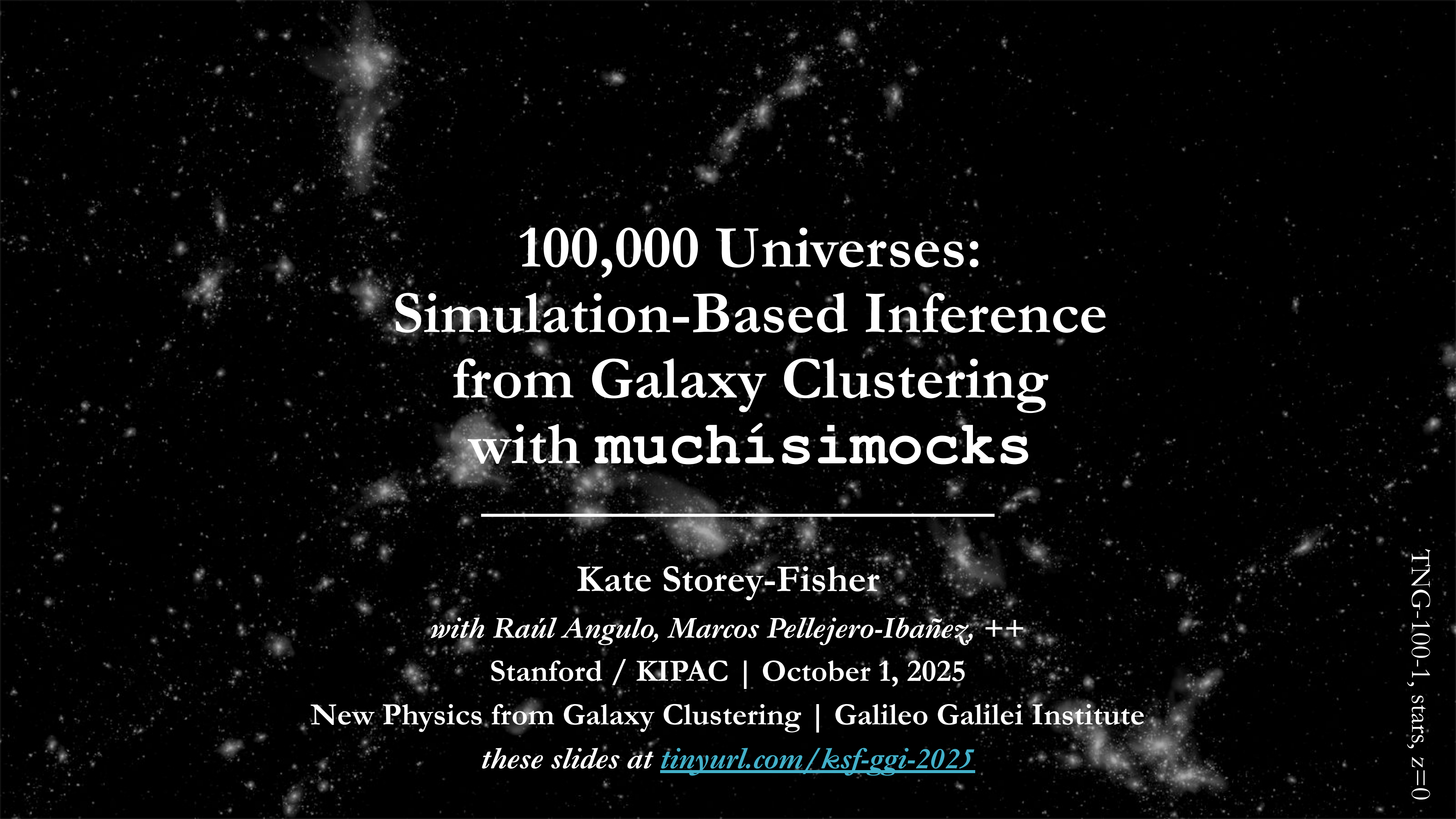


TNG-100-1, stars, z=0

100,000 Universes: Simulation-Based Inference from Galaxy Clustering with `muchísimocks`

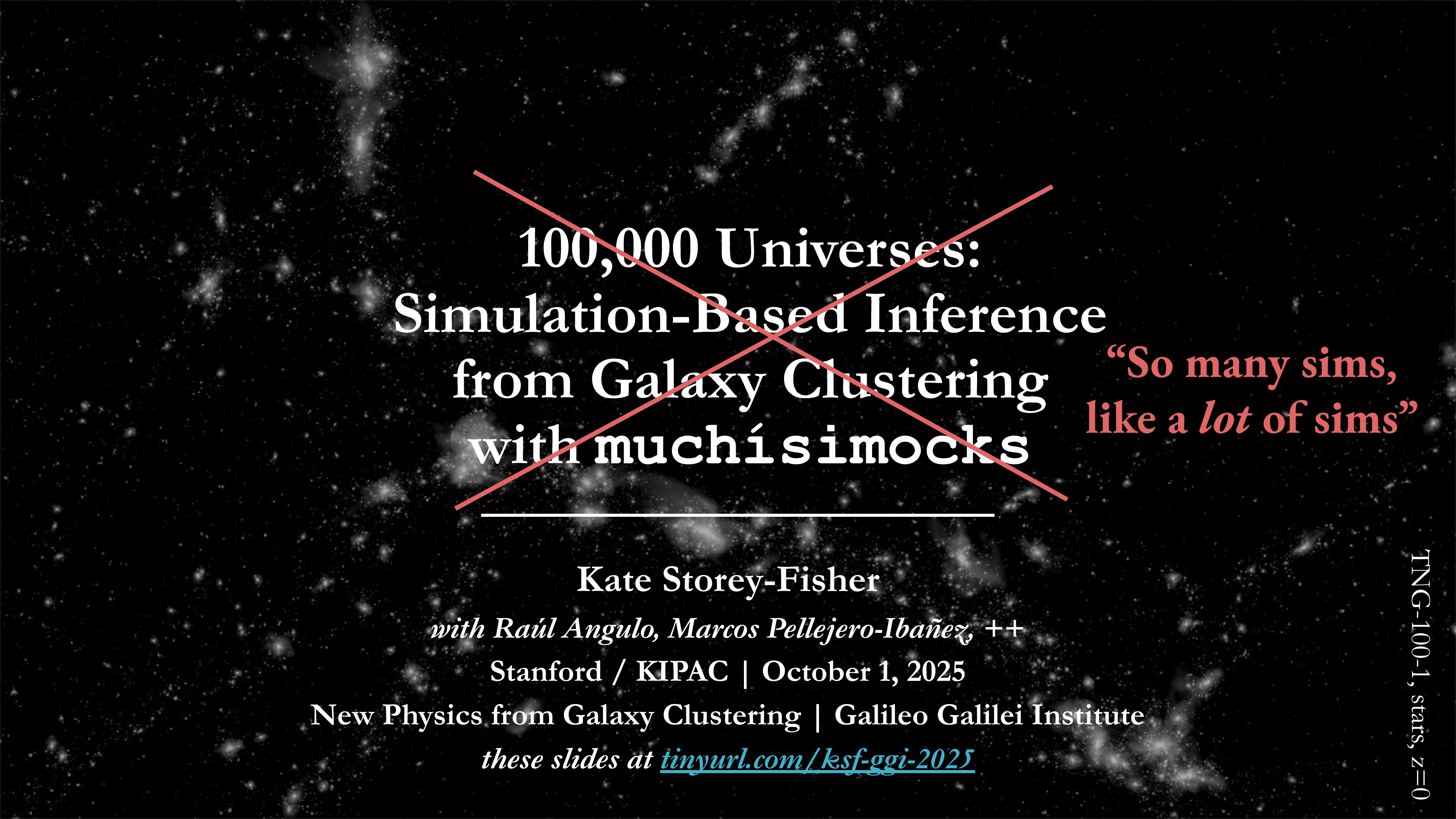
Kate Storey-Fisher

with Raúl Angulo, Marcos Pellejero-Ibañez, ++

Stanford / KIPAC | October 1, 2025

New Physics from Galaxy Clustering | Galileo Galilei Institute

these slides at tinyurl.com/ksf-ggi-2025



~~100,000 Universes: Simulation-Based Inference from Galaxy Clustering with `muchísimocks`~~

“So many sims, like a *lot* of sims”

Kate Storey-Fisher

with Raúl Angulo, Marcos Pellejero-Ibañez, ++

Stanford / KIPAC | October 1, 2025

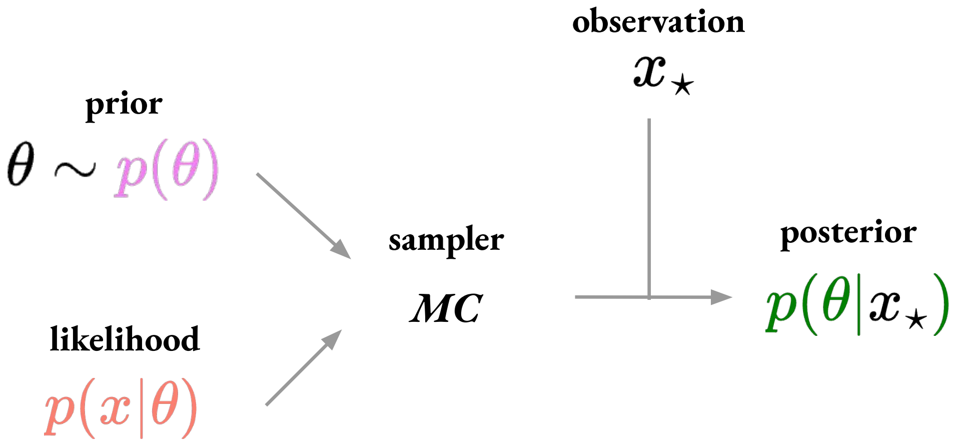
New Physics from Galaxy Clustering | Galileo Galilei Institute

these slides at tinyurl.com/ksf-ggi-2025

TNG-100-1, stars, z=0

Likelihood-based inference

- Requires theoretical modeling of statistics
- Hard to model high-order statistics
- Hard to model impact of systematics
- Typically assumes Gaussian likelihood



observation

$x_\star$

prior

$\theta \sim p(\theta)$

sampler

MC

posterior

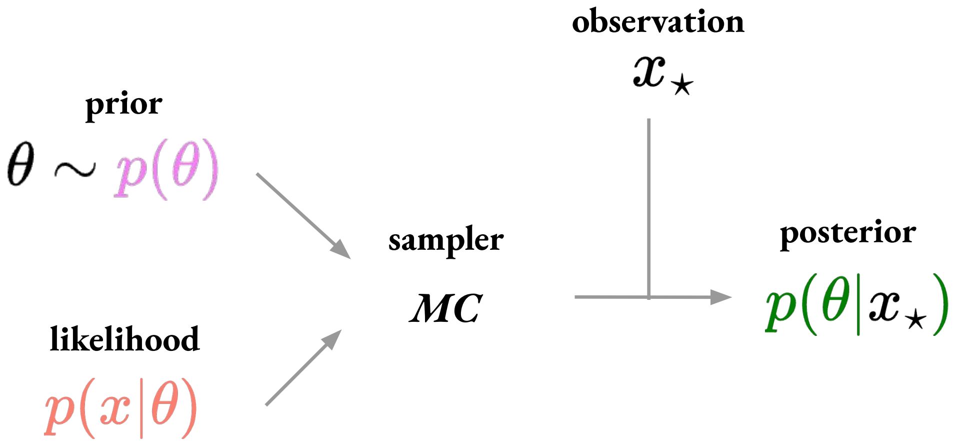
$p(\theta|x_\star)$

likelihood

$p(x|\theta)$

Likelihood-based inference

- Requires theoretical modeling of statistics
- Hard to model high-order statistics
- Hard to model impact of systematics
- Typically assumes Gaussian likelihood



observation

$x_\star$

prior

$\theta \sim p(\theta)$

sampler

MC

posterior

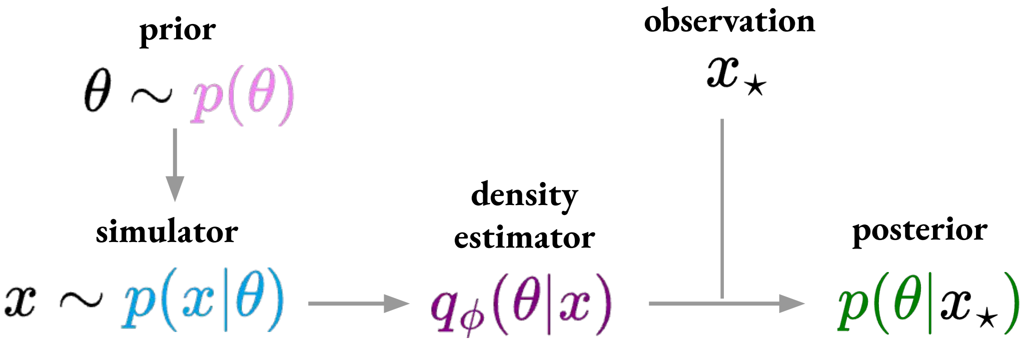
$p(\theta|x_\star)$

likelihood

$p(x|\theta)$

Simulation-based inference

- Only requires forward model of data
- Naturally good for high-order statistics
- Can forward-model systematics
- Does not assume Gaussian likelihood



prior

$\theta \sim p(\theta)$

observation

$x_\star$

simulator

$x \sim p(x|\theta)$

density estimator

$q_\phi(\theta|x)$

posterior

$p(\theta|x_\star)$

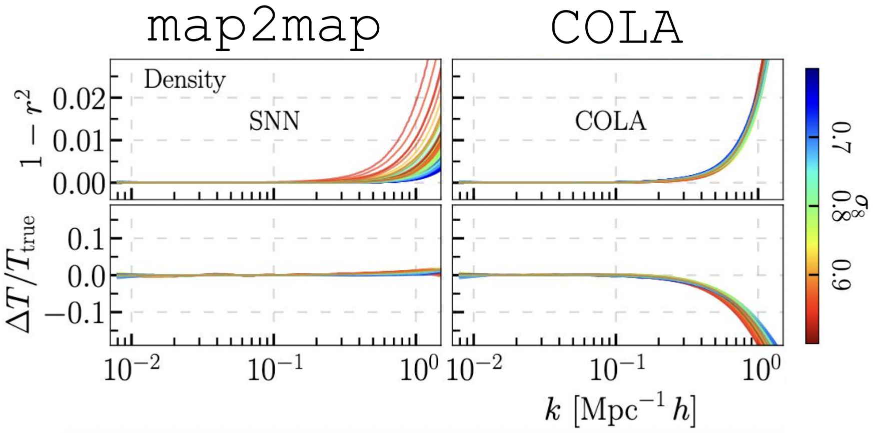
SBI for galaxy clustering limited by cost of training simulations

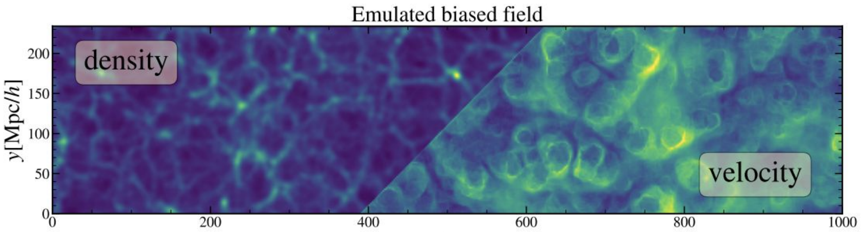
The `muchisimocks` forward model: the matter distribution

Jamieson+2022 (2206.04594)

We use the **`map2map` full-field emulator** (Jamieson+2022): Predicts the *N*-body displacement and velocity fields from an initial ZA approximation.

- Trained on the `Quijote` simulations: 1000 simulations, 5 cosmological parameters, 1 (Gpc/h)^3, 512^3 particles
- %-level accuracy on *P(k)* down to *k*=1 Mpc/*h*
- Takes ~4 minutes per emulated field





Pellejero-Ibañez+2023 (2307.09134)

The muchísimocks forward model: galaxy bias

Zennaro+2023
(2101.12187)

We use the **hybrid Lagrangian bias expansion** to model how galaxies populate the matter distribution:

Model galaxy density as function of matter density, expanding to 2nd order:

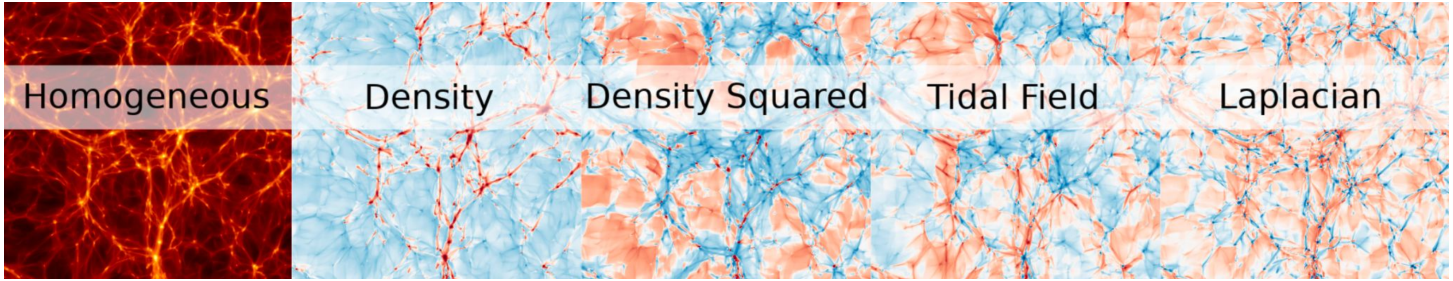
$$\delta_g(\boldsymbol{q}) = 1 + b_1\delta_L(\boldsymbol{q}) + b_2\left[\delta_L^2(\boldsymbol{q}) - \langle\delta^2\rangle\right] + b_{s^2}\left[s^2(\boldsymbol{q}) - \langle s^2\rangle\right] + b_\Delta\Delta^2\delta(\boldsymbol{q})$$

Advect Lagrangian to Eulerian space, using simulations to compute the displacement fie$\Psi(\boldsymbol{q})$:

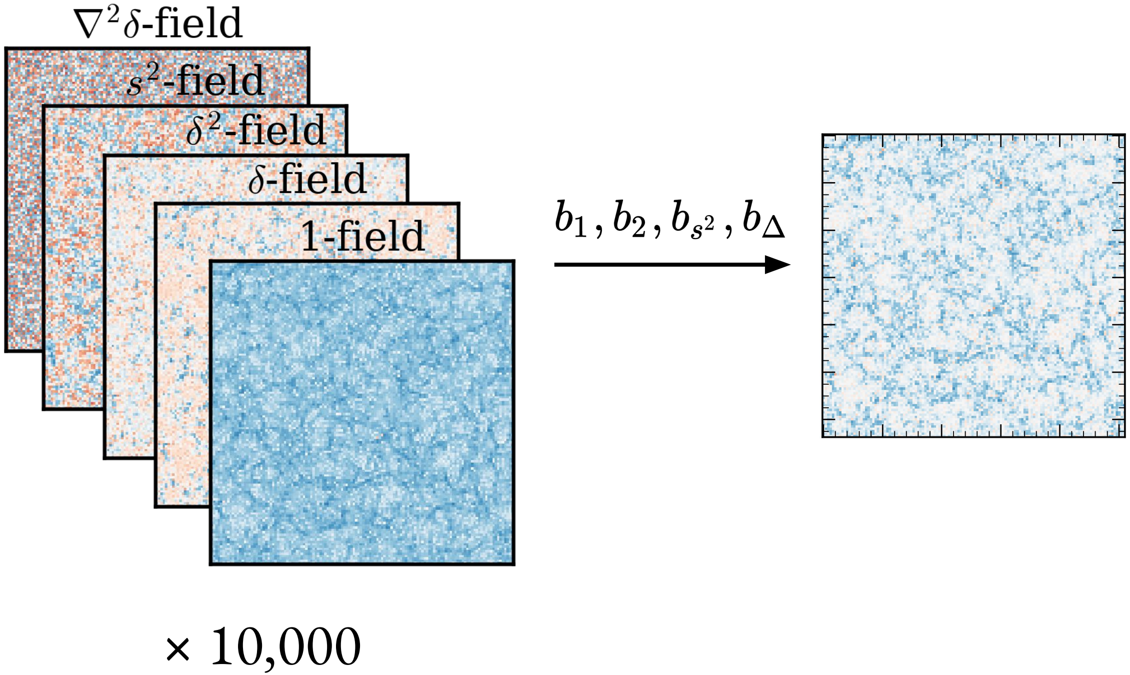
$$\boldsymbol{x} = \boldsymbol{q} + \Psi(\boldsymbol{q})$$

Our model parameters are the bias coefficients,

$$b_1, b_2, b_{s^2}, b_\Delta$$



`muchísimocks`: a library of 3D biased tracer fields



○ 10,000 sets of bias fields, with cosmologies spanning a Latin hypercube of σ_8, Ω_m, h, Ω_b

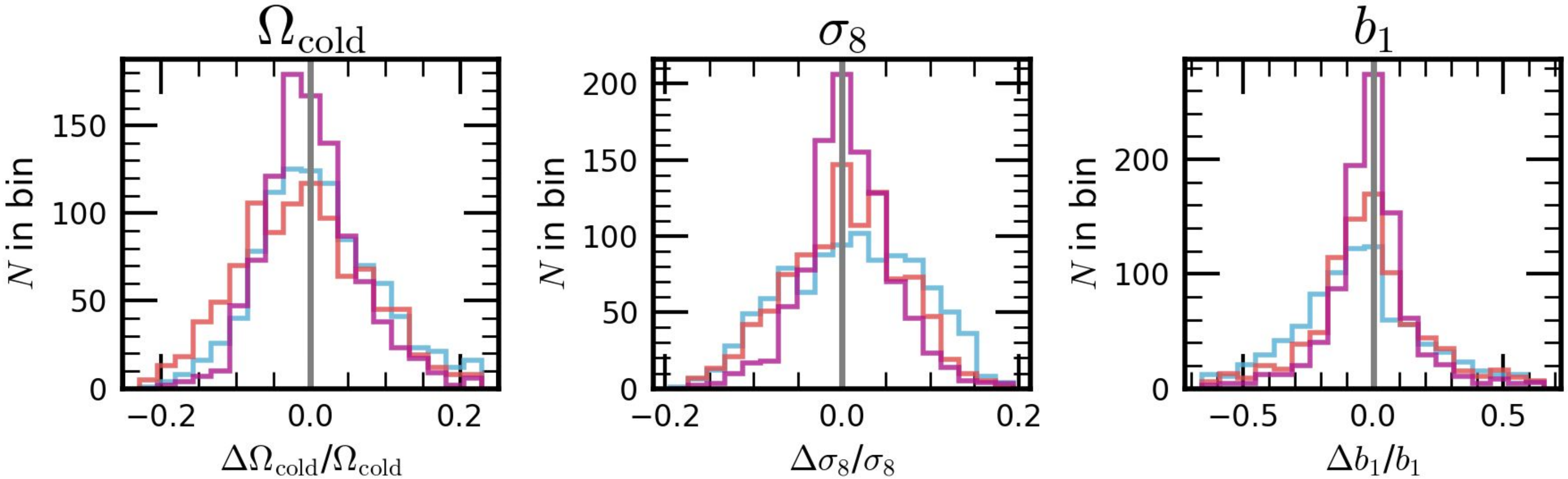
○ Can combine bias fields with choice of 2nd order bias parameters to construct density field

○ Each $(1\ h^{-1}\mathrm{Gpc})^3$, 512^3 resolution → cut small-k modes to get 128^3

○ ~1 TB of mocks (storage-limited); will be publicly available

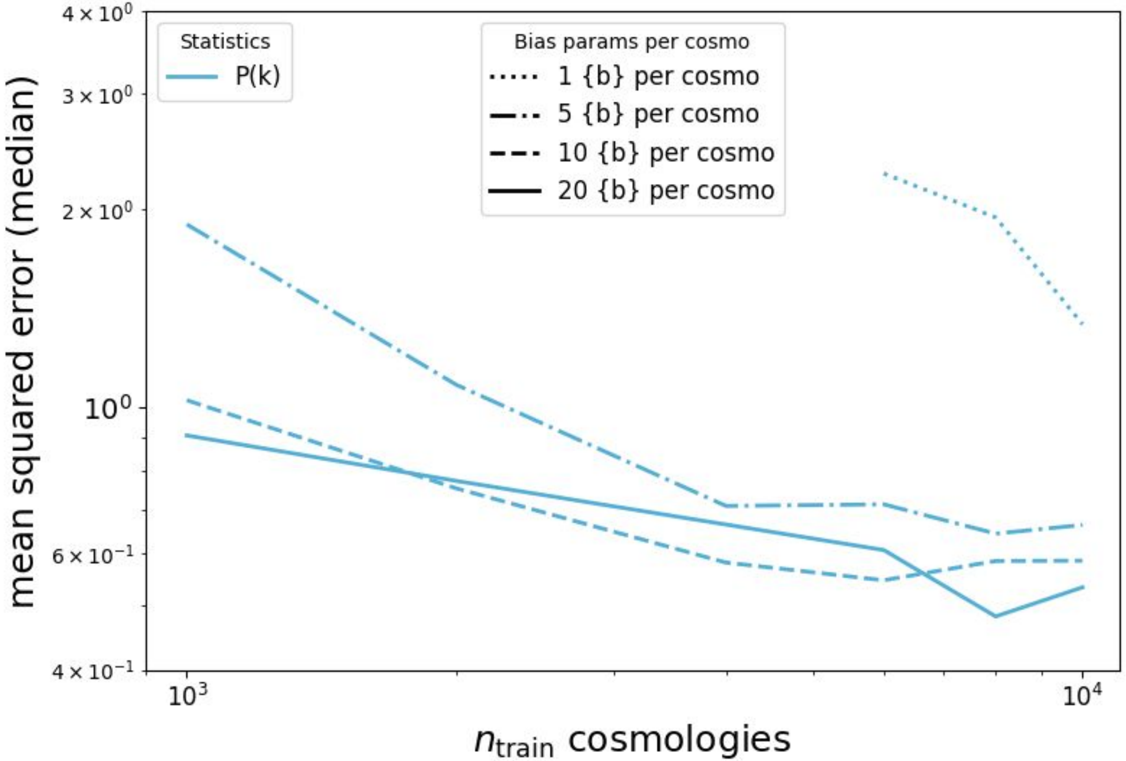
Simulation-based inference recovery test with *P(k)* & bispectrum

training: 10,000 cosmologies, 20x {*b*}
testing: 1,000 mocks spanning training set



Both the power spectrum and bispectrum give unbiased cosmological and bias parameters, and including the bispectrum tightens constraints by ~1.4x on Ω_m and ~2x on σ_8.

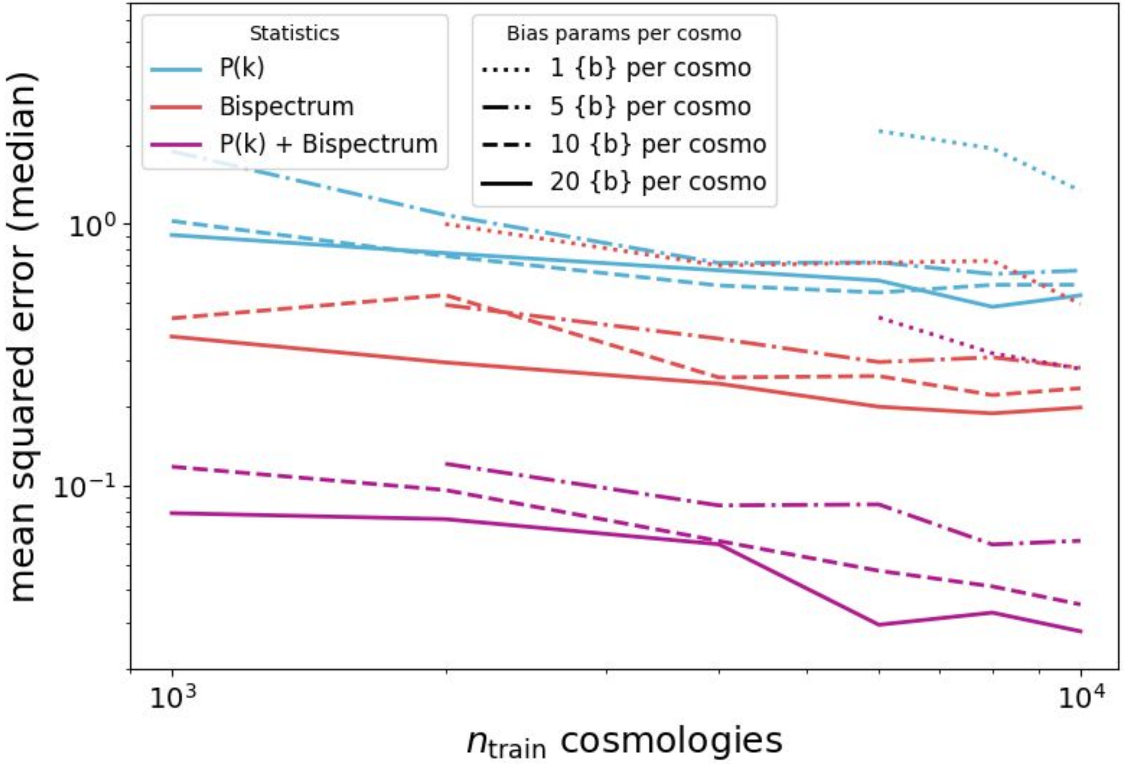
Inference results: Convergence with size of training data



The accuracy & precision improves significantly from 1→20 bias parameter sets per cosmology, and from 1k→10k cosmologies, with ~convergence around 6k cosmologies.

preliminary

Inference results: Convergence with size of training data

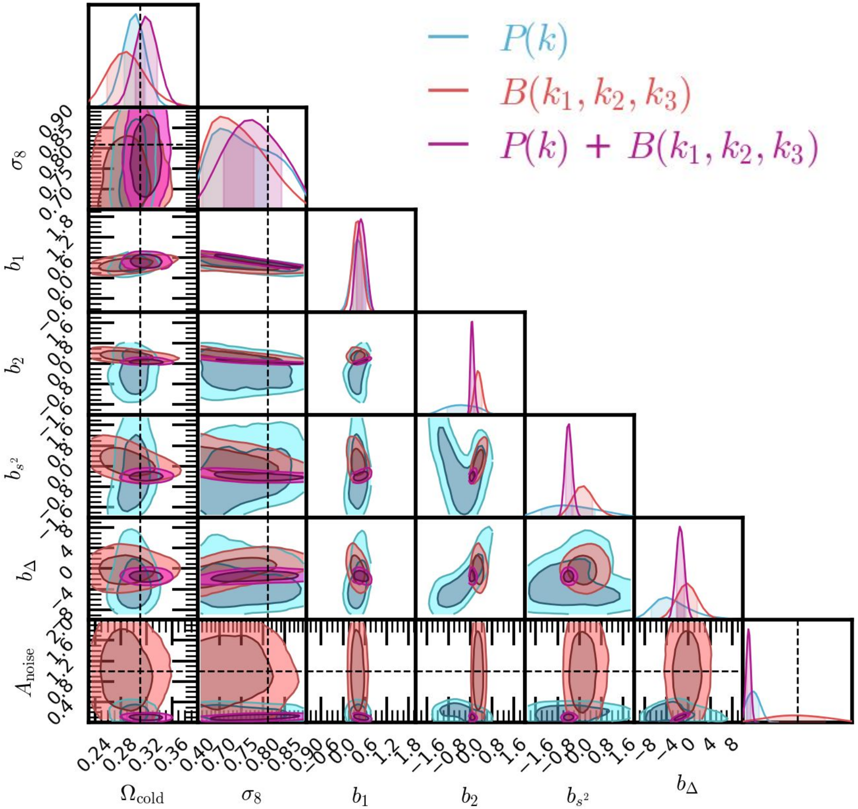


The accuracy & precision improves significantly from 1→20 bias parameter sets per cosmology, and from 1k→10k cosmologies, with ~convergence around 6k cosmologies.

preliminary

Out-of-distribution test: BACCO high-res N-body + SHAMe galaxies

training: 10,000 cosmologies, 20x $\{b\}$



The model is robust to out-of-distribution test data, recovering unbiased cosmological parameters (with the noise parameter likely absorbing some model mismatch).

kstoreyf
@stanford.edu

@kstoreyf

cosmo.nyu.edu/ksf

Summary & Outlook

- We present the **`muchísimocks`** library of 10,000 density fields: uses a **field-level emulator** of *N*-body simulations to span cosmological parameter space, and the **hybrid Lagrangian bias expansion** for galaxy bias space; the approach is a path forward for simulation-limited galaxy clustering analyses.

- We trained a **simulation-based inference** pipeline on the power spectrum and bispectrum, finding good convergence with the number of cosmologies and bias parameter sets (200,000 total training sims). Tests on an out-of-distribution SHAMe mock demonstrate robustness.

- `muchísimocks` can be utilized for a range of analyses, such as full-field inference, testing beyond-standard statistics, velocity reconstruction, etc *[reach out with ideas!]* - publicly available soon!

- *Upcoming/in-progress:* Redshift-space, scale dependence, bias expansion checks, bispectrum `map2map` checks, ++

Extra Slides

TNG-100-1, stars, z=0

map2map accuracy

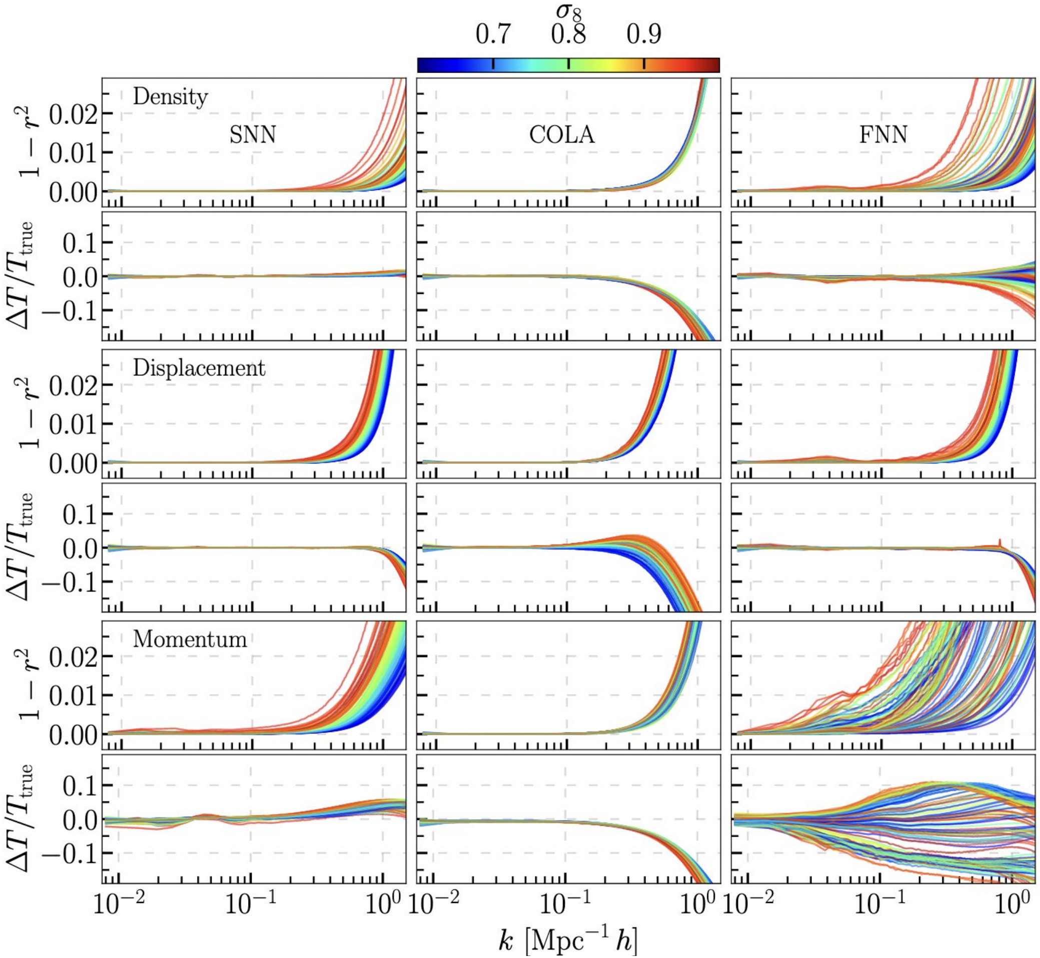
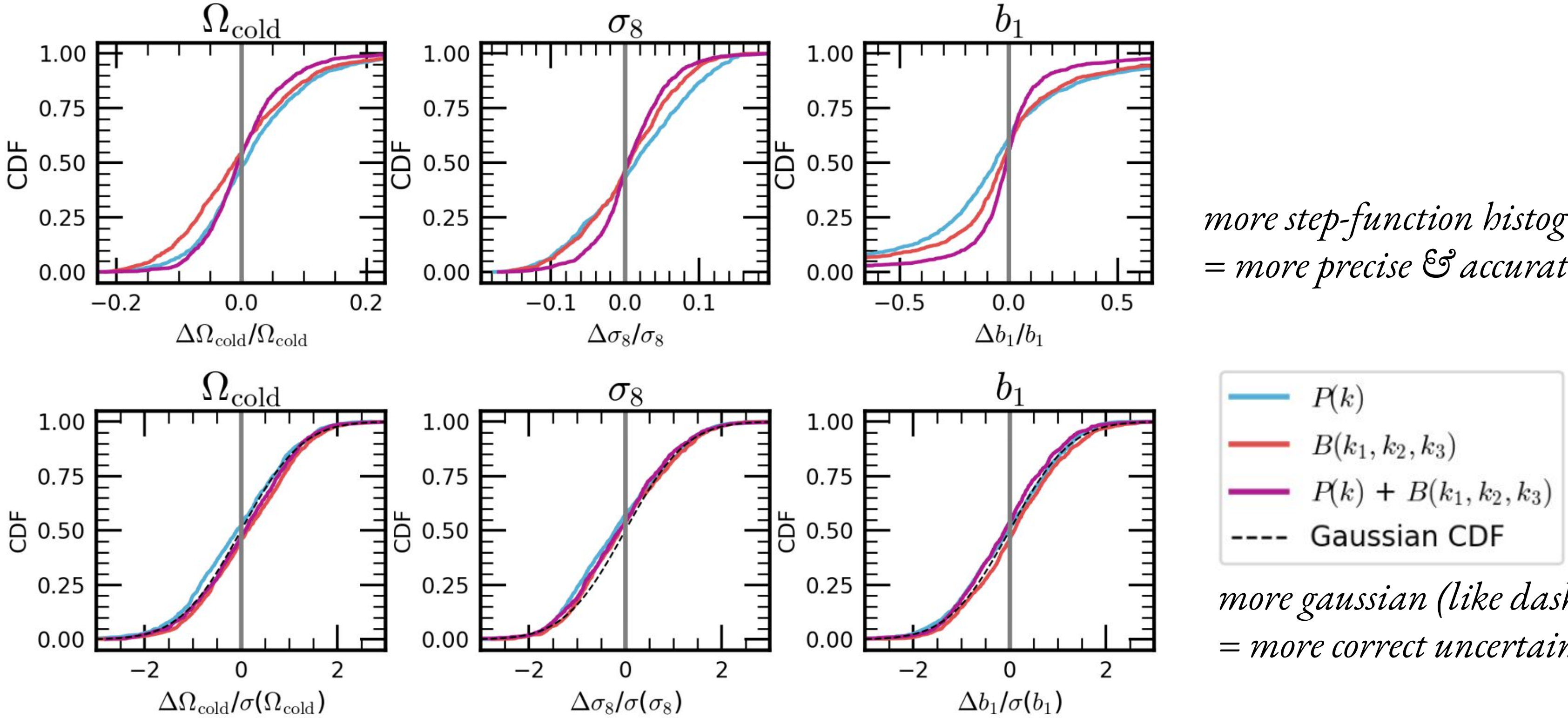


Fig. 1. Power spectra errors for the styled neural network (left), COLA (middle), and the fiducial neural network (right). Each pair of rows shows the stochasticities and transfer function errors for the Eulerian density field power spectra (top two), the Lagrangian displacement field power spectra (middle two), and the Eulerian momentum field (bottom two) power spectra. The color of each curve corresponds to its value of σ_8 according to the color bar at the top.

Jamieson+2022
(2206.04594)

Simulation-based inference results: *P(k)* & bispectrum, 20x{*b*}

coverage test: test set spanning training set space (cosmo & bias varied)

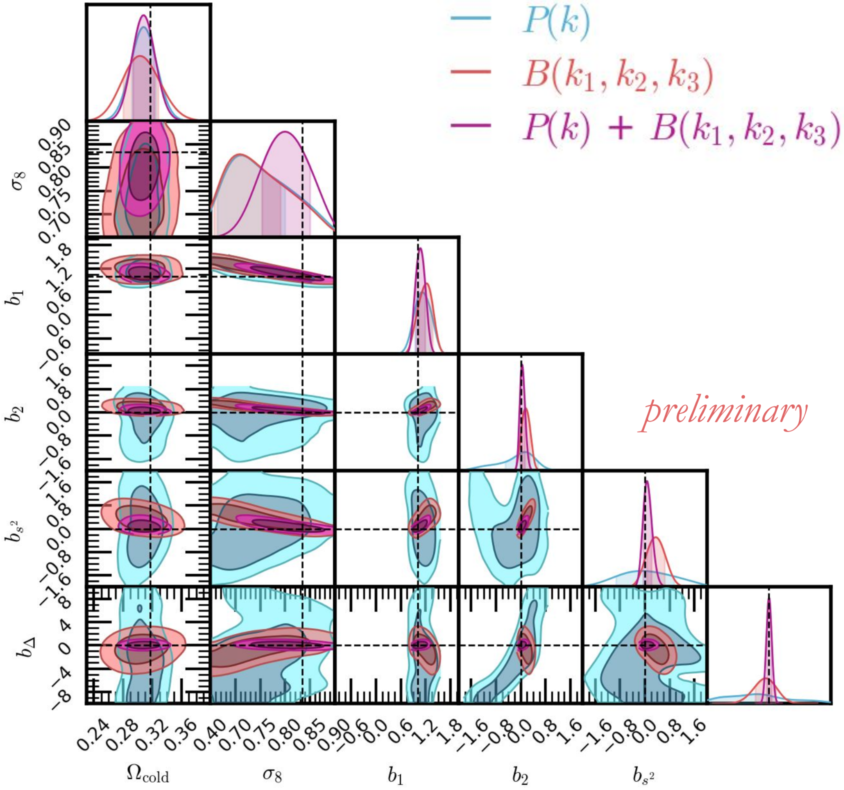


more step-function histogram
= more precise & accurate

more gaussian (like dashed line)
= more correct uncertainties

Inference recovery test results

recovery test: mean of 1000 fixed cosmo mocks (20x {b})



preliminary

Simulation-based inference: neural posterior estimation

Train a neural network to *approximate* the posterior by minimizing the Kullback-Leibler divergence between it and the true posterior, with trainable parameters ϕ:

$$\min_\phi D_{\mathrm{KL}}\left(p(\theta|x)p(x) \| q_\phi(\theta|x)p(x)\right) = \min_\phi \int p(\theta, x) \log \frac{p(\theta|x)}{q_\phi(\theta|x)} d\theta dx$$

integral→sum because we have discrete samples

$$\approx \min_\phi \sum_i \log p(\theta_i|x_i) - \log q_\phi(\theta_i|x_i)$$

the true posterior is independent of NN parameters

$$\approx \min_\phi \sum_i - \log q_\phi(\theta_i|x_i)$$

minimizing the negative log-likelihood is the same as maximizing the log-likelihood

$$\approx \max_\phi \sum_i \log q_\phi(\theta_i|x_i)$$

Simulation-based inference: density estimation

We use a *normalizing flow* as our model for approximating the posterior, specifically a *masked autoregressive flow*.

a base variable $\mathbf{z}$ is sampled from a Gaussian

the f's are conditioned on the observable x

the flow thus learns to transform noise into samples from the posterior

$$q_\phi(\theta|x) = \mathcal{N}(\mathbf{z}_0|\mathbf{0},\mathbf{I}) \prod_{t=1}^{T} \left|\det\left(\frac{\partial f_t}{\partial \mathbf{z}_{t-1}}\right)\right|^{-1}$$

the $\mathbf{z}$'s are transformed via invertible functions f (NN layers) to resemble the parameters θ

$$\mathbf{z}_0 \xrightarrow{f_1} \mathbf{z}_1 \xrightarrow{f_2} \mathbf{z}_2 \xrightarrow{\cdots} \mathbf{z}_T = \boldsymbol{\theta}$$