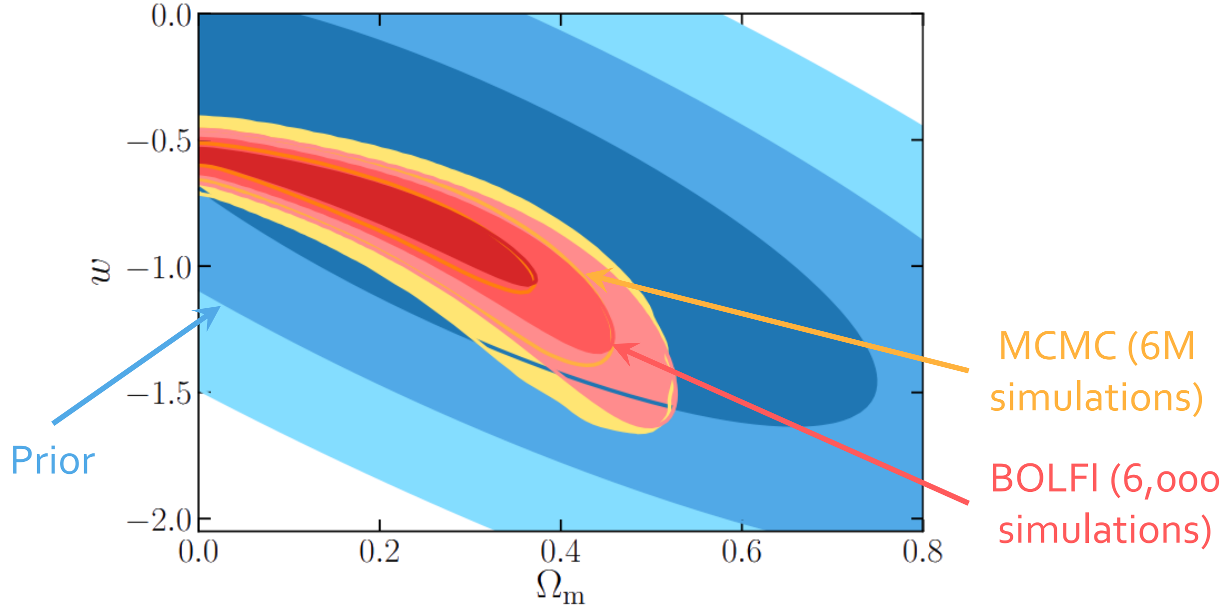
$$d_\mathrm{FR}(\tilde{\omega}, \tilde{\omega}_\mathrm{O}) \equiv \sqrt{(\tilde{\omega} - \tilde{\omega}_\mathrm{O})^\intercal \, \mathbf{F}_0 (\tilde{\omega} - \tilde{\omega}_\mathrm{O})}$$

FL, 2209.11057

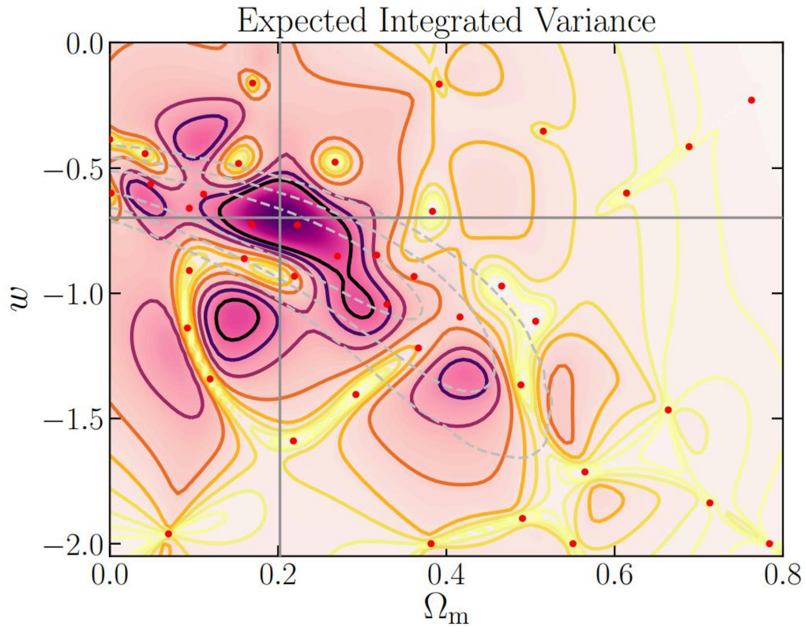
Dealing with expensive simulators in ILI problems:
The BOLFI algorithm (*Bayesian Optimisation for Likelihood-Free Inference*)

- The simulator will typically be extremely expensive (*N*-body simulation, halo finding, complex observational effects). We can typically afford O(10,000) evaluations.
- Emulation of the data model is not the only option.

Re-analysis of the JLA supernovae data:



- BOLFI (*Bayesian Optimisation for Likelihood-Free Inference*) uses an acquisition function to place expensive simulations in the parameter space.
- The optimal acquisition function for implicit inference can be derived: the Expected Integrated Variance.



FL, 1805.07152

Conclusion: a science-ready statistical framework for arbitrary probes of galaxy surveys

- A novel two-step implicit likelihood inference approach, combining SELFI and BOLFI, 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/BOLFI):
  - 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.
  - BOLFI uses active acquisition to deal with expensive simulators.

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

References:

- Leclercq 2018, 1805.07152, *Bayesian optimisation for likelihood-free cosmological inference*
- Leclercq *et al.* 2019, 1902.10149, *Primordial power spectrum and cosmology from black-box galaxy surveys*
- Leclercq 2022, 2209.11057, *Simulation-based inference of Bayesian hierarchical models while checking for model misspecification*
- Hoellinger & Leclercq, in prep.

https://pyselfi.florent-leclercq.eu: publicly available implementation of SELFI
https://aquila-consortium.org