Simulation-based inference for higher-order galaxy clustering statistics

Beatriz Tucci
Max Planck Institute for Astrophysics (MPA)

with Fabian Schmidt, Nhat-Minh Nguyen, Ivana Babić, Ivana Nikolac, Andrija Kostić, Martin Reinecke

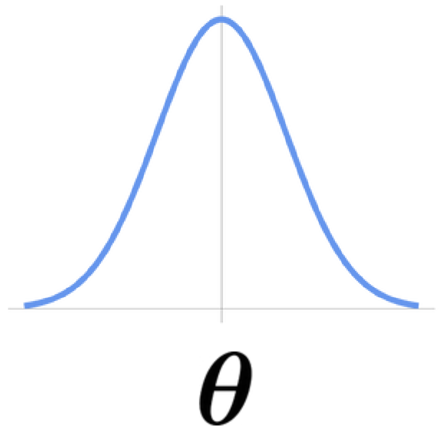
New Physics from Galaxy Clustering III
Parma, 2024

Part 0

Inferring the cosmological parameters:

standard techniques & challenges

Bayesian inference



Posterior Likelihood Prior

$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{D}) \propto \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

E.g., assuming that the data vector is Gaussian distributed:

$$-2\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) = \left(\mathbf{D} - \mathbf{T}(\boldsymbol{\theta})\right)\cdot\mathbf{C}^{-1}\cdot\left(\mathbf{D} - \mathbf{T}(\boldsymbol{\theta})\right)$$

Data vector

Covariance of the data vector

Theoretical prediction of data vector

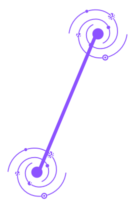
Case study:

inferring the cosmological parameter σ_8

Inferring σ_8 with the power-spectrum

$$T(\boldsymbol{\theta})$$

$$P_g(k) = \langle \delta_g(\boldsymbol{k}) \delta_g(\boldsymbol{k}') \rangle'$$



$$\delta_g(\boldsymbol{k}) = b_1 \delta(\boldsymbol{k}) + \varepsilon(\boldsymbol{k})$$

$$P_g^{\mathrm{tree}}(k) = b_1^2 P_L(k) + P_\varepsilon$$

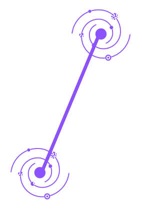
$$P_L(k) = \langle \delta^{(1)}(\boldsymbol{k}) \delta^{(1)}(\boldsymbol{k}') \rangle'$$

$$\propto \sigma_8^2$$

Inferring σ_8 with the power-spectrum

$$T(\boldsymbol{\theta})$$

$$P_g(k) = \langle \delta_g(\boldsymbol{k}) \delta_g(\boldsymbol{k}') \rangle'$$



$$\delta_g(\boldsymbol{k}) = b_1 \delta(\boldsymbol{k}) + \varepsilon(\boldsymbol{k})$$

$$P_g^{\text{tree}}(k) = b_1^2 P_L(k) + P_\varepsilon$$

$$P_L(k) = \langle \delta^{(1)}(\boldsymbol{k}) \delta^{(1)}(\boldsymbol{k}') \rangle'$$

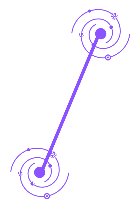
$$\propto \sigma_8^2$$

Bias parameter and σ_8 are degenerated in the tree-level galaxy power-spectrum

Inferring σ_8 with the power-spectrum

$$T(\boldsymbol{\theta})$$

$$P_g(k) = \langle \delta_g(\boldsymbol{k}) \delta_g(\boldsymbol{k}') \rangle'$$



$$\delta_g(\boldsymbol{k}) = b_1 \delta(\boldsymbol{k}) + \varepsilon(\boldsymbol{k})$$

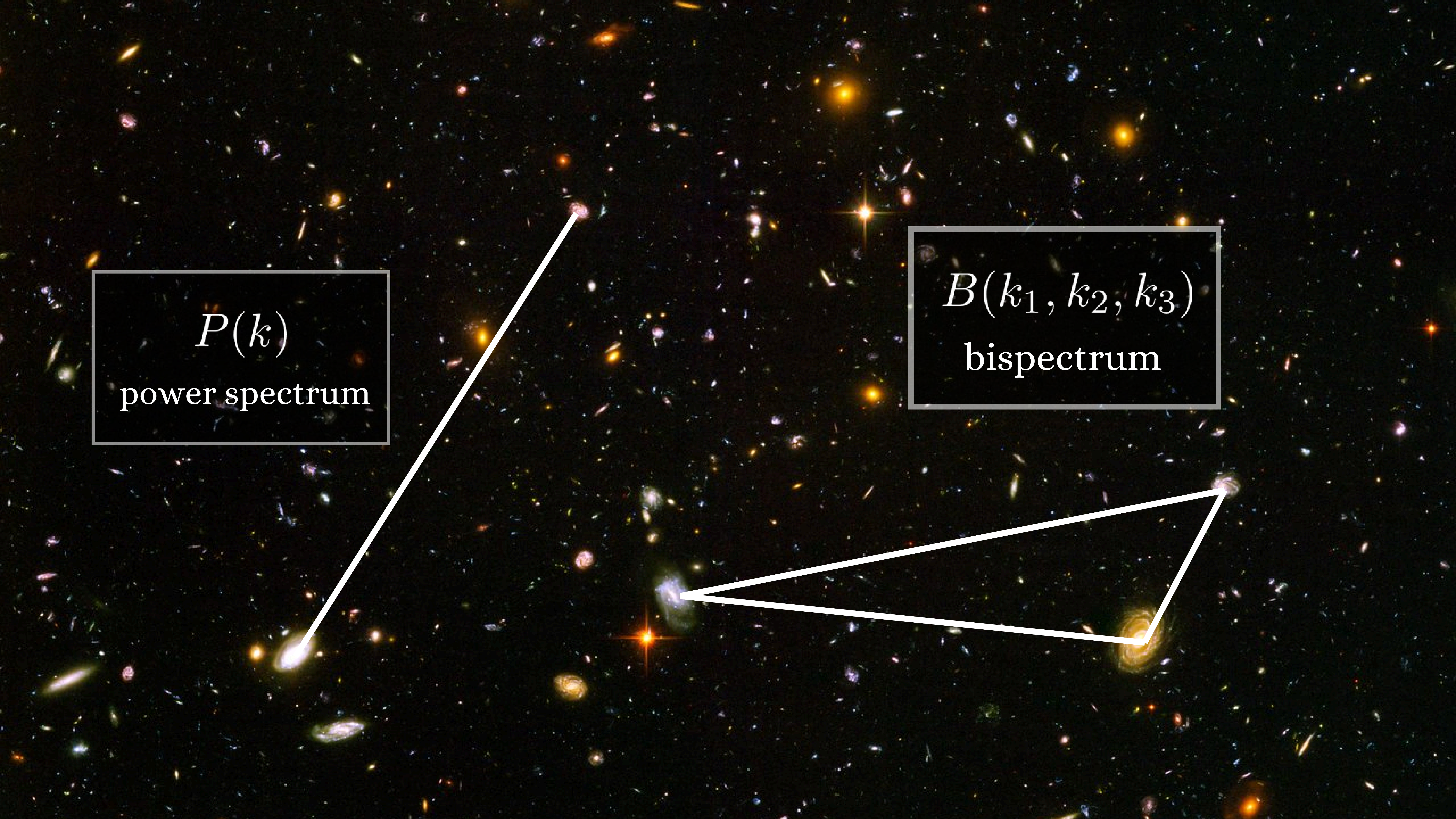
$$P_g^{\mathrm{tree}}(k) = b_1^2 P_L(k) + P_\varepsilon$$

$$P_L(k) = \langle \delta^{(1)}(\boldsymbol{k}) \delta^{(1)}(\boldsymbol{k}') \rangle'$$

$$\propto \sigma_8^2$$

Bias parameter and σ_8 are degenerated in the tree-level galaxy power-spectrum

How to break this degeneracy?


$P(k)$
power spectrum
$B(k_1, k_2, k_3)$
bispectrum

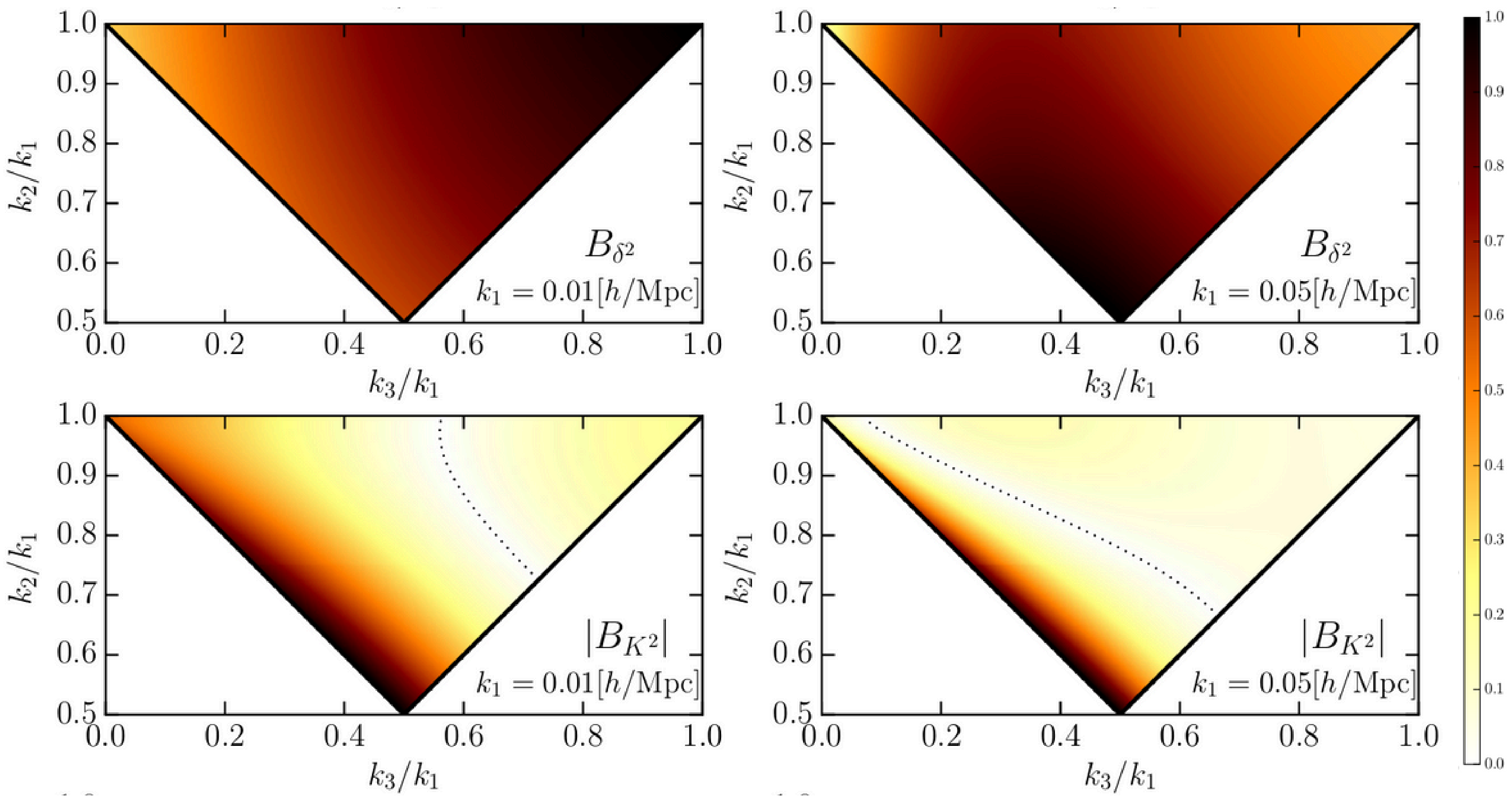
Degeneracy breaking with bispectrum

$$B_g^{\rm tree}(k_1,k_2,k_3) \supset b_1^2\left[b_2 B_{\delta^2}(k_1,k_2,k_3) + 2b_{K^2}B_{K^2}(k_1,k_2,k_3)\right]$$

$$B_{\delta^2}(k_1,k_2,k_3) = P_L(k_1)P_L(k_2) + 2 \text{ perm.}$$

$$B_{K^2}(k_1,k_2,k_3) = \left([\hat{\boldsymbol{k}}_1\cdot\hat{\boldsymbol{k}}_2]^2 - \frac{1}{3}\right)P_L(k_1)P_L(k_2) + 2 \text{ perm.}$$

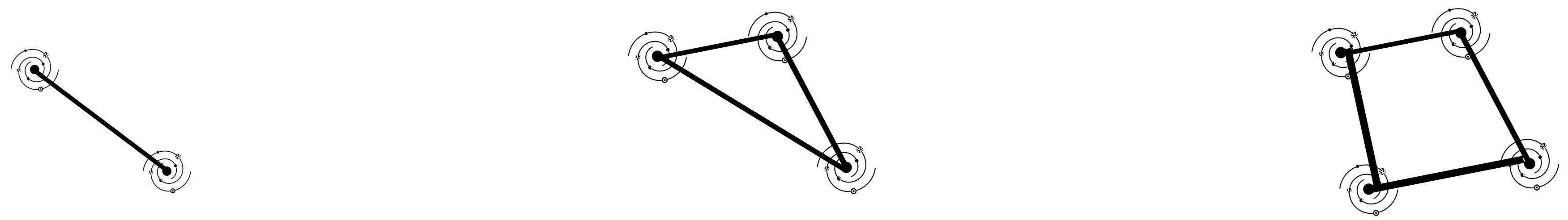
$$P_L(k) = \langle\delta^{(1)}(\boldsymbol{k})\delta^{(1)}(\boldsymbol{k}')\rangle' \propto \sigma_8^2$$



Adapted from Desjacques, Jeong & Schmidt (2016)

Inferring the cosmological parameters: **challenges**

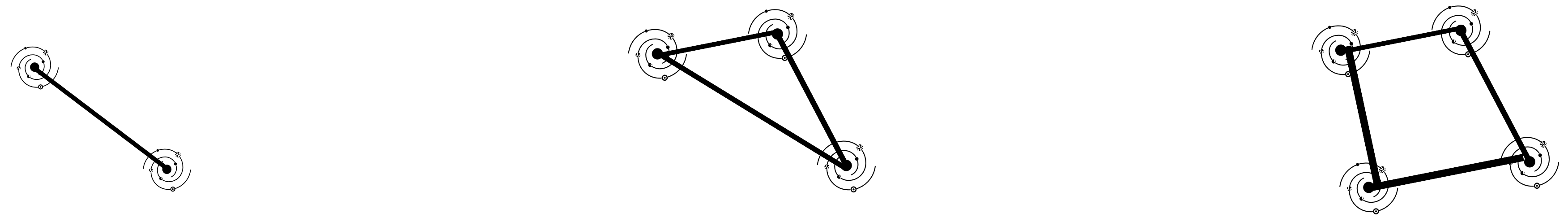
power-spectrum bispectrum trispectrum **Increasing complexity**



… n

Inferring the cosmological parameters: **challenges**

power-spectrum bispectrum trispectrum **Increasing complexity**



... n

$D = 18$ $D = 714$

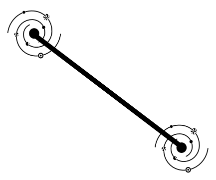
$$k_{\rm max} = 0.12h\,{\rm Mpc}^{-1}$$

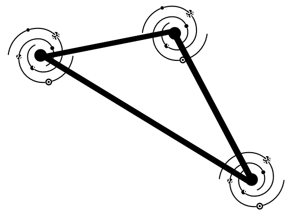
$$\Delta k = 2k_f$$

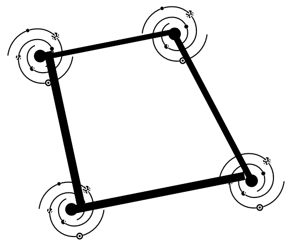
$$L = 2000h^{-1}{\rm Mpc}$$

Inferring the cosmological parameters: **challenges**

power-spectrum — bispectrum — trispectrum — **Increasing complexity**







... n

$D = 18$ $\qquad D = 714$ $\qquad D = 15\,093$

$$k_{\max} = 0.12 h\,\mathrm{Mpc}^{-1}$$
$$\Delta k = 2k_f$$
$$L = 2000 h^{-1}\mathrm{Mpc}$$

Inferring the cosmological parameters: **challenges**

power-spectrum bispectrum trispectrum **Increasing complexity**

Analytical approximations Estimation Modelling

$$-2\ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) = \left(\mathbf{D} - \mathbf{T}(\boldsymbol{\theta})\right) \cdot \mathbf{C}^{-1} \cdot \left(\mathbf{D} - \mathbf{T}(\boldsymbol{\theta})\right)$$

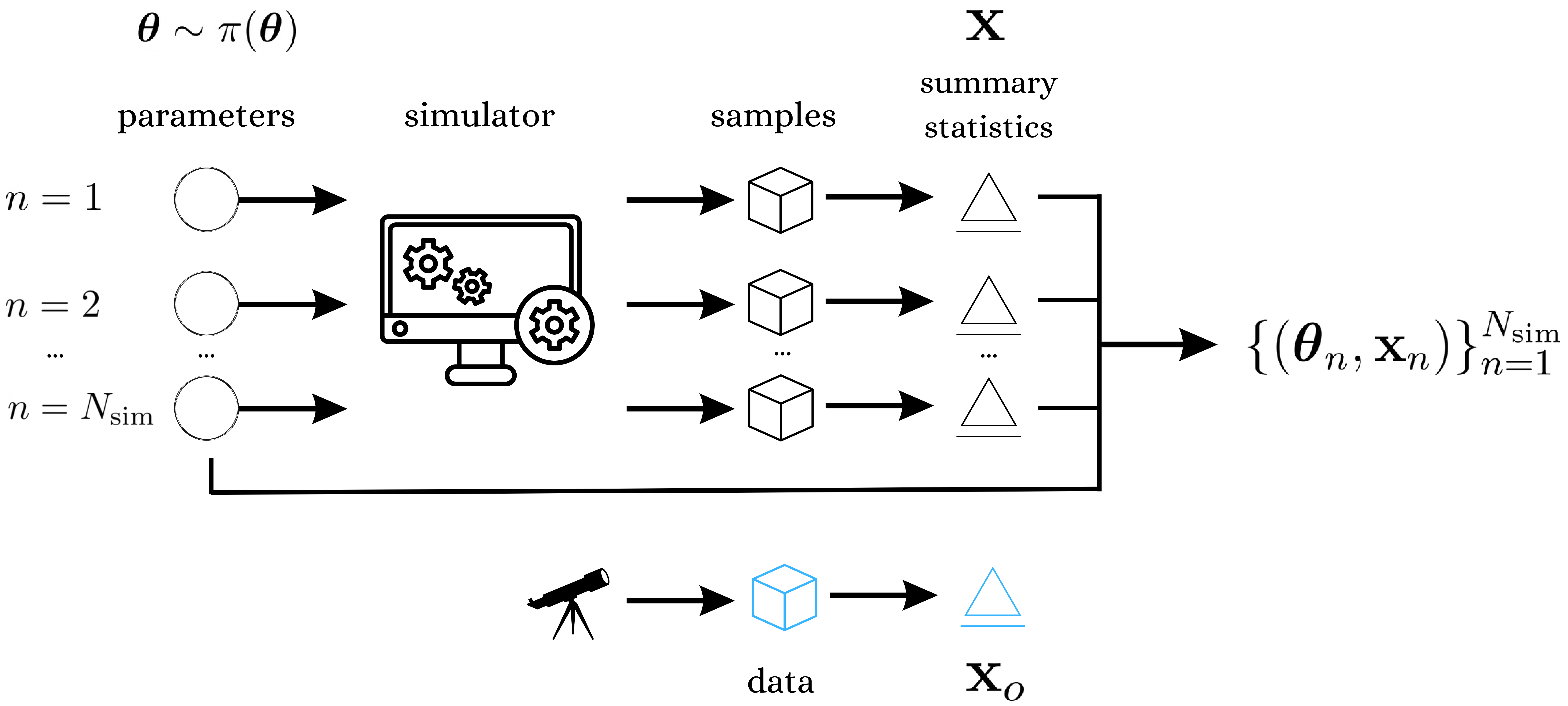
Measurements

Part I

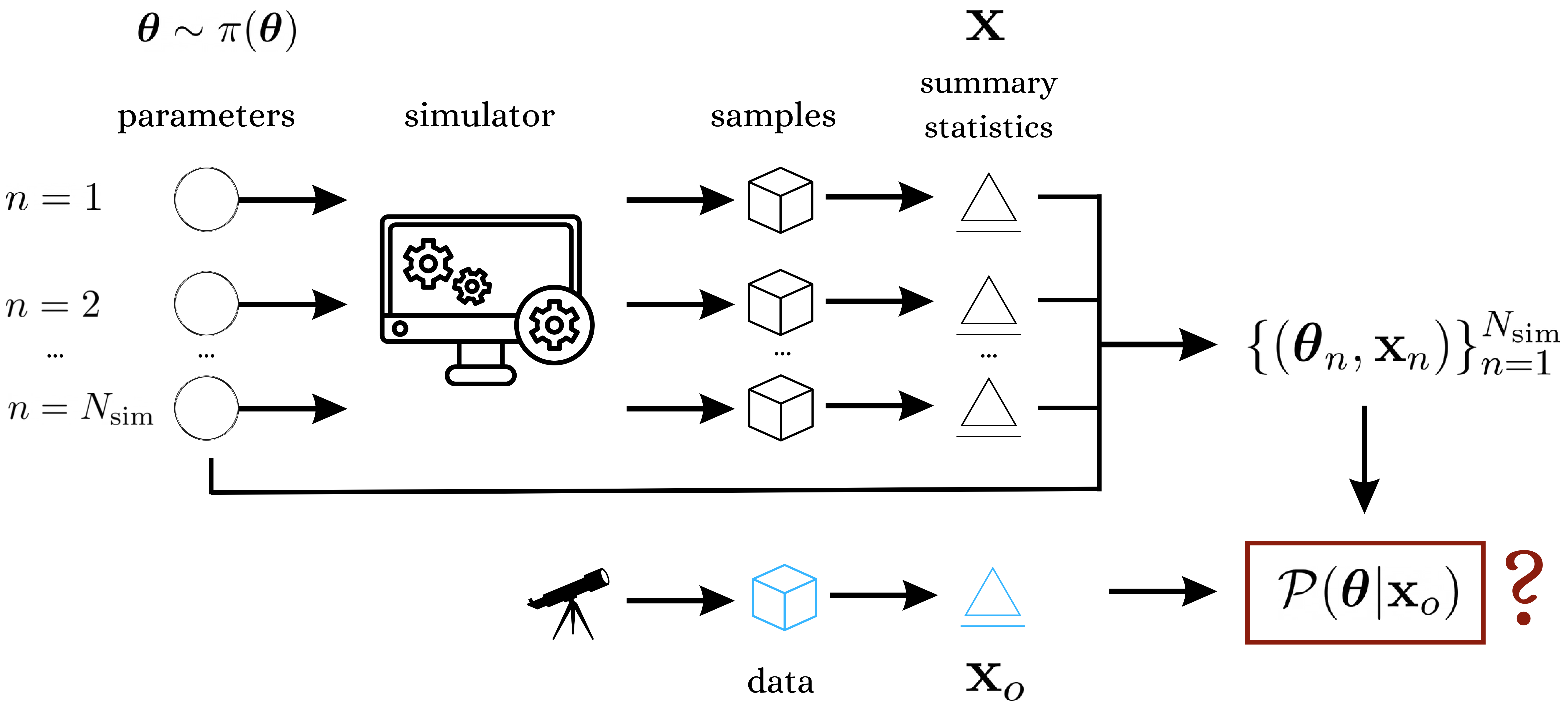
Simulation-based inference

(SBI)

Simulation-based inference

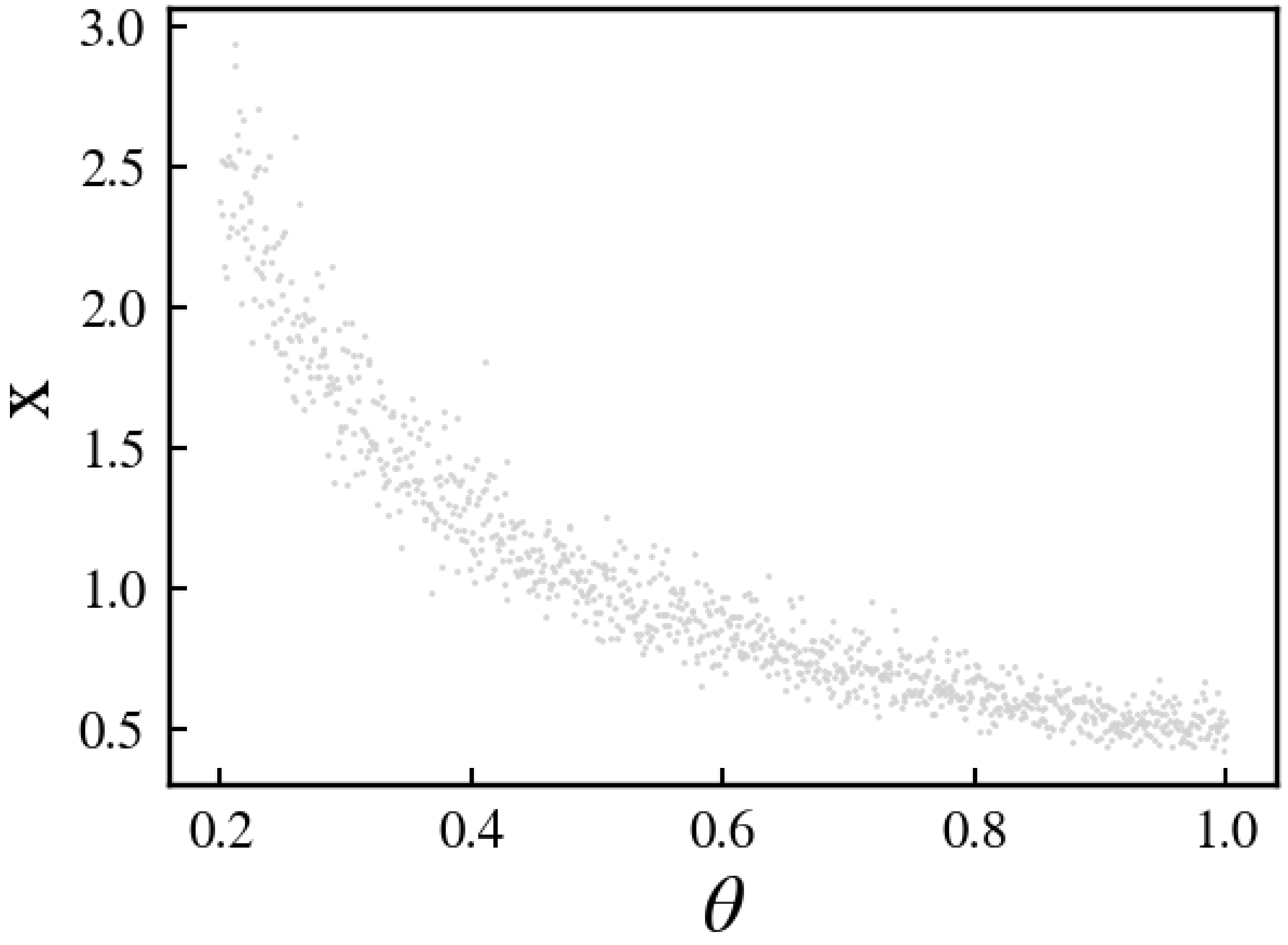


Simulation-based inference



Simulation-based inference

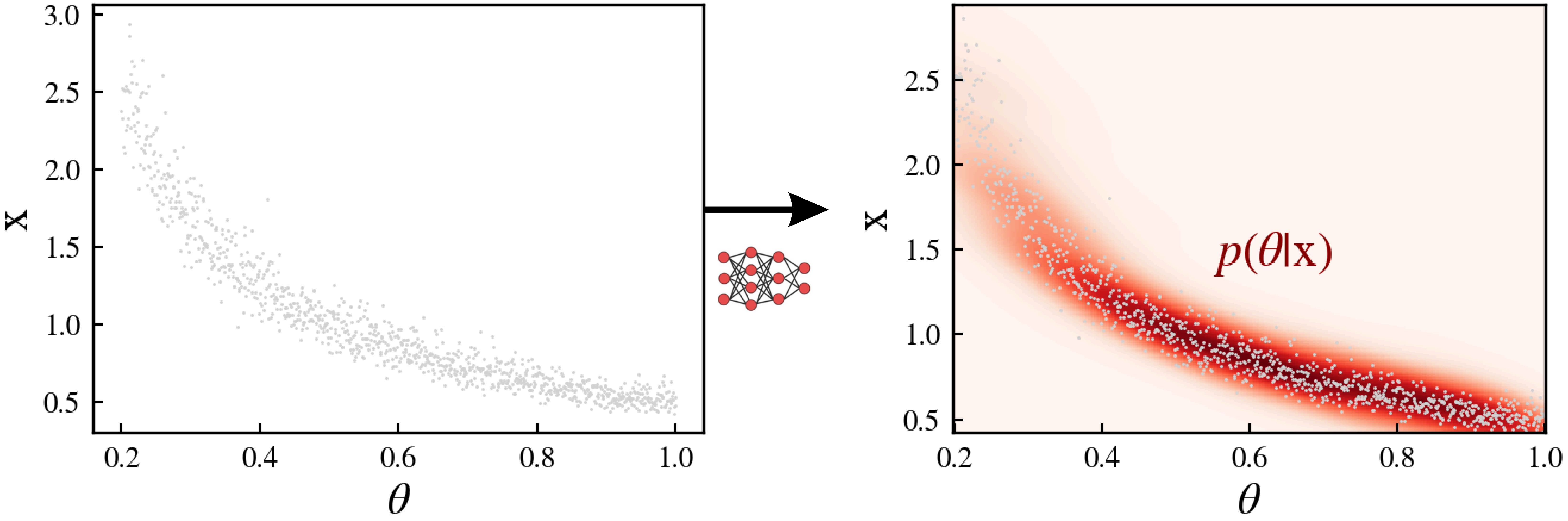
$$\{(\boldsymbol{\theta}_n, \mathbf{x}_n)\}_{n=1}^{N_{\text{sim}}}$$



Simulation-based inference

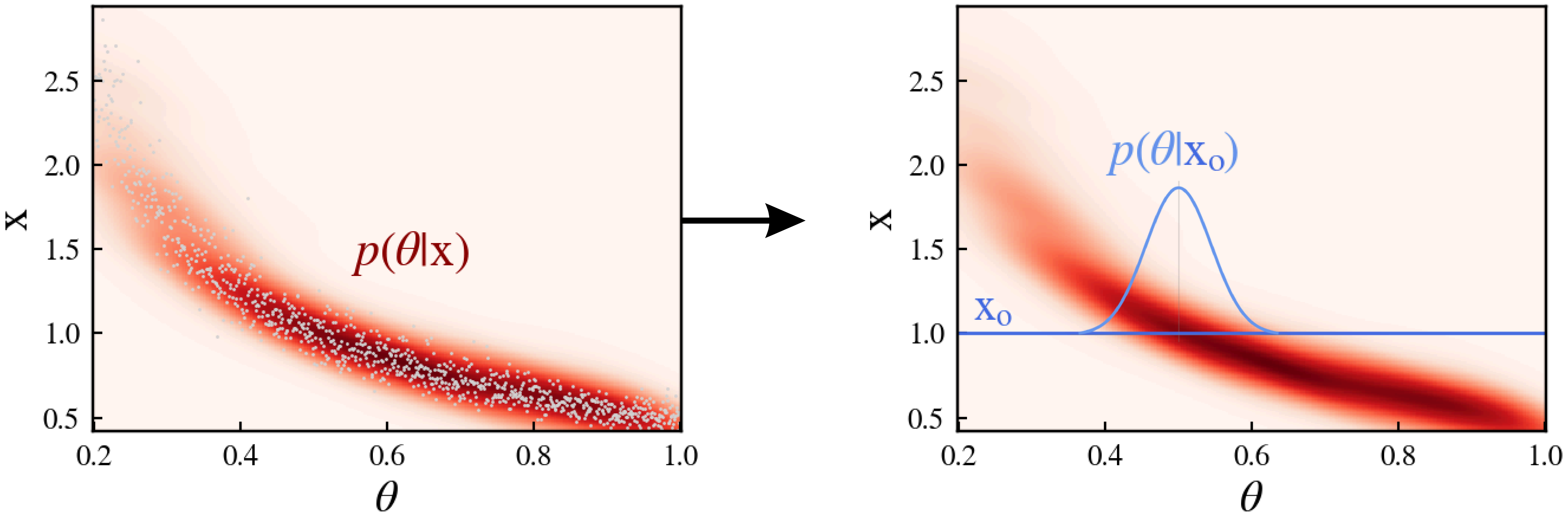
$\{(\boldsymbol{\theta}_n, \mathbf{x}_n)\}_{n=1}^{N_{\text{sim}}}$

Neural Posterior Estimation (NPE)

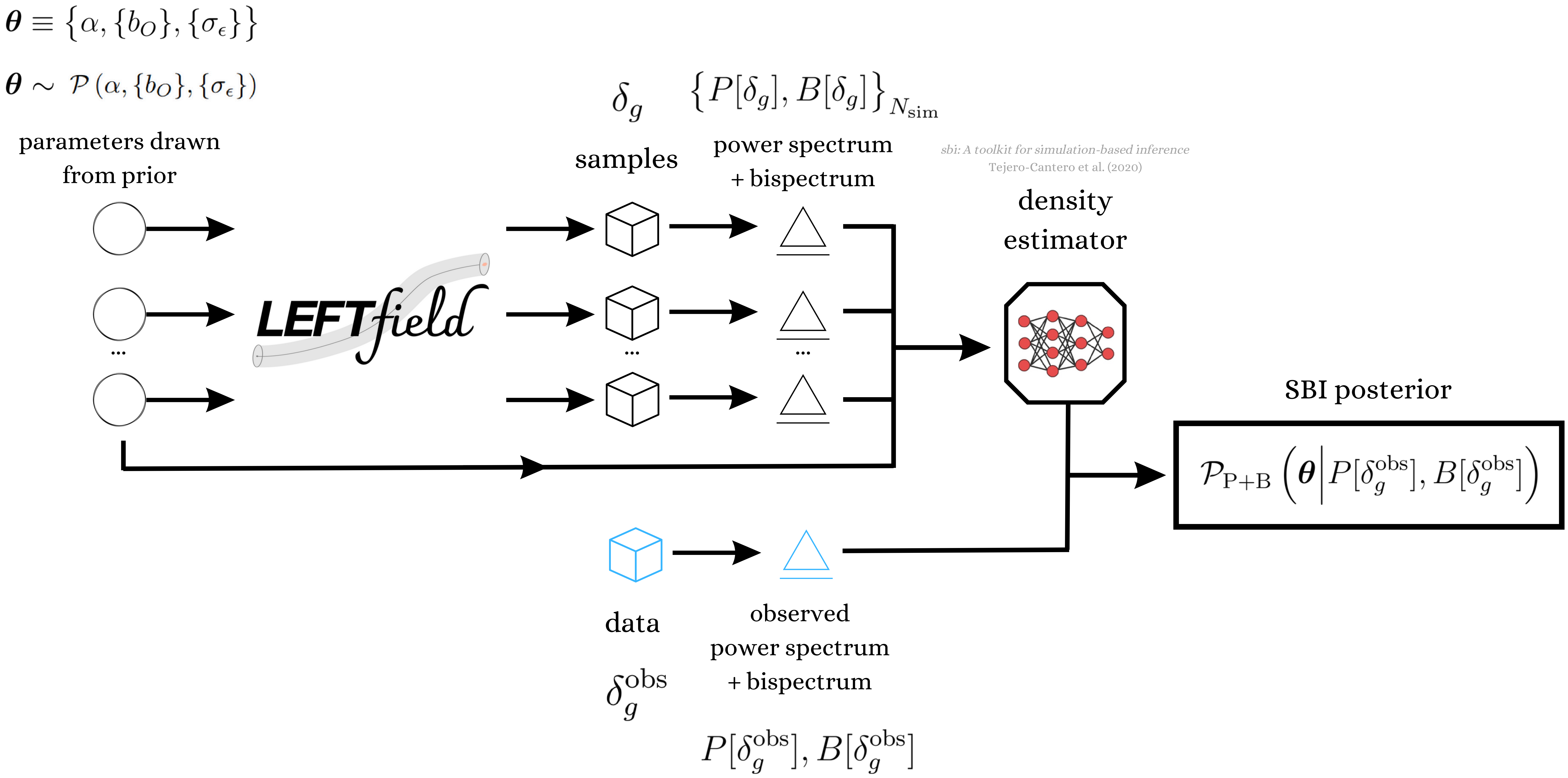


Simulation-based inference

Posterior



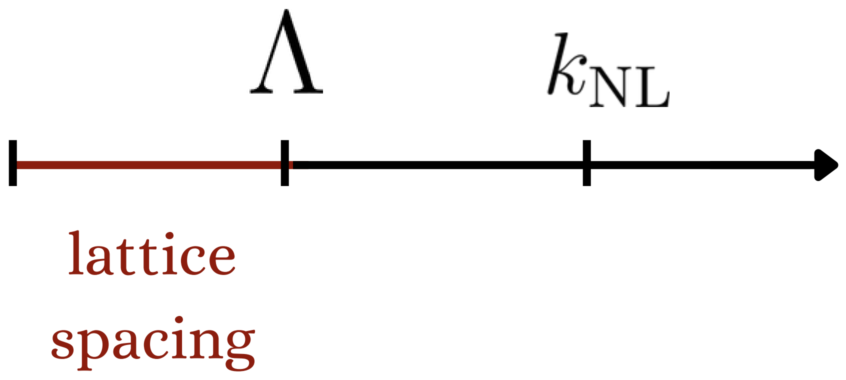
Simulation-based inference for galaxy clustering



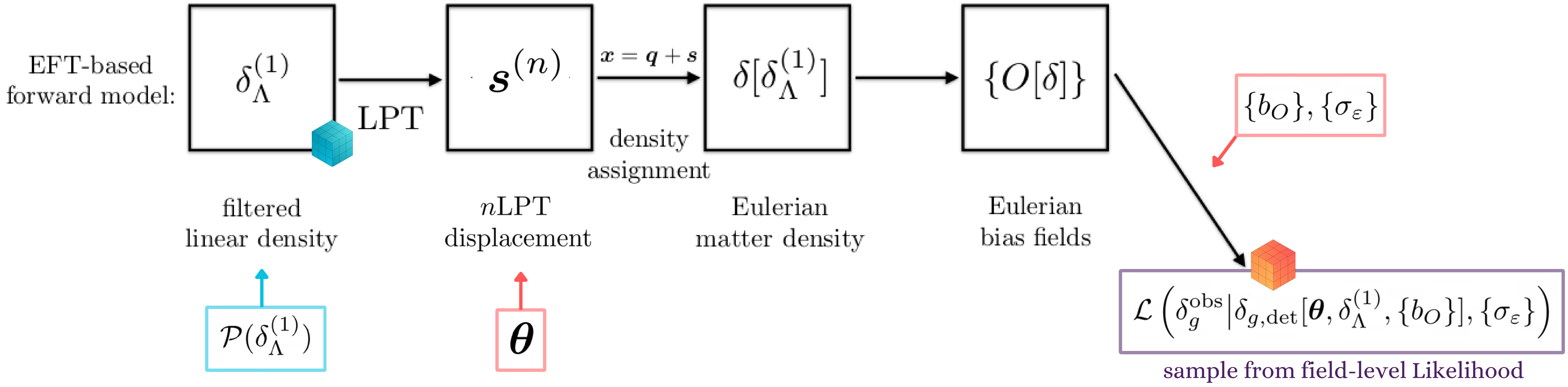
The forward model

based on the EFTofLSS & the bias expansion

- A **fast** forward model based on the EFTofLSS that solves the gravitational evolution of all modes in a lattice up to the cutoff scale
- nLPT and incorporates bias and stochastic parameters, marginalizing over reasonable models of galaxy formation
- Easier to deal with redshift space, masks and systematic effects

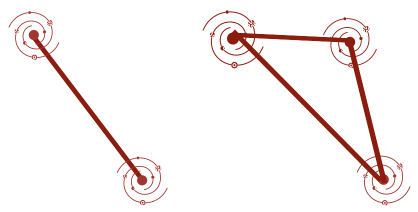


The forward model



An n-th order Lagrangian Forward Model for Large-Scale Structure

Schmidt (2021)



Testing SBI on Euclid-like mock data

Breaking degeneracy between σ_8 and bias parameters with the galaxy power-spectrum and bispectrum

Tucci & Schmidt (2024)

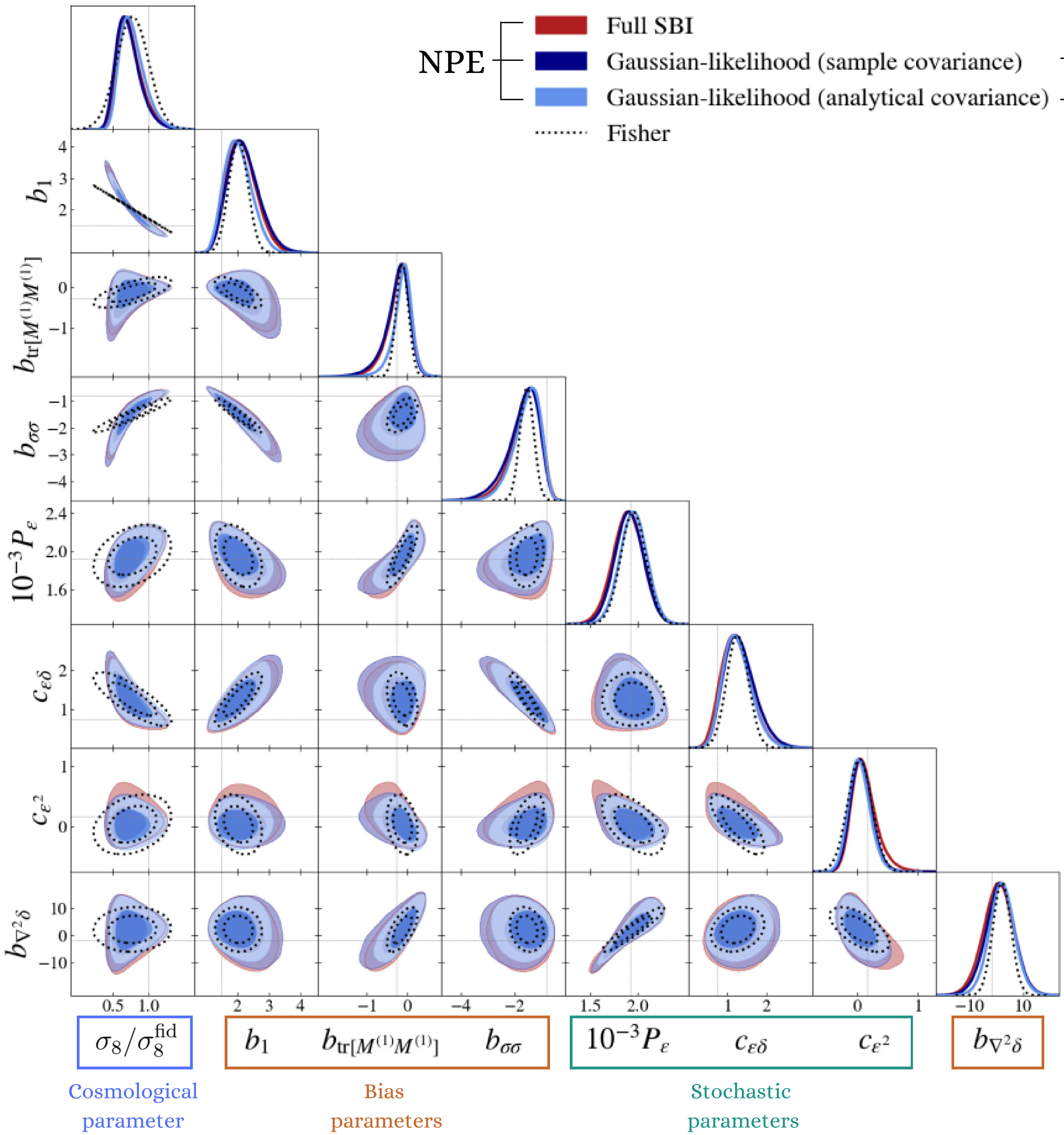
JCAP

Cosmological constraints

$N_{\rm sim} = 10^5$

$k_{\rm max} = \Lambda = 0.1 h{\rm Mpc}^{-1}$

$D = N_{\rm bin} + N_{\rm tri} = 33$

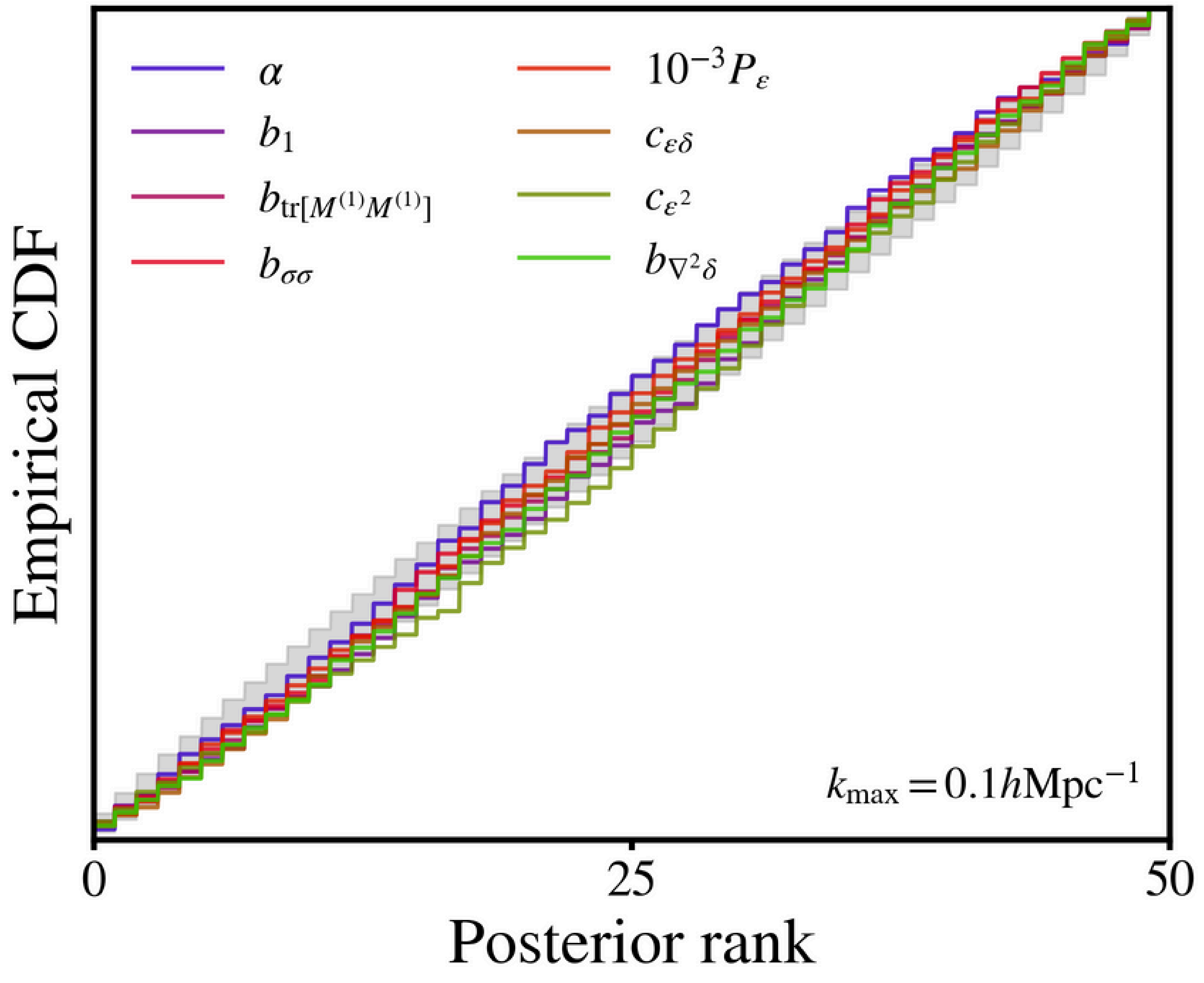


Gaussian-likelihood

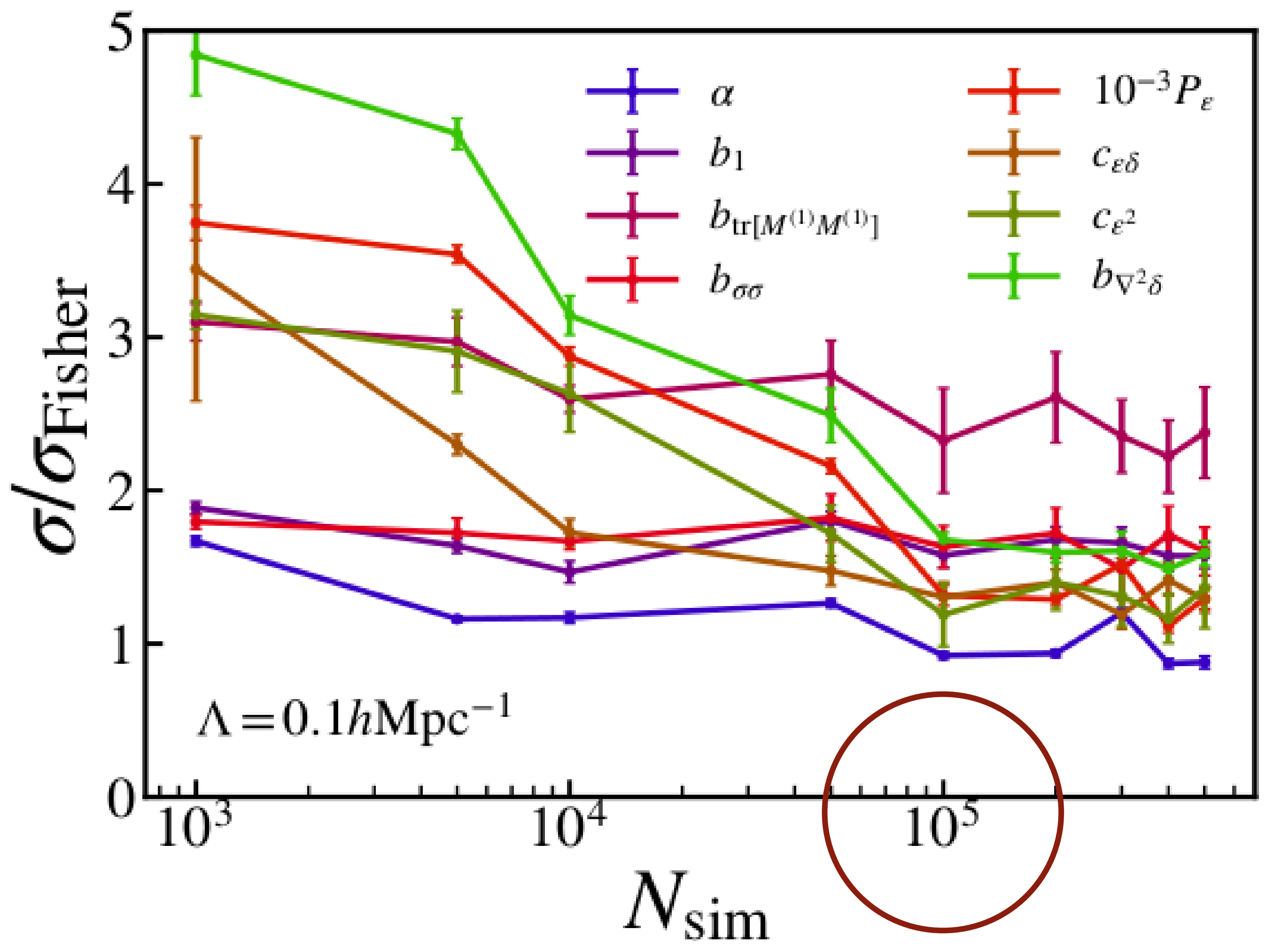
$\mathbf{x}_n \sim \mathcal{N}\left(\langle \mathbf{x}_n \rangle, {\rm Cov}[\mathbf{x}_o]\right)$

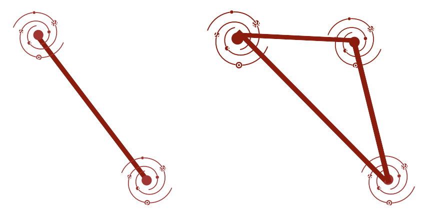
Tests of inference

Simulation-based calibration



Convergence





SBI on dark-matter halos

Breaking degeneracy between σ_8 and bias parameters with the galaxy power-spectrum and bispectrum

Nguyen, Schmidt, **Tucci** et al. (2024)

PRL (accepted)

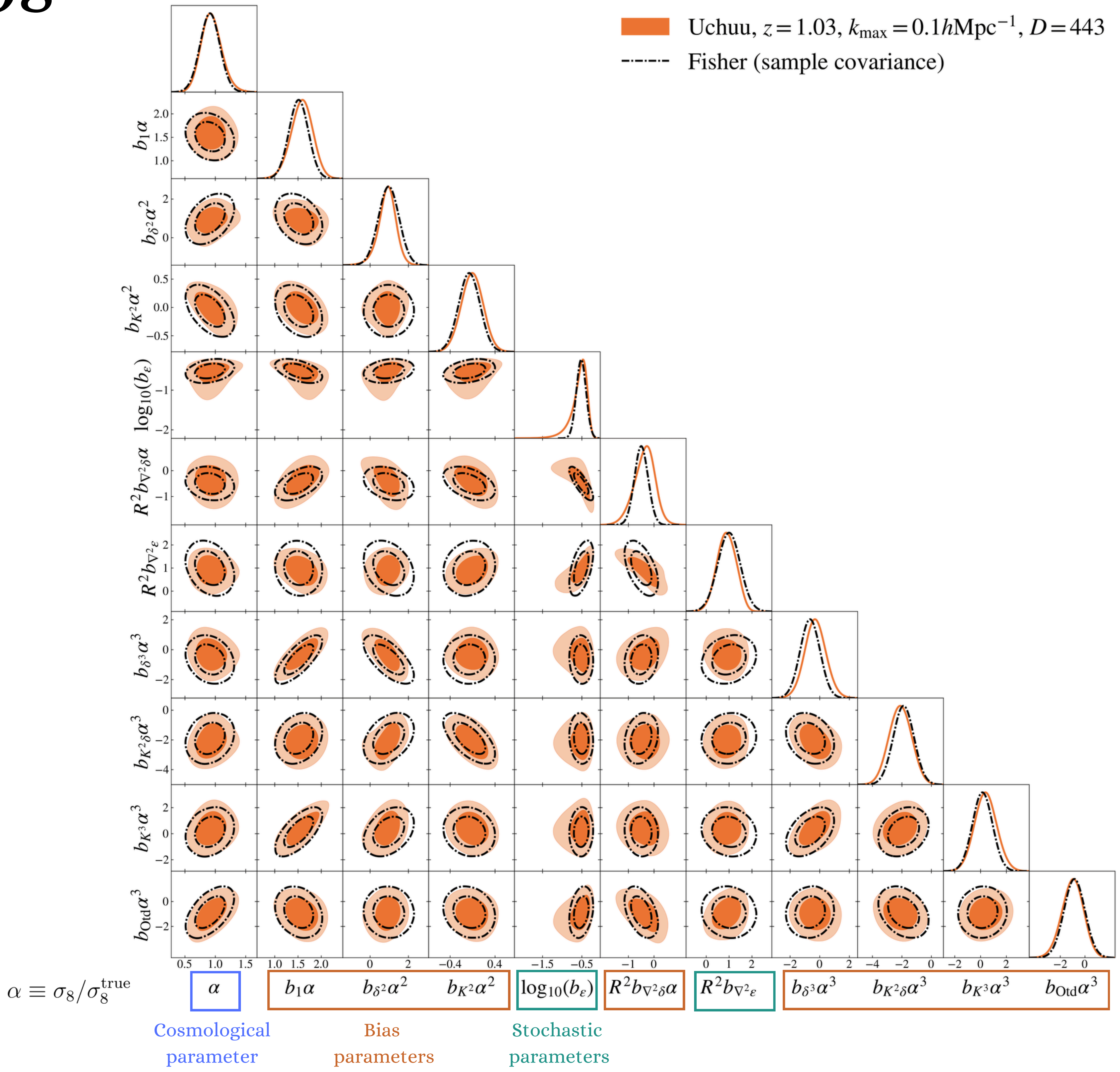
Inference setup: halo samples

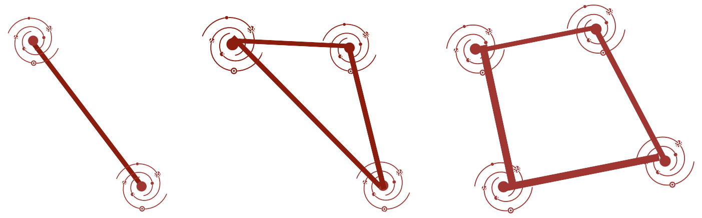
	SNG	Uchuu
Redshift	$z = 0.50$	$z = 1.03$
$V\ [h^{-3}\mathrm{Mpc}^3]$	2000^3	2000^3
$\bar{n}_g\ [h^3\mathrm{Mpc}^{-3}]$	1.3×10^{-3}	3.6×10^{-3}

Two scale cuts:

$k_{\max} = 0.1h\mathrm{Mpc}^{-1}$ & $k_{\max} = 0.12h\mathrm{Mpc}^{-1}$

SBI on halos


Uchuu, $z=1.03$, $k_{\max}=0.1h\mathrm{Mpc}^{-1}$, $D=443$
Fisher (sample covariance)
$\alpha \equiv \sigma_8/\sigma_8^{\mathrm{true}}$
α
$b_1\alpha$
$b_{\delta^2}\alpha^2$
$b_{K^2}\alpha^2$
$\log_{10}(b_\varepsilon)$
$R^2 b_{\nabla^2\delta}\alpha$
$R^2 b_{\nabla^2\varepsilon}$
$b_{\delta^3}\alpha^3$
$b_{K^2\delta}\alpha^3$
$b_{K^3}\alpha^3$
$b_{\mathrm{Otd}}\alpha^3$
Cosmological parameter
Bias parameters
Stochastic parameters

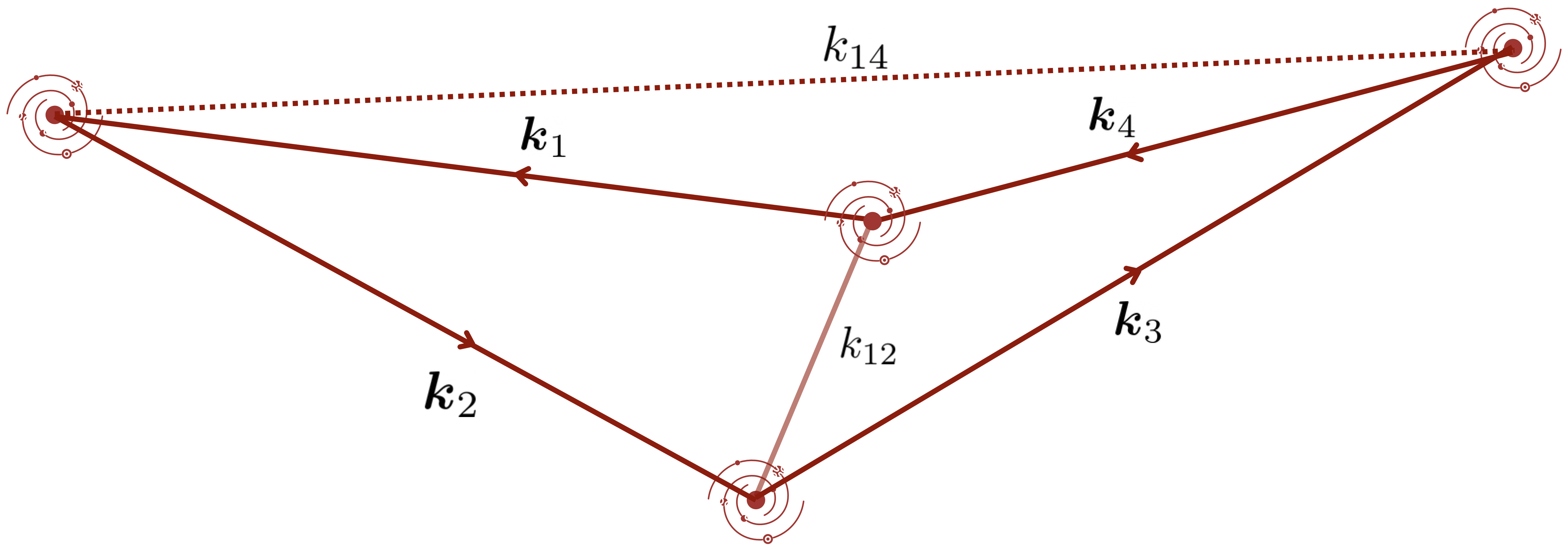


What if we add the galaxy **trispectrum**?

Breaking degeneracy between σ_8 and bias parameters with power-spectrum, bispectrum and trispectrum on dark-matter halos

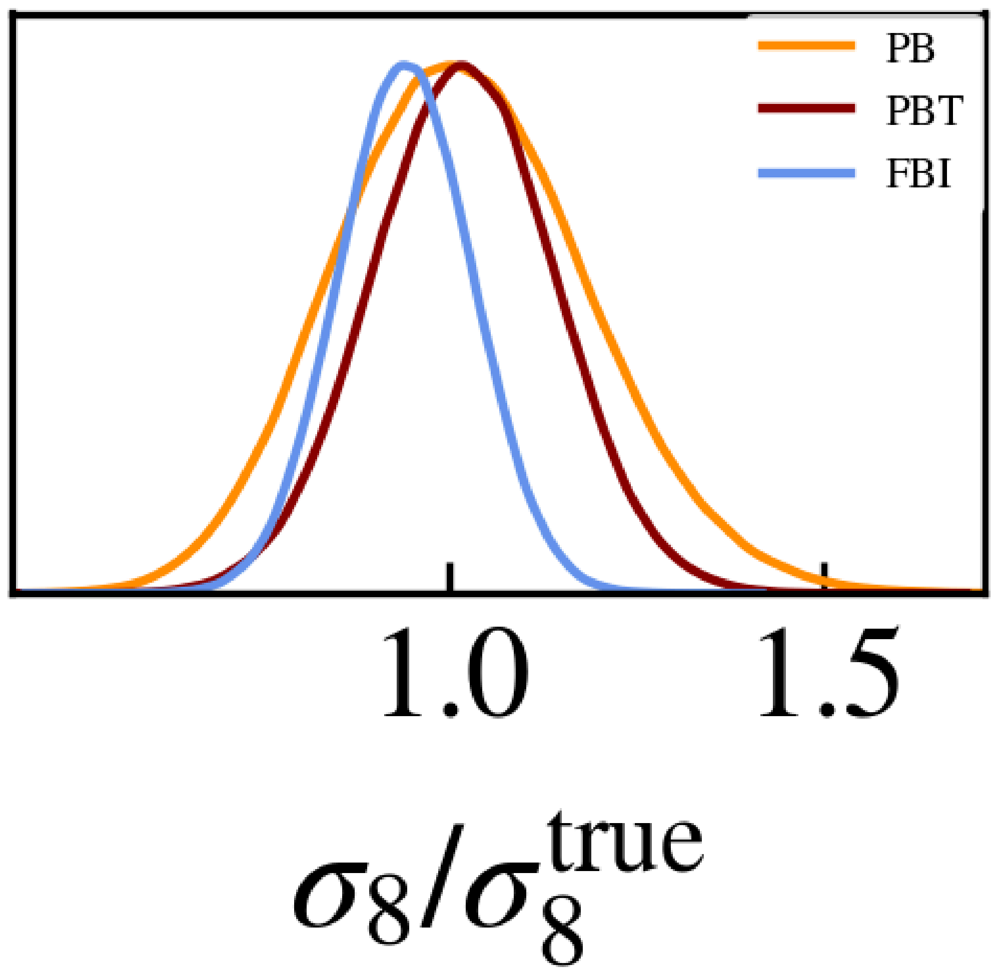
Tucci & Schmidt (in prep.)

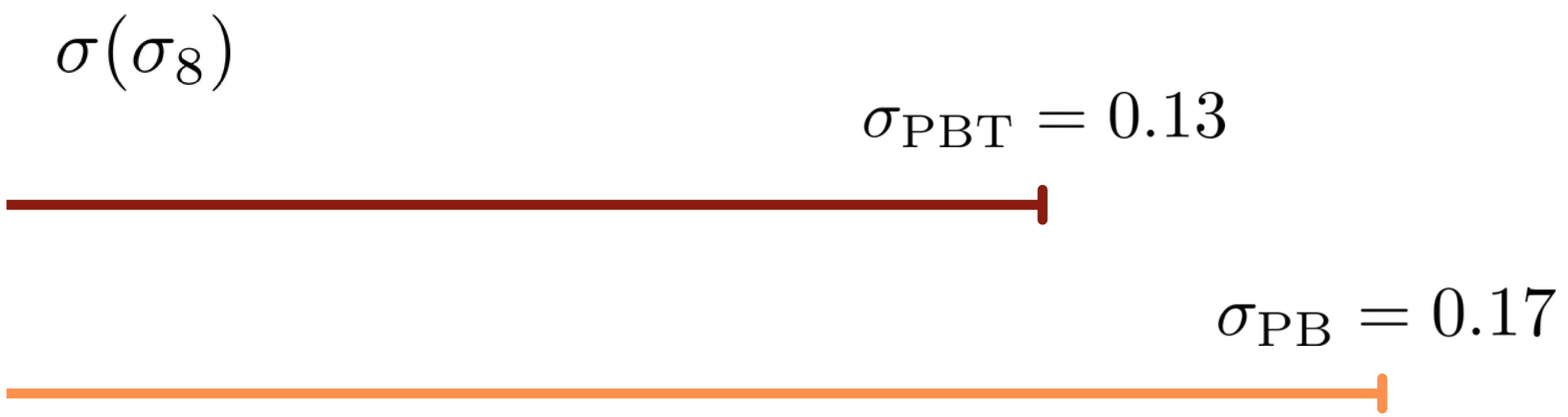
Trispectrum: the estimator



Jung+23, Coulton+23, Goldstein+24

Trispectrum: **preliminary** results





$k_{\mathrm{max}} = 0.1 h\,\mathrm{Mpc}^{-1}$

Uchuu halos at z=1

Brute force approach:
10^6 simulations

SBI with LEFTfield: Conclusions

- Robust analysis with EFTofLSS and bias expansion
- LEFTfield allows for **fast** analysis in **cosmological volumes** with **convergence** and posterior **diagnostics** tests
- Need order of 10^5 simulations for convergence (investigating how we can improve that)
- SBI allows for cosmological inference using **trispectrum**, which is **unfeasible** with standard inference techniques
- No need to assume Gaussian likelihood, explicit loop or covariance calculations

Inferring the cosmological parameters: **challenges**

power-spectrum bispectrum trispectrum **Increasing complexity**

Analytical approximations Estimation Modelling

$$-2\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) = (\mathbf{D}-\mathbf{T}(\boldsymbol{\theta}))\cdot\mathbf{C}^{-1}\cdot(\mathbf{D}-\mathbf{T}(\boldsymbol{\theta}))$$

SBI Measurements

Inferring the cosmological parameters: **challenges**

power-spectrum bispectrum trispectrum **Increasing complexity**

where to stop?
is there a better way?

Analytical approximations

Estimation

Modelling

$$-2\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) = (\mathbf{D}-\mathbf{T}(\boldsymbol{\theta}))\cdot\mathbf{C}^{-1}\cdot(\mathbf{D}-\mathbf{T}(\boldsymbol{\theta}))$$

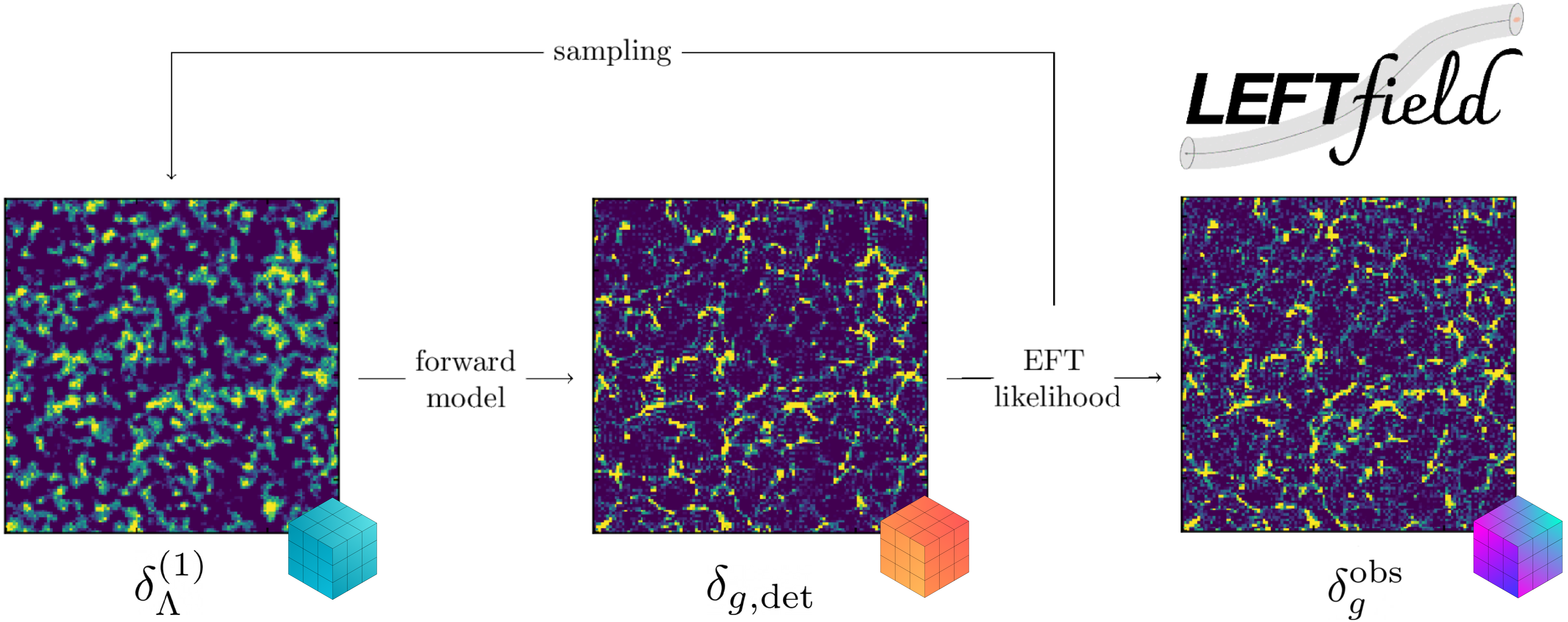
SBI

Measurements

Part II

Field-level Bayesian inference

(FBI)

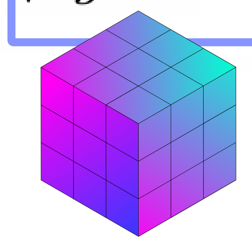


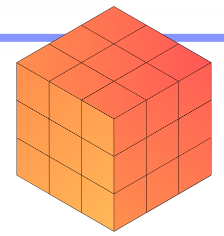
Credits: Julia Stadler

Field level Likelihood

Mode by mode
data and theory
comparison!

$$\ln\mathcal{L}\left(\delta_g^{\mathrm{obs}}\middle|\delta_{g,\mathrm{det}}[\boldsymbol{\theta},\delta_\Lambda^{(1)},\{b_O\}],\{\sigma_\varepsilon\}\right) = -\frac{1}{2}\sum_{k<k_{\max}}\left[\frac{1}{\sigma_\varepsilon^2(k)}\boxed{\left|\delta_g^{\mathrm{obs}}(\boldsymbol{k})-\delta_{g,\mathrm{det}}[\boldsymbol{\theta},\delta_\Lambda^{(1)},\{b_O\}](\boldsymbol{k})\right|^2}+\ln[2\pi\sigma_\varepsilon^2(k)]\right]$$

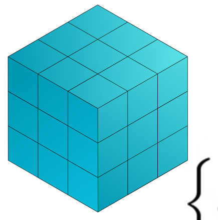




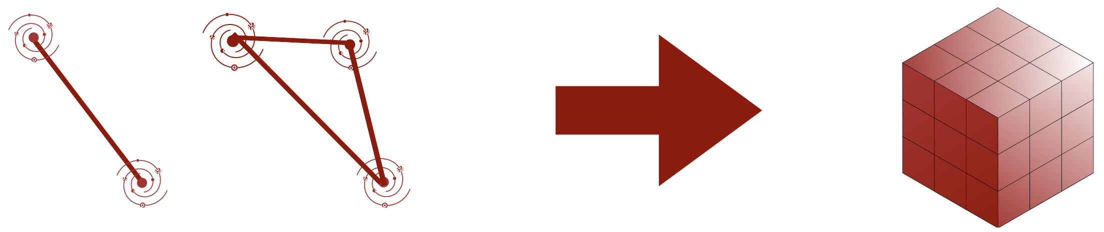
$\downarrow$ HMC

$$\boxed{\mathcal{P}\left(\boldsymbol{\theta},\boxed{\delta_\Lambda^{(1)}},\{b_O\},\{\sigma_\varepsilon\}\middle|\delta_g^{\mathrm{obs}}\right)}$$

Full posterior
including initial
conditions!



$\left\{\delta_{\Lambda,i}^{(1)}\right\}_{i=1}^{N_g^\Lambda}$



How much information is retained at the galaxy density field?

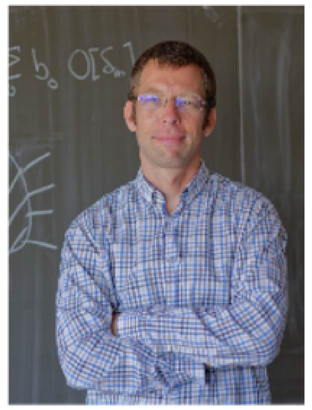
Breaking degeneracy between σ_8 and bias parameters on dark-matter halos

Nguyen, Schmidt, **Tucci** et al. (2024)
PRL (accepted)

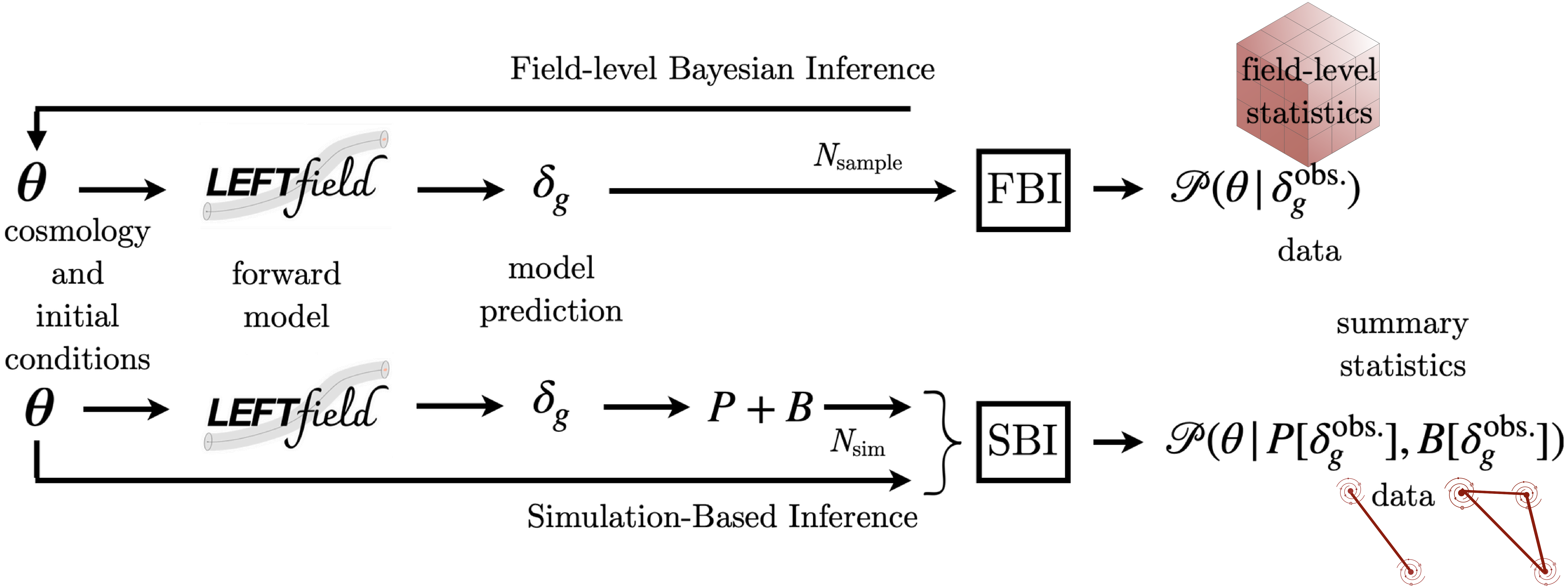
3rd order bias expansion

$$
\boxed{\begin{aligned} O_{\mathrm{det}} &\in \left[\delta, \delta^2, K^2, \delta^3, K^3, \delta K^2, O_{\mathrm{td}}, \nabla^2\delta\right] \\ O_{\mathrm{stoch}} &\in \left[\varepsilon, \nabla^2\varepsilon\right] \end{aligned}}
$$

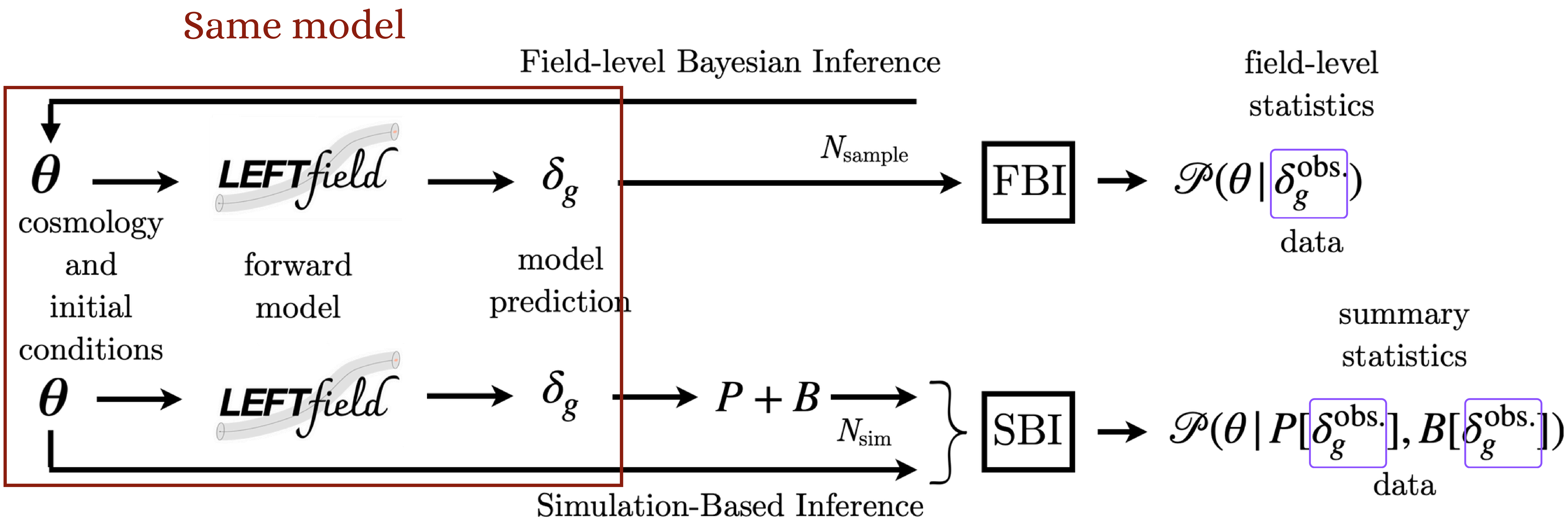
Nhat-Minh Nguyen
(IPMU)



Fabian Schmidt
(MPA)



Apples-to-apples comparison

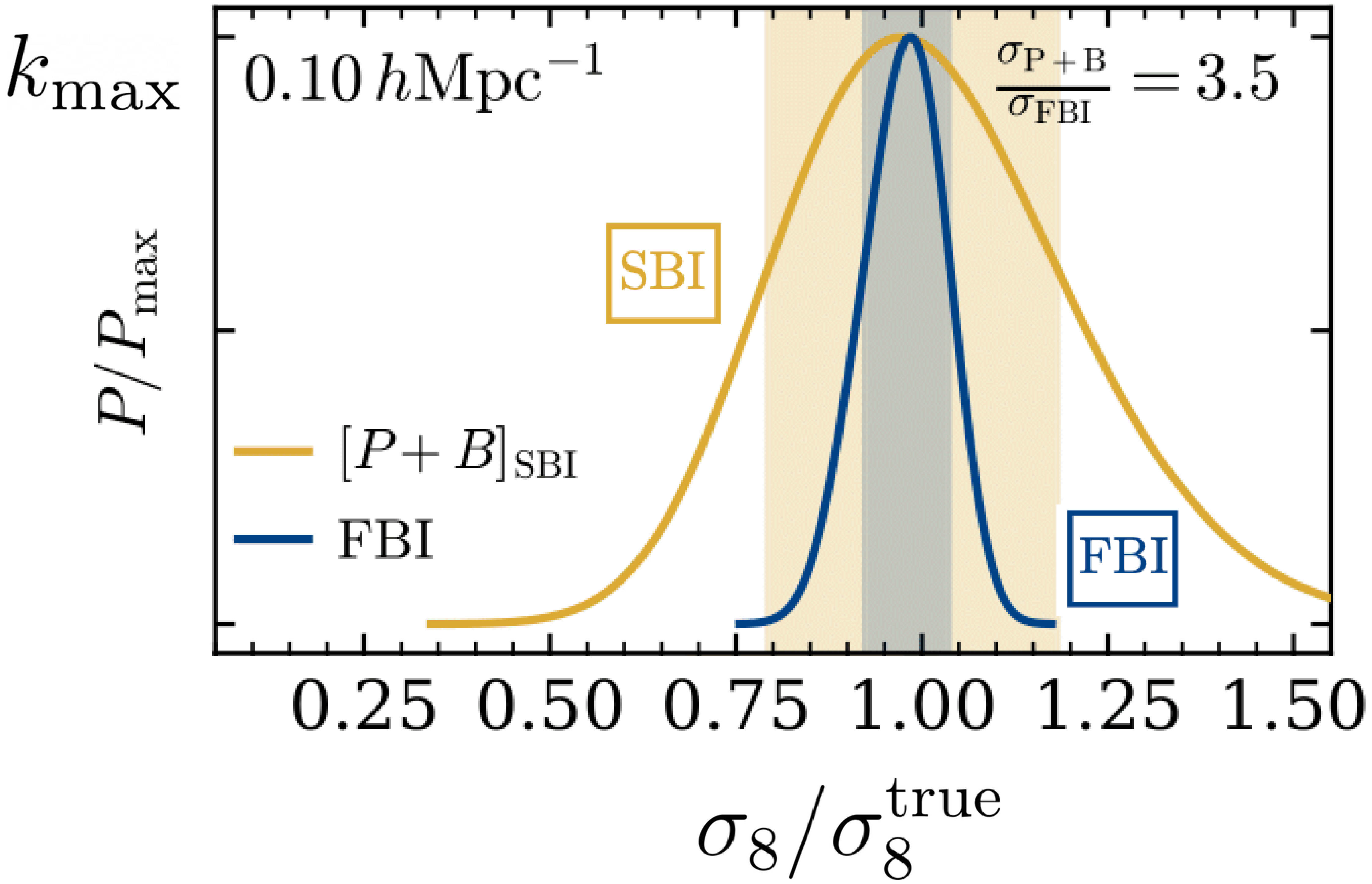


Same halos

Same scale cuts

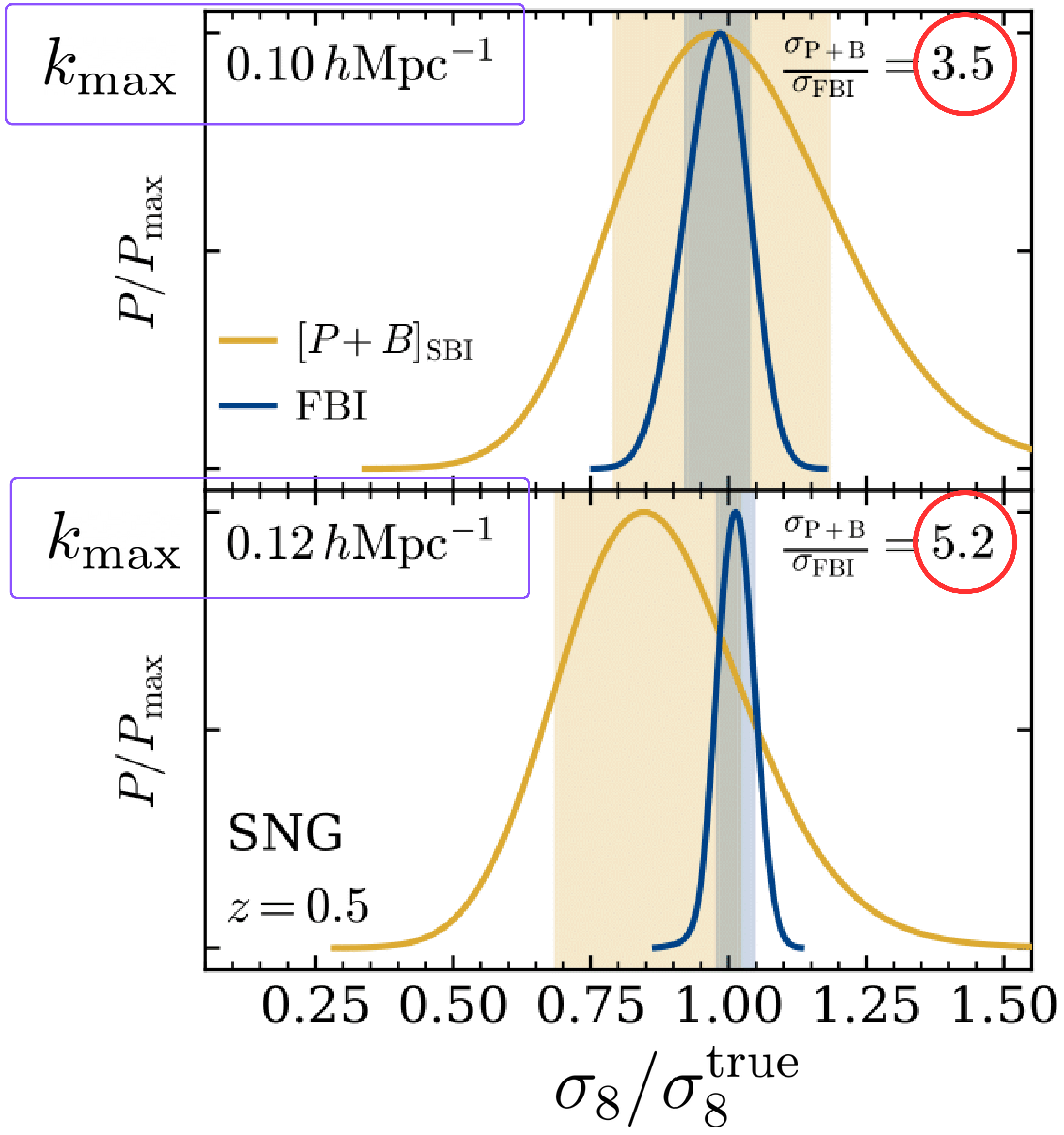
A lot of reliable information at the field-level!

SNG halos



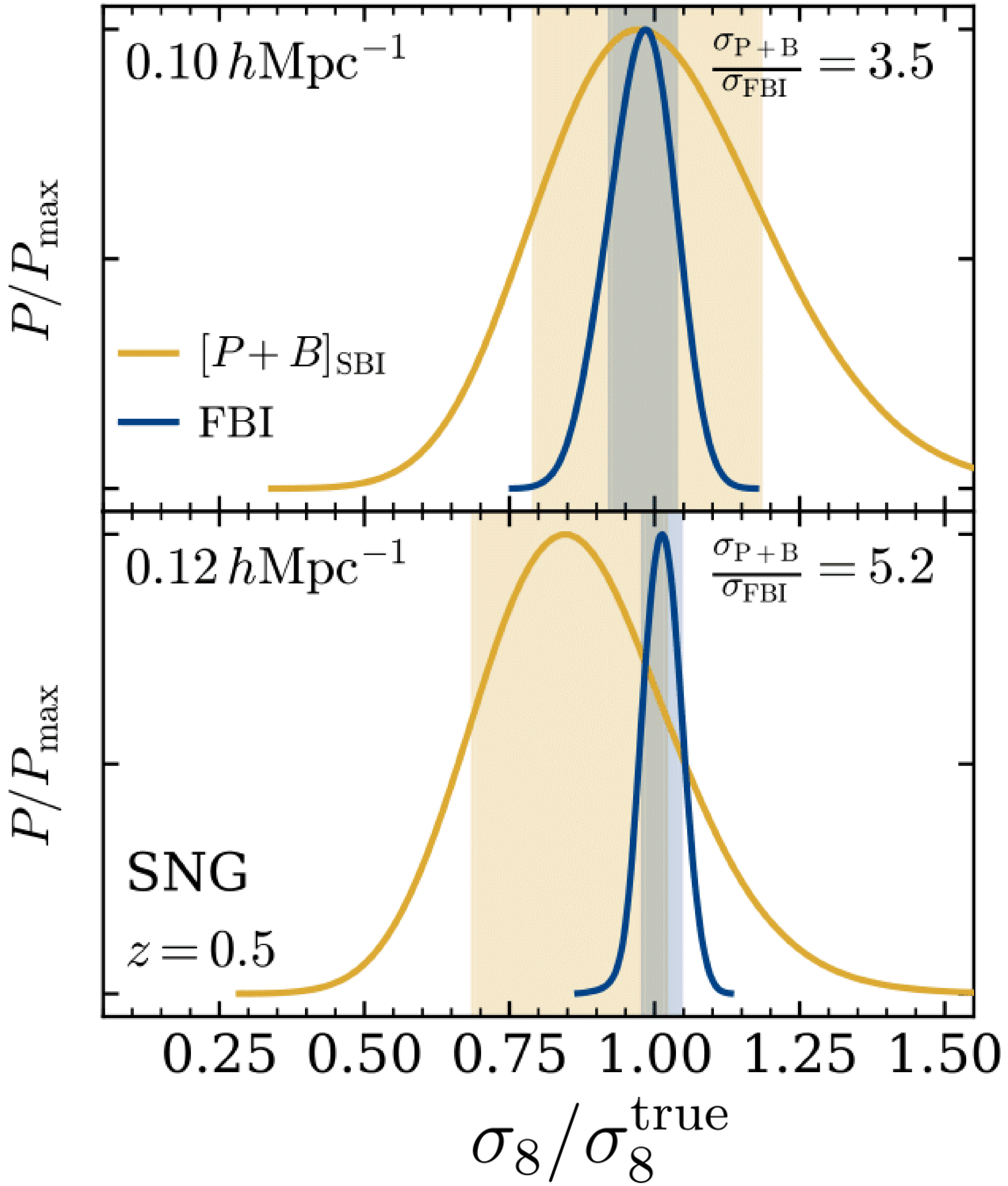
3.5 improvement factor!

SNG halos

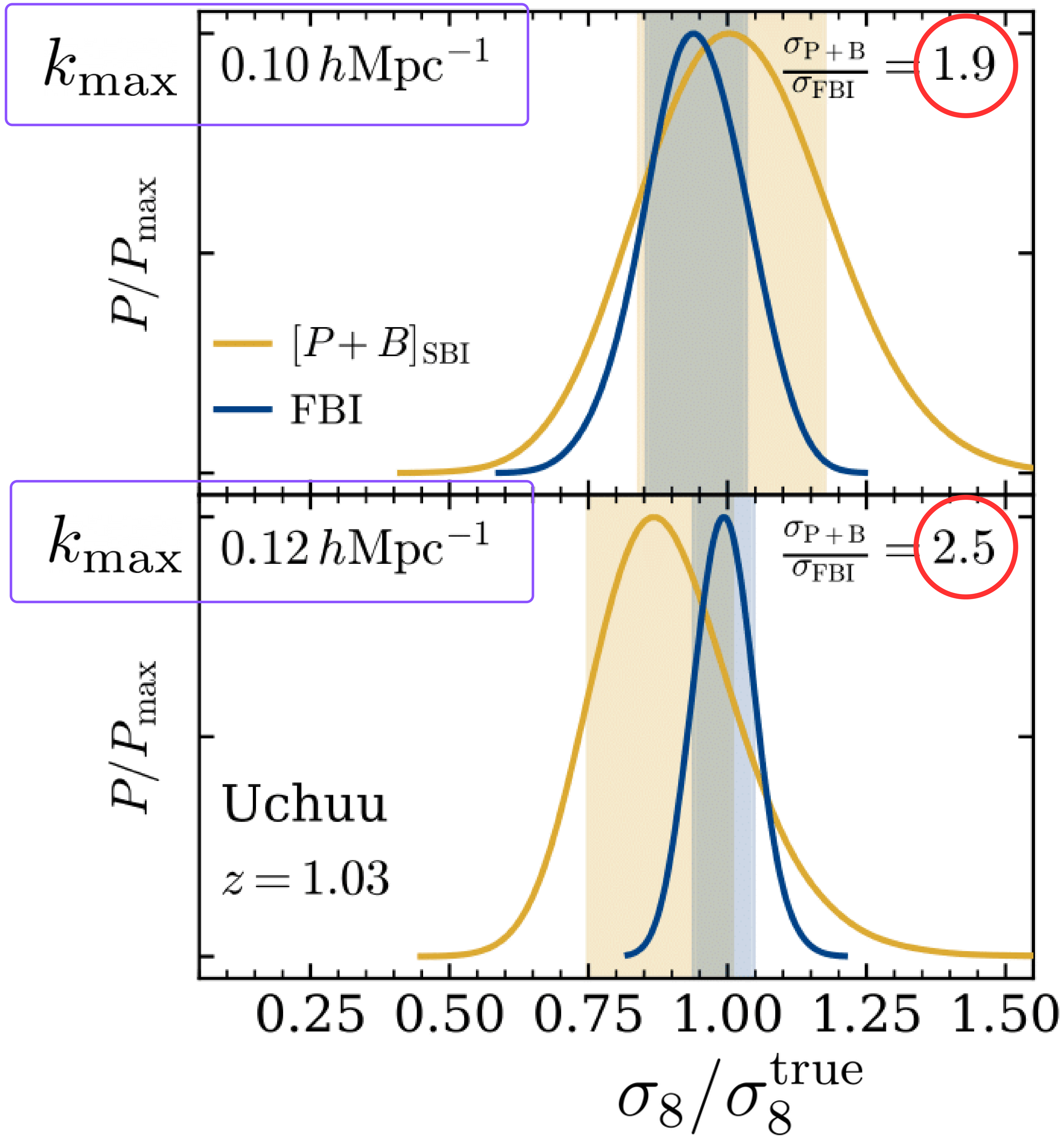


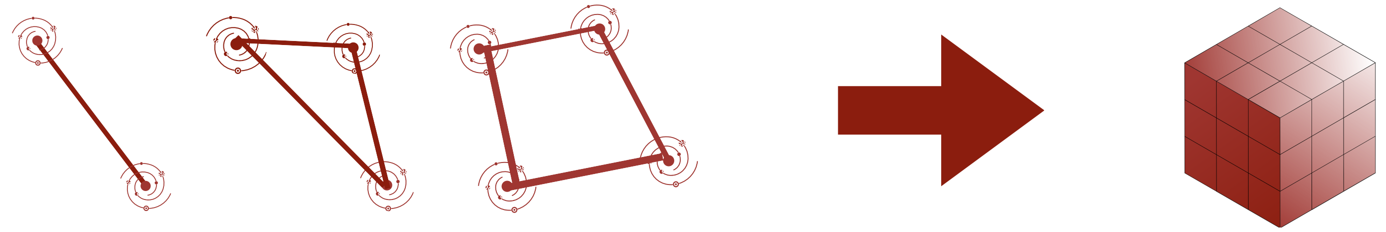
$k_{\rm max}$ $0.10\,h\mathrm{Mpc}^{-1}$
$\frac{\sigma_{\rm P+B}}{\sigma_{\rm FBI}} = 3.5$
$[P+B]_{\rm SBI}$
FBI
$k_{\rm max}$ $0.12\,h\mathrm{Mpc}^{-1}$
$\frac{\sigma_{\rm P+B}}{\sigma_{\rm FBI}} = 5.2$
SNG
$z = 0.5$
$P/P_{\rm max}$
0.25 0.50 0.75 1.00 1.25 1.50
$\sigma_8/\sigma_8^{\rm true}$

SNG halos



Uchuu halos

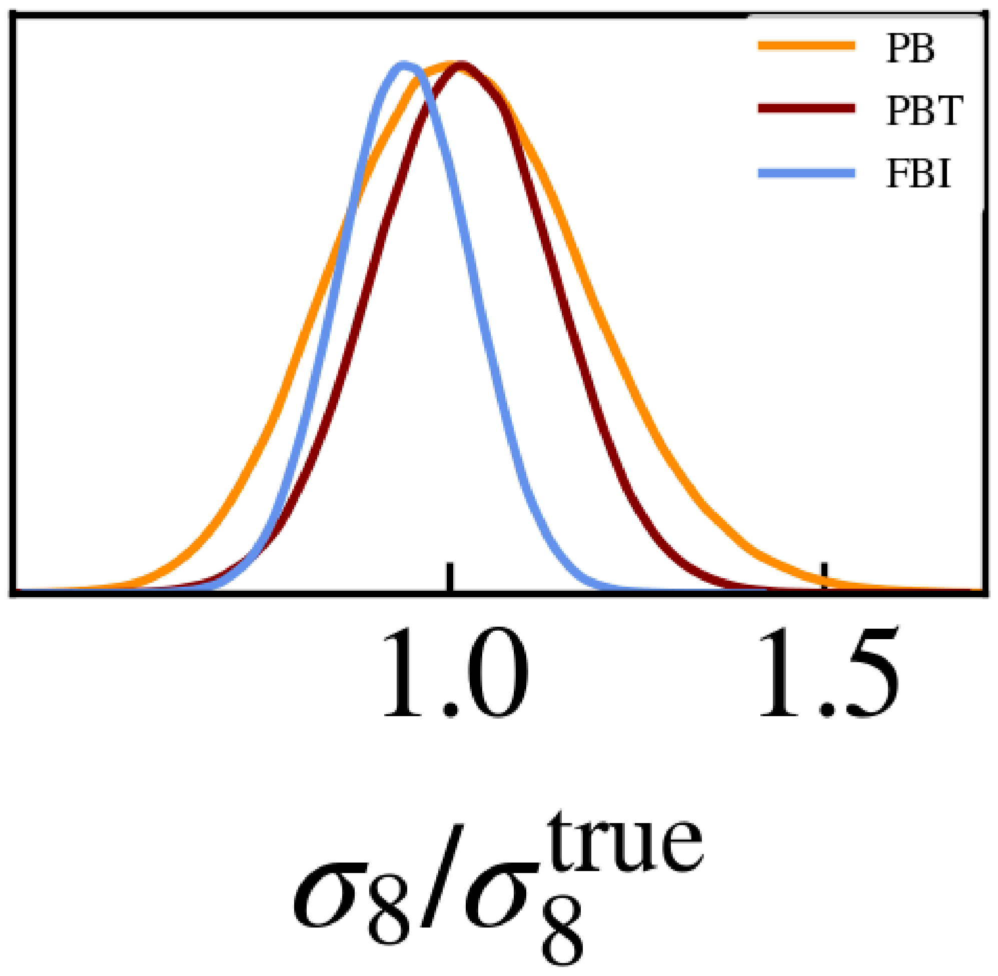


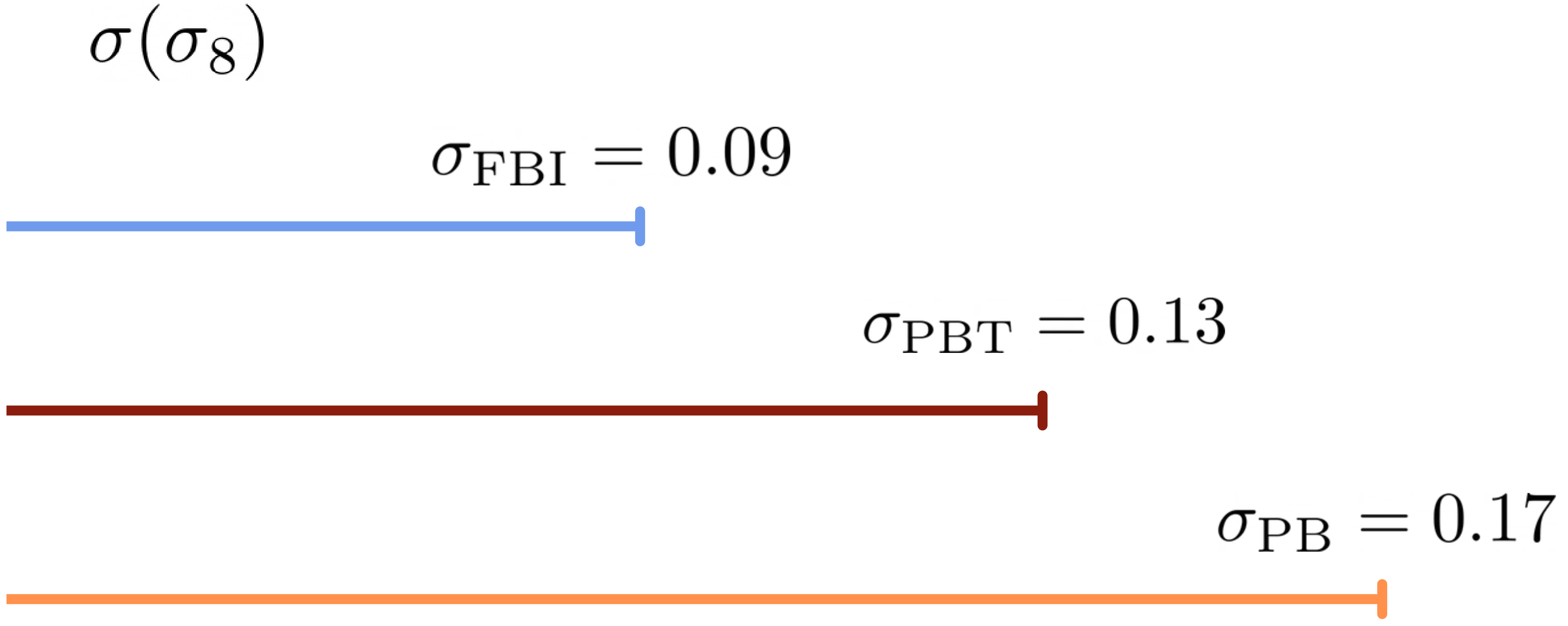


What if we add the **trispectrum**?

Tucci & Schmidt (in prep.)

Trispectrum: **preliminary** results





$k_{\mathrm{max}} = 0.1h\,\mathrm{Mpc}^{-1}$

Uchuu halos at z=1

Brute force approach:
10^6 simulations

Conclusion & Next Steps

- We demonstrated to have **unbiased** and **accurate** results from halo catalogs using LEFTfield for SBI and FBI
- **Apple-to-apple comparison** of field-level inference and SBI shows that there is a lot of **reliable** information beyond 2+3(+4)-point functions in the 3D maps of galaxies

Next steps to connect with observations:

- Include more observational effects
- Expand the cosmological parameter space
- Explore summaries in SBI

MAX-PLANCK-INSTITUT
FÜR ASTROPHYSIK

LEFTfield

Beatriz Tucci

tucci@mpa-garching.mpg.de

On the Bispectrum stochasticity

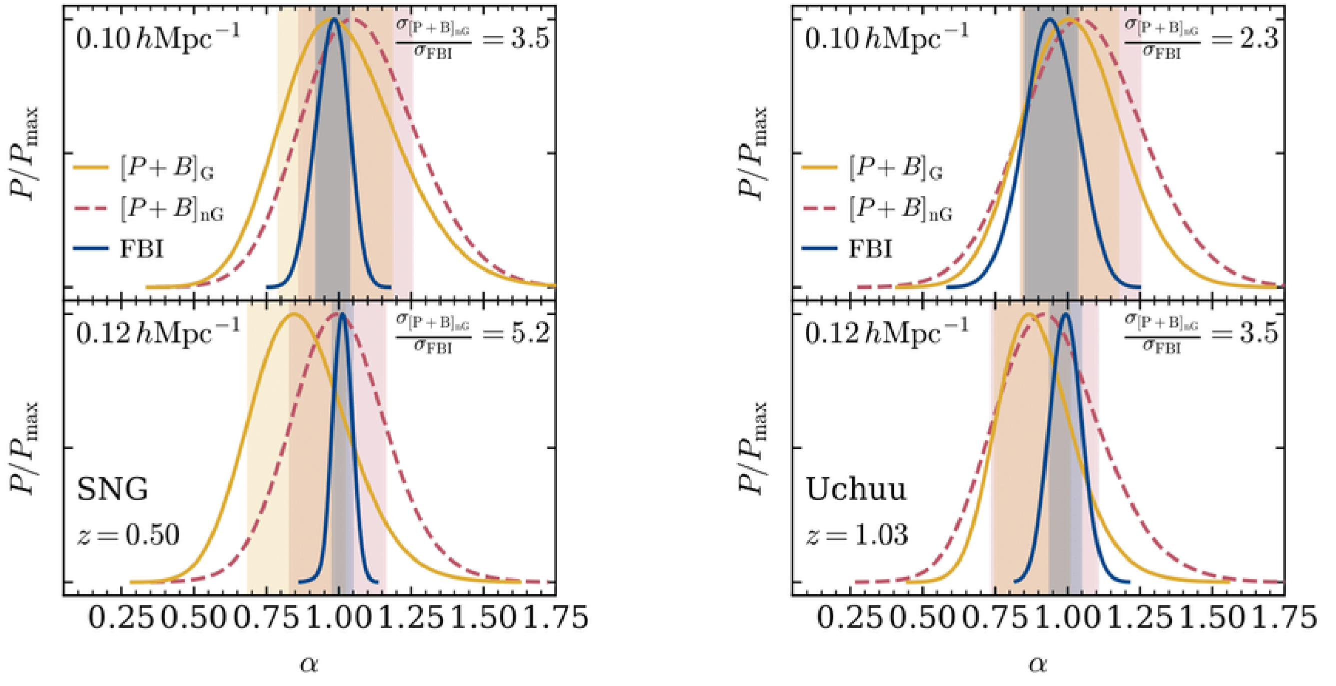
Usual Perturbation Theory

$$\langle \delta_g(k_1)\delta_g(k_2)\delta_g(k_3)\rangle'^{\rm LO}_{\rm stoch} = \boxed{B_\varepsilon} + 2b_1 \boxed{P_{\varepsilon\varepsilon_\delta}} (P_{\rm m}(k_1) + 2 \text{ perm.})$$

Perturbative Forward Model

$$\langle \delta_g(k_1)\delta_g(k_2)\delta_g(k_3)\rangle'^{\rm LO}_{\rm stoch} = \boxed{6c_\varepsilon^{\rm NG} P_\varepsilon^2} + 2b_1 P_\varepsilon \boxed{\sigma_{\varepsilon\delta}} (P_{\rm m}(k_1) + 2 \text{ perm.})$$

$$\boxed{\delta_g(\mathbf{x},\tau) = \delta_{g,\rm det}(\mathbf{x},\tau) + \varepsilon(\mathbf{x},\tau) + \sigma_{\varepsilon\delta}(\tau)\varepsilon(\mathbf{x},\tau)\delta(\mathbf{x},\tau) + c_\varepsilon^{\rm NG}(\tau)\varepsilon^2(\mathbf{x},\tau)} \qquad \varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon^2)$$



$$O_{\rm det} \in [\delta, \delta^2, K^2, O_{\rm td}, \nabla^2\delta]$$
$$O_{\rm stoch} \in [\varepsilon, \varepsilon\delta, \varepsilon^2, \nabla^2\varepsilon]$$

Neural Density Estimation

How to train the model? (For example, NLE)

$$\mathbb{E}_{p(\boldsymbol{\theta})}\left[D_{\mathrm{KL}}\left[\underbrace{p(\mathbf{x}|\boldsymbol{\theta})}_{\text{target density}} \,||\, q_{\underbrace{\phi}_{\substack{\text{neural network}\\ \text{trainable parameters}}}}(\mathbf{x}|\boldsymbol{\theta})\right]\right] = \int d\boldsymbol{\theta}\, p(\boldsymbol{\theta}) \int d\mathbf{x}\, p(\mathbf{x}|\boldsymbol{\theta}) \log\left(\frac{p(\mathbf{x}|\boldsymbol{\theta})}{q_\phi(\mathbf{x}|\boldsymbol{\theta})}\right)$$

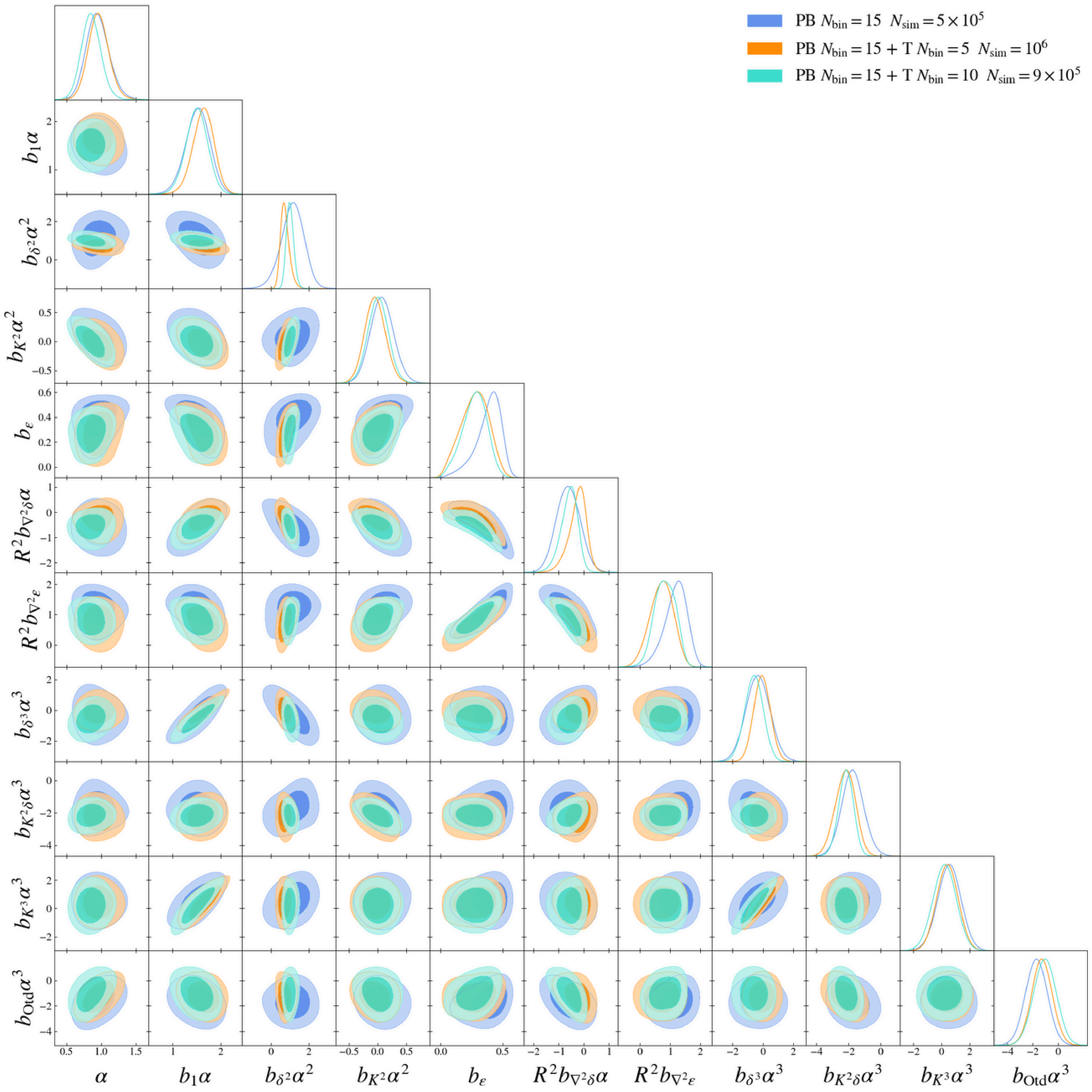
$$= \int d\boldsymbol{\theta}\, d\mathbf{x}\, p(\boldsymbol{\theta}, \mathbf{x}) \log\left(\frac{p(\mathbf{x}|\boldsymbol{\theta})}{q_\phi(\mathbf{x}|\boldsymbol{\theta})}\right)$$

$$= -\mathbb{E}_{p(\boldsymbol{\theta},\mathbf{x})}\left[\log q_\phi(\mathbf{x}|\boldsymbol{\theta})\right] + \text{const.}$$

$$\approx -\frac{1}{N_{\text{sim}}} \sum_{n=1}^{N_{\text{sim}}} \log q_\phi(\mathbf{x}_n|\boldsymbol{\theta}_n) + \text{const.}\,,$$

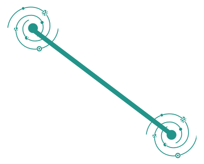
$$\boxed{\{(\boldsymbol{\theta}_n, \mathbf{x}_n)\}_{n=1}^{N_{\text{sim}}}}$$

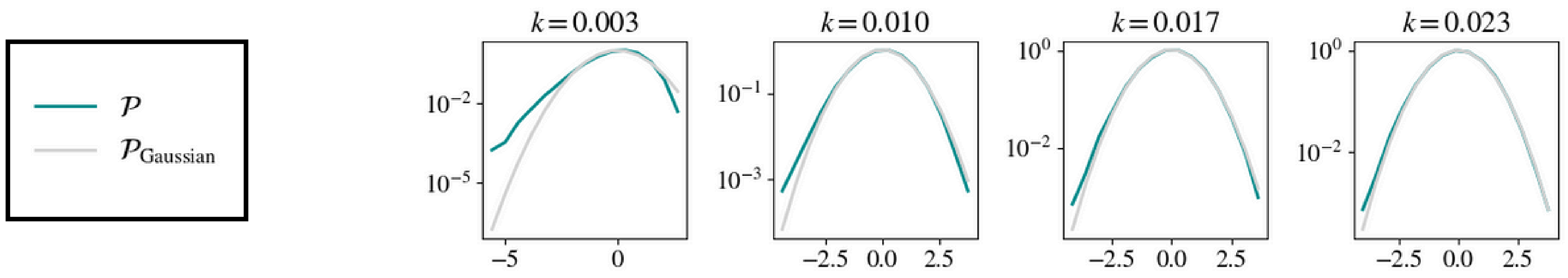
PB vs PBT


PB $N_{\rm bin}=15$ $N_{\rm sim}=5\times10^5$
PB $N_{\rm bin}=15$ + T $N_{\rm bin}=5$ $N_{\rm sim}=10^6$
PB $N_{\rm bin}=15$ + T $N_{\rm bin}=10$ $N_{\rm sim}=9\times10^5$
α
$b_1\alpha$
$b_{\delta^2}\alpha^2$
$b_{K^2}\alpha^2$
b_ϵ
$R^2 b_{\nabla^2\delta}\alpha$
$R^2 b_{\nabla^2\epsilon}$
$b_{\delta^3}\alpha^3$
$b_{K^2\delta}\alpha^3$
$b_{K^3}\alpha^3$
$b_{\rm Otd}\alpha^3$

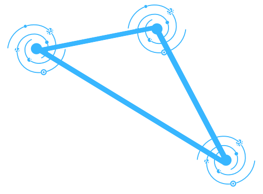
On the Gaussianity assumption of the n-point functions

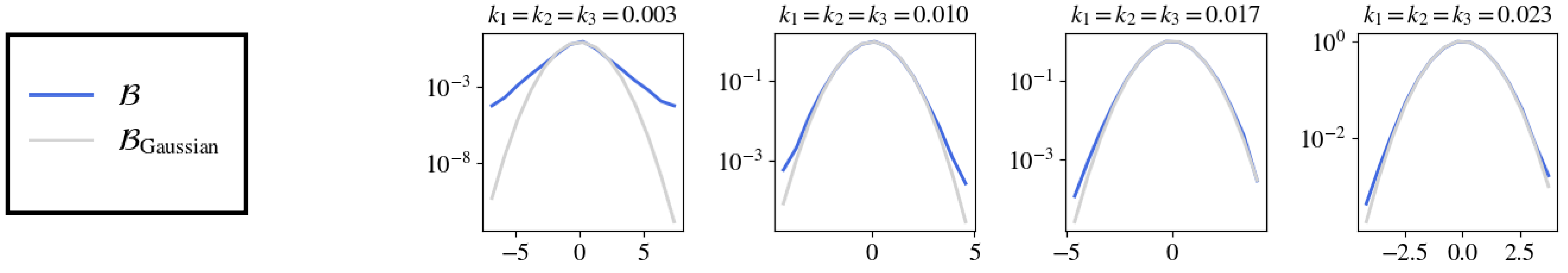
power-spectrum



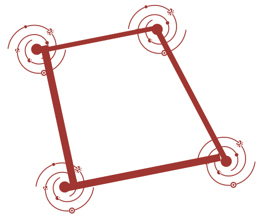


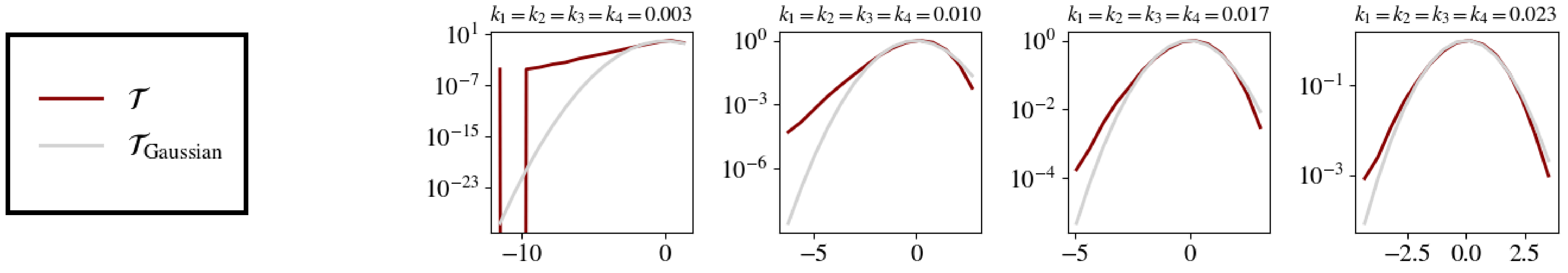
bispectrum





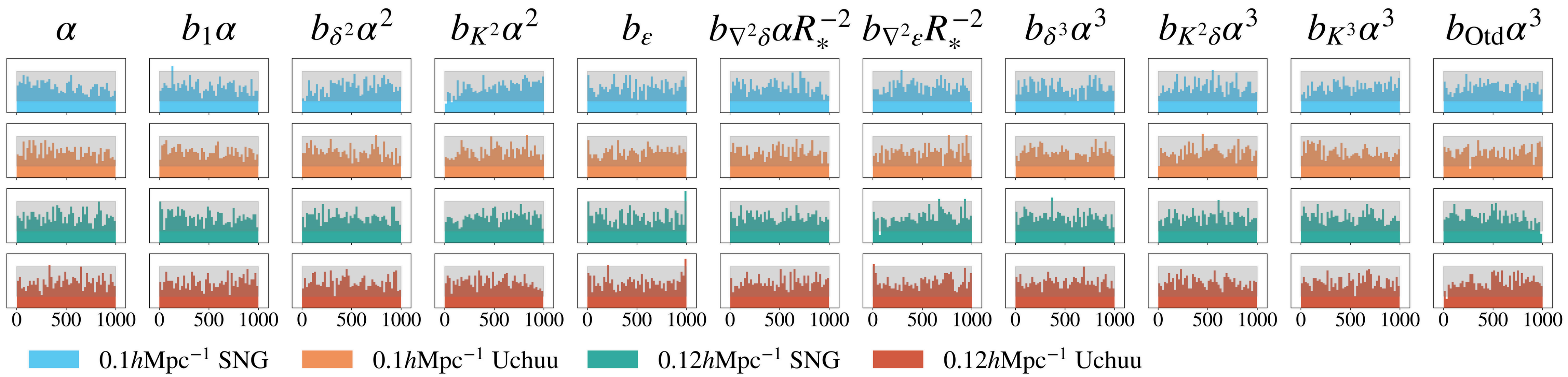
trispectrum



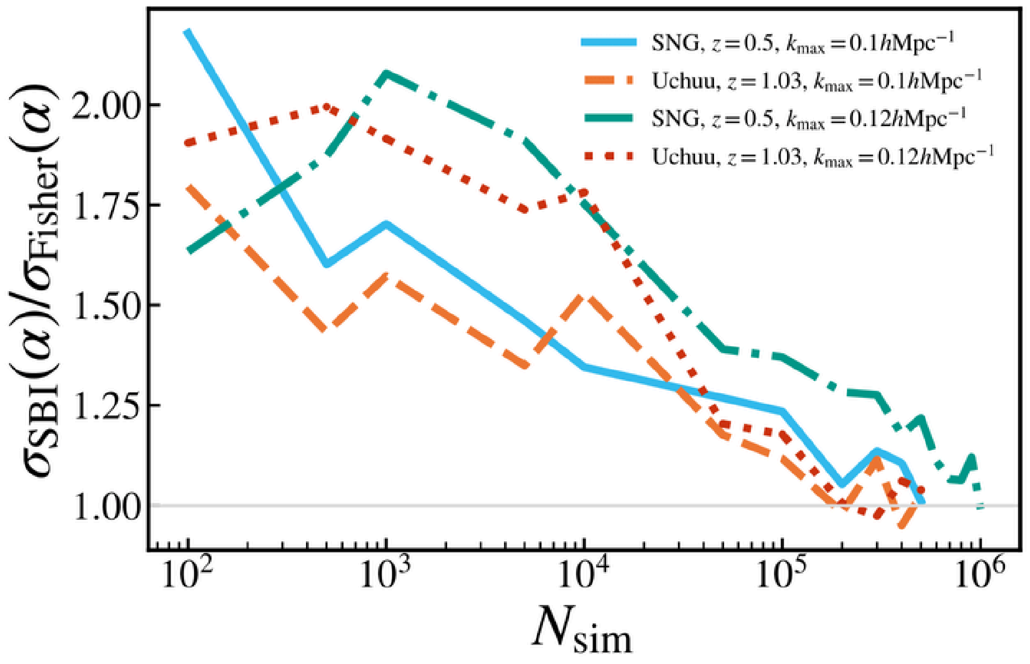


SBI posterior diagnostics

Simulation-based calibration (as in Talts et al. 2018)

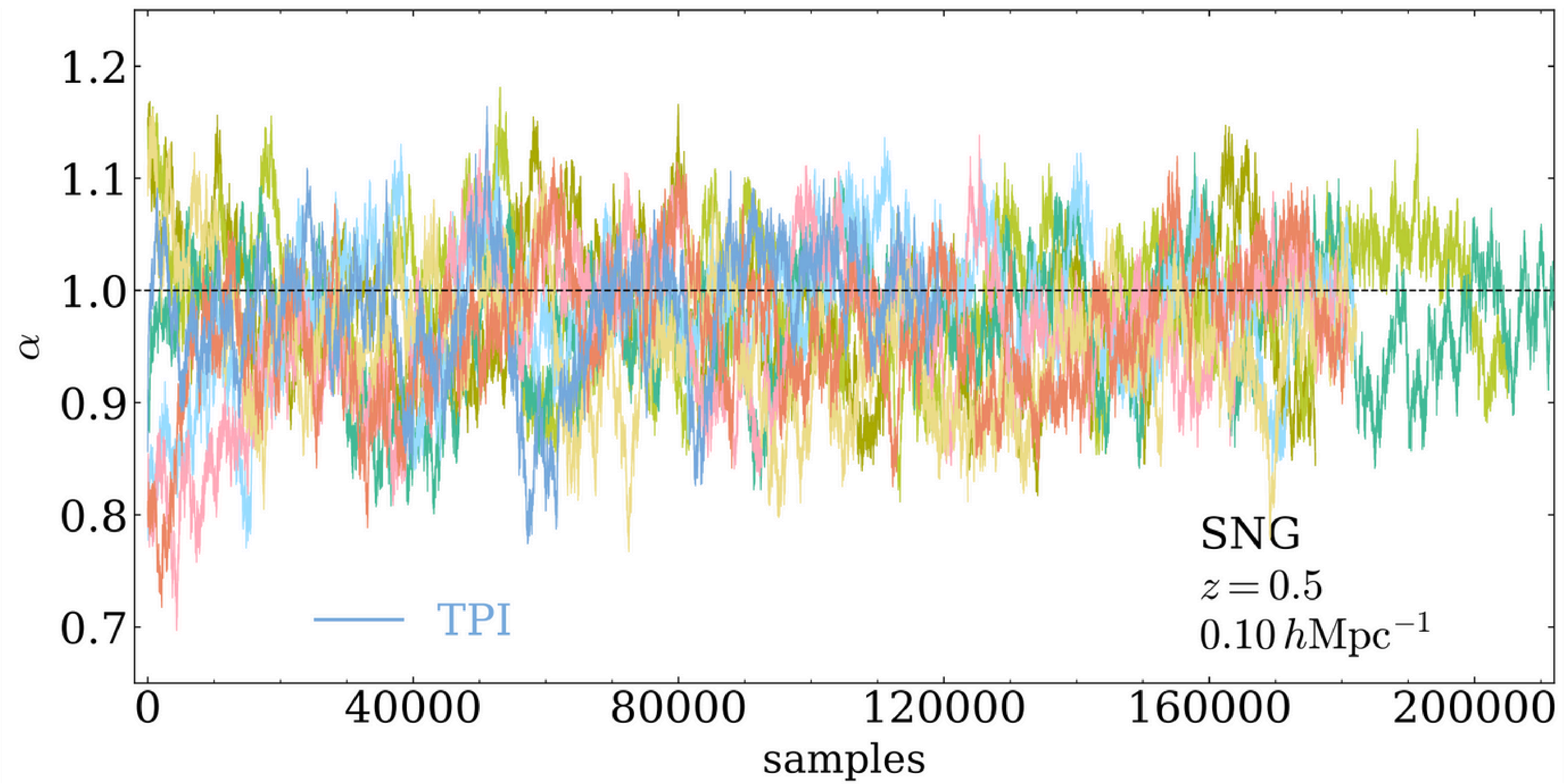


Convergence with respect to simulation budget

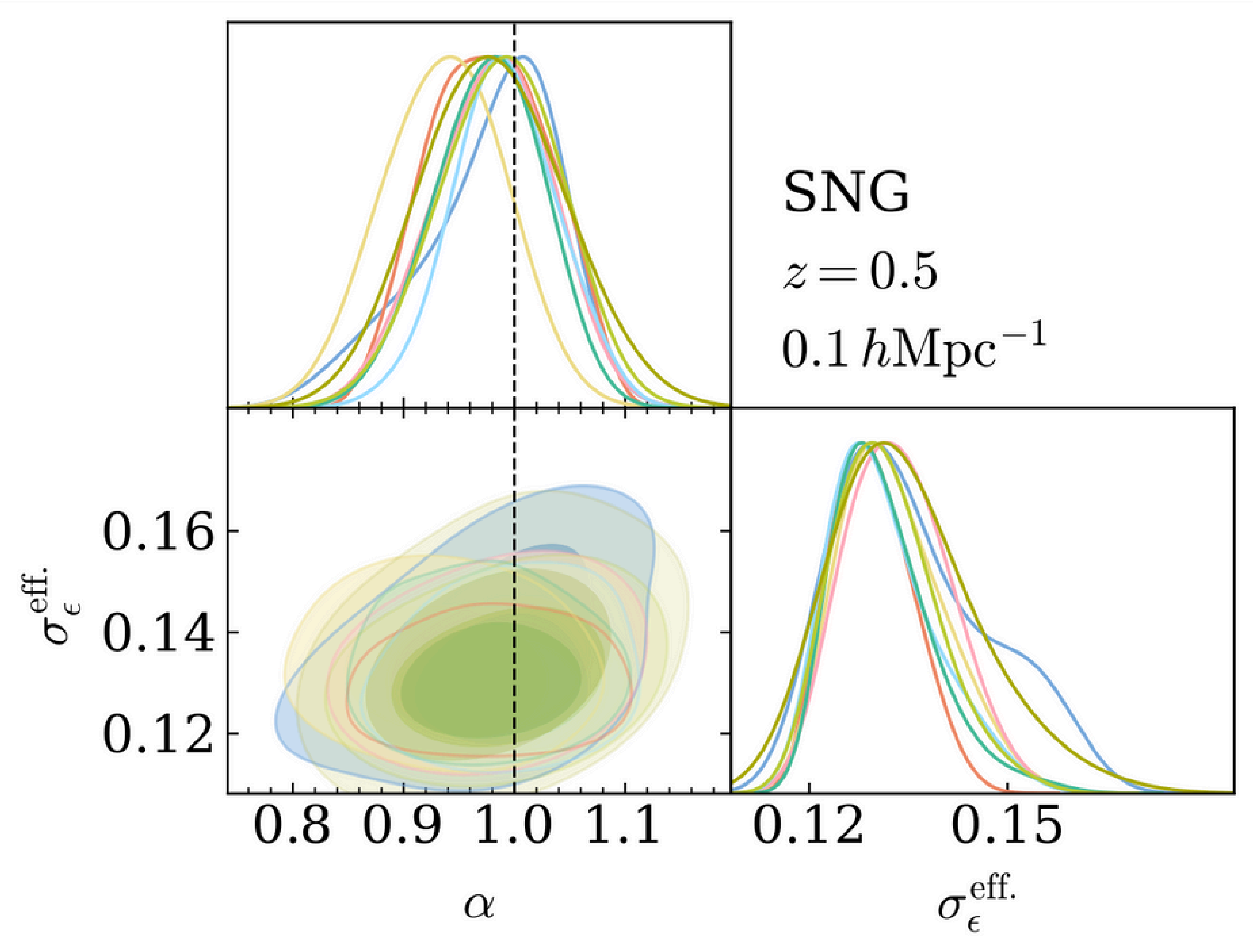


FBI: posterior diagnostics

MCMC convergence



Posterior consistency



The field-level galaxy likelihood

Schmidt et al. (2019)
Cabass & Schmidt (2020)

$$\mathcal{P}[\delta_g|\boldsymbol{\theta}] = \int \mathcal{D}\delta_{\rm in}\mathcal{P}[\delta_{\rm in}] \int \mathcal{D}\varepsilon\mathcal{P}[\varepsilon]\,\mathcal{P}[\delta_g|\boldsymbol{\theta},\delta_{\rm in},\varepsilon]$$

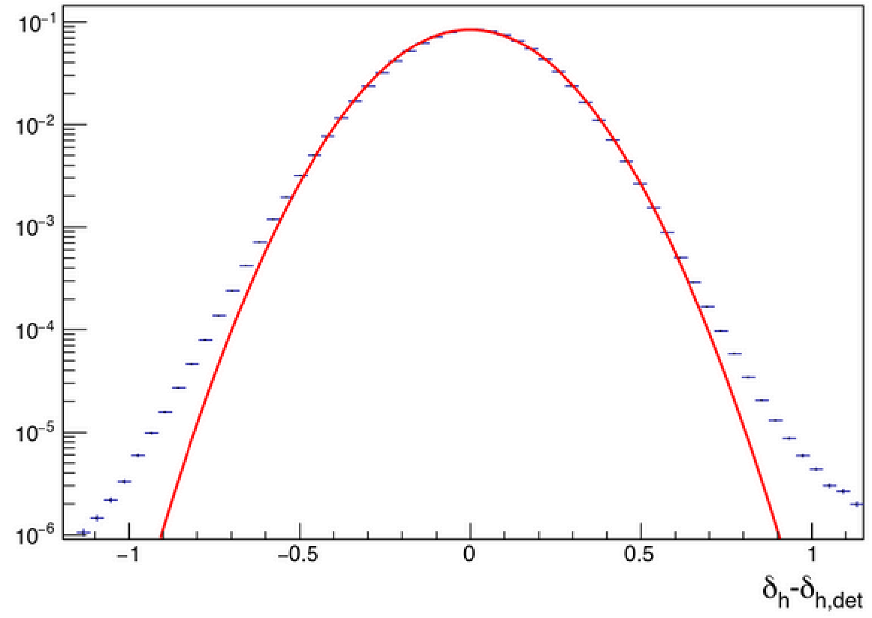
$$= \int \mathcal{D}\delta_{\rm in}\mathcal{P}[\delta_{\rm in}] \int \mathcal{D}\varepsilon\mathcal{P}[\varepsilon]\,\delta_D(\delta_g - \delta_{g,\rm det}[\boldsymbol{\theta},\delta_{\rm in}] - \varepsilon)$$

Assume Gaussian stochasticity

$$\mathcal{P}[\varepsilon] \propto \exp\left[-\frac{1}{2}\int_{\boldsymbol{k}} \frac{|\varepsilon(\boldsymbol{k})|^2}{P_\varepsilon(k)}\right]$$

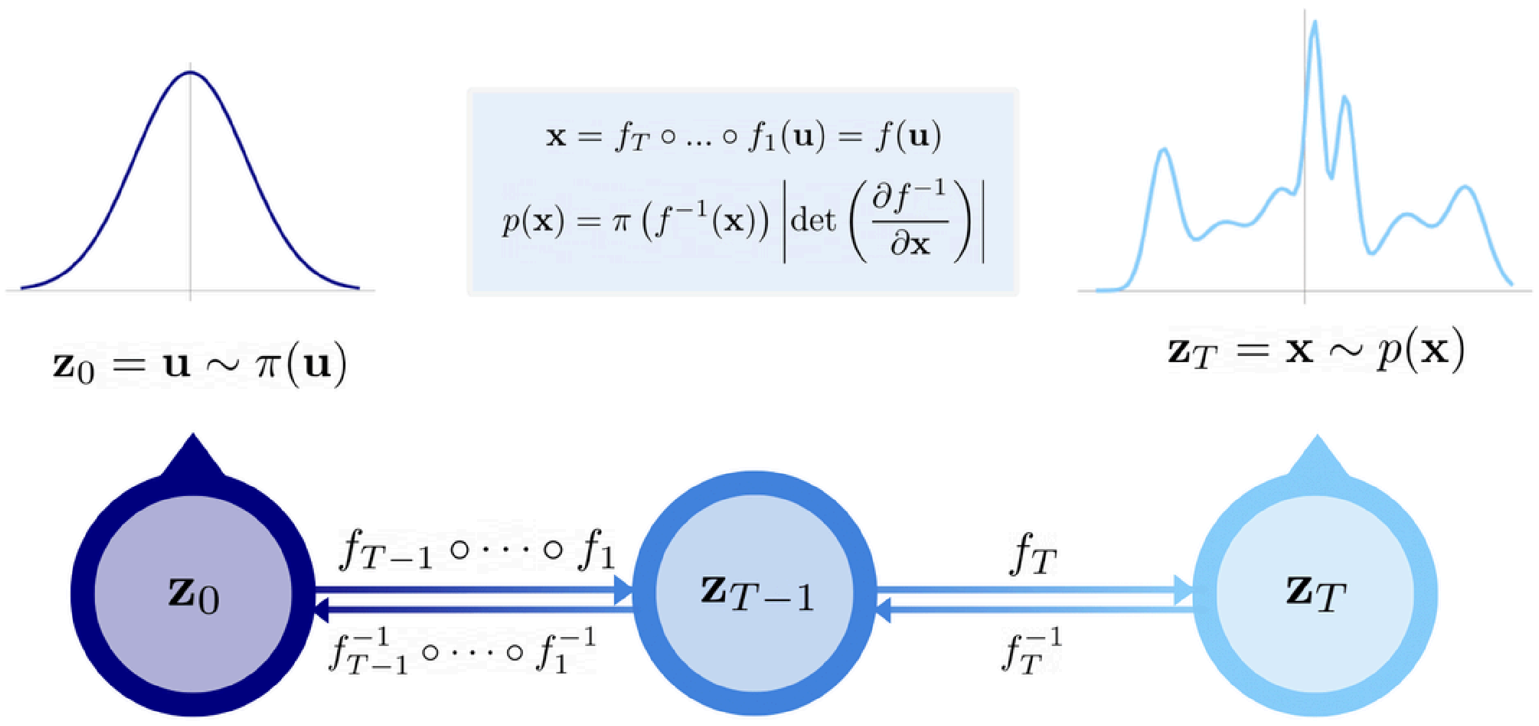
$$P_\varepsilon(k) \equiv \langle\varepsilon(\boldsymbol{k})\varepsilon(-\boldsymbol{k})\rangle'$$

$$\delta_{\rm h} = \delta_{\rm h,det} + \delta_{\rm h,stoch}$$



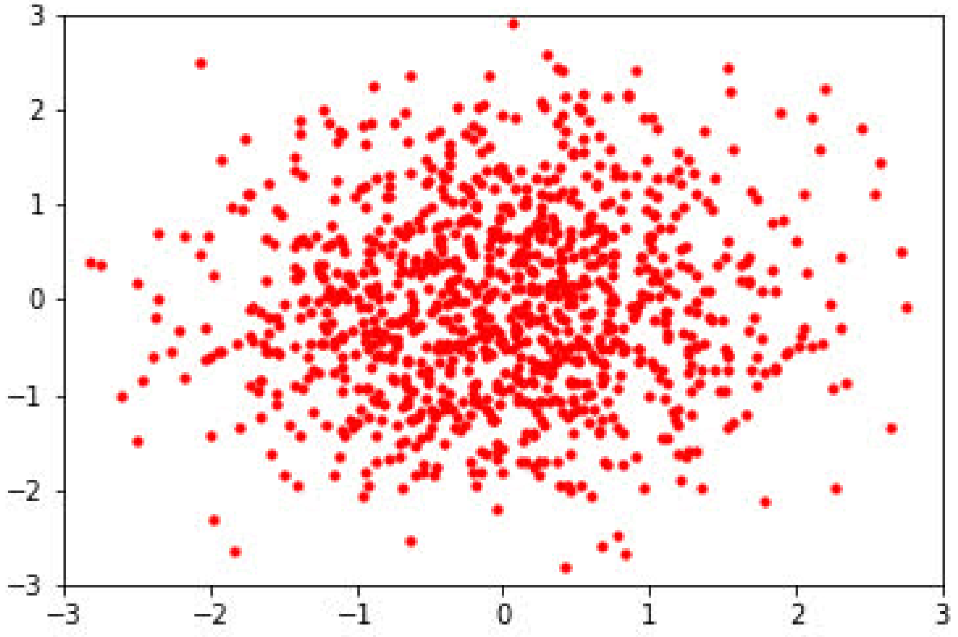
Schmidt et al. (2020)

Normalizing Flows



Tucci, Schmidt (2023)

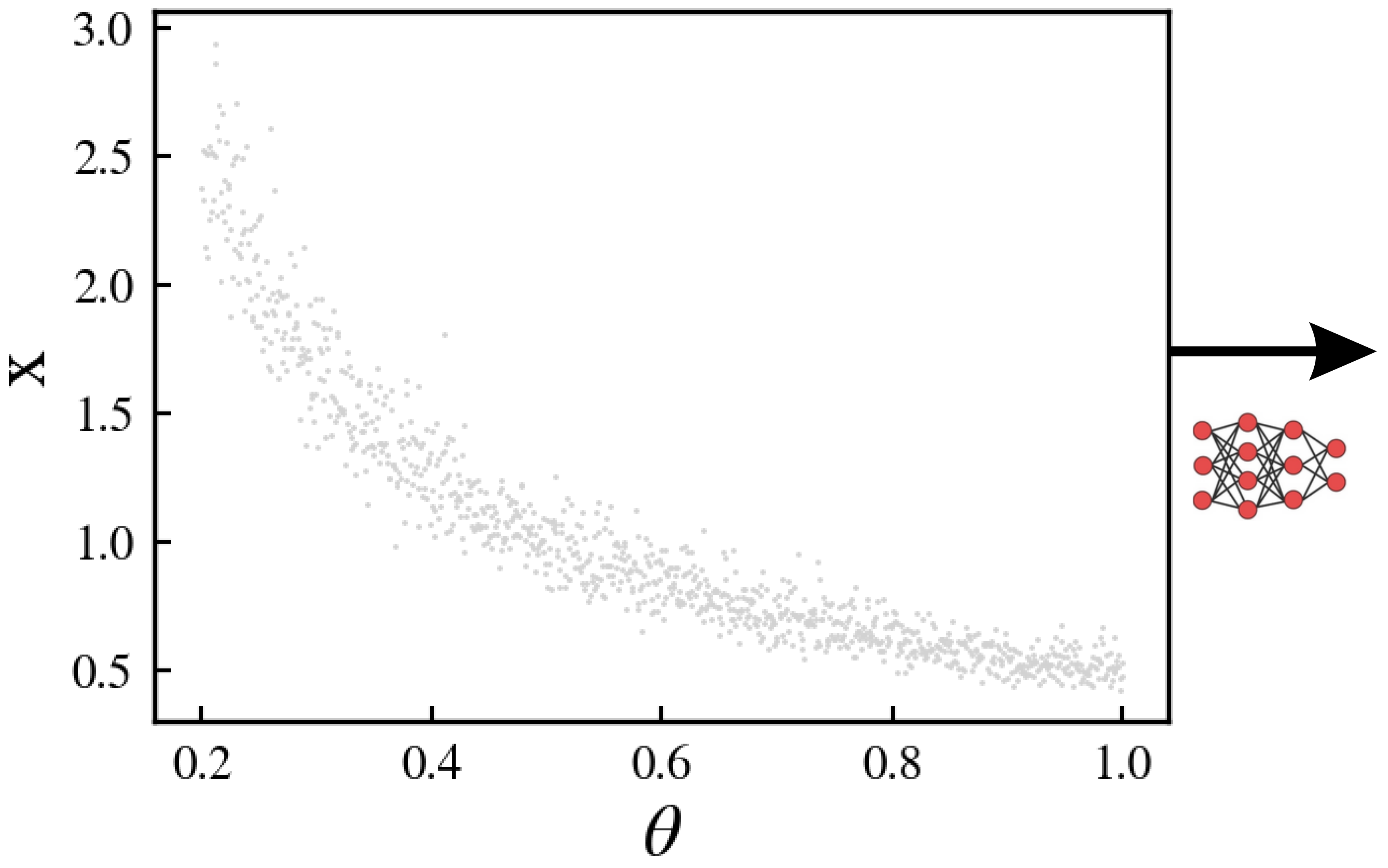
$$\boxed{q_\phi(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_0|\mathbf{0}, \mathbf{I}) \prod_{t=1}^{T} \left|\det\left(\frac{\partial f_t}{\partial \mathbf{z}_{t-1}}\right)\right|^{-1}}$$



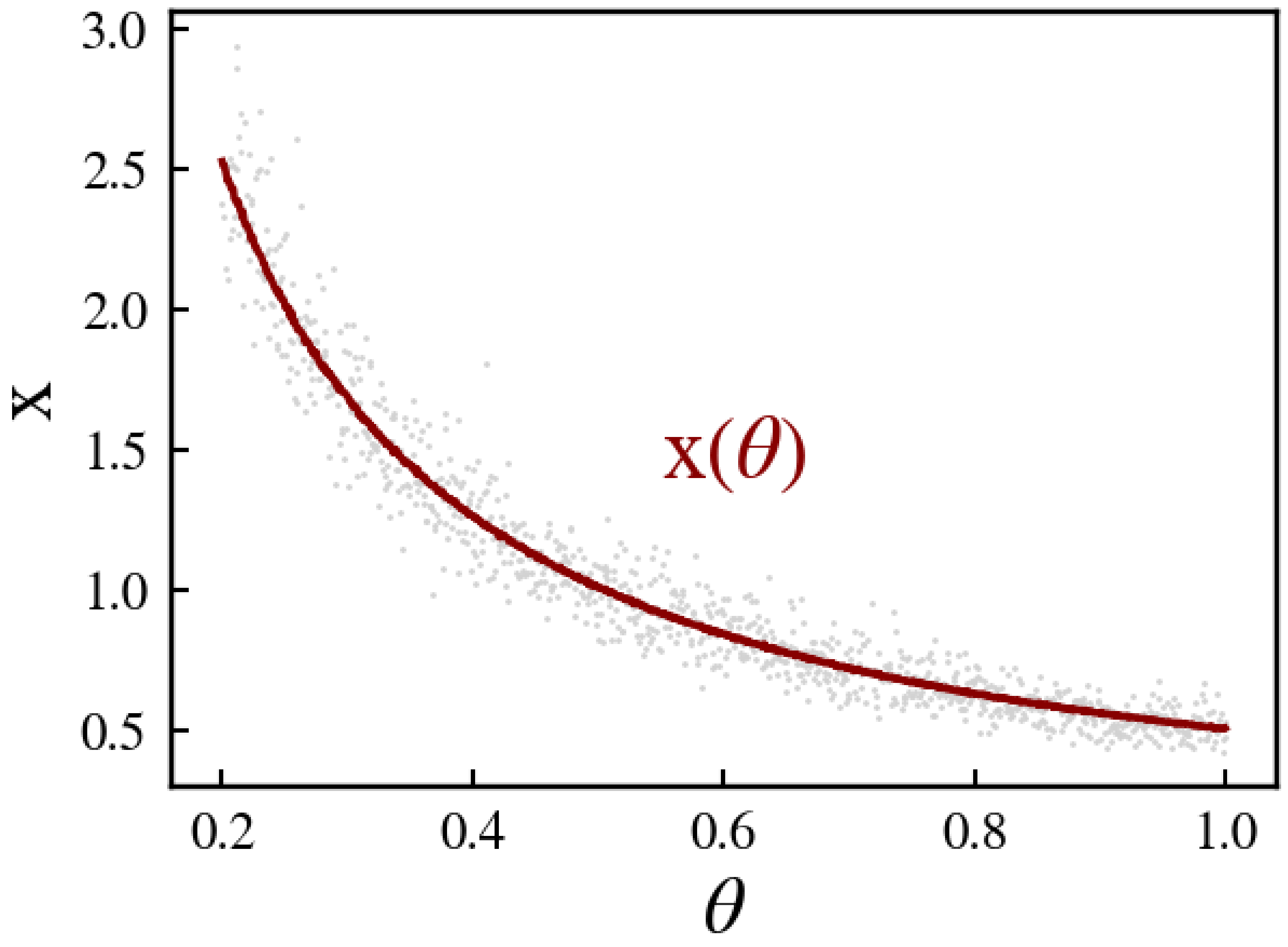
Credits: Miles Cranmer

Simulation-based inference

$$\{(\boldsymbol{\theta}_n, \mathbf{x}_n)\}_{n=1}^{N_{\text{sim}}}$$



Summary statistics emulators



This is not what we are doing!

Simulation-based calibration (SBC)

How to check if the obtained posterior uncertainty is correct?

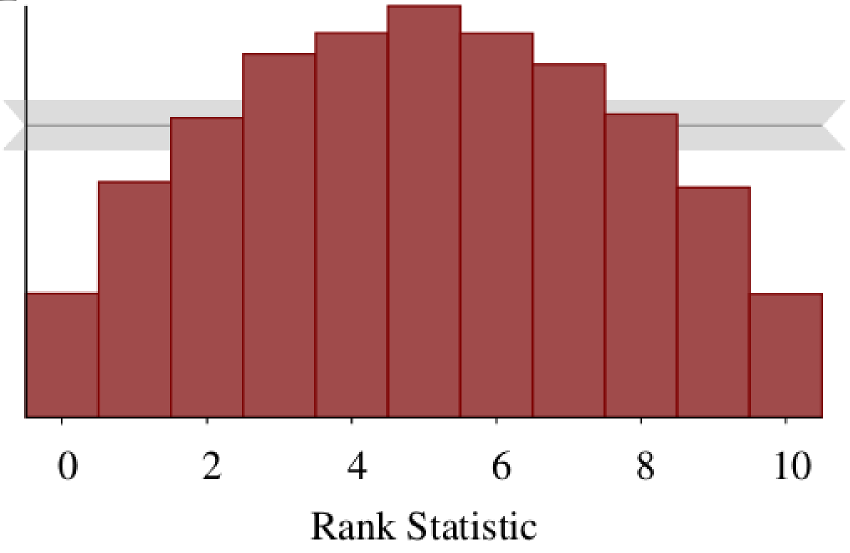
$$\mathbf{x}_o^i = \text{simulator}(\boldsymbol{\theta}_o^i)$$

$$\{\hat{\boldsymbol{\theta}}\}_i \sim \hat{p}(\boldsymbol{\theta}|\mathbf{x}_o^i)$$

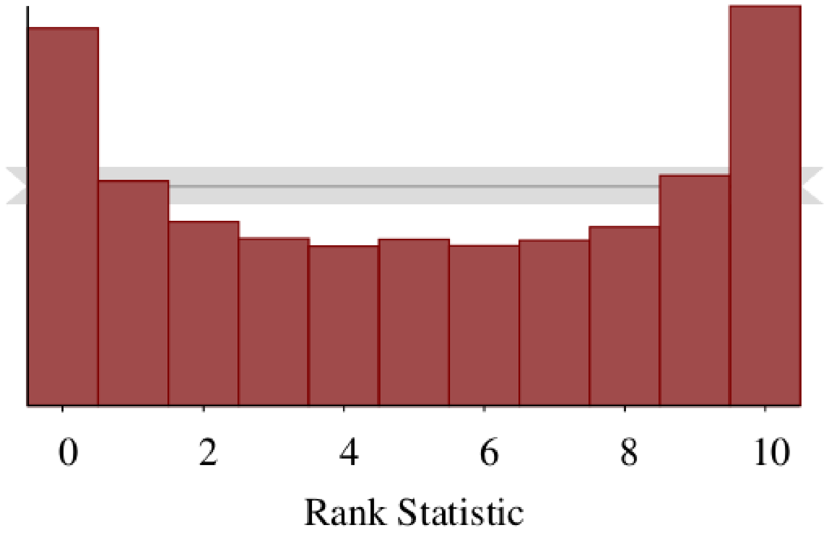
$$\hat{\theta}_1 < \hat{\theta}_2 < \ldots < \hat{\theta}_{\text{rank}} < \theta_o^i < \ldots < \hat{\theta}_{\text{Nsamples}}$$

Ranks should be uniformly distributed if the posterior is well calibrated

Underconfident posterior



Superconfident posterior



Beyond 2-point mock data challenge

Krause, ..., Nguyen, Schmidt+ (2024)
arXiv:2405.02252