Fast cosmological inference with iterative emulators

Supranta S. Boruah
University of Arizona

Cosmology from Home, 2022

Based on 2203.06124

In collaboration with Tim Eifler, Vivian Miranda, Sai Krishanth

THE UNIVERSITY OF ARIZONA

About me

- Current: Postdoc at University of Arizona
- Past: University of Waterloo, IIT Kanpur
- Research interest:
    - Fast inference methods using ML (This talk)
    - Bayesian field-level inference
    - Direct probes of peculiar velocity

Current and upcoming cosmological surveys

Surveys:

KIDS, HSC, eBOSS

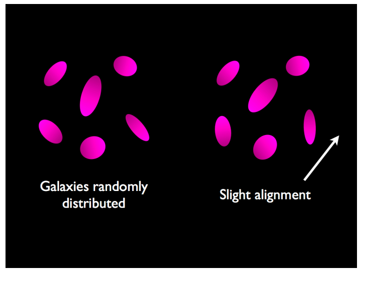
Stage-III

2020

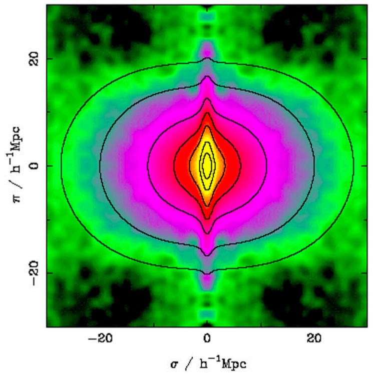
2024

2027

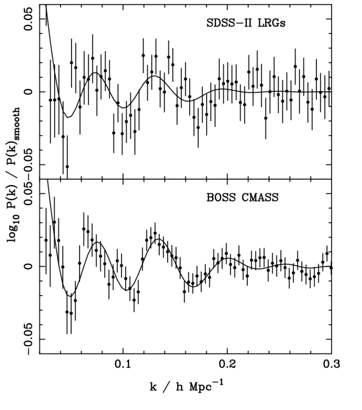
Cosmological probes:



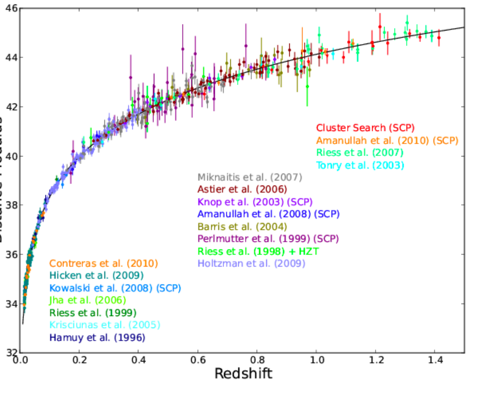
Weak Lensing



RSD



BAO



SNe

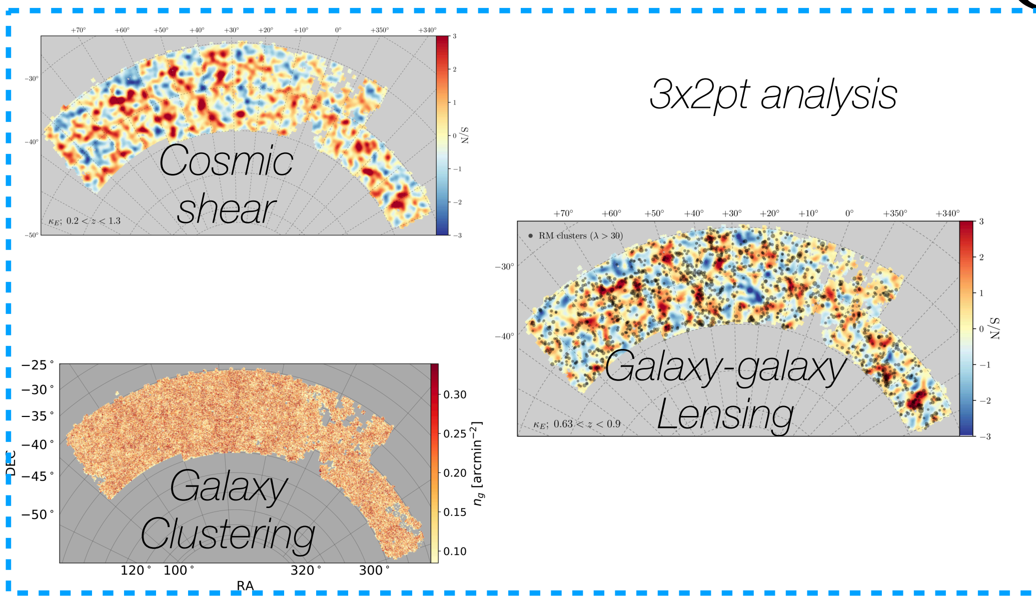
.....

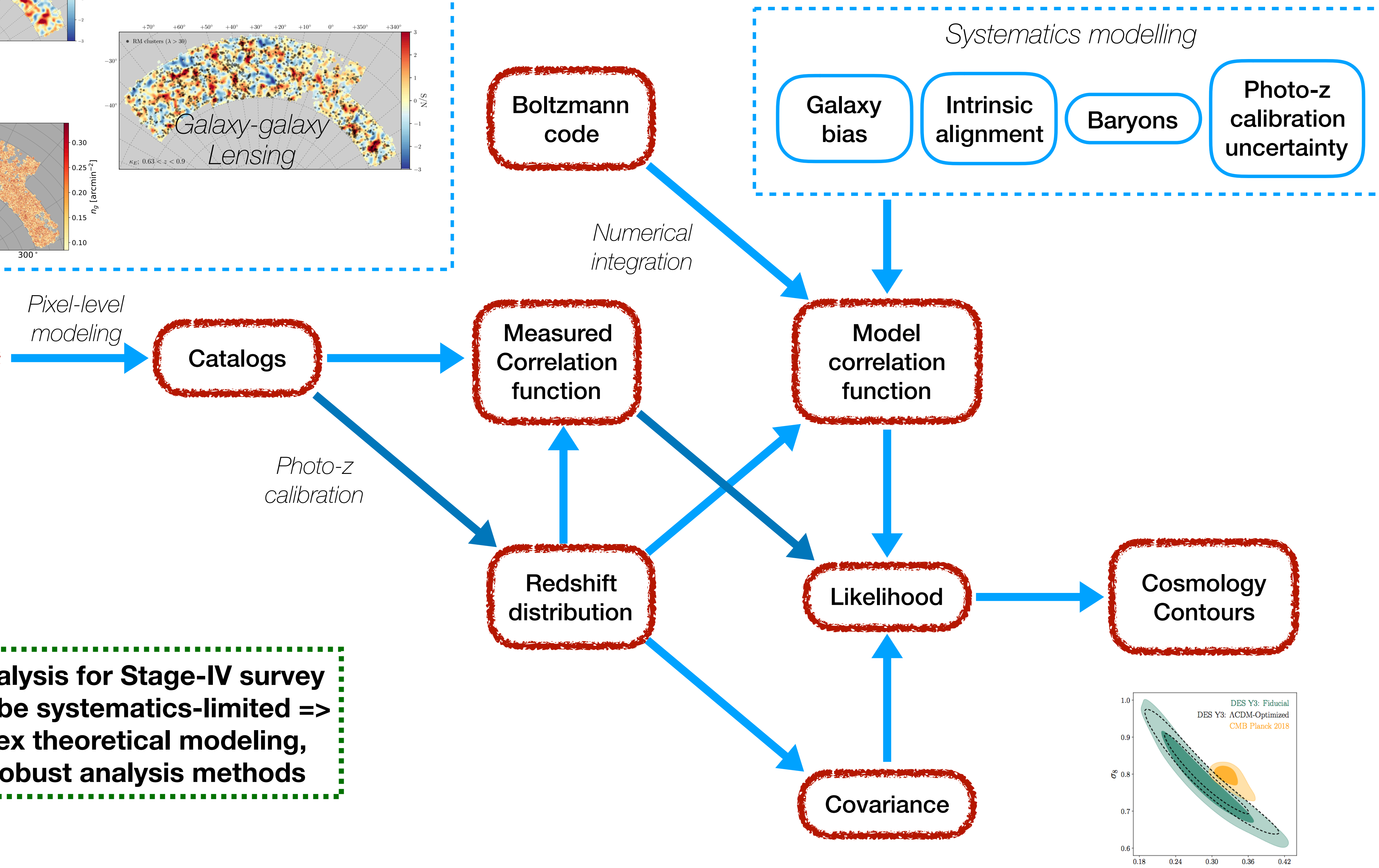
Science return:

- Nature of dark energy
- Mass of neutrino
- Primordial non-Gaussianity
- ....

Images: CFHTLens, NASA, SDSS, SCP,

Connecting theory to data



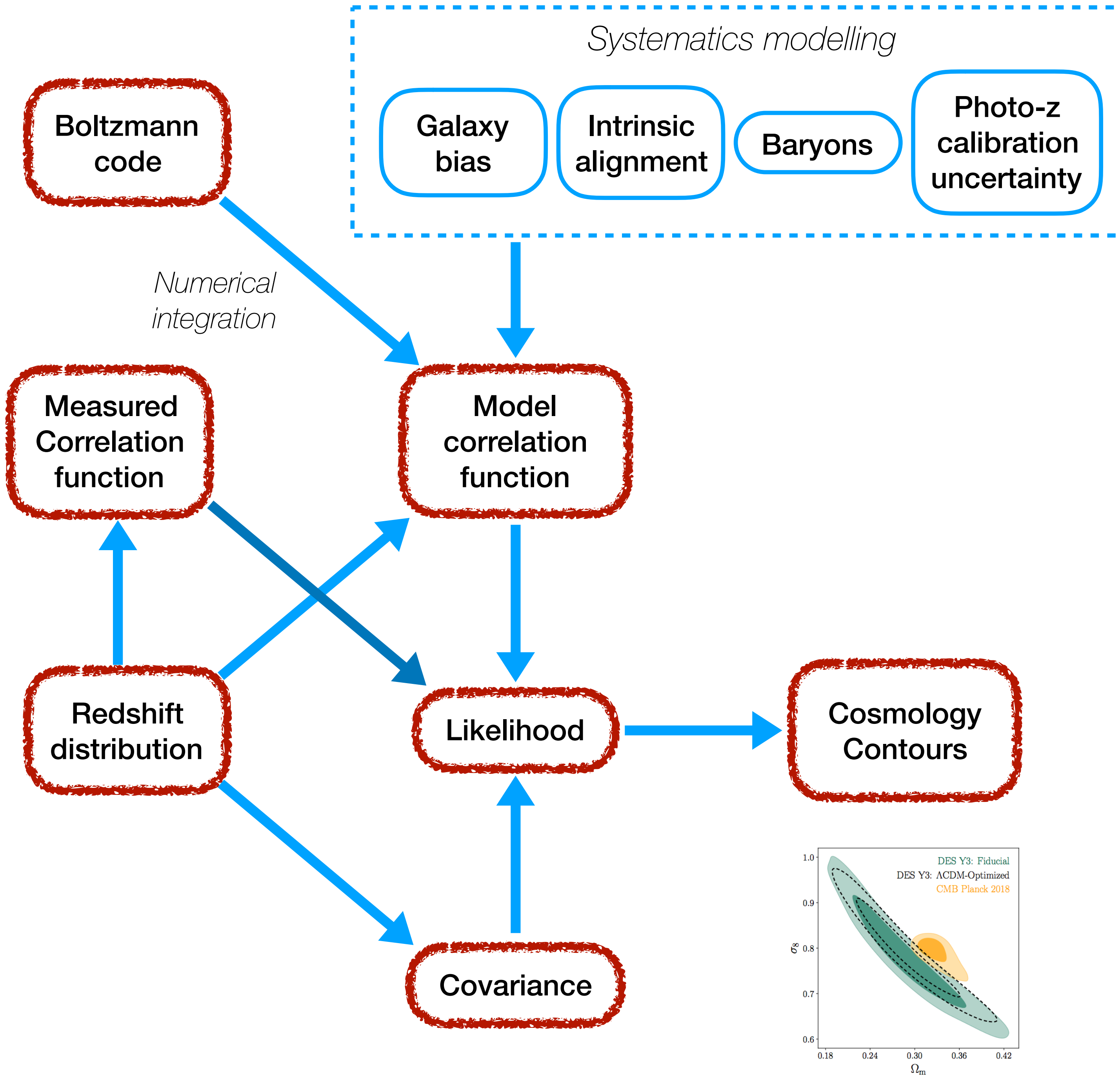


Data analysis for Stage-IV survey likely to be systematics-limited => complex theoretical modeling, Need robust analysis methods

Images: DES, NASA

Inference process

- Inference using MCMC or nested sampling algorithms
- MCMC runs are expensive, e.g, each LSST 3x2 point analysis require $\mathcal{O}$(10k CPUh)
- Robust data analysis will necessitate **thousands** of simulated MCMC runs



Images: DES, NASA

Simulated likelihood analyses

Quantifying the impact of different systematic effects

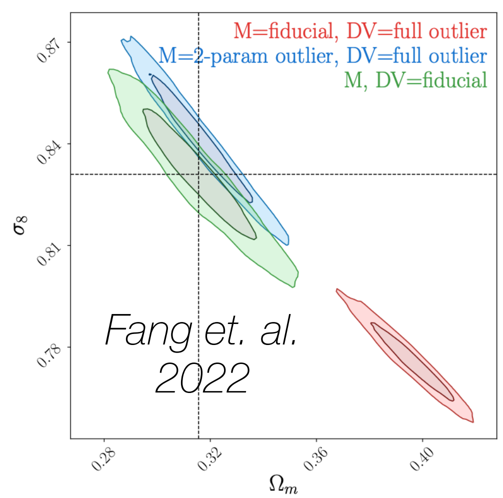
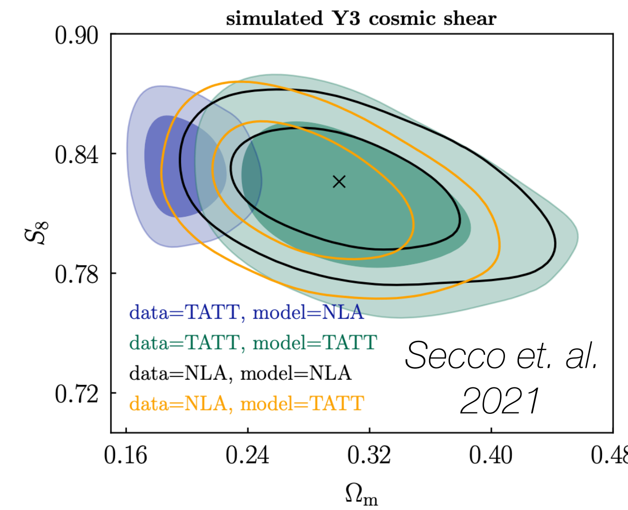
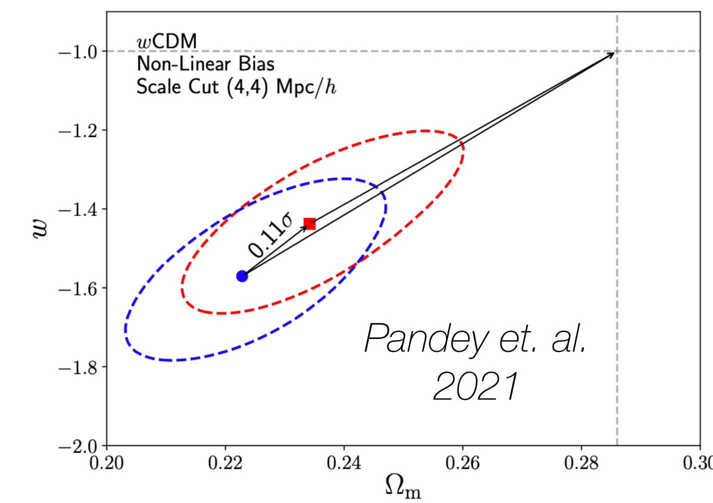


Photo-z bias/outliers

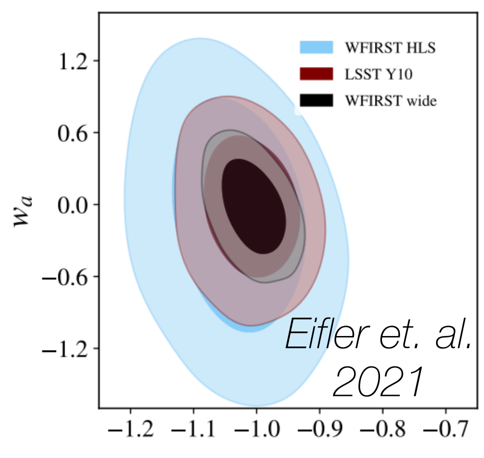


Intrinsic alignment



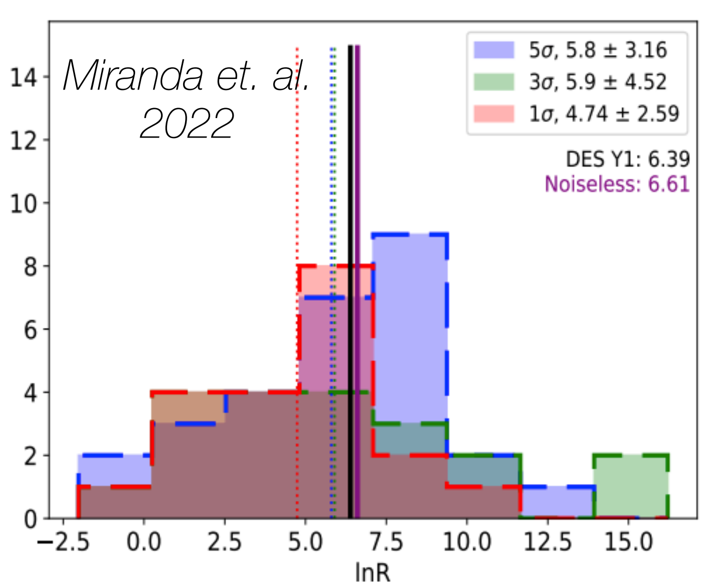
Galaxy bias

Optimizing survey strategy



Different survey strategy (e.g, deep / wide) impacts the science return.

Assessing tensions

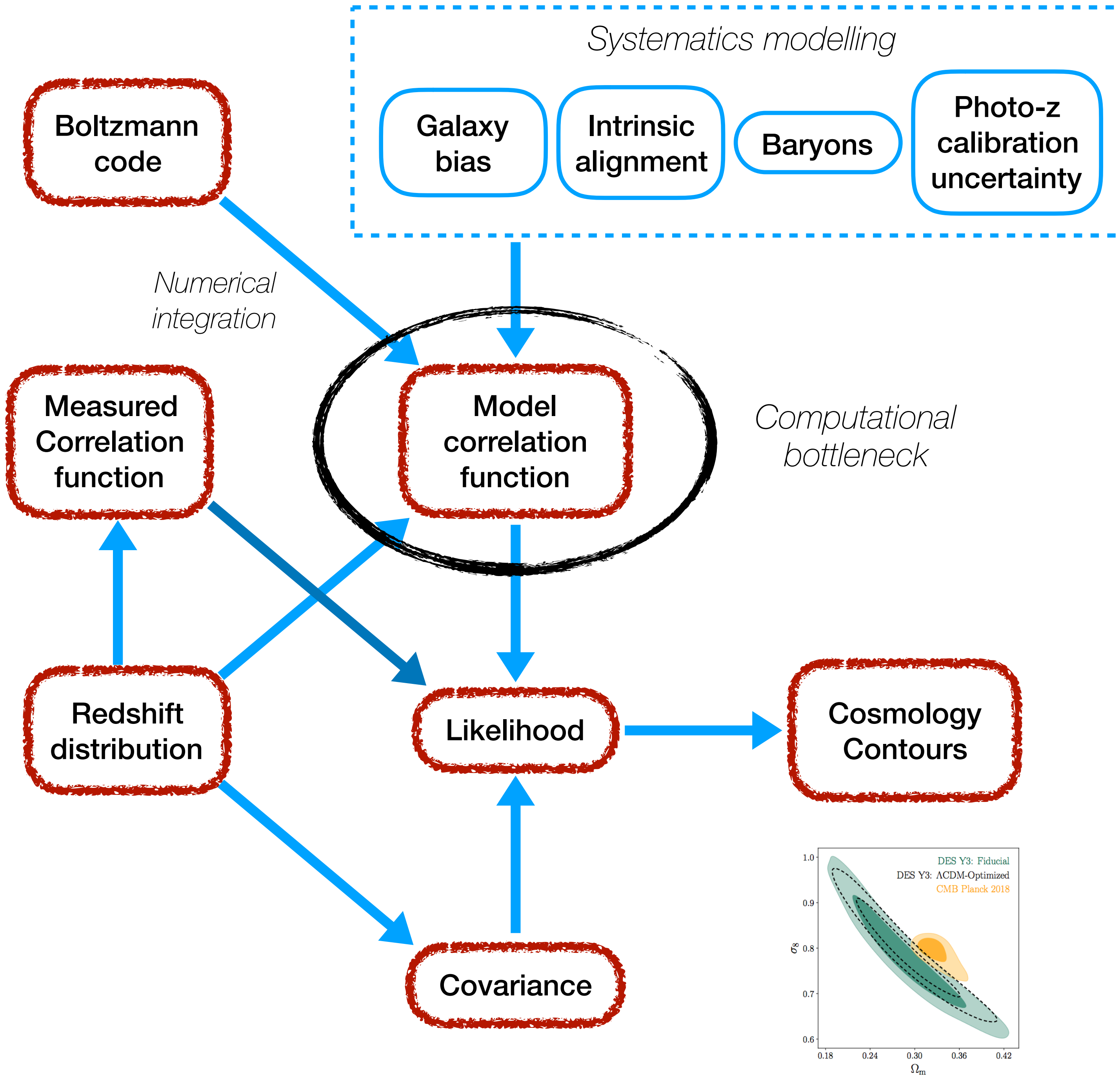


Each of these require hundreds/thousands of MCMC runs.

Computational cost prohibitive.

Challenge I: slow evaluation of model data vector

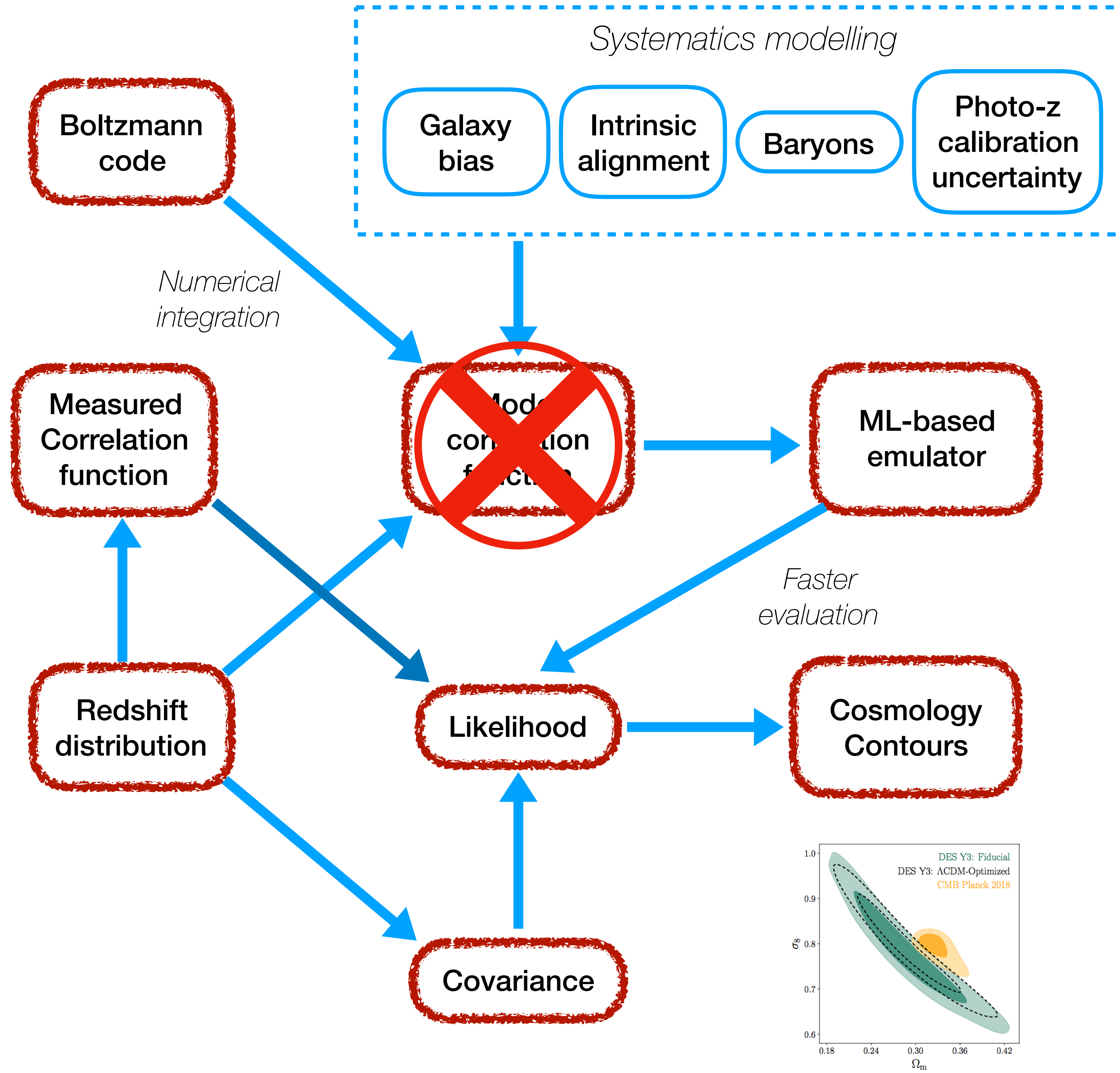
- Boltzmann codes / numerical integration are expensive.
- $\mathcal{O}(10^5 - 10^6)$ evaluations required in each MCMC analysis.



Images: DES, NASA

Challenge I: slow evaluation of model data vector

- Boltzmann codes / numerical integration are expensive.
- $\mathcal{O}(10^5 - 10^6)$ evaluations required in each MCMC analysis.
- _Alternative_: Use an emulator to speed-up data vector evaluation.
- After the initial computation of the data vectors, each call in MCMC very fast



Images: DES, NASA

Challenge II: Emulation in high dimensions

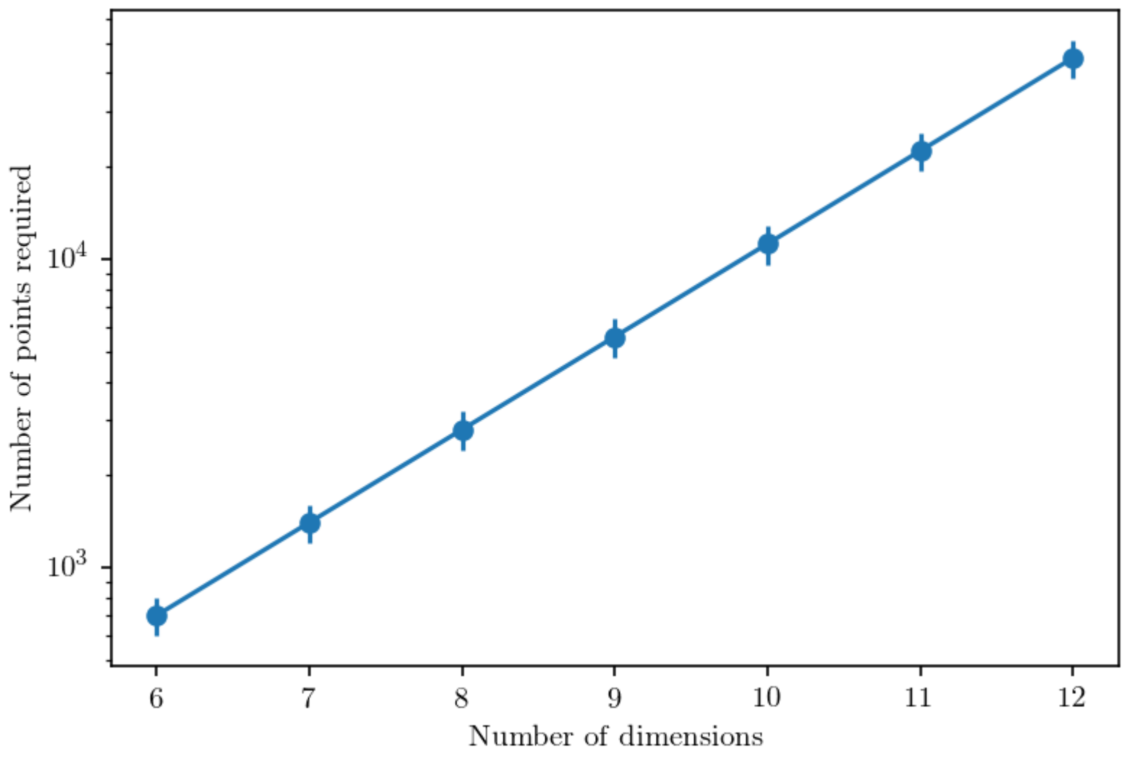
- Stage-IV analyses systematics-limited => more complex systematics modeling

- $\sim \mathcal{O}(\gtrsim 50)$ dimensional inference

- High dimensional emulation is hard (Curse of dimensionality)

- Numerical experiment:
  - Create LH sample within $[-4\sigma, 4\sigma]$ of a D-dim Gaussian
  - Determine # of points required to sample to get 50 points within 3σ ball

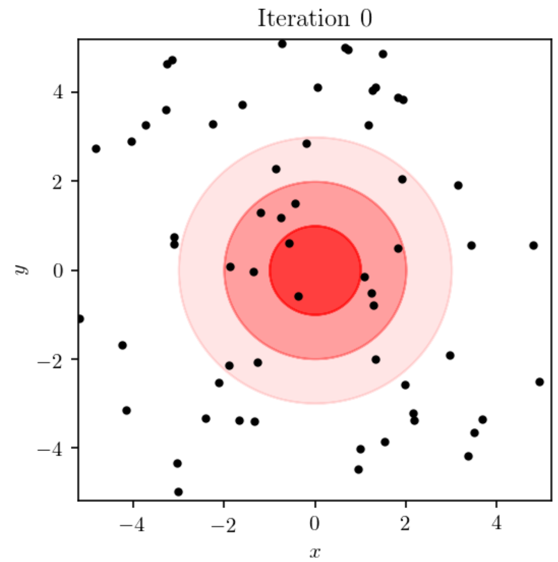
Emulation in high dimension is non-trivial!

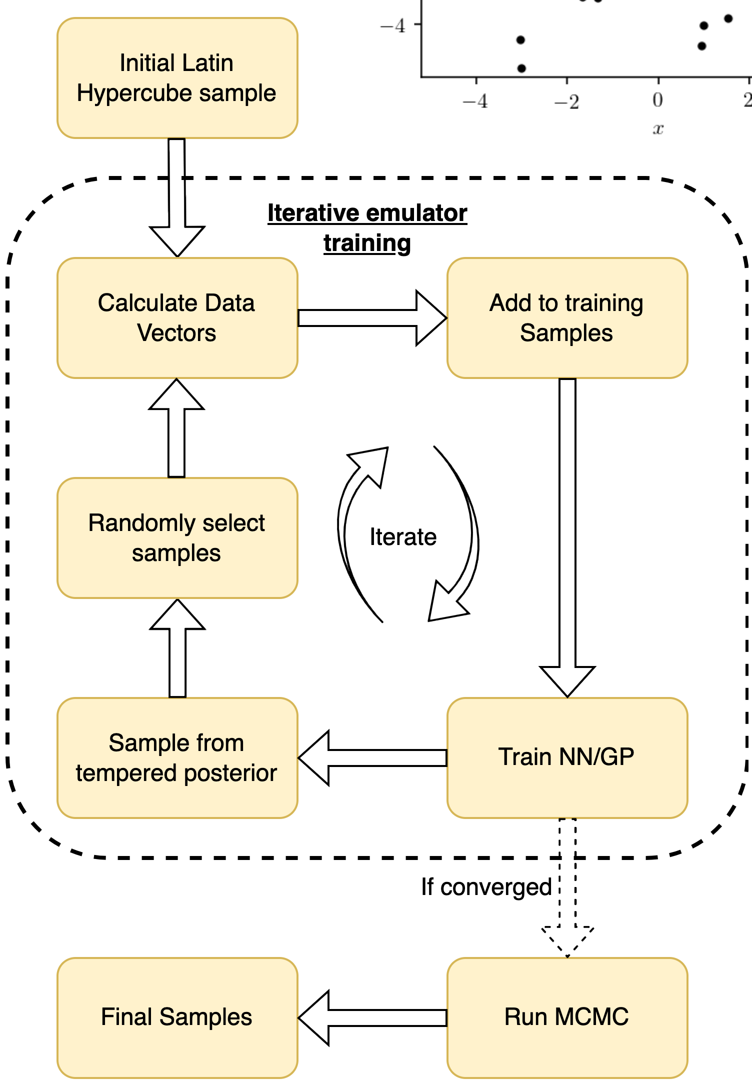


Need:

- $\mathcal{O}(10^5)$ *points for 12 dimensions*
- $\mathcal{O}(10^7)$ *points for 20 dimensions*

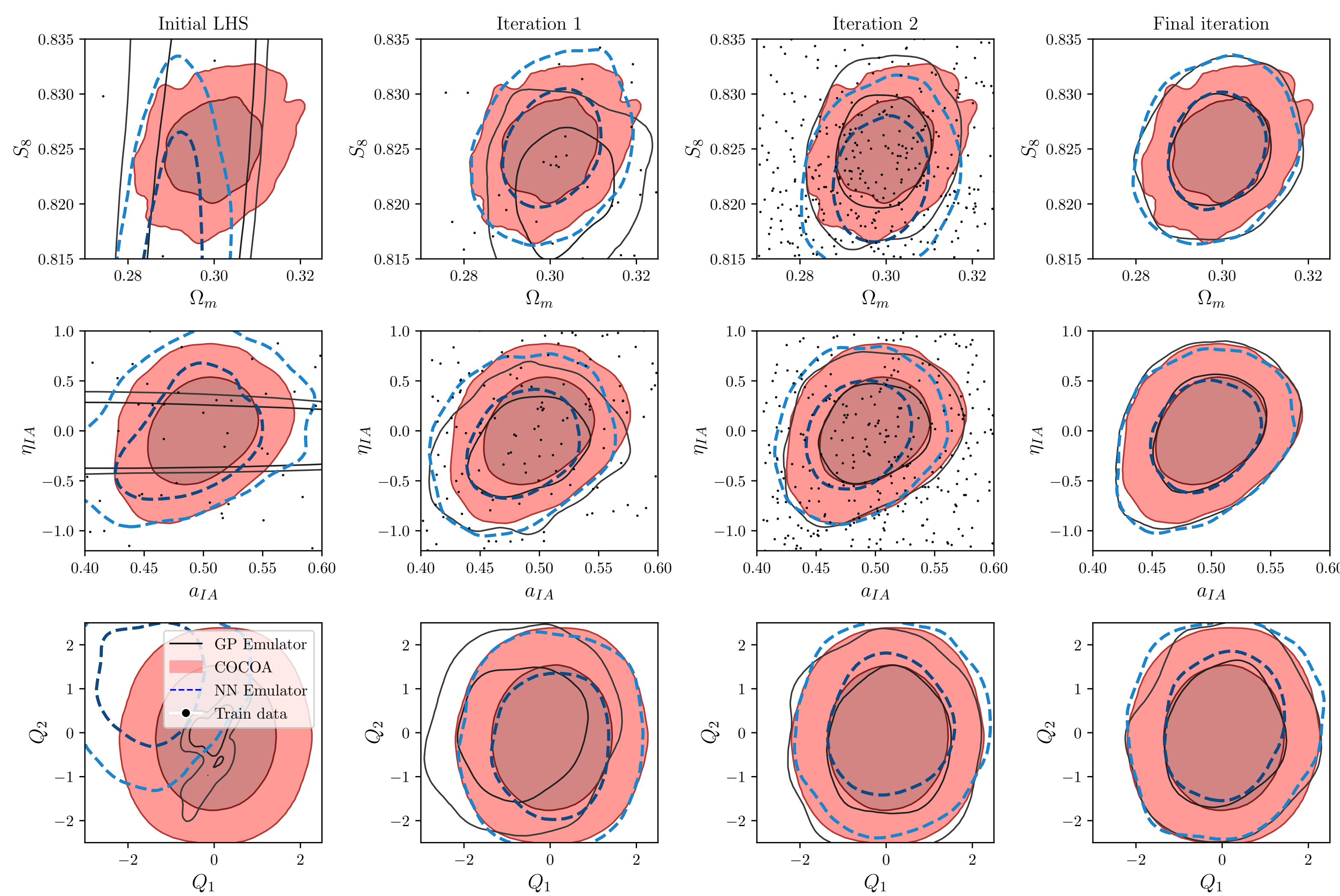
Iterative Emulator to the rescue



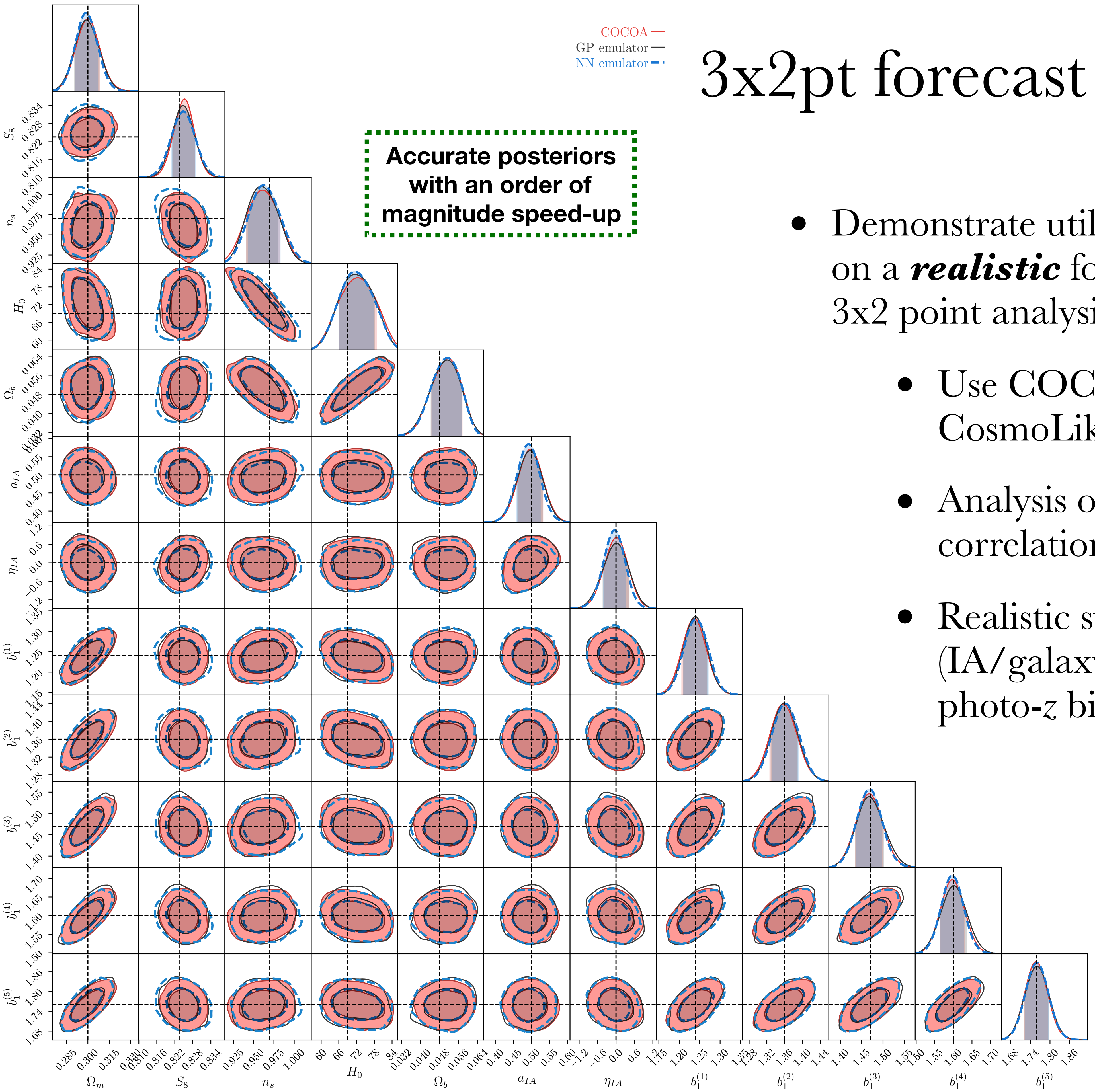


Motivation: **For inference, we only need high-fidelity emulation in the high posterior region**

Idea: **Use an iterative emulator to focus into the high posterior region, adding samples from the high posterior region**



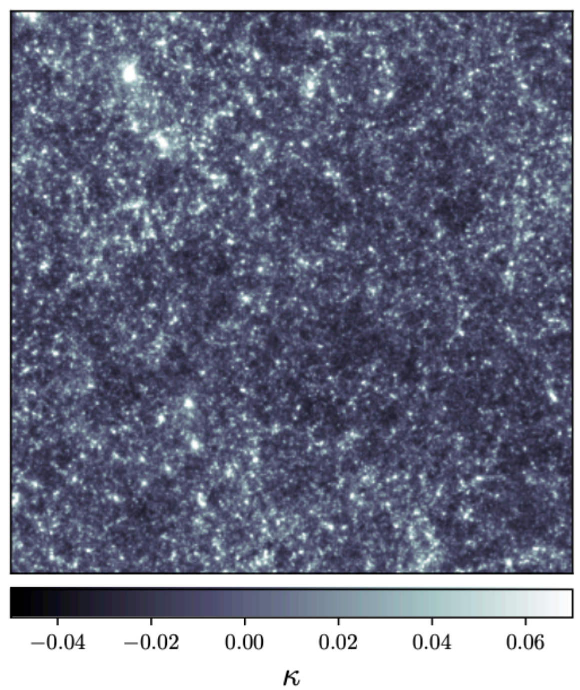
3x2pt forecast for LSST-Y1



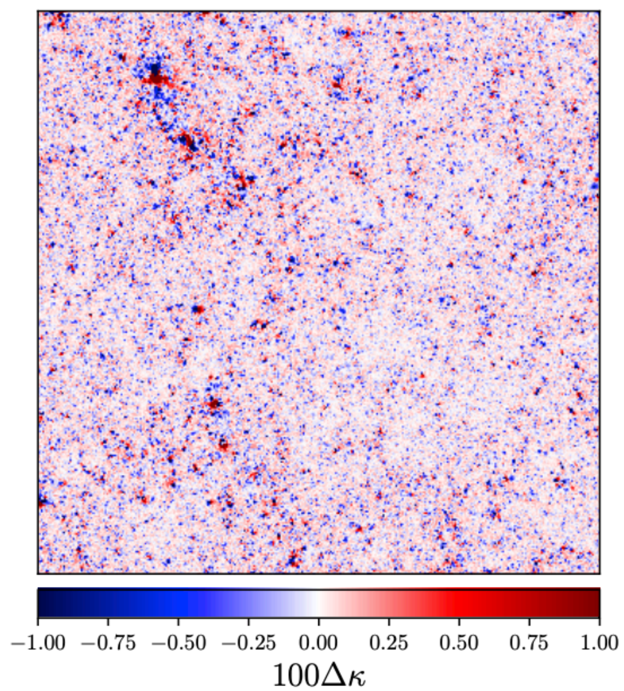
- Demonstrate utility of the emulator on a ***realistic*** forecast of LSST-Y1 3x2 point analysis:
    - Use COCOA (COBAYA/ CosmoLike joint architecture)
    - Analysis of real space correlation functions
    - Realistic systematics modeling (IA/galaxy bias/baryons/ photo-*z* bias)

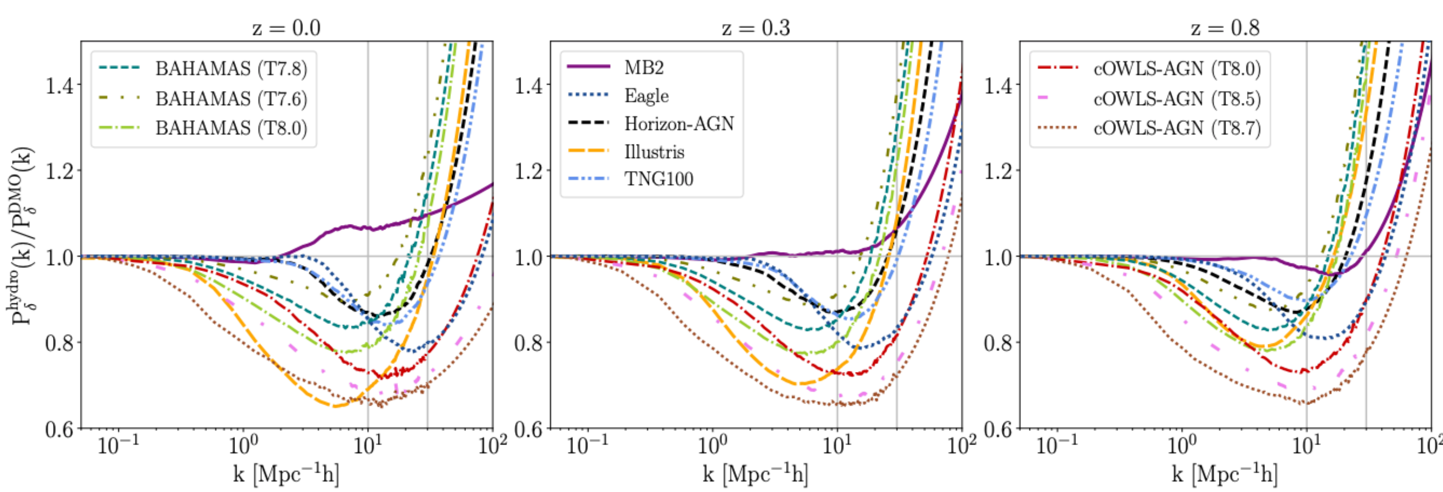
Case study: Impact of baryons on 3x2pt analysis

KappaTNG mass maps



Osato, Liu, '21



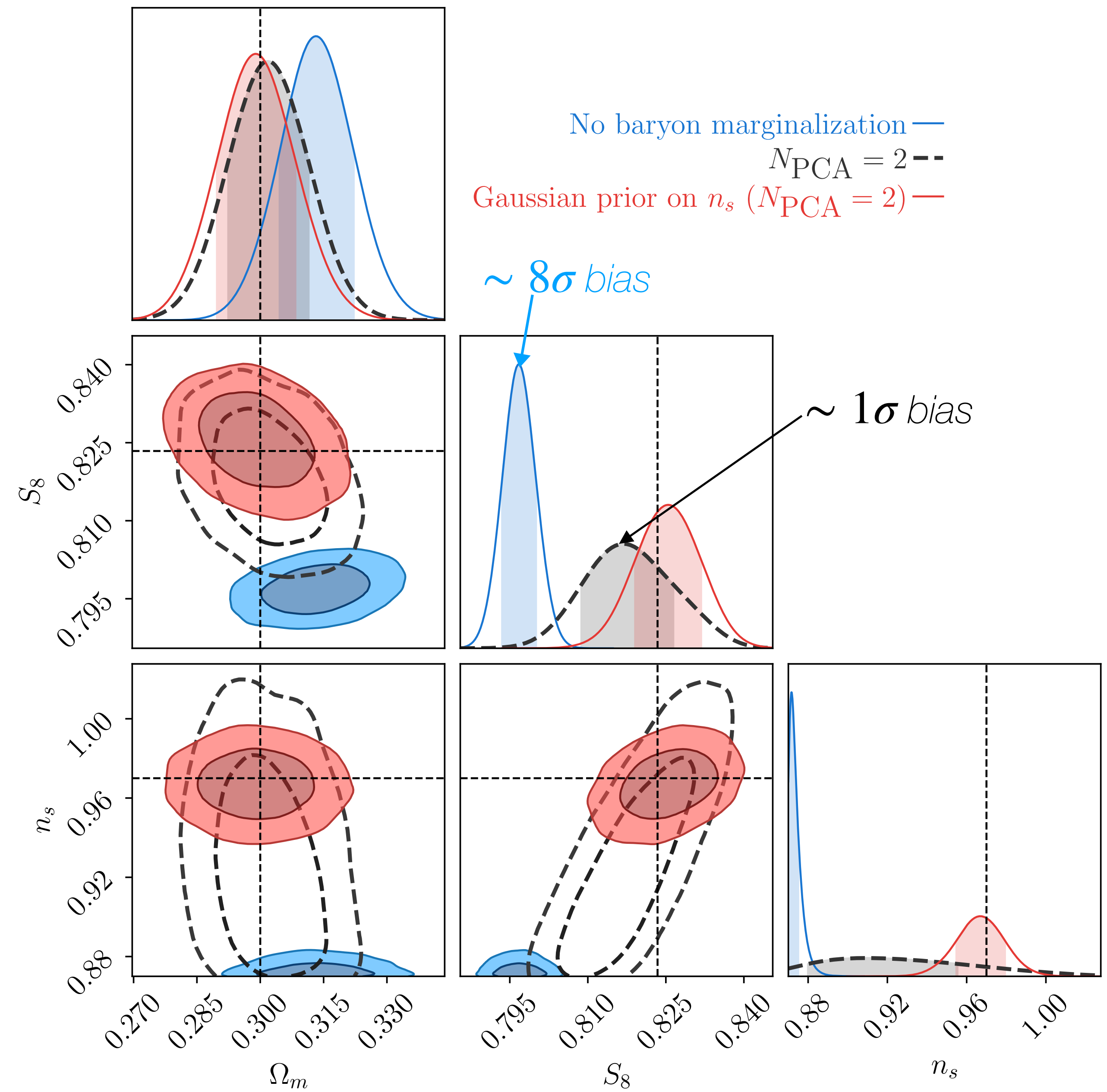


Huang+, 21

- Baryons impact the distribution of matter on small scales
- Usually mitigated by cutting small scales in the analysis
- Need simulated analyses to determine scale cuts/ analysis choice to mitigate baryonic effects

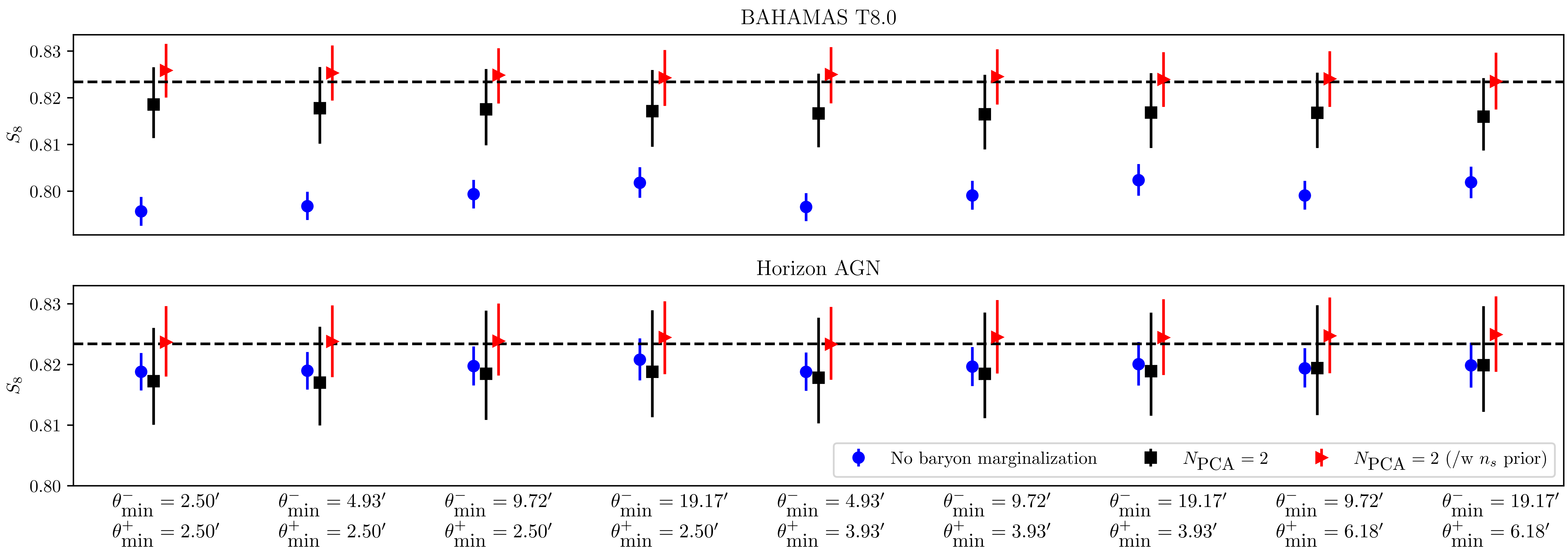
Case study: Impact of baryons on 3x2pt analysis

- Study different scale-cuts by reusing trained emulator
- Use scales up to 2.5 arcmin
- Highly biased for BAHAMAS (T8.0) w/o any mitigation scheme
- Mitigated using a PCA marginalization approach
- Impose prior on n_s to eliminate residual biases



Baryon marginalization assuming BAHAMAS (T8.0) contaminated data vector

(Extremely) Fast scale-cut analyses



- 5 minutes wall-time per analysis. $\mathcal{O}(1000)$ speed-up!
- Need one trained emulator for: *i)* Different scale cuts, *ii)* Different priors, *iii)* Different observed data vector, *iv)* Sub-space of parameter space.

Summary

- Challenges for data analysis of upcoming surveys:
  - ✴ Computation cost for inference is prohibitive
  - ✴ Emulation in high-dimensions is non-trivial
- We designed an iterative scheme that leads to fast and accurate inference. Demonstrated on LSST-Y1 3x2 point forecast.
- Ideal for:
  - ✴ Quantifying impact of different systematics
  - ✴ Optimizing survey strategy
  - ✴ Assessing tensions