Opportunities and Challenges of SBI
for galaxy clustering

Changhoon Hahn

Princeton University

changhoon.hahn@princeton.edu

changhoonhahn.github.io

Sesto — New Strategies for Extracting Cosmology from Galaxy Surveys— July 3, 2024

what is simulation-based inference?

opportunities for simulation-based inference?

challenges for simulation-based inference?

what is simulation-based inference?

opportunities for simulation-based inference?

challenges for simulation-based inference?

goal: infer the *posterior of cosmological parameters* given observations

$$p(\theta \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \theta)\, p(\theta)}{p(\mathbf{X})}$$

Bayes

goal: infer the *posterior of cosmological parameters* given observations

likelihood *prior*

$$p(\theta \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \theta)\, p(\theta)}{p(\mathbf{X})}$$

posterior *evidence*

goal: infer the *posterior of cosmological parameters* given observations

$$p(\theta \,|\, \mathbf{X}) \propto p(\mathbf{X} \,|\, \theta)\, p(\theta)$$

lets ignore evidence since it's independent of θ

goal: infer the *posterior of cosmological parameters* given observations

$$p(\theta \,|\, \mathbf{X}) \propto p(\mathbf{X} \,|\, \theta)\, p(\theta)$$

$$\log p(\mathbf{X} \,|\, \theta) = \log \mathscr{L} = [\, (m(\theta) - \mathbf{X})^T\, \mathbf{C}^{-1}\, (m(\theta) - \mathbf{X})\,] + \log(2\pi)^{-k/2}\, |\, \mathbf{C}\, |^{-1/2}$$

m = *your favorite theory model for* $\mathbf{X}$

$\mathbf{C}$ = *covariance matrix from mocks*

standard likelihood in cosmology

what is **simulation-based inference**?

$$\mathbf{X}' \sim F(\theta')$$

what is **simulation-based inference**?

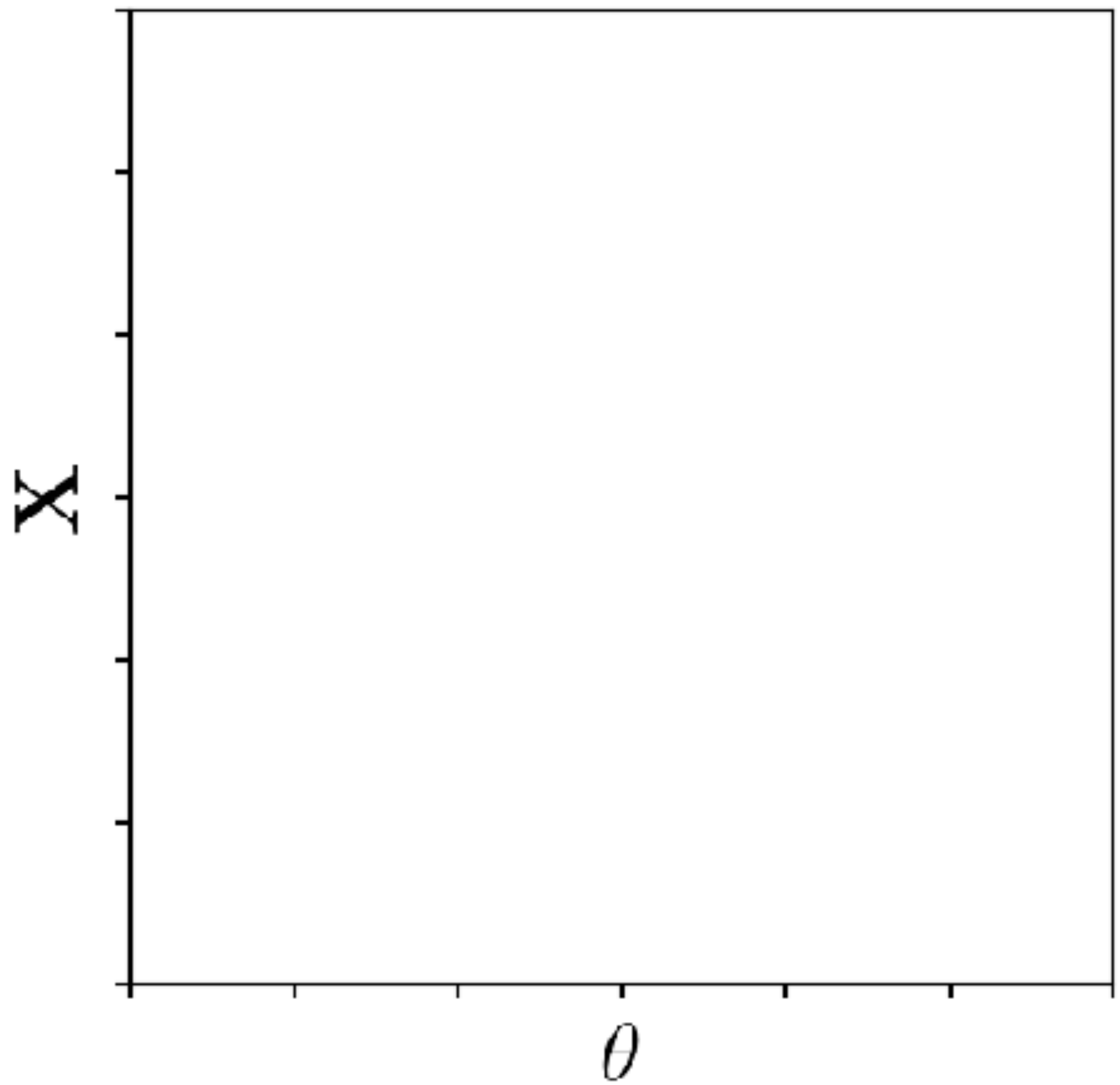
some stochastic forward model/simulator

$$\mathbf{X}' \sim F(\theta')$$

what is **simulation-based inference**?

some stochastic forward model/simulator

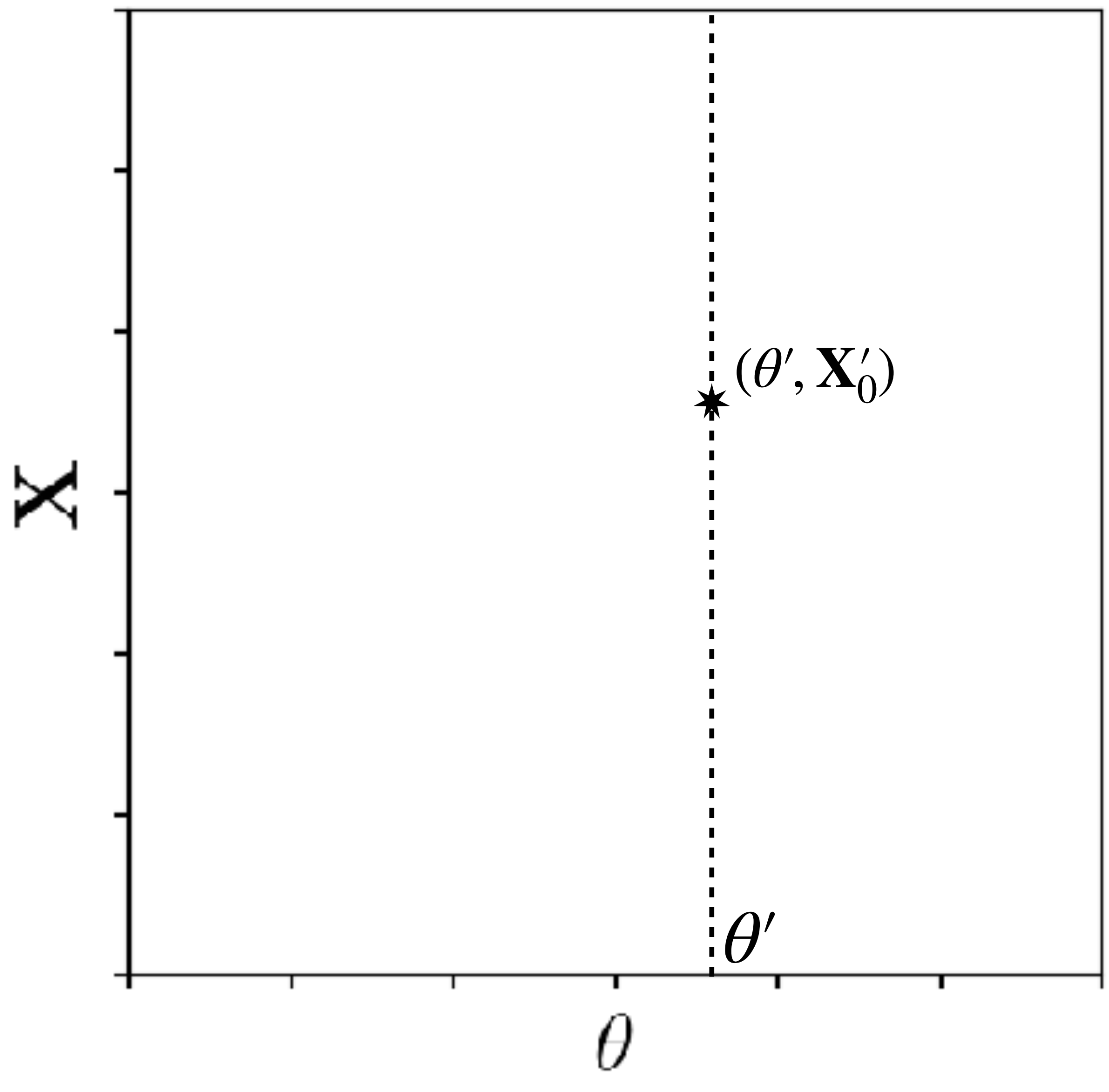
$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

some stochastic forward model/simulator

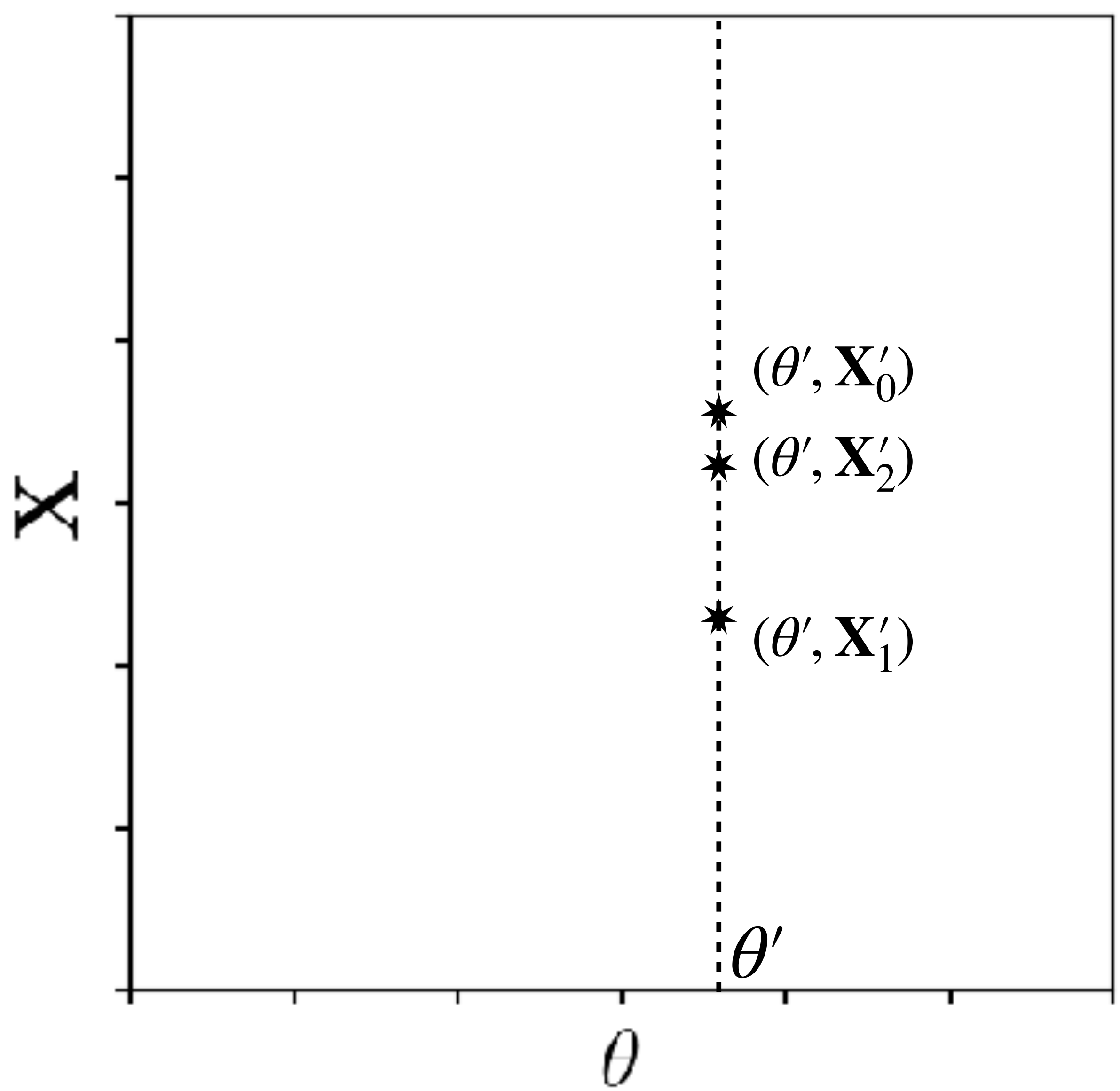
$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

some stochastic forward model/simulator

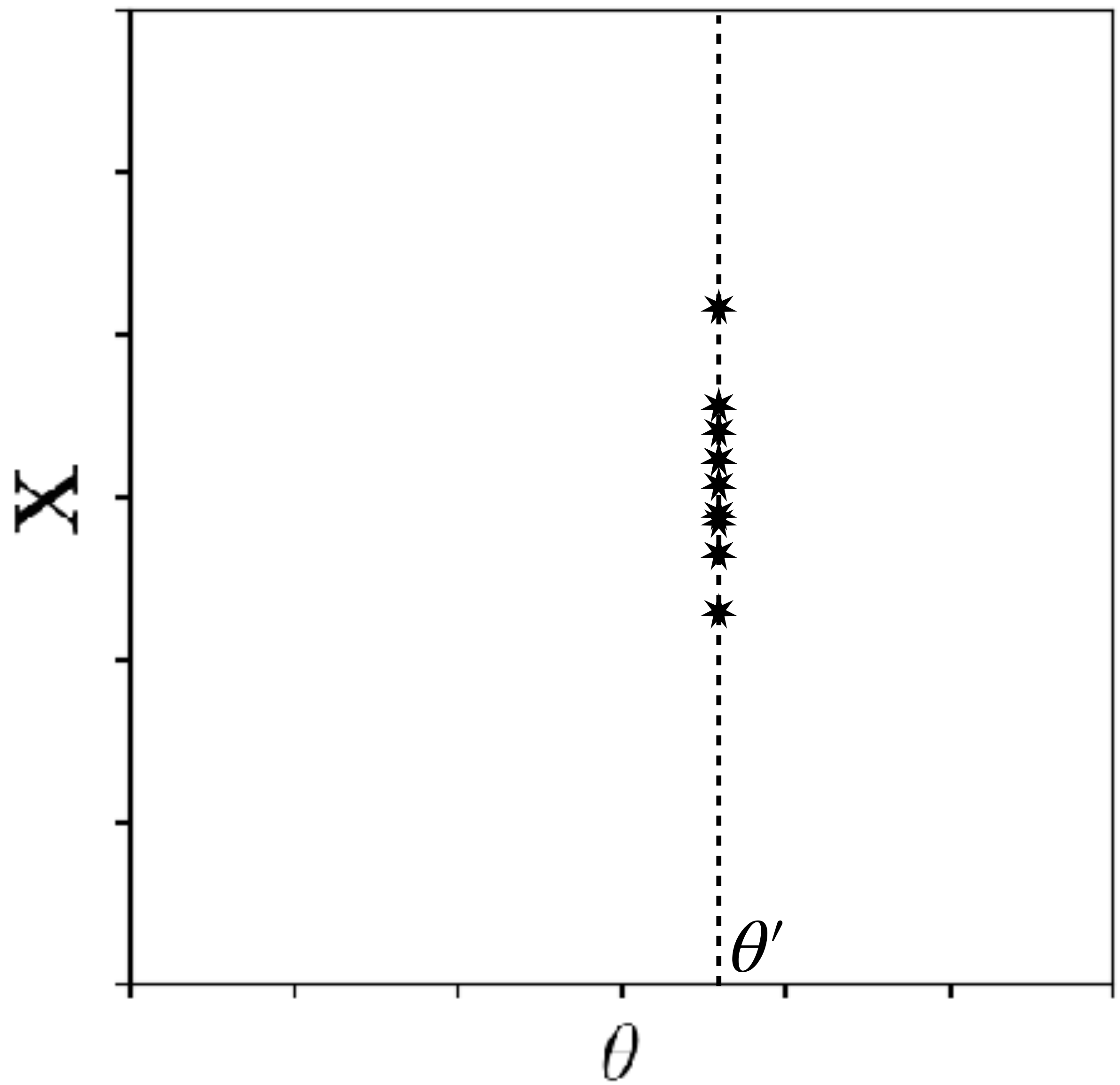
$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

some stochastic forward model/simulator

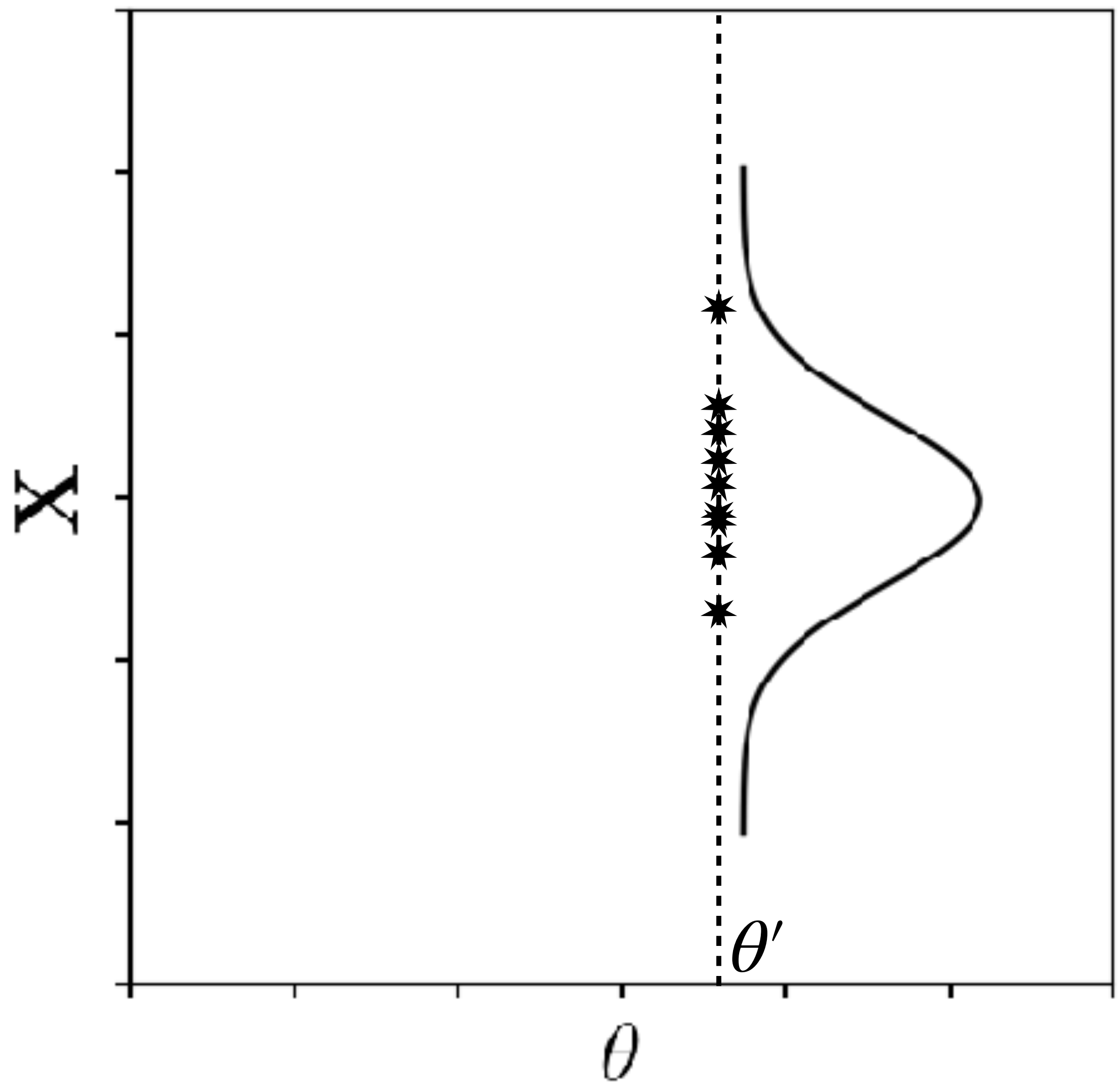
$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

the forward model/simulator implicitly defines our likelihood

$$\mathbf{X}' \sim F(\theta') \qquad \equiv \qquad \mathbf{X}' \sim p(\mathbf{X} \mid \theta')$$



what is **simulation-based inference**?

the forward model/simulator implicitly defines our likelihood

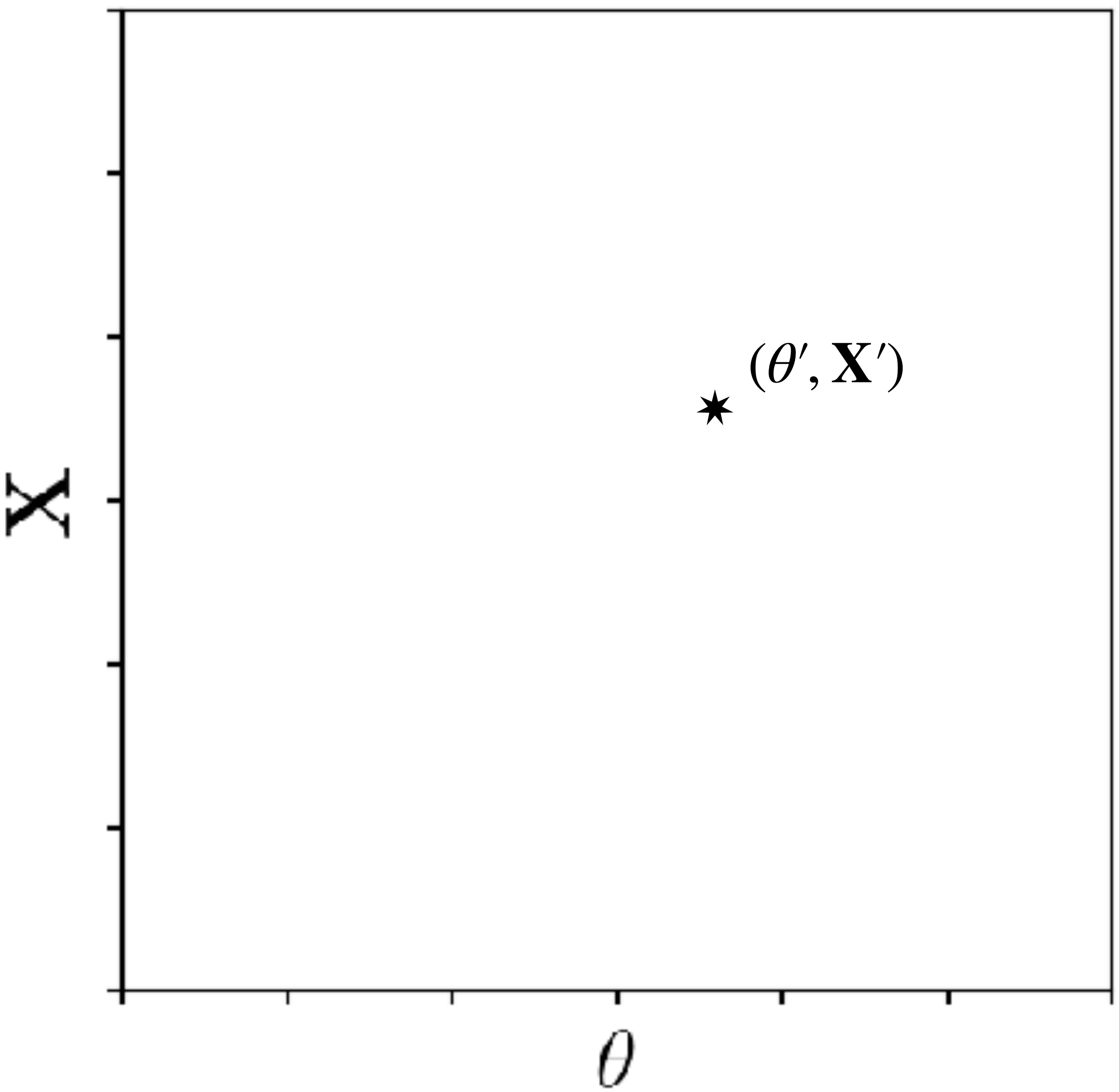
$$\mathbf{X}' \sim F(\theta') \quad \equiv \quad \mathbf{X}' \sim p(\mathbf{X}\,|\,\theta')$$

1. sample the prior

$$\theta' \sim p(\theta)$$

2. run simulator

$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

the forward model/simulator implicitly defines our likelihood

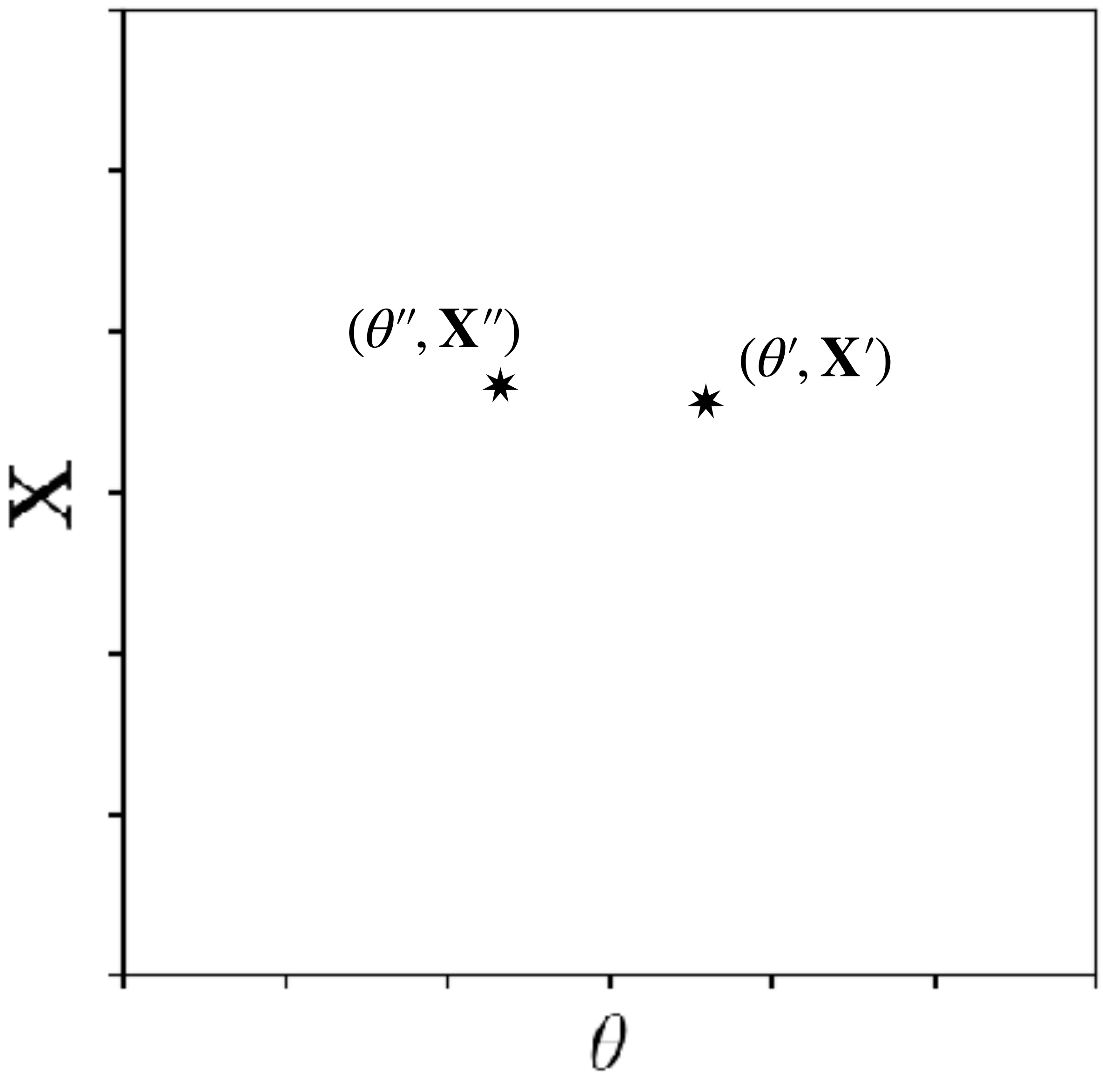
$$\mathbf{X}' \sim F(\theta') \quad \equiv \quad \mathbf{X}' \sim p(\mathbf{X} \,|\, \theta')$$

1. sample the prior

$$\theta' \sim p(\theta)$$

2. run simulator

$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

the forward model/simulator implicitly defines our likelihood

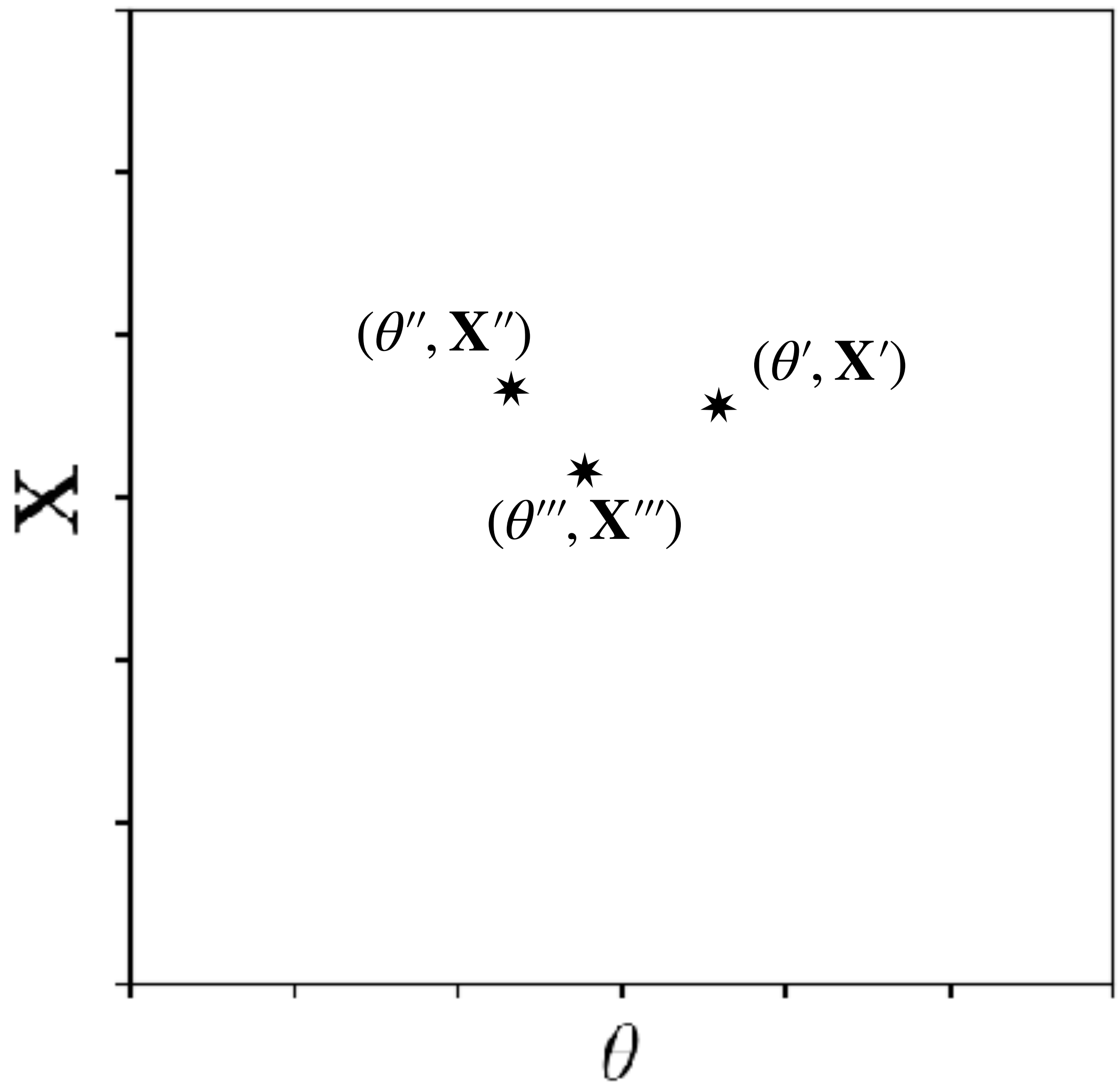
$$\mathbf{X}' \sim F(\theta') \quad \equiv \quad \mathbf{X}' \sim p(\mathbf{X} \mid \theta')$$

1. sample the prior

$$\theta' \sim p(\theta)$$

2. run simulator

$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

the forward model/simulator implicitly defines our likelihood

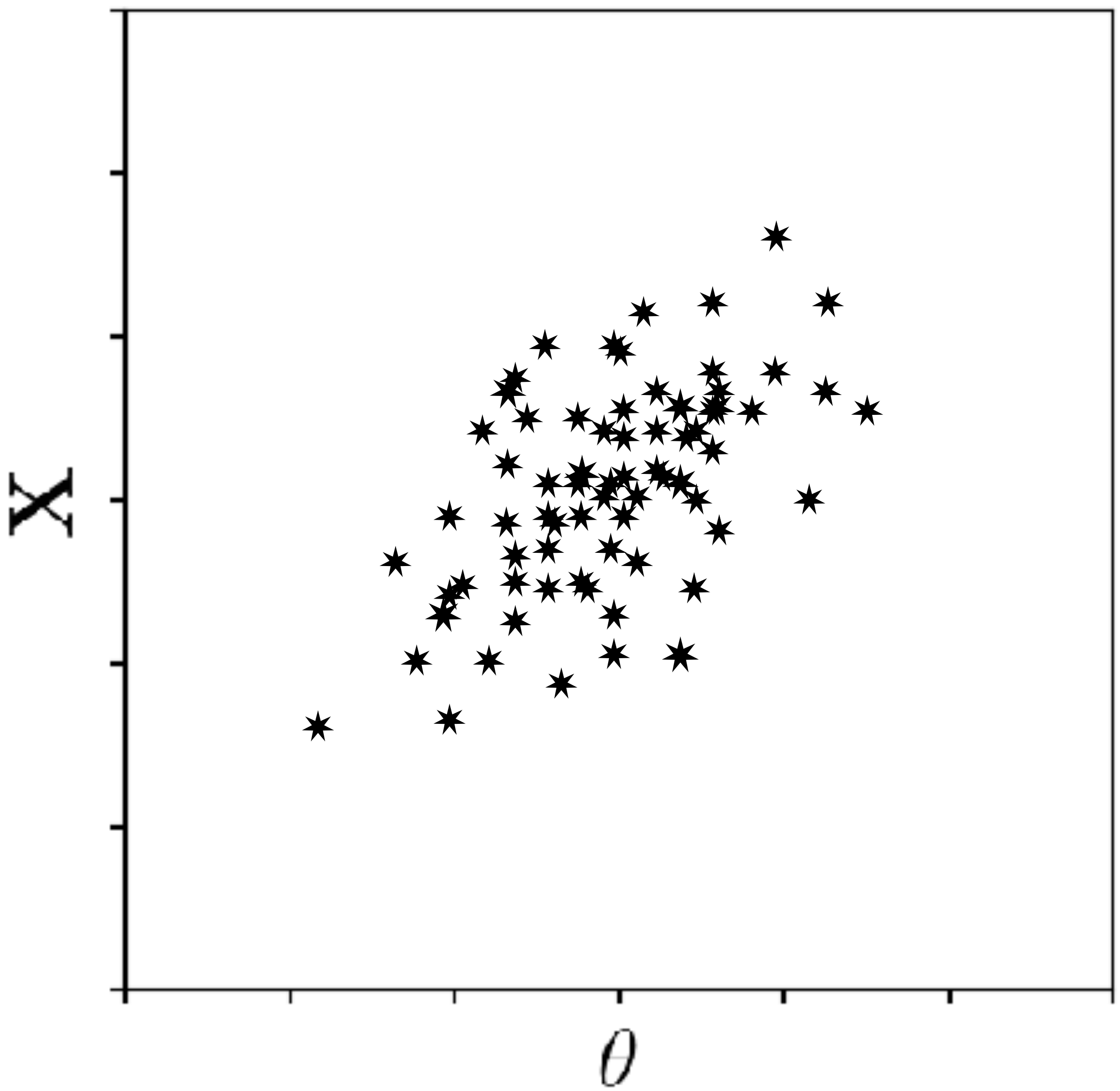
$$\mathbf{X}' \sim F(\theta') \quad \equiv \quad \mathbf{X}' \sim p(\mathbf{X} \mid \theta')$$

1. sample the prior

$$\theta' \sim p(\theta)$$

2. run simulator

$$\mathbf{X}' \sim F(\theta')$$



what is **simulation-based inference**?

the forward model/simulator implicitly defines our likelihood

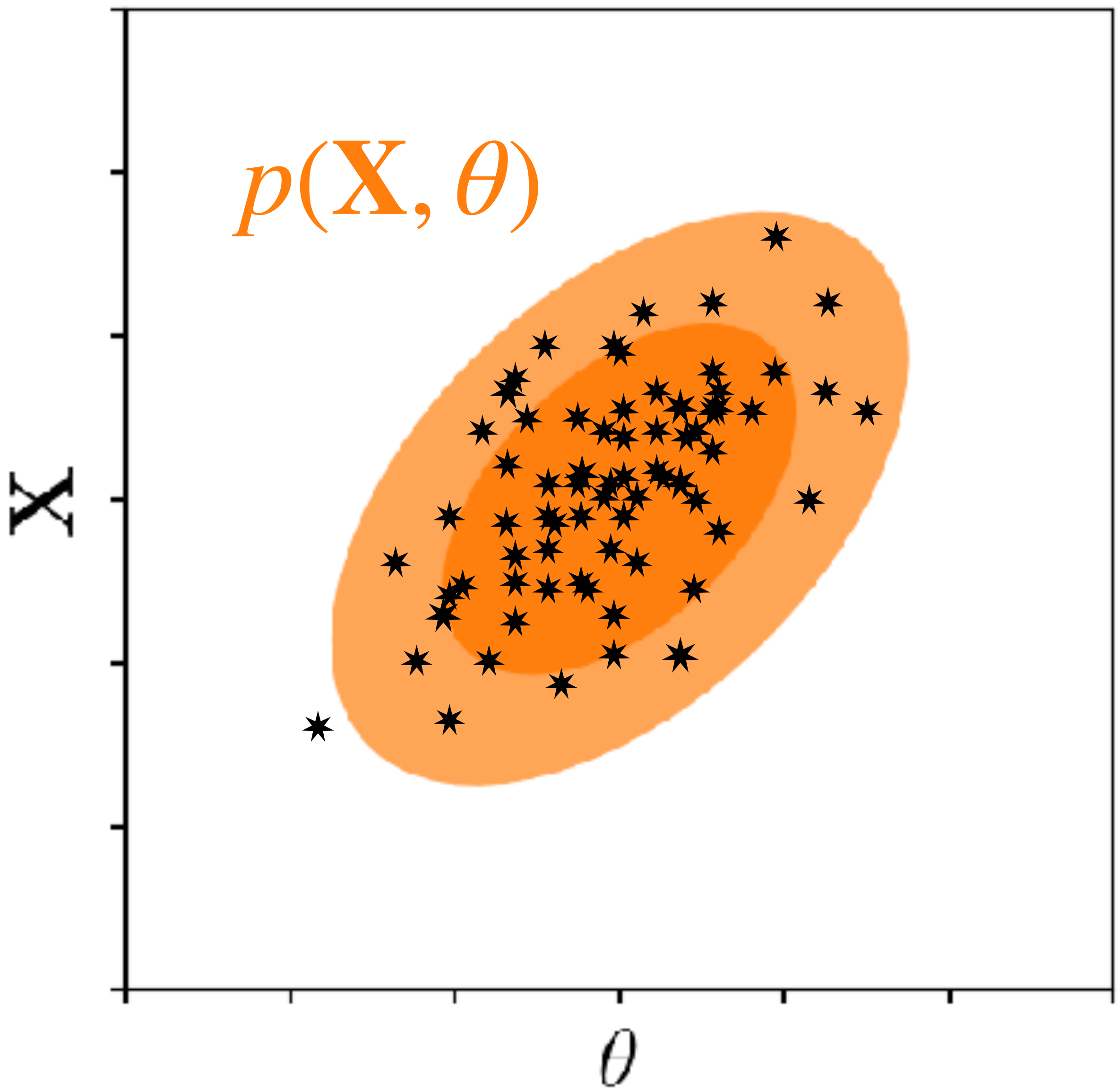
$$\mathbf{X}' \sim F(\theta') \quad \equiv \quad \mathbf{X}' \sim p(\mathbf{X} \,|\, \theta')$$

1. sample the prior

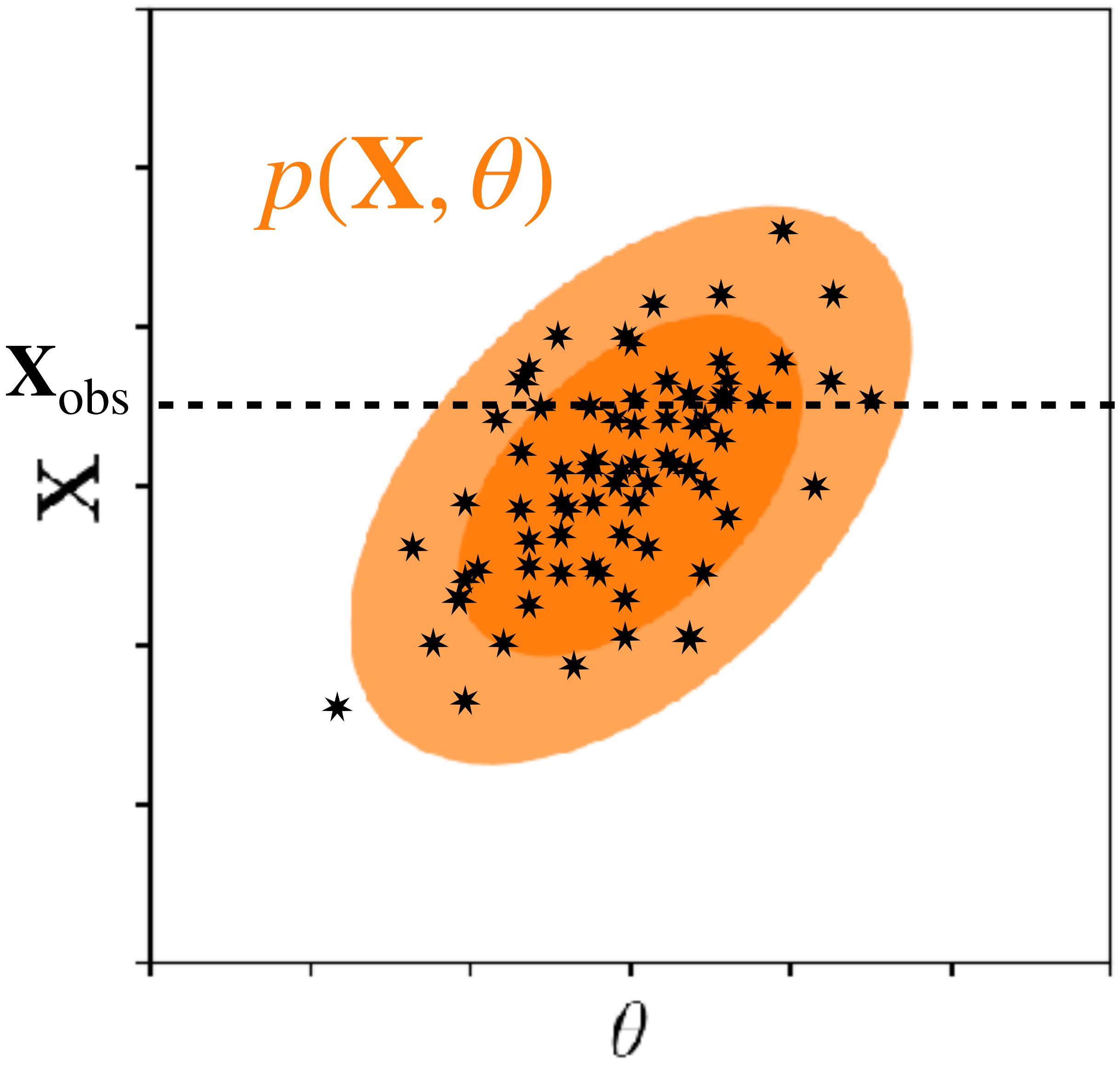
$$\theta' \sim p(\theta)$$

2. run simulator

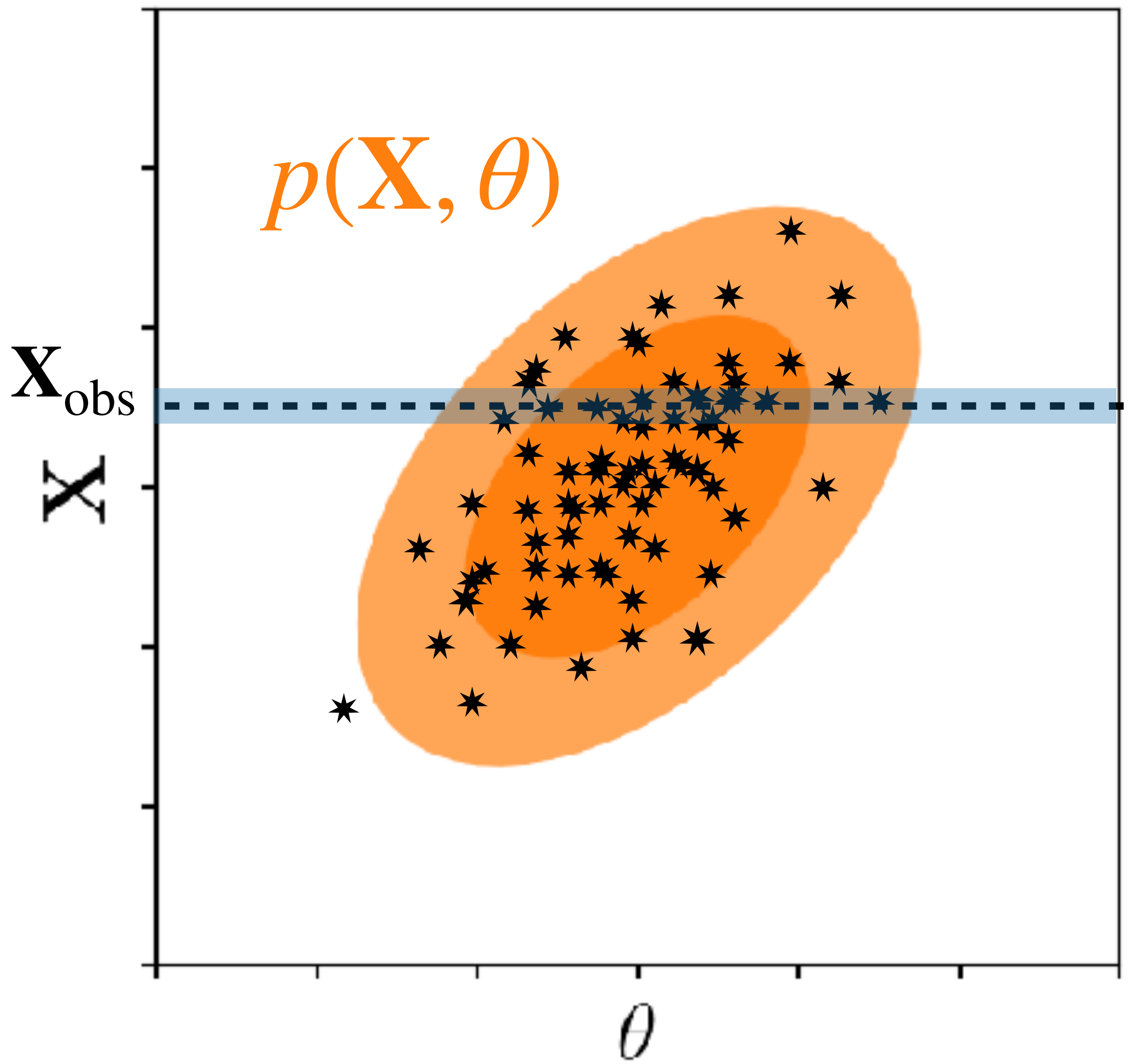
$$\mathbf{X}' \sim F(\theta')$$



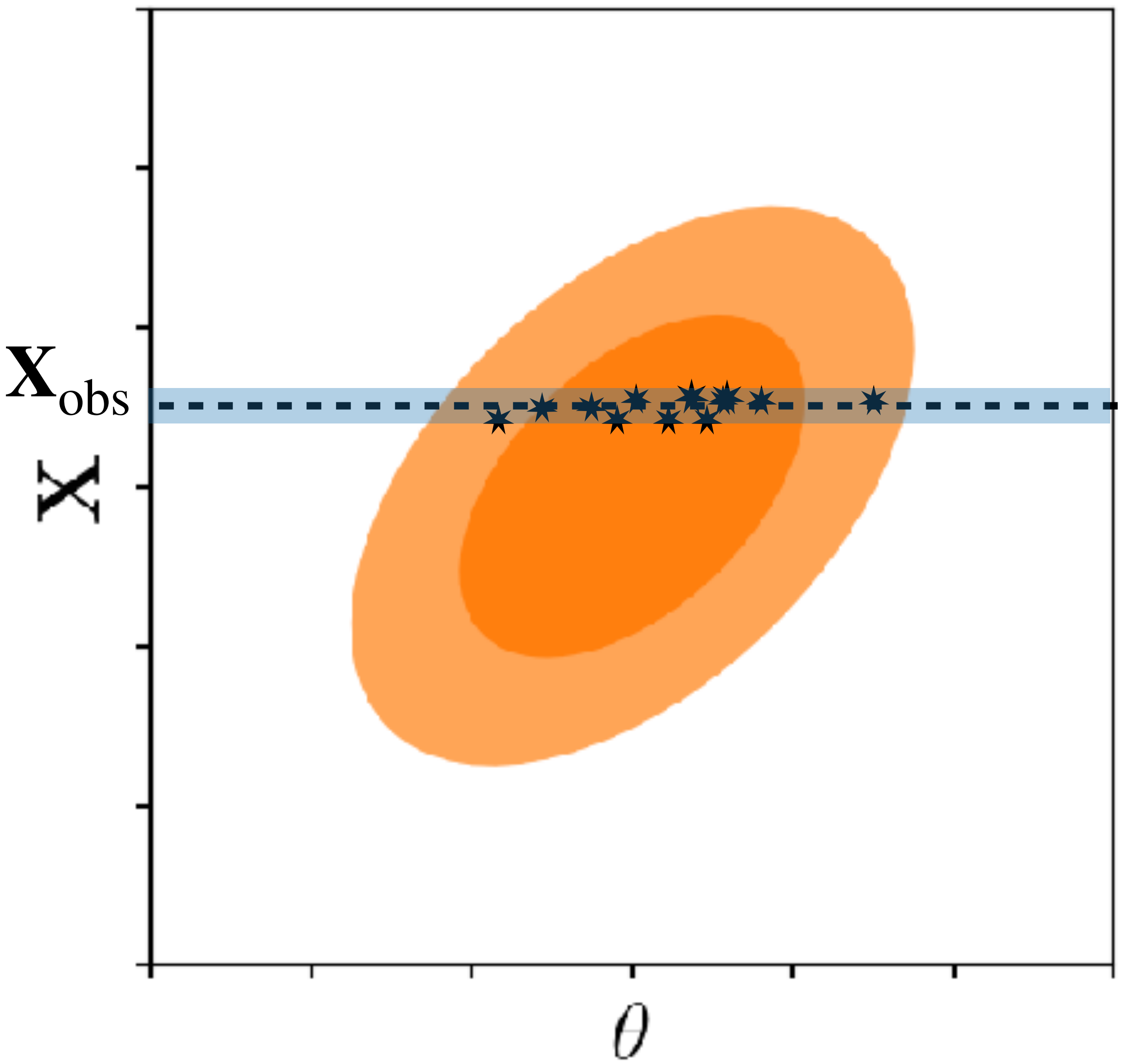
what is **simulation-based inference**?



what is **simulation-based inference**?

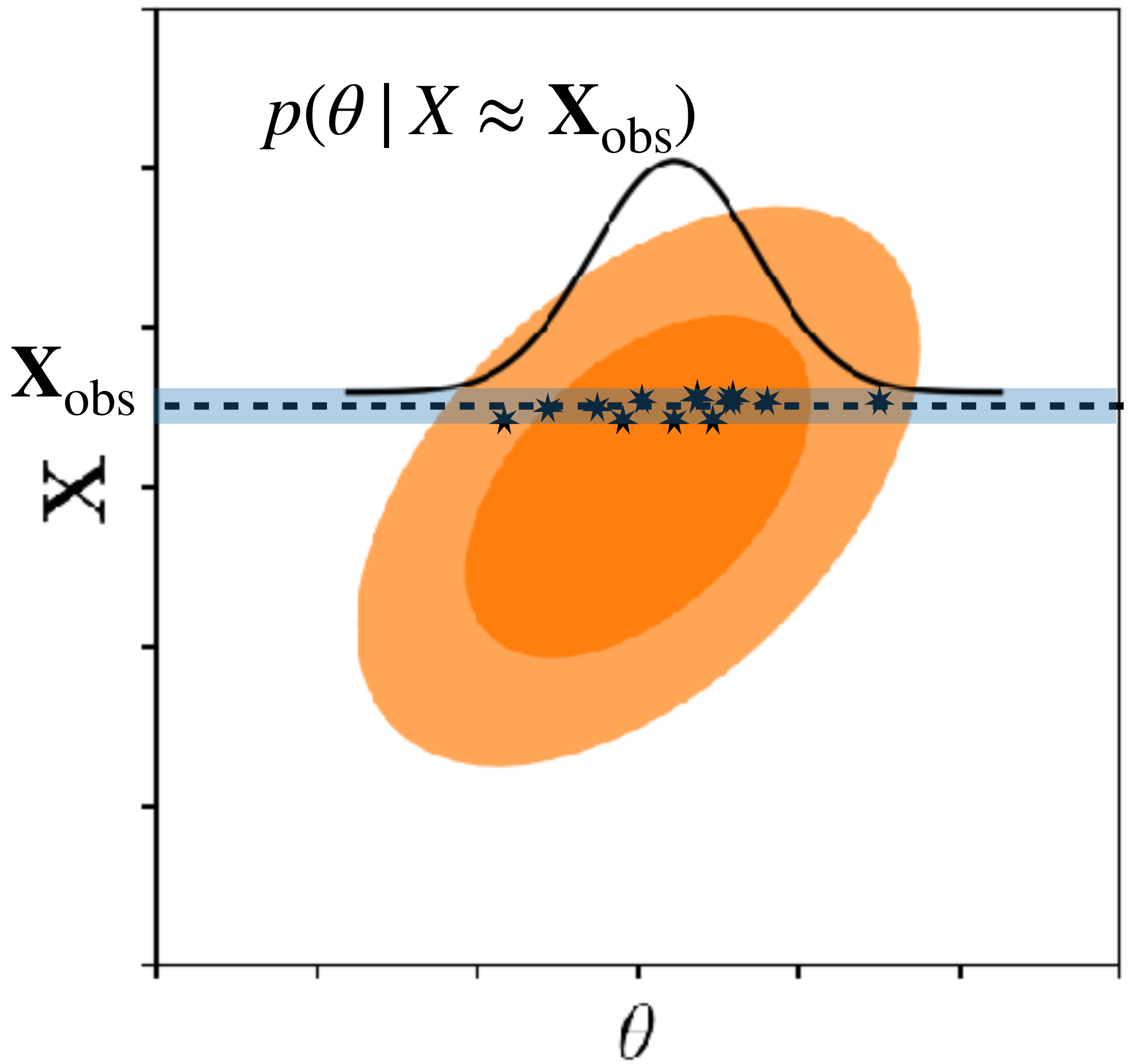


what is **simulation-based inference**?



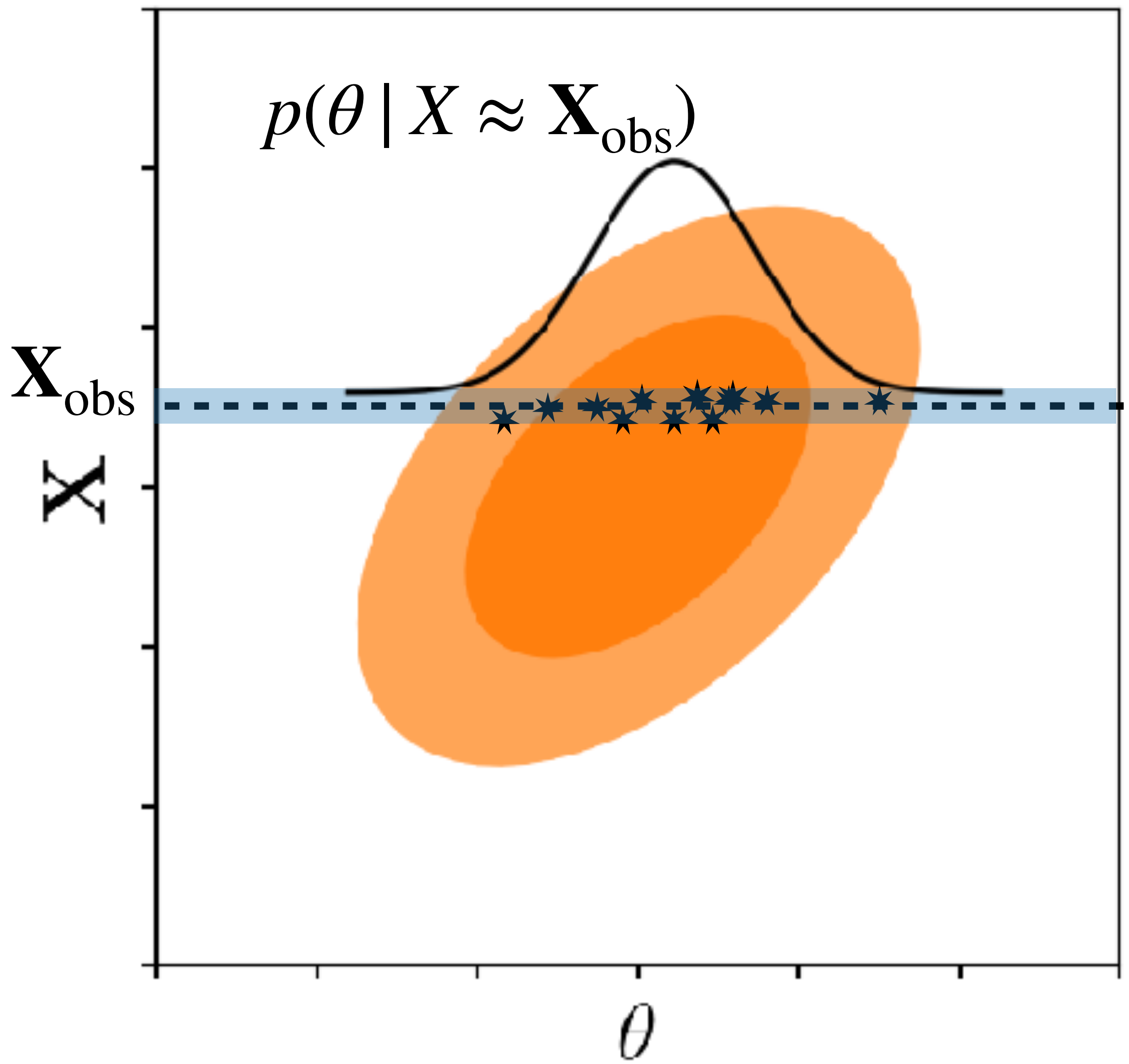
what is **simulation-based inference**?

$$p(\theta \,|\, \mathbf{X}_{\mathrm{obs}}) \approx p(\theta \,|\, X \approx \mathbf{X}_{\mathrm{obs}})$$

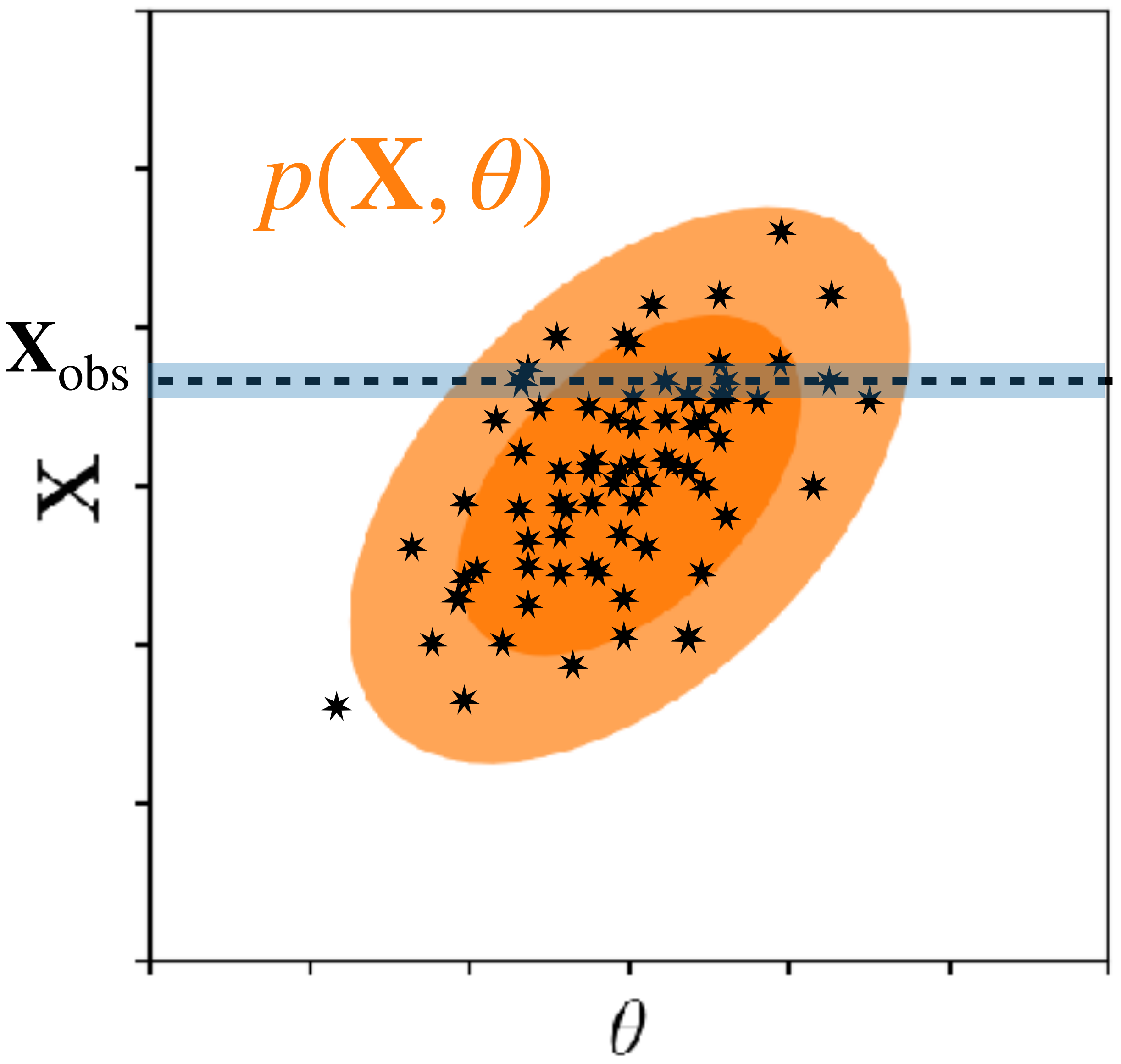


what is **simulation-based inference**?

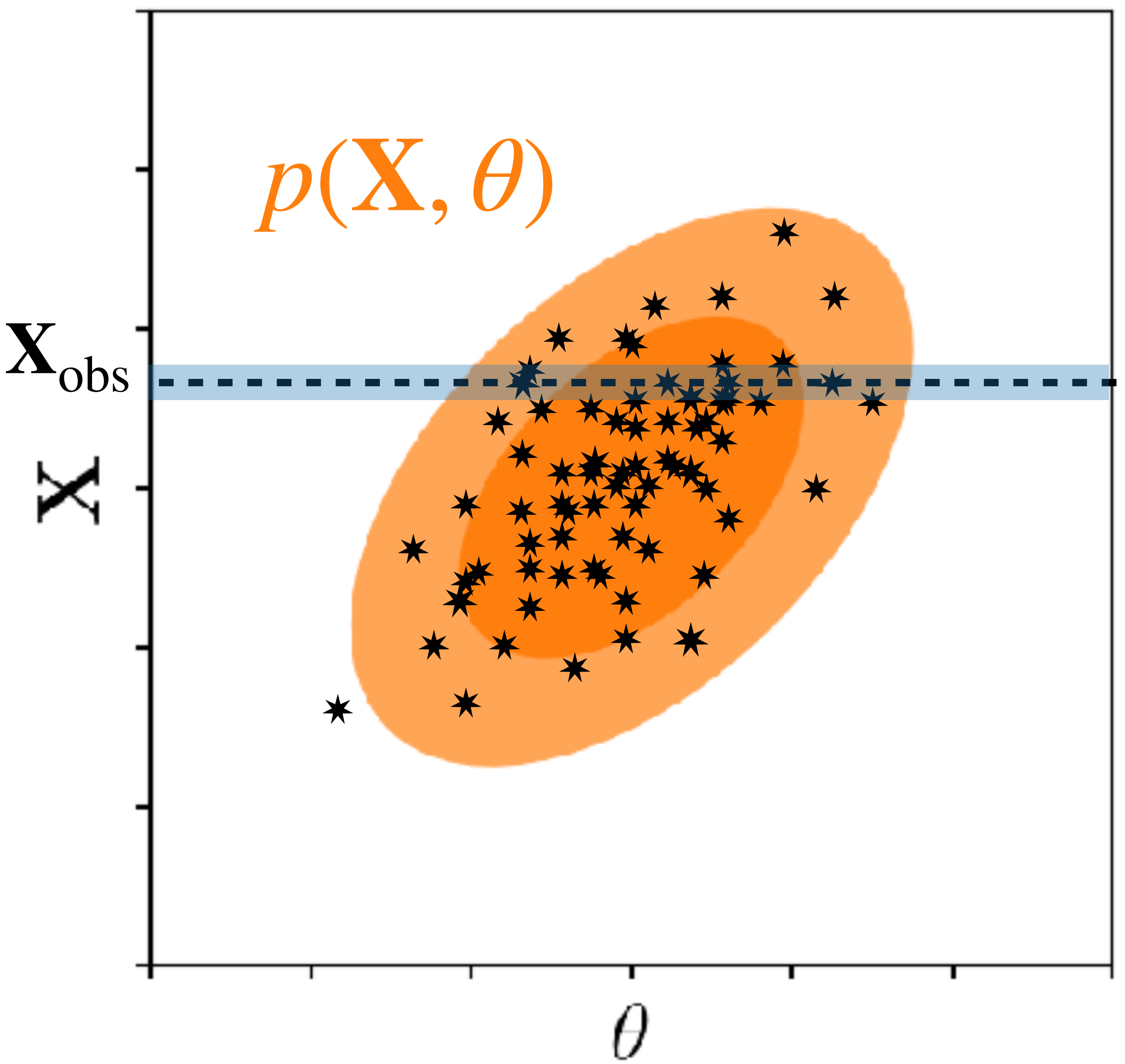
$$p(\theta \mid \mathbf{X}_{\text{obs}}) \approx p(\theta \mid X \approx \mathbf{X}_{\text{obs}})$$



simulation-based inference *in practice*

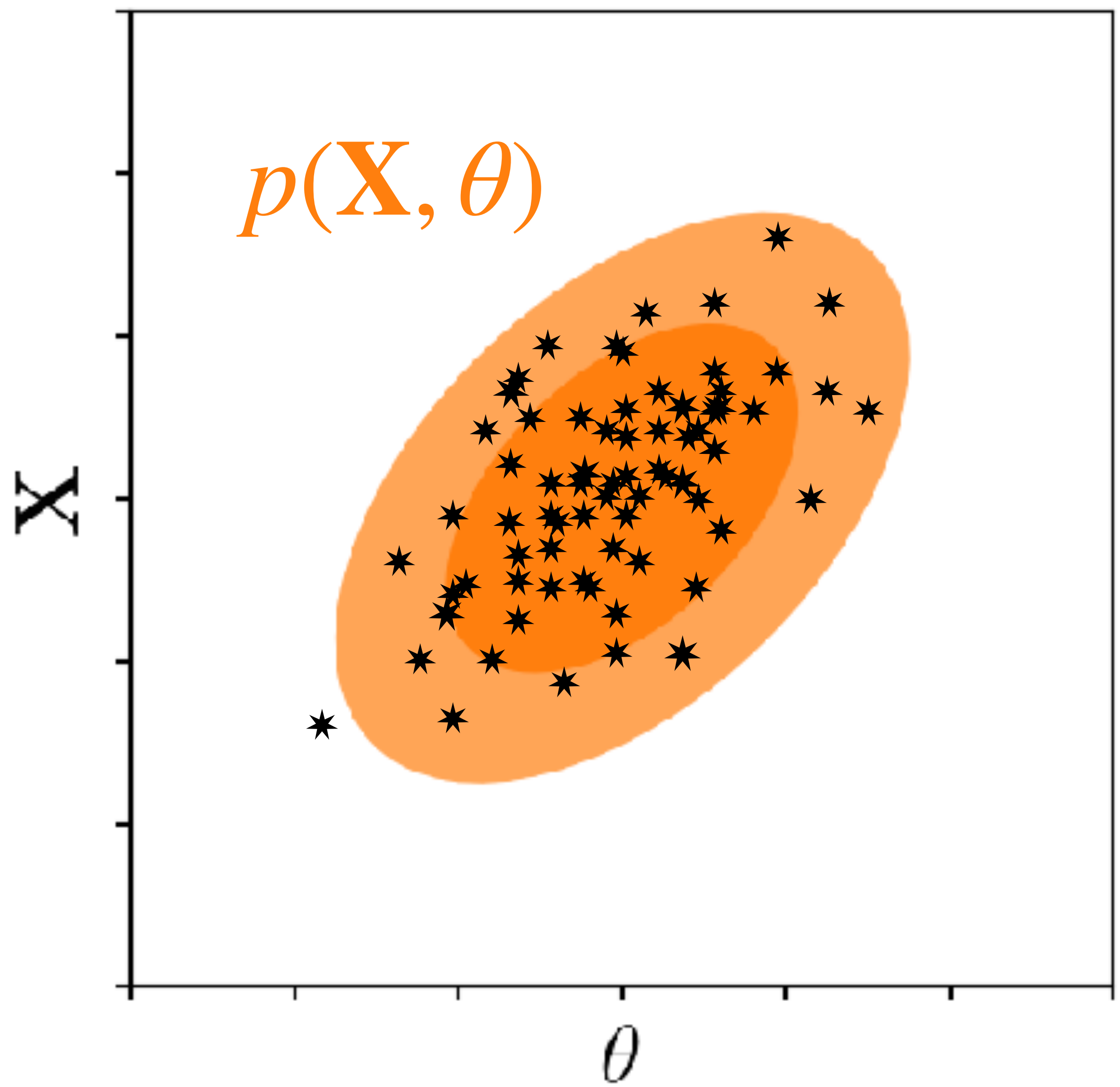


simulation-based inference *in practice*



approximate bayesian computation is often *infeasible*

simulation-based inference *in practice* — density estimation



can we estimate $p(\theta \mid \mathbf{X})$ from $\mathbf{X}' \sim F(\theta)$?

$\sim p(\mathbf{X} \mid \theta)$

estimate $p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

estimate $p(\theta \mid \mathbf{X}) \approx q_{\phi}(\theta \mid \mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

some model q with free parameters ϕ

**Gaussian Mixture Models,*
Independent Component Analysis, neural density estimators…

$$\text{estimate } p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X}) \qquad \text{from } \{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta) \text{ ?}$$

we can determine ϕ *by*

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

Kullback-Leibler divergence (a.k.a. relative entropy)

$$\text{estimate } p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X}) \qquad \text{from } \{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta) \text{ ?}$$

we can determine ϕ by

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

> *"it quantifies how much more surprising it would be to observe data from P if you were assuming it was coming from Q"*

> *"the amount of additional information required to encode samples from P using the distribution Q instead of P"*

Kullback-Leibler divergence (a.k.a. relative entropy)

estimate $p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

we can determine ϕ by

$$\min_{\phi} \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

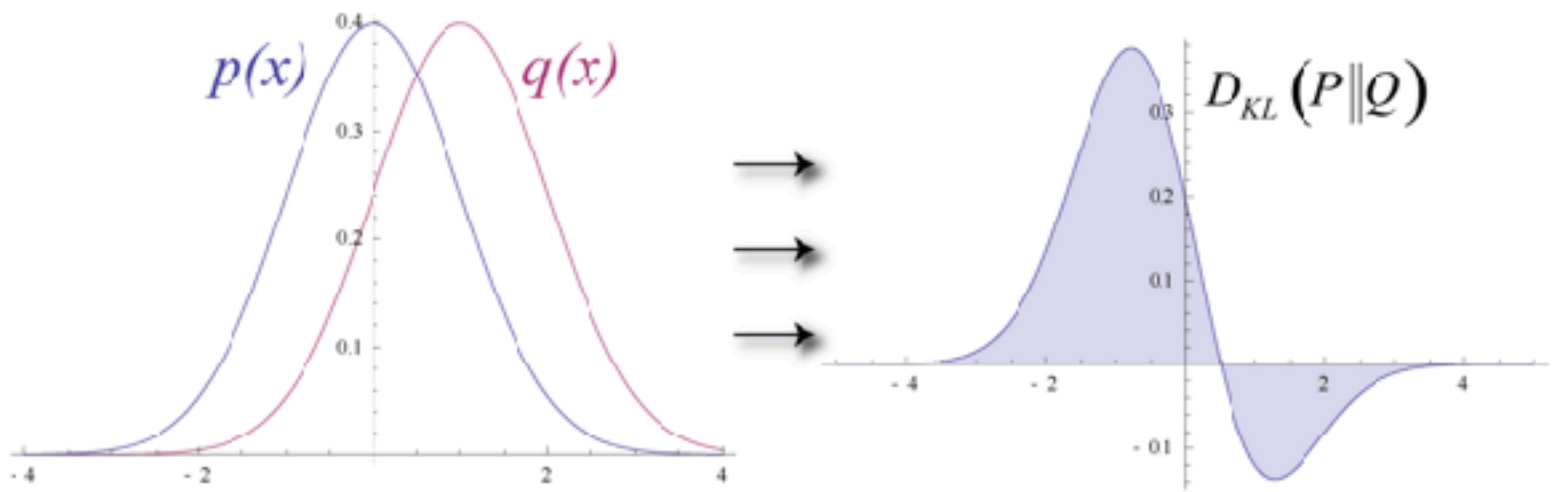


image: wikipedia

$$\text{estimate } p(\theta \,|\, \mathbf{X}) \approx q_\phi(\theta \,|\, \mathbf{X}) \qquad \text{from } \{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta) \text{ ?}$$

we can determine ϕ *by*

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \,|\, \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \,|\, \mathbf{X})\, p(\mathbf{X})\,)$$

$$= \min_\phi \; \int p(\theta \,|\, \mathbf{X})\, p(\mathbf{X}) \;\log \frac{p(\theta \,|\, \mathbf{X})\, p(\mathbf{X})}{q_\phi(\theta \,|\, \mathbf{X})\, p(\mathbf{X})}$$

Kullback-Leibler divergence (a.k.a. relative entropy)

$$\text{estimate } p(\theta\,|\,\mathbf{X}) \approx q_\phi(\theta\,|\,\mathbf{X}) \qquad \text{from } \{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)\ ?$$

we can determine ϕ by

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta\,|\,\mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta\,|\,\mathbf{X})\, p(\mathbf{X})\,)$$

$$= \min_\phi \int \underbrace{p(\theta\,|\,\mathbf{X})\, p(\mathbf{X})}_{p(\mathbf{X}, \theta)} \log \frac{p(\theta\,|\,\mathbf{X})\, p(\mathbf{X})}{q_\phi(\theta\,|\,\mathbf{X})\, p(\mathbf{X})}$$

estimate $p(\theta \,|\, \mathbf{X}) \approx q_\phi(\theta \,|\, \mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

we can determine ϕ *by*

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \,|\, \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \,|\, \mathbf{X})\, p(\mathbf{X})\,)$$

$$= \min_\phi \int p(\theta \,|\, \mathbf{X})\, p(\mathbf{X}) \; \log \frac{p(\theta \,|\, \mathbf{X})\, p(\mathbf{X})}{q_\phi(\theta \,|\, \mathbf{X})\, p(\mathbf{X})}$$

$$\approx \min_\phi \sum_{(\mathbf{X}', \theta') \sim p(\mathbf{X}, \theta)} \log \frac{p(\theta' \,|\, \mathbf{X}')}{q_\phi(\theta' \,|\, \mathbf{X}')}$$

$$\text{estimate } p(\theta \,|\, \mathbf{X}) \approx q_\phi(\theta \,|\, \mathbf{X}) \qquad \text{from } \{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta) \text{ ?}$$

we can determine ϕ by

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \,|\, \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \,|\, \mathbf{X})\, p(\mathbf{X})\,)$$

$$= \min_\phi \int p(\theta \,|\, \mathbf{X})\, p(\mathbf{X}) \; \log \frac{p(\theta \,|\, \mathbf{X})\, p(\mathbf{X})}{q_\phi(\theta \,|\, \mathbf{X})\, p(\mathbf{X})}$$

$$\approx \min_\phi \sum_{(\mathbf{X}', \theta') \sim p(\mathbf{X}, \theta)} \log \frac{p(\theta' \,|\, \mathbf{X}')}{q_\phi(\theta' \,|\, \mathbf{X}')}$$

estimate $p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X}) \qquad$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

we can determine ϕ *by*

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

$$= \min_\phi \int p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \; \log \frac{p(\theta \mid \mathbf{X})\, p(\mathbf{X})}{q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})}$$

$$\approx \min_\phi \sum_{(\mathbf{X}',\theta') \sim p(\mathbf{X},\theta)} \log \frac{p(\theta' \mid \mathbf{X}')}{q_\phi(\theta' \mid \mathbf{X}')}$$

$$= \min_\phi \sum_{(\mathbf{X}',\theta') \sim p(\mathbf{X},\theta)} - \log \; q_\phi(\theta' \mid \mathbf{X}')$$

$$\text{estimate } p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X}) \qquad \text{from } \{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta) \text{ ?}$$

we can determine ϕ *by*

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

$$= \min_\phi \int p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \; \log \frac{p(\theta \mid \mathbf{X})\, p(\mathbf{X})}{q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})}$$

$$\approx \min_\phi \sum_{(\mathbf{X}', \theta') \sim p(\mathbf{X}, \theta)} \log \frac{p(\theta' \mid \mathbf{X}')}{q_\phi(\theta' \mid \mathbf{X}')}$$

$$= \max_\phi \sum_{(\mathbf{X}', \theta') \sim p(\mathbf{X}, \theta)} \log \, q_\phi(\theta' \mid \mathbf{X}')$$

estimate $p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

we can determine ϕ by

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

$$\approx \max_\phi \sum_{(\mathbf{X}', \theta') \sim p(\mathbf{X}, \theta)} \log \, q_\phi(\theta' \mid \mathbf{X}')$$

$q_\phi(\theta \mid \mathbf{X})$ is guaranteed to converge to $p(\theta \mid \mathbf{X})$ if

q_ϕ is flexibly expressive

$N \to \infty$ samples from $p(\mathbf{X}, \theta)$

successful optimization

Sebastian Wagner-Carena

estimate $p(\theta\,|\,\mathbf{X}) \approx q_\phi(\theta\,|\,\mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

we can determine ϕ *by*

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta\,|\,\mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta\,|\,\mathbf{X})\, p(\mathbf{X})\,)$$

$$\approx \max_\phi \sum_{(\mathbf{X}',\theta')\sim p(\mathbf{X},\theta)} \log\, q_\phi(\theta'\,|\,\mathbf{X}')$$

$q_\phi(\theta\,|\,\mathbf{X})$ is guaranteed to converge to $p(\theta\,|\,\mathbf{X})$ if

q_ϕ *is flexibly expressive*

$N \to \infty$ *samples from* $p(\mathbf{X}, \theta)$

successful optimization

proofs in arxiv:1605.06376

estimate $p(\theta \mid \mathbf{X}) \approx q_\phi(\theta \mid \mathbf{X})$ from $\{(\theta', \mathbf{X}')\} \sim p(\mathbf{X}, \theta)$?

we can determine ϕ by

$$\min_\phi \; D_{\mathrm{KL}}(\, p(\theta \mid \mathbf{X})\, p(\mathbf{X}) \;\|\; q_\phi(\theta \mid \mathbf{X})\, p(\mathbf{X})\,)$$

$$\approx \max_\phi \sum_{(\mathbf{X}', \theta') \sim p(\mathbf{X}, \theta)} \log\, q_\phi(\theta' \mid \mathbf{X}')$$

$q_\phi(\theta \mid \mathbf{X})$ is guaranteed to converge to $p(\theta \mid \mathbf{X})$ if

*q_ϕ is flexibly expressive**

$N \to \infty$ samples from $p(\mathbf{X}, \theta)$

successful optimization

**normalizing flows, diffusion,*
(your favorite neural density estimation)

if you’ve ever trained a neural network, *you’ve done neural posterior estimation*

if you've ever trained a neural network, *you've done neural posterior estimation*

$$\mathbf{X} \xrightarrow{F_\phi} \theta$$

train a neural network F_ϕ *to take input* $\mathbf{X}$ *and predict* θ

if you've ever trained a neural network, *you've done neural posterior estimation*

$$\mathbf{X} \xrightarrow{F_\phi} \theta$$

train a neural network F_ϕ *to take input* $\mathbf{X}$ *and predict* θ *by minimizing the mean squared error:*

$$\min_\phi \sum_{(\mathbf{X}', \theta')} (F_\phi(\mathbf{X}') - \theta')^2$$

if you've ever trained a neural network, *you've done neural posterior estimation*

$$\mathbf{X} \xrightarrow{F_\phi} \theta$$

train a neural network F_ϕ *to take input* $\mathbf{X}$ *and predict* θ *by minimizing the mean squared error:*

$$\min_\phi \sum_{(\mathbf{X}',\theta')} (F_\phi(\mathbf{X}') - \theta')^2$$

$$= \max_\phi \sum_{(\mathbf{X}',\theta')} -(F_\phi(\mathbf{X}') - \theta')^2$$

if you've ever trained a neural network, *you've done neural posterior estimation*

$$\mathbf{X} \xrightarrow{F_\phi} \theta$$

train a neural network F_ϕ *to take input* $\mathbf{X}$ *and predict* θ *by minimizing the mean squared error:*

$$\min_\phi \sum_{(\mathbf{X}',\theta')} (F_\phi(\mathbf{X}') - \theta')^2$$

$$= \max_\phi \sum_{(\mathbf{X}',\theta')} \log \exp - (F_\phi(\mathbf{X}') - \theta')^2$$

if you've ever trained a neural network, *you've done neural posterior estimation*

$$\mathbf{X} \xrightarrow{F_\phi} \theta$$

train a neural network F_ϕ *to take input* $\mathbf{X}$ *and predict* θ *by minimizing the mean squared error:*

$$\min_\phi \sum_{(\mathbf{X}',\theta')} (\, F_\phi(\mathbf{X}') - \theta' \,)^2$$

$$= \max_\phi \sum_{(\mathbf{X}',\theta')} \log \; \exp - \frac{(F_\phi(\mathbf{X}') - \theta')^2}{2\sigma^2} + \text{const}(\sigma)$$

if you've ever trained a neural network, *you've done neural posterior estimation*

$$\mathbf{X} \xrightarrow{F_\phi} \theta$$

train a neural network F_ϕ *to take input* $\mathbf{X}$ *and predict* θ *by minimizing the mean squared error:*

$$\min_\phi \sum_{(\mathbf{X}',\theta')} \left(F_\phi(\mathbf{X}') - \theta' \right)^2$$

$$= \max_\phi \sum_{(\mathbf{X}',\theta')} \log \boxed{\exp - \frac{(F_\phi(\mathbf{X}') - \theta')^2}{2\sigma^2}} + \text{const}(\sigma)$$

gaussian $q_\phi(\theta \mid \mathbf{X})$

see also moment networks (Jeffrey & Wandelt 2020)

if you've ever done Bayesian inference in cosmology, *you've likely done simulation-based inference*

$$\log p(\mathbf{X}\,|\,\theta) = \log \mathscr{L} = [\,(m(\theta) - \mathbf{X})^T\,\mathbf{C}^{-1}\,(m(\theta) - \mathbf{X})\,] + \log(2\pi)^{-k/2}\,|\,\mathbf{C}\,|^{-1/2}$$

m = *your favorite theory model for* $\mathbf{X}$

$\mathbf{C}$ = *covariance matrix from mocks*

if you've ever done Bayesian inference in cosmology, *you've likely done simulation-based inference*

$$\log p(\mathbf{X}\,|\,\theta) = \log \mathscr{L} = [\,(m(\theta) - \mathbf{X})^T\, \mathbf{C}^{-1}\,(m(\theta) - \mathbf{X})\,] + \log(2\pi)^{-k/2}\,|\,\mathbf{C}\,|^{-1/2}$$

m = *your favorite theory model for* $\mathbf{X}$

$\mathbf{C}$ = *covariance matrix from mocks*

$$\equiv$$

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

SBI with Gaussian described by $m(\theta)$ *and* $\mathbf{C}$

if you've ever done Bayesian inference in cosmology, *you've likely done simulation-based inference*

$$\log p(\mathbf{X} \mid \theta) = \log \mathscr{L} = [\,(m(\theta) - \mathbf{X})^T\, \mathbf{C}^{-1}\, (m(\theta) - \mathbf{X})\,] + \log(2\pi)^{-k/2}\,|\,\mathbf{C}\,|^{-1/2}$$

m = your favorite theory model for $\mathbf{X}$

$\mathbf{C}$ = covariance matrix from mocks

$$\equiv$$

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

SBI with Gaussian described by $m(\theta)$ and $\mathbf{C}$

simulation-based!

what is simulation-based inference?

opportunities for simulation-based inference?

challenges for simulation-based inference?

what is simulation-based inference?

why simulation-based inference* for galaxy clustering?

challenges for simulation-based inference?

**state-of-the-art SBI (e.g. neural posterior estimation)*

misconception: standard inference is more "rigorous" than SBI

$$\log p(\mathbf{X}\,|\,\theta) = \log \mathscr{L} = [\,(m(\theta) - \mathbf{X})^T\, \mathbf{C}^{-1}\,(m(\theta) - \mathbf{X})\,] + \log(2\pi)^{-k/2}\,|\,\mathbf{C}\,|^{-1/2}$$

m = *your favorite theory model for* $\mathbf{X}$

$\mathbf{C}$ = *covariance matrix from mocks*

$$\equiv$$

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

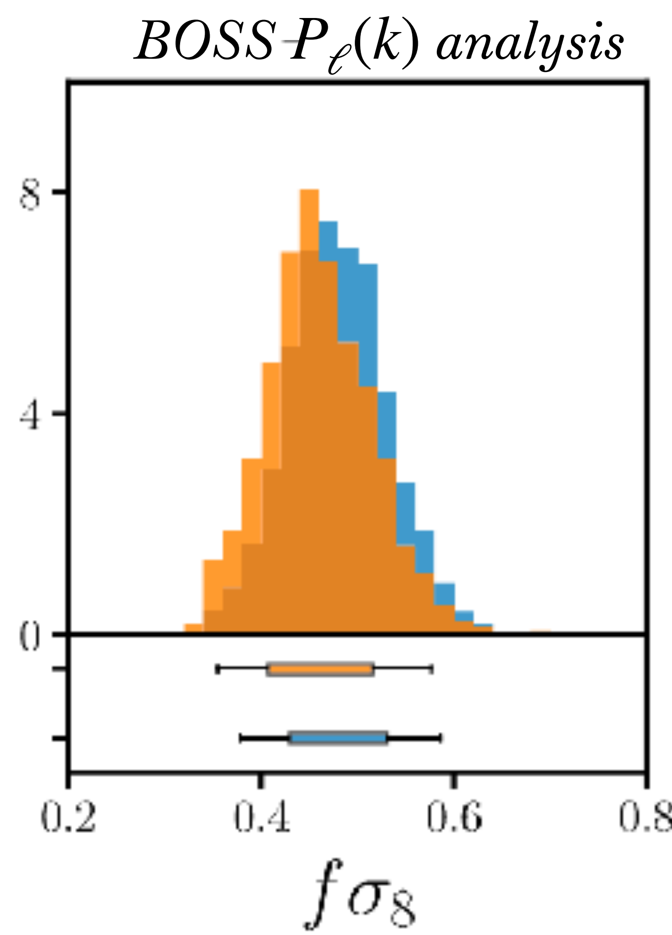
SBI with Gaussian described by $m(\theta)$ *and* $\mathbf{C}$

assumptions!

assumptions: Gaussian likelihood with cosmology independent covariance matrix from approximate mocks

Hahn et al.(2019);
see also Sellentin & Heavens (2017), Sellentin et al. (2017)

assumptions: *Gaussian likelihood* with cosmology independent covariance matrix from approximate mocks

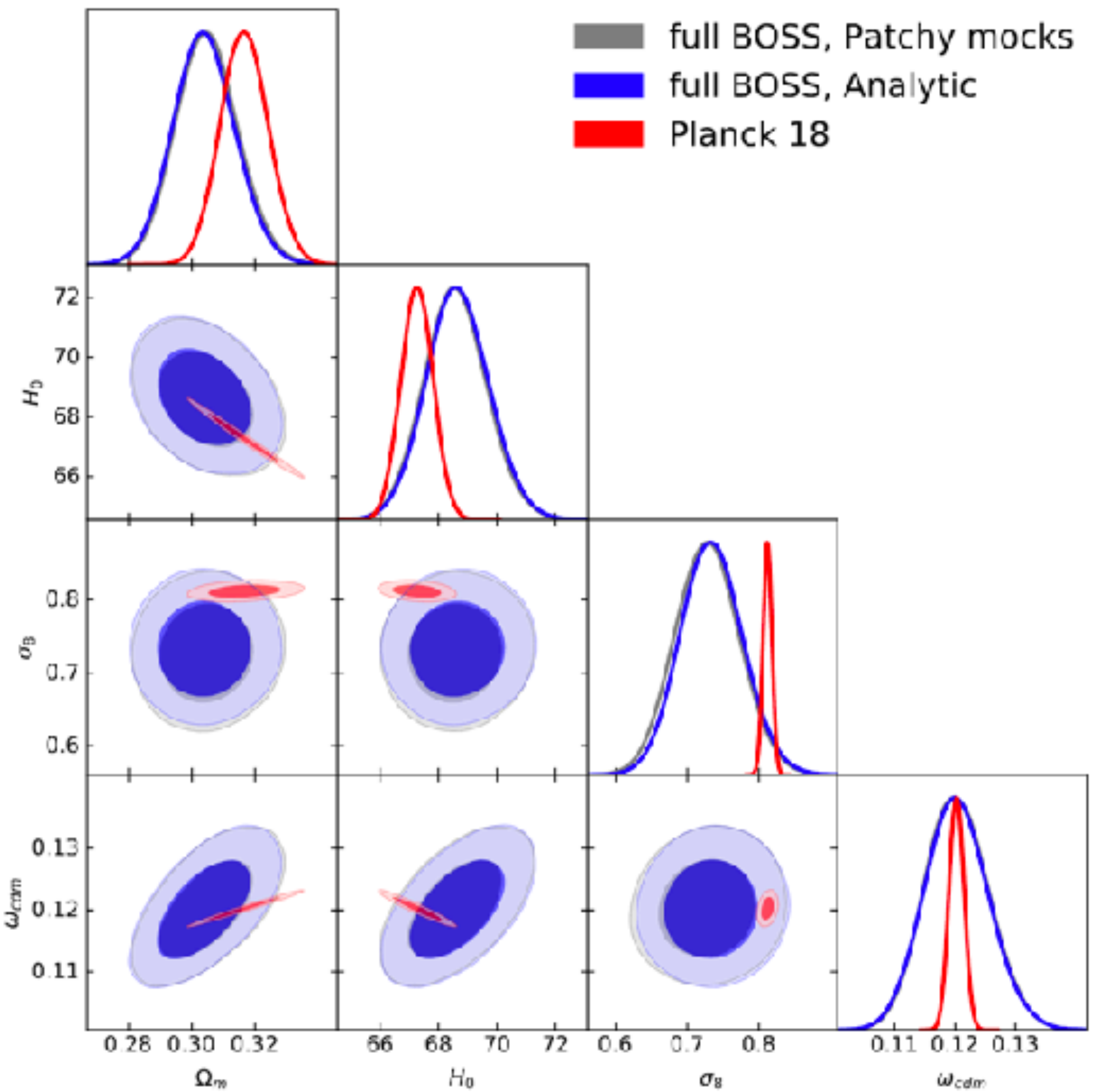


Gaussian $\mathscr{L}$ *(Beutler et al. 2017)*

non-Gaussian $\mathscr{L}$

Hahn et al.(2019);
*see also Sellentin & Heavens (2017), Sellentin et al. (2017), but **Benedict's** talk*

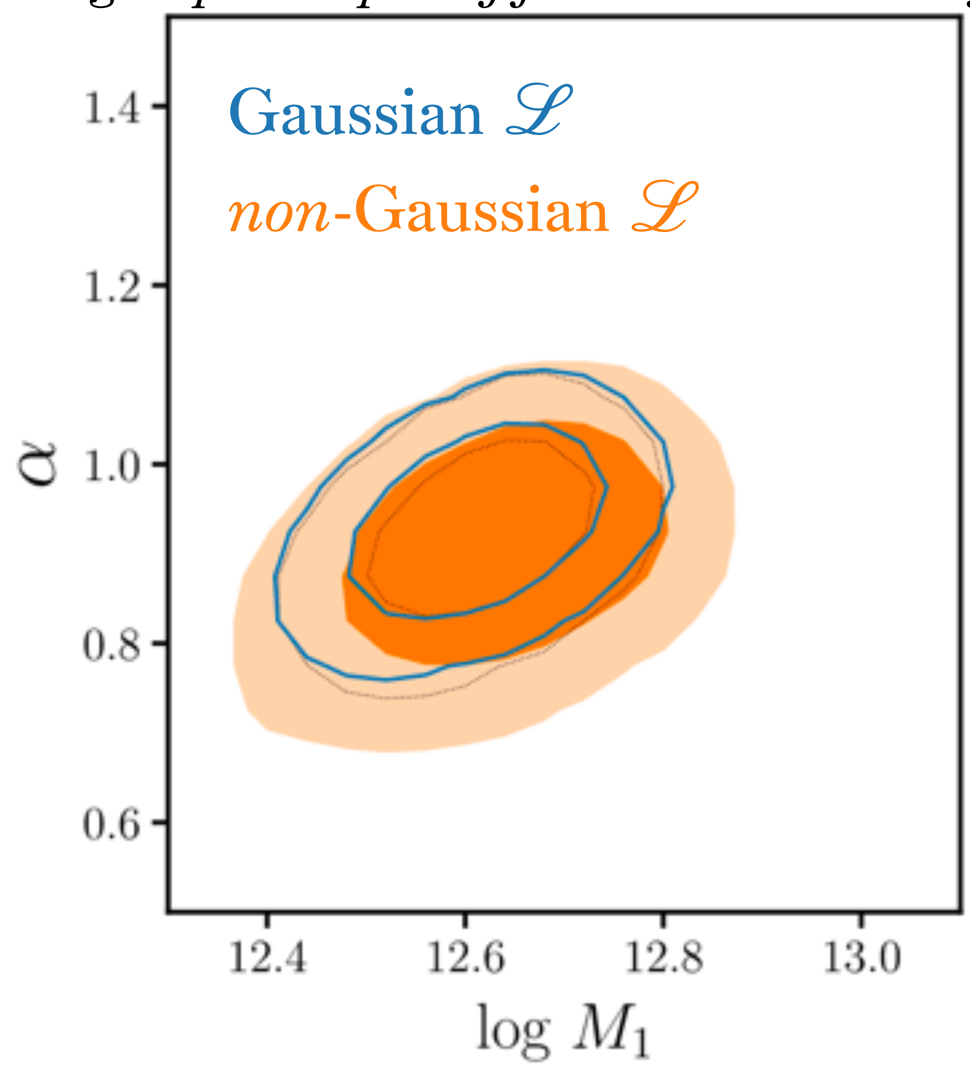
assumptions: Gaussian likelihood with *cosmology independent covariance matrix* from approximate mocks



Wadekar et al.(2020)
but see also ***Alessandra's talk***

assumptions: Gaussian likelihood with cosmology independent covariance matrix from approximate mocks

group multiplicity function HOD analysis

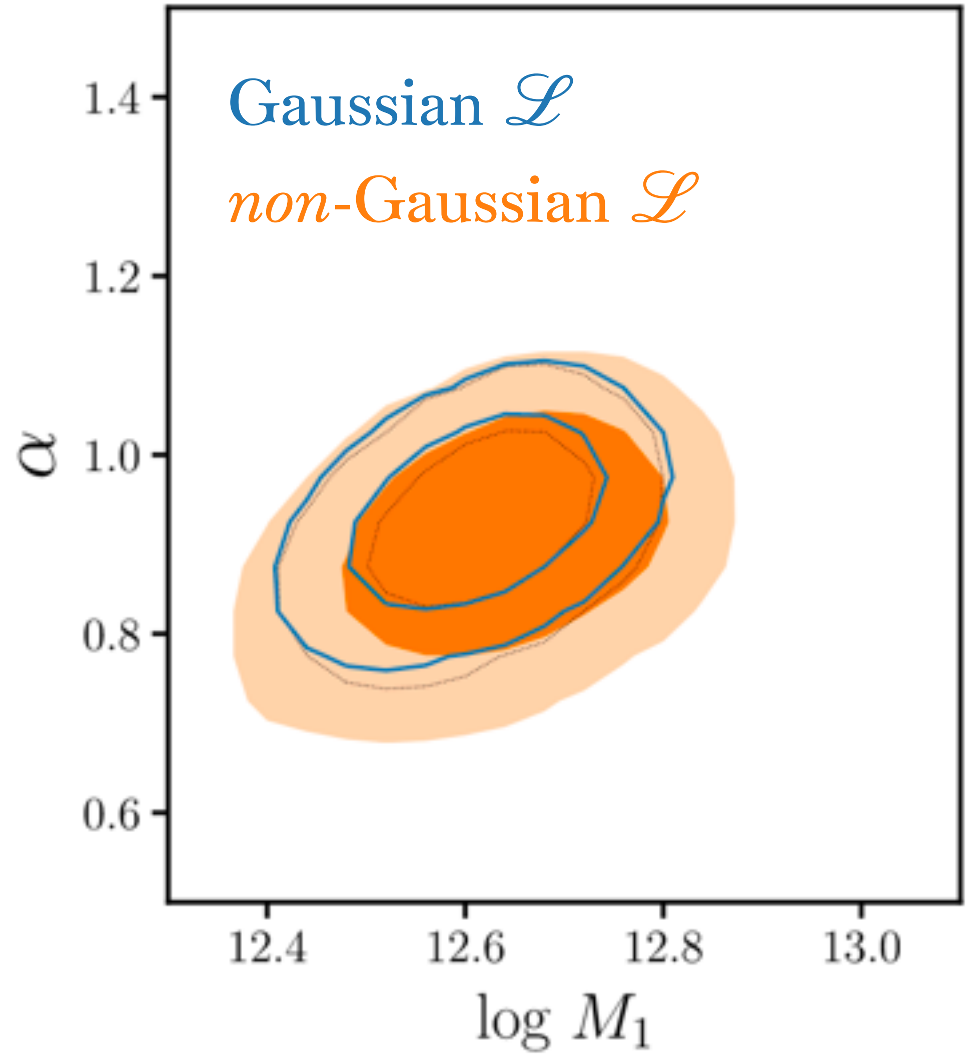


fair assumptions for **beyond 2-pt analyses**?

Hahn et al.(2019)

assumptions: Gaussian likelihood with cosmology independent covariance matrix from approximate mocks

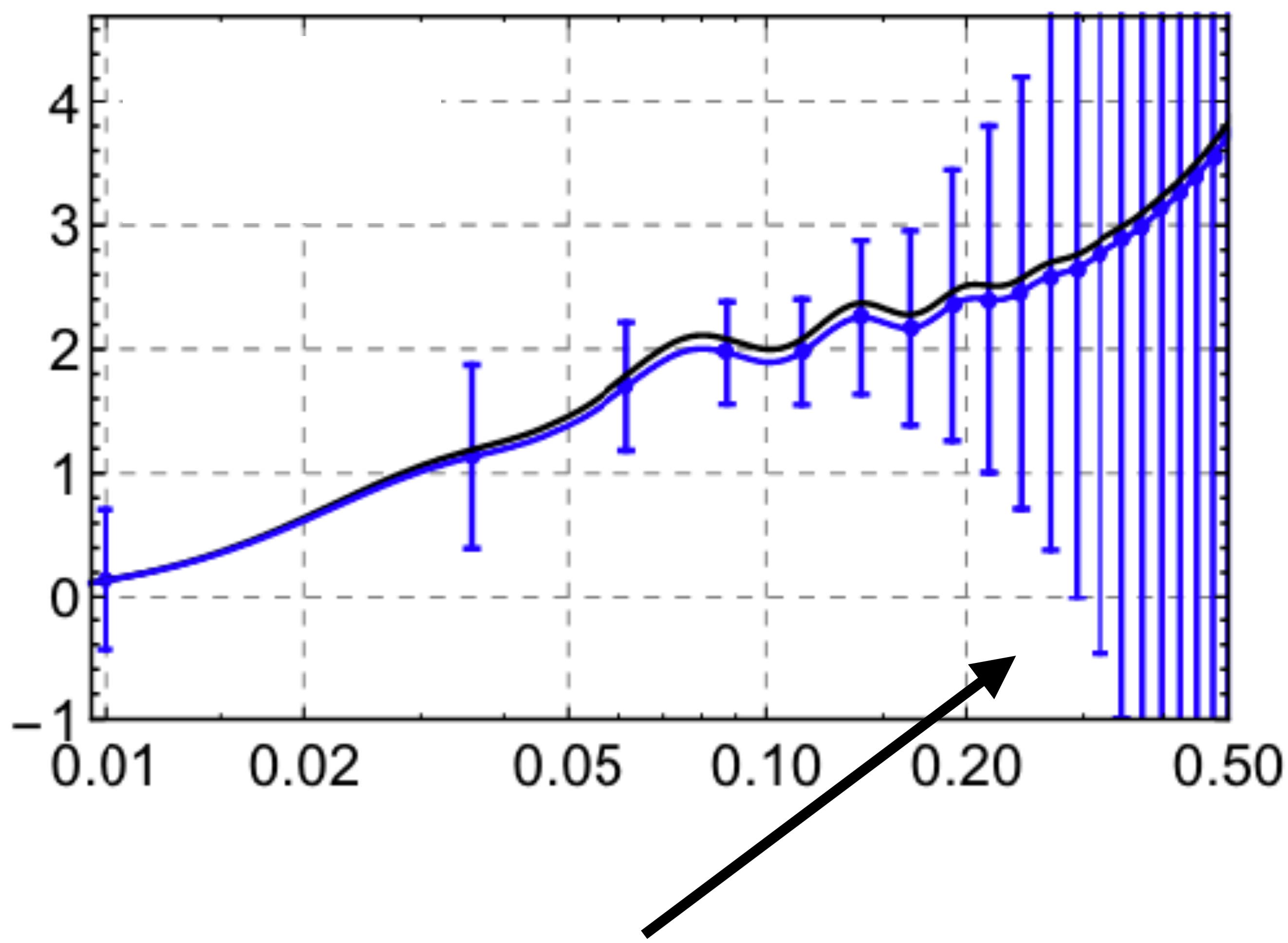
group multiplicity function HOD analysis



SBI can *relax* these assumptions by learning the likelihood from the forward model

advantages of SBI: leveraging simulations to access *additional cosmological information*

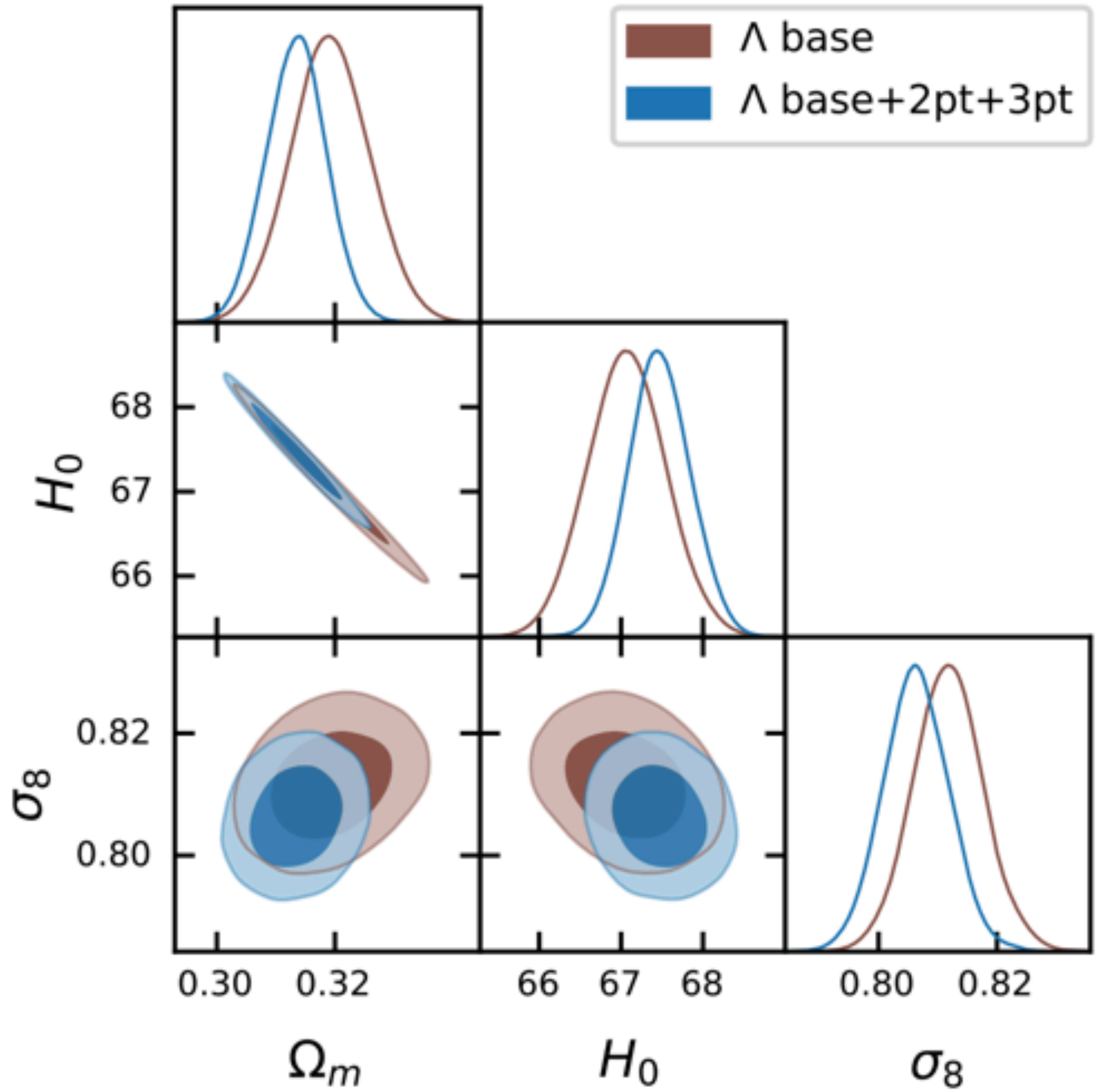
advantages of SBI: leveraging simulations to access *additional cosmological information*



theoretical uncertainties for perturbation theory

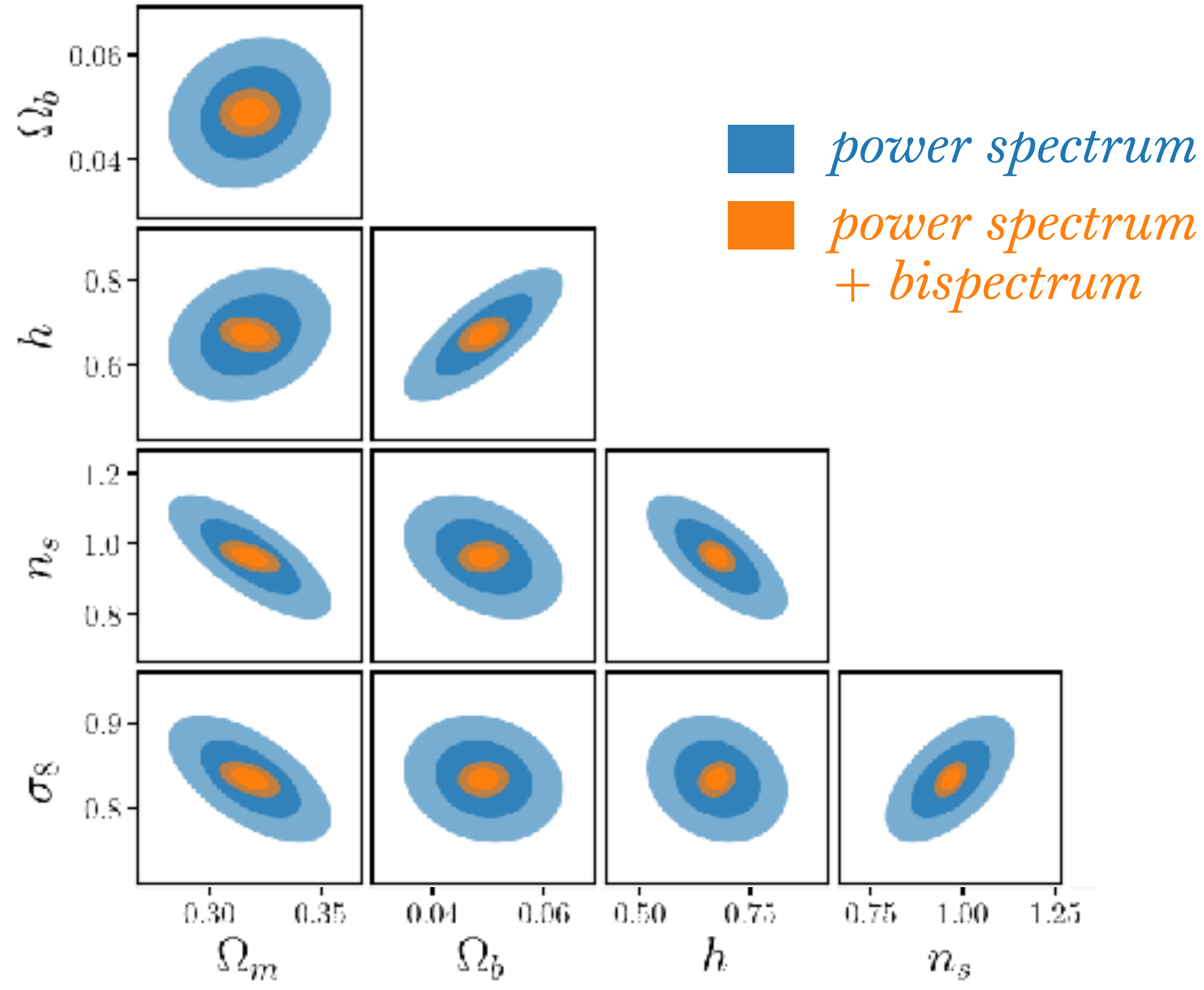
*Chudaykin & Ivanov (2019); see also **Pierre's talk***

advantages of SBI: leveraging simulations to access *additional cosmological information*



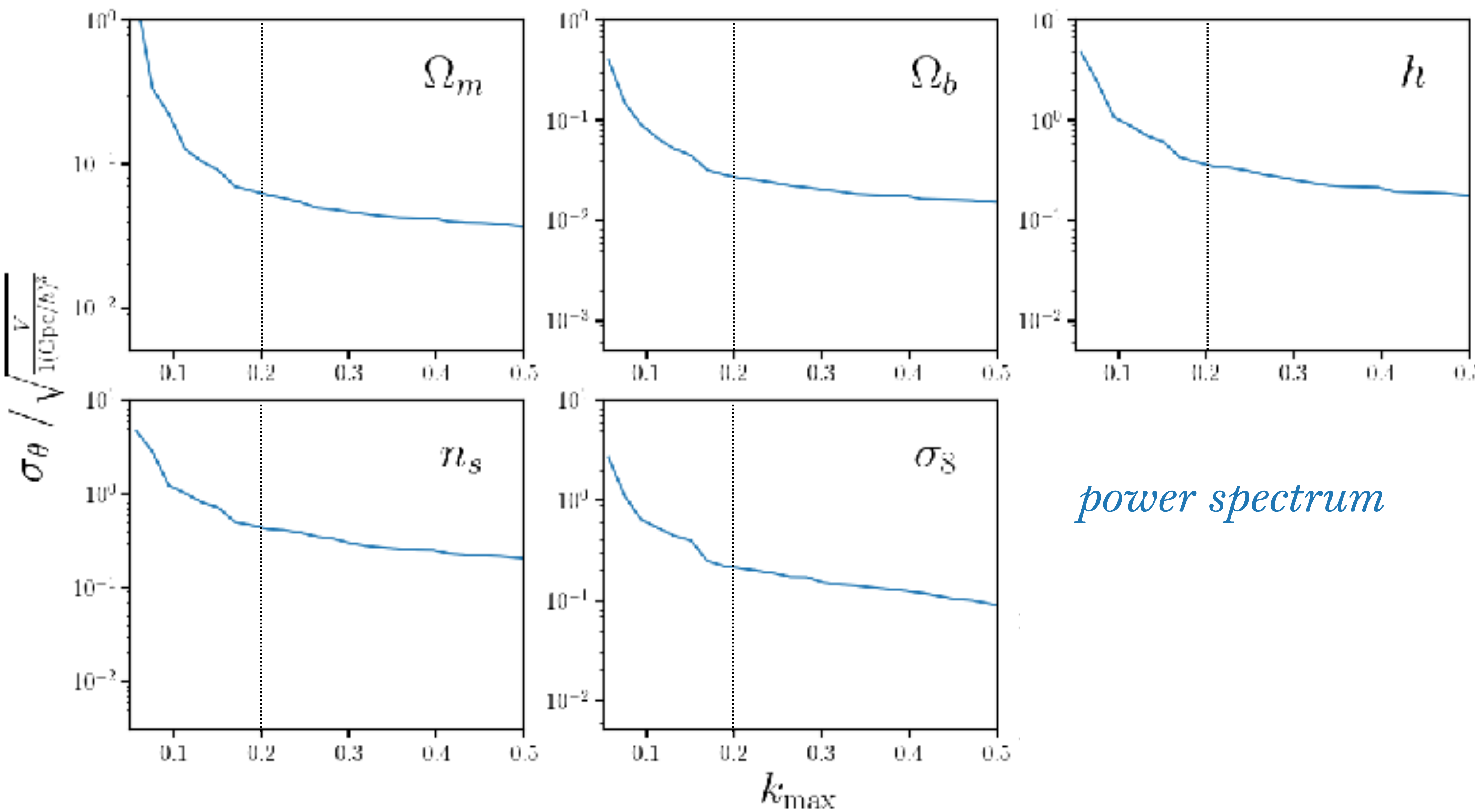
*Spaar & **Zhang** (2023)*
*see also Scoccimarro et al.(2001); Verde et al.(2022); **Gil-Marin** et al.(2017); Philcox & Ivanov (2022); Ivanov et al.(2023); D'Amico et al.(2024);*

advantages of SBI: leveraging simulations to access *additional cosmological information*



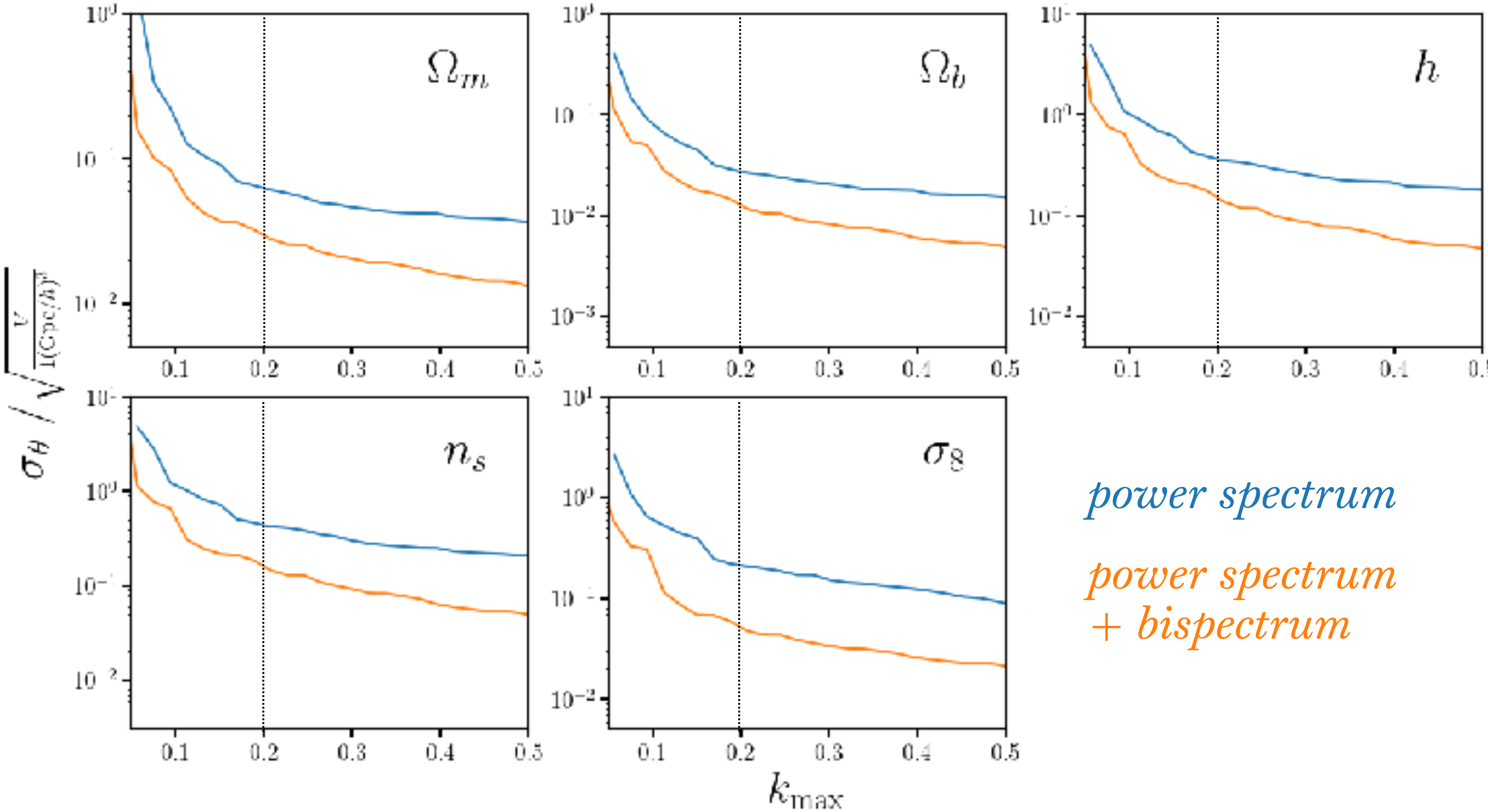
Quijote and Molino forecasts:
Hahn et al. (2020), Hahn & Villaescusa-Navarro (2021)
see also ***Lado's talk***

advantages of SBI: leveraging simulations to access *additional cosmological information*



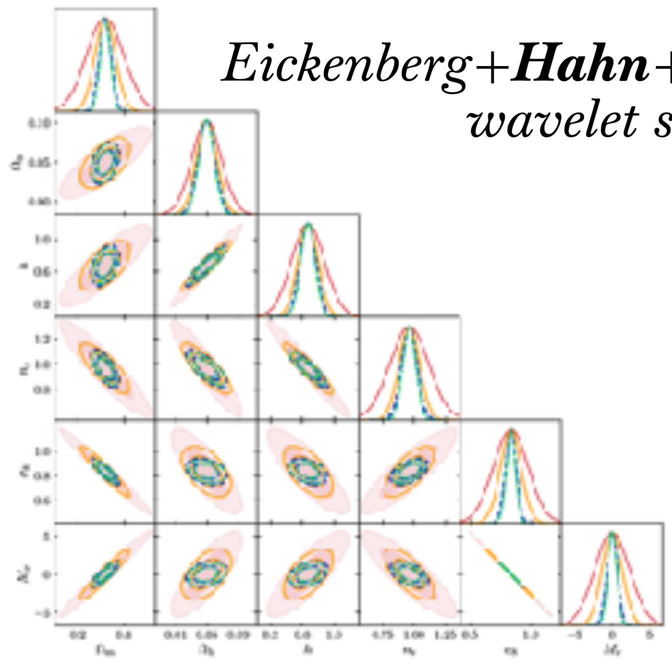
Quijote and Molino forecasts:
Hahn et al. (2020), Hahn & Villaescusa-Navarro (2021)
see also ***Lado's talk***

advantages of SBI: leveraging simulations to access *additional cosmological information on non-linear scales*

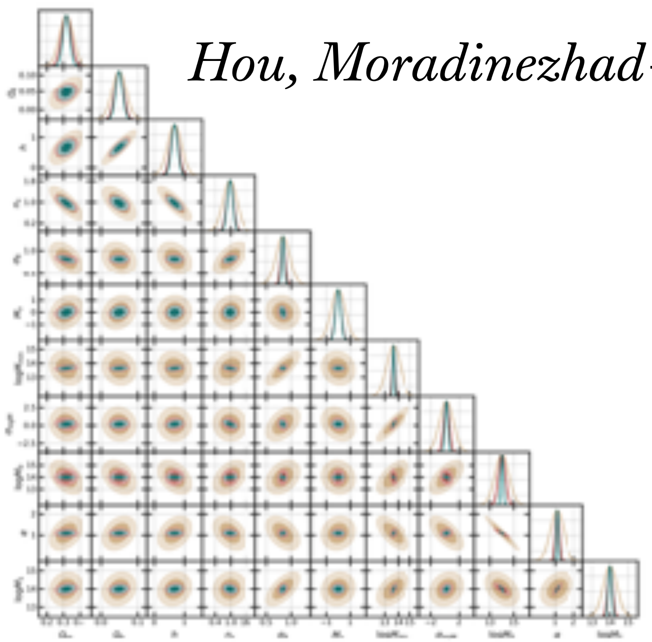


Quijote and Molino forecasts
Hahn, Villaescusa-Navarro, Castorina, Scoccimaro (2020)
Hahn & Villaescusa-Navarro (2021)

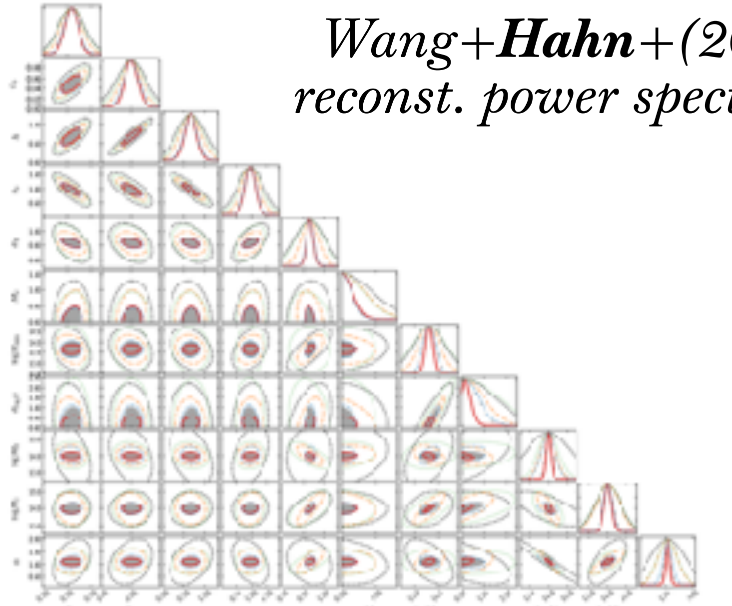
advantages of SBI: leveraging simulations to access *additional cosmological information from beyond standard statistics*



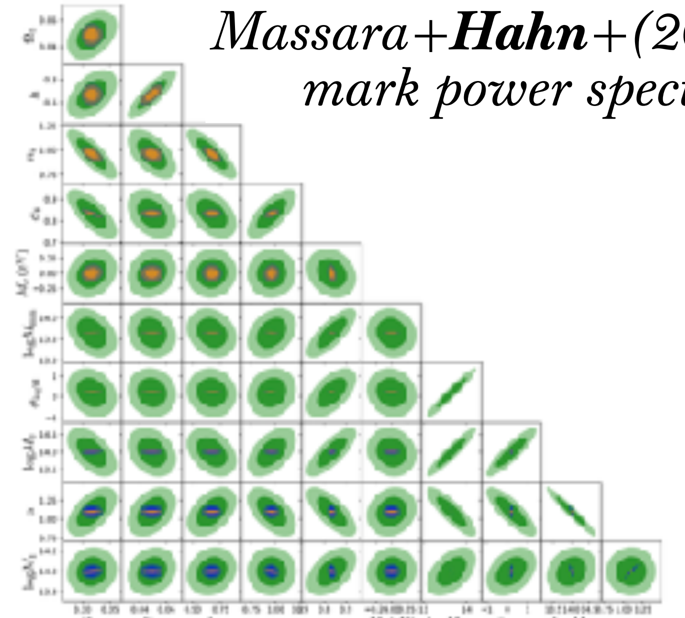
*Eickenberg+**Hahn**+(2022)*
wavelet statistics



*Hou, Moradinezhad+**Hahn**+(2022)*
skew spectra



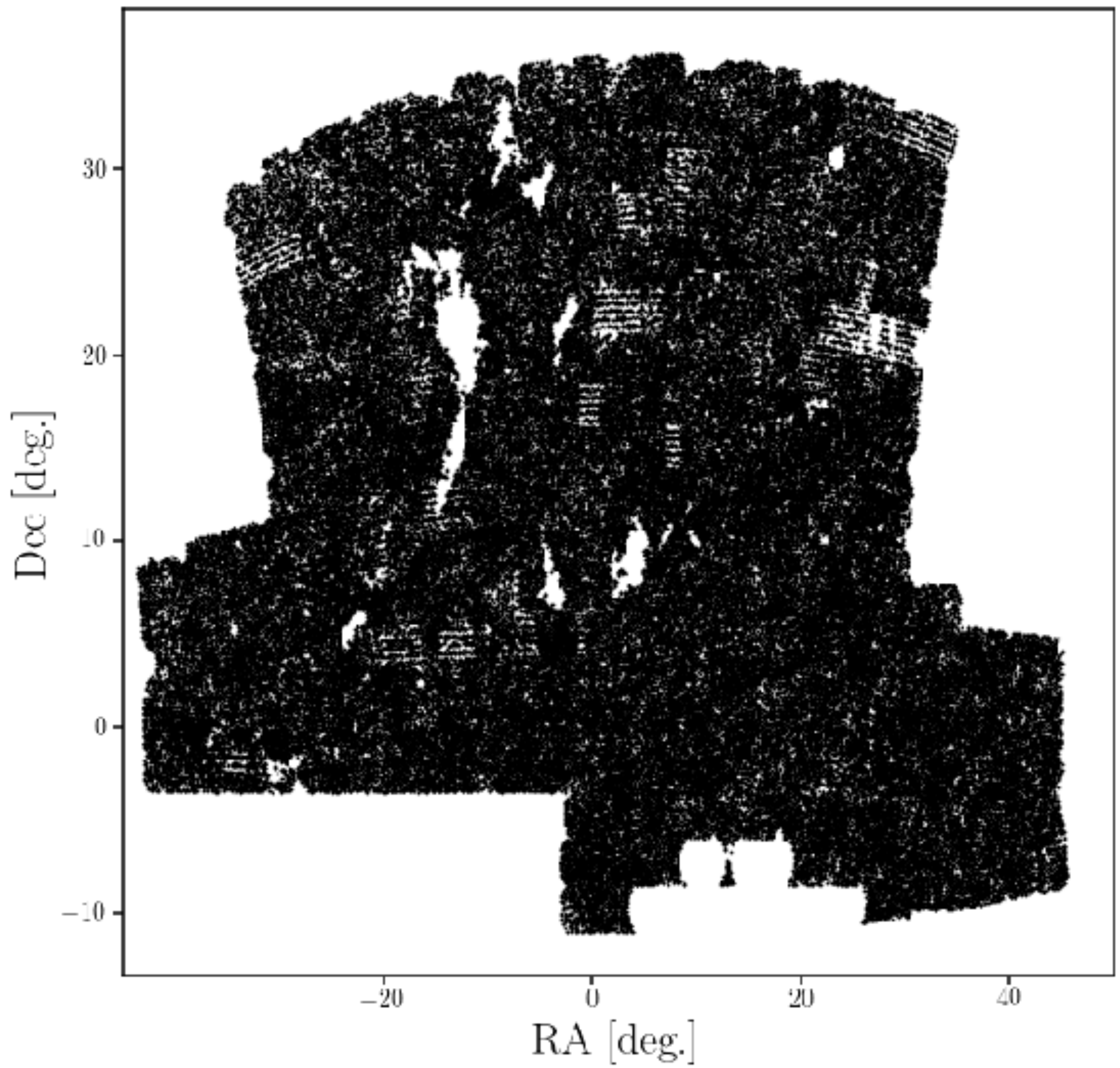
*Wang+**Hahn**+(2024)*
reconst. power spectrum



*Massara+**Hahn**+(2022)*
mark power spectrum

advantages of SBI: leveraging simulations to account for *observational systematics*

advantages of SBI: leveraging simulations to account for *observational systematics* — complex masks

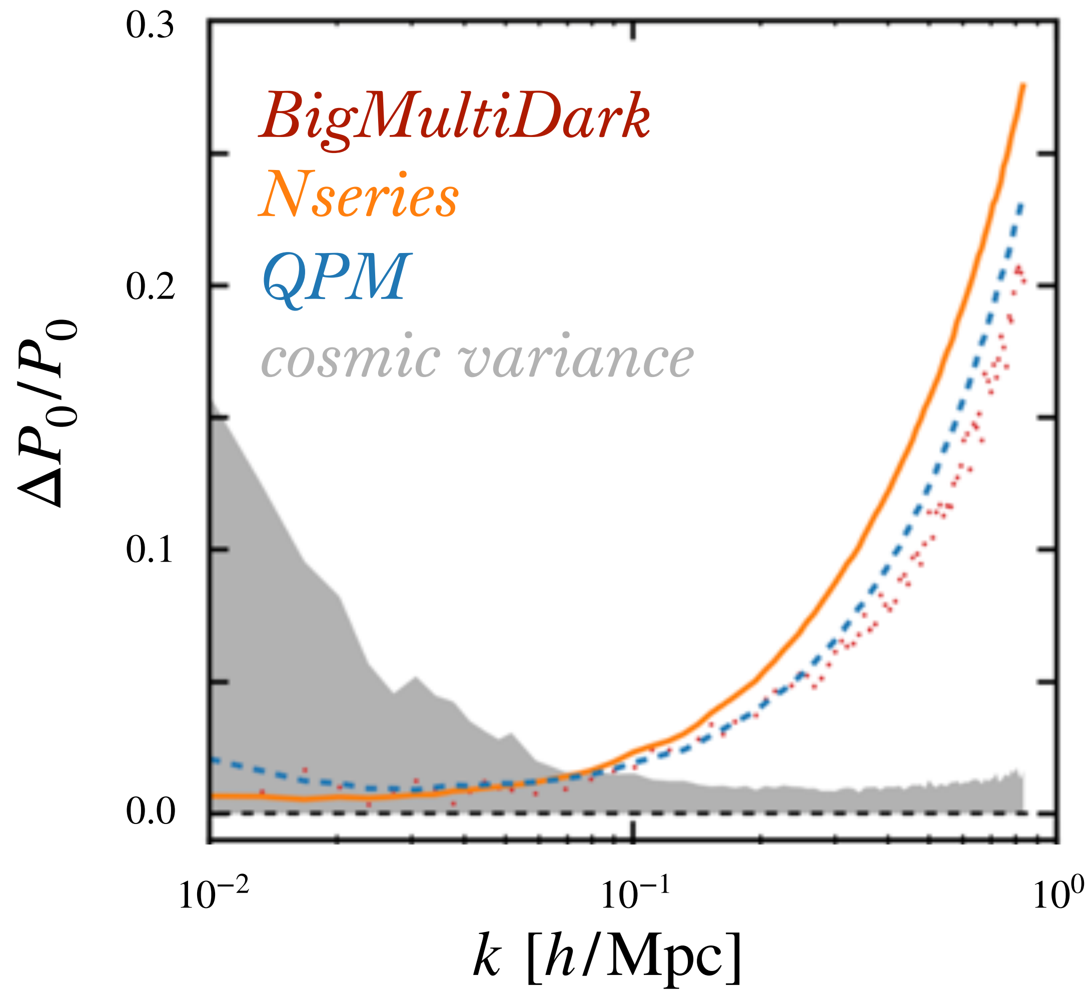


SDSS-III: BOSS CMASS Southern Galactic Cap

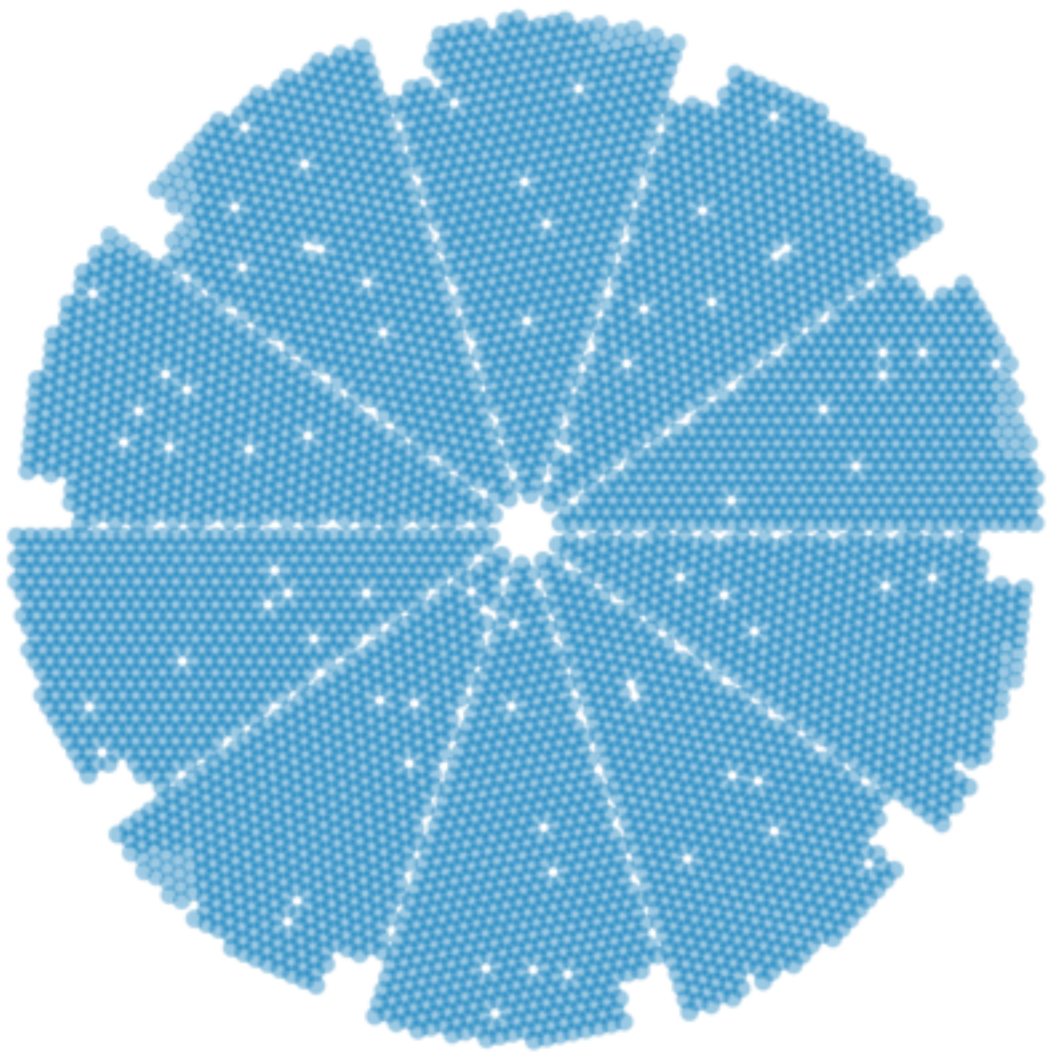
advantages of SBI: leveraging simulations to account for *observational systematics* — fiber collisions

focal plane of SDSS

advantages of SBI: leveraging simulations to account for *observational systematics* — fiber collisions

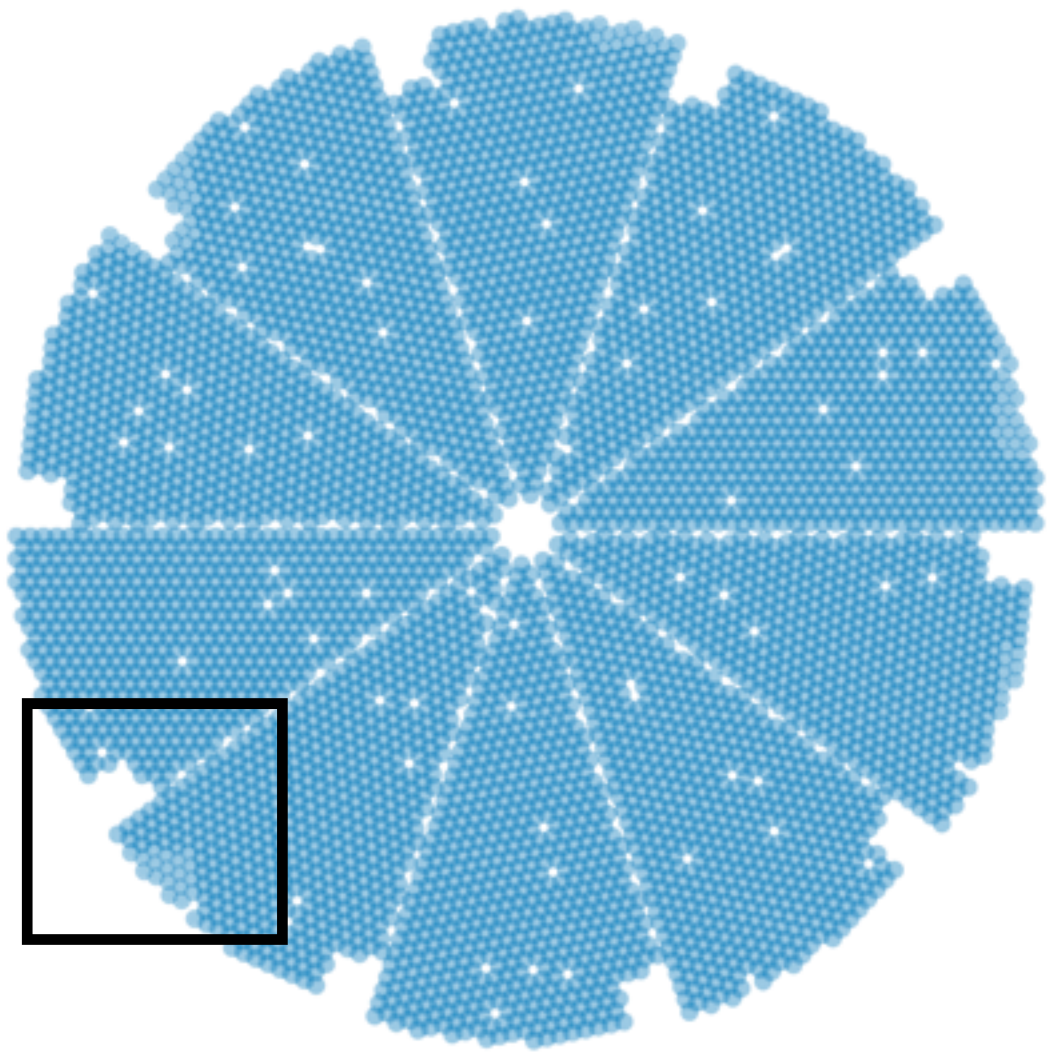


advantages of SBI: leveraging simulations to account for *observational systematics* — fiber collisions

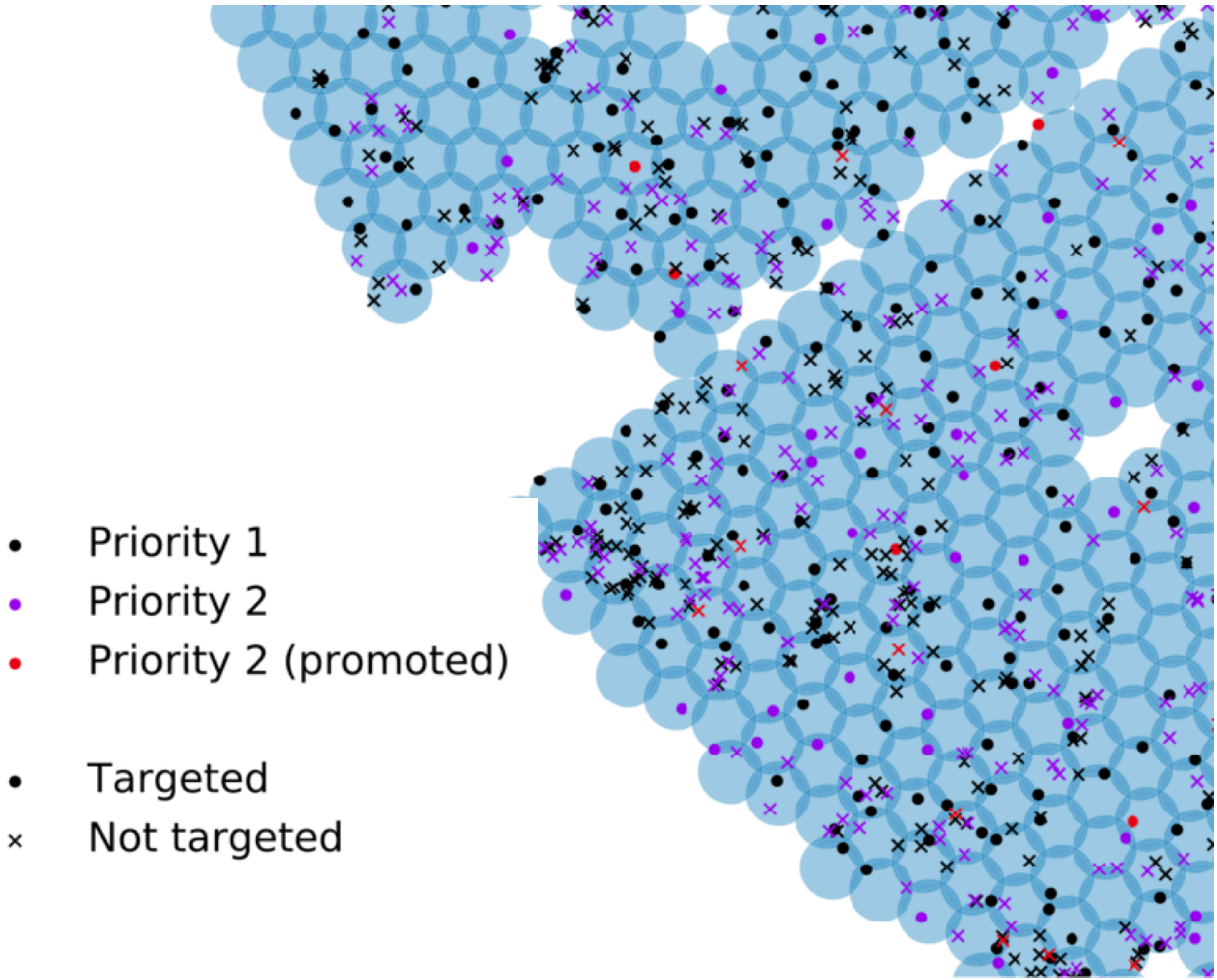


DESI tile — Smith et al.(2019); ***Hector's talk***

advantages of SBI: leveraging simulations to account for *observational systematics* — fiber collisions

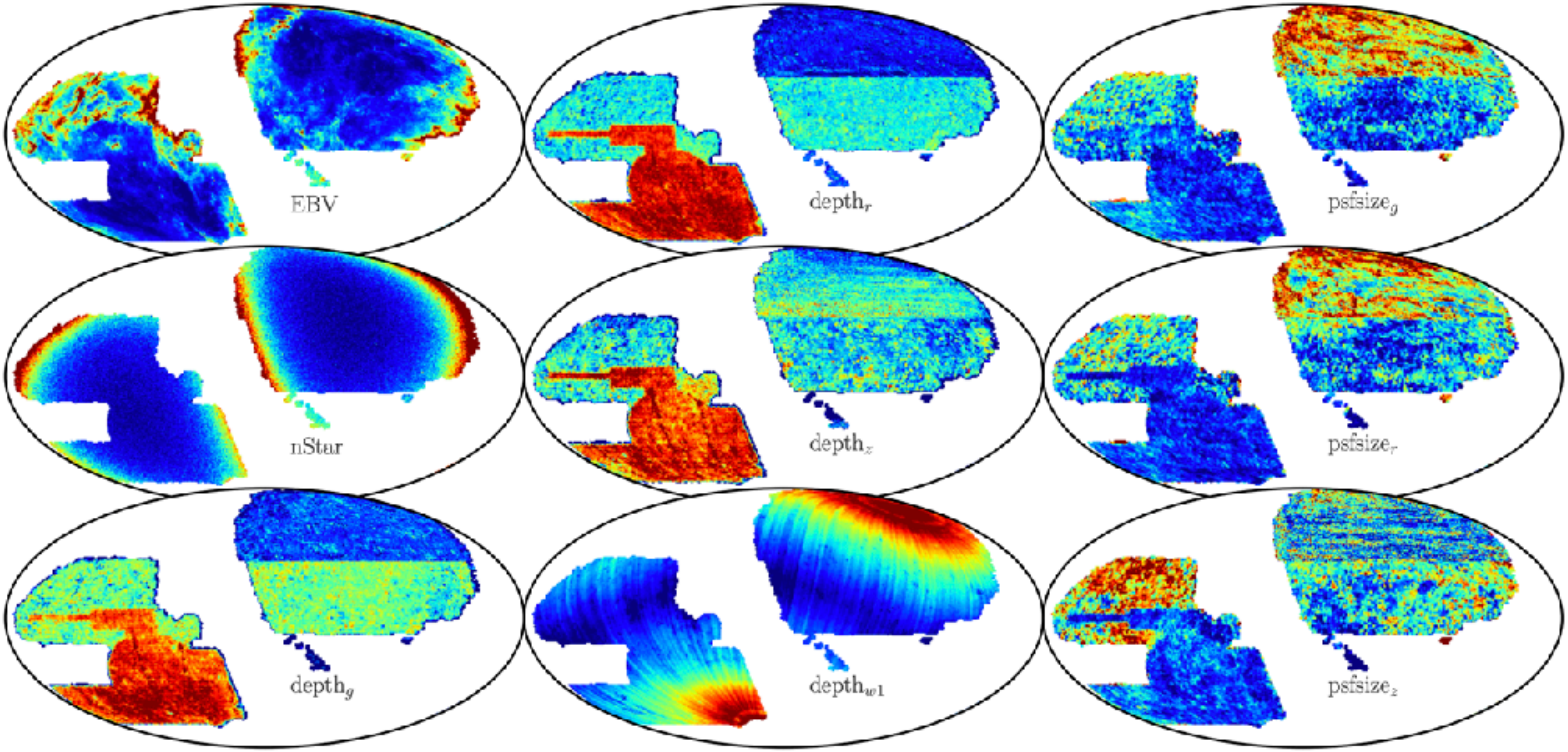


DESI tile — Smith et al.(2019)



Smith et al.(2019)

advantages of SBI: leveraging simulations to account for *observational systematics* — imaging systematics



DESI LRG imaging systematics (Rezaie et al. 2024)
see also ***Rongpu's talk***

what is simulation-based inference?

simulation-based inference* *in action*

challenges for simulation-based inference?

**state-of-the-art SBI (e.g. neural posterior estimation)*

SIMBIG

Simulation-**B**ased **I**nference of **G**alaxies

ChangHoon Hahn
Princeton Univ.
(spokesperson)

Michael
Eickenberg
CCM Flatiron

Shirley Ho
CCA Flatiron

Jiamin Hou
Univ. of Florida

Liam Parker
Princeton Univ.

Pablo Lemos
MILA

Elena Massara
UWaterloo

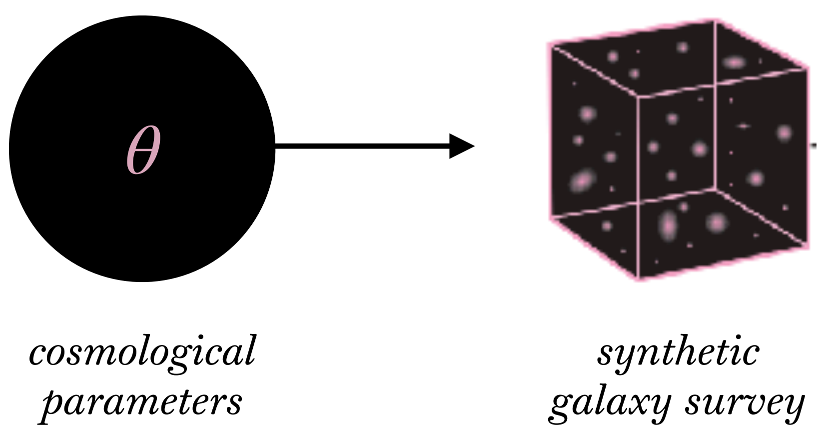
Chirag Modi
CCA CCM
Flatiron

Azadeh
Moradinezhad
Univ. de Genève

Bruno Régaldo-
Saint Blancard
CCM Flatiron

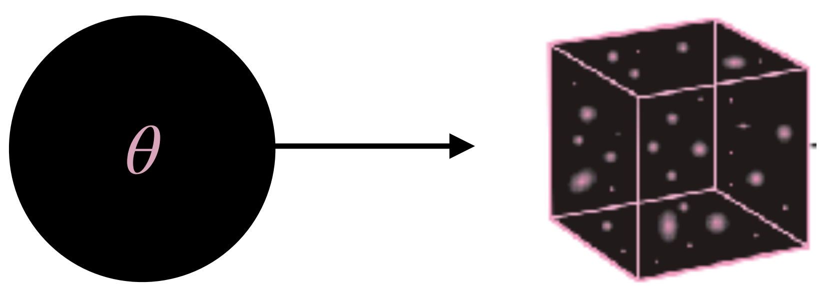
https://changhoonhahn.github.io/simbig/

SIMBIG — 1. *generating training data of synthetic observations*



SDSS-III: BOSS

SimBIG — 1. *generating training data of synthetic observations*

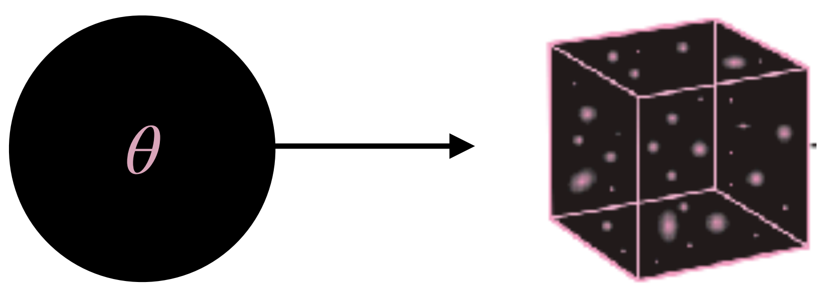


cosmological parameters

synthetic galaxy survey

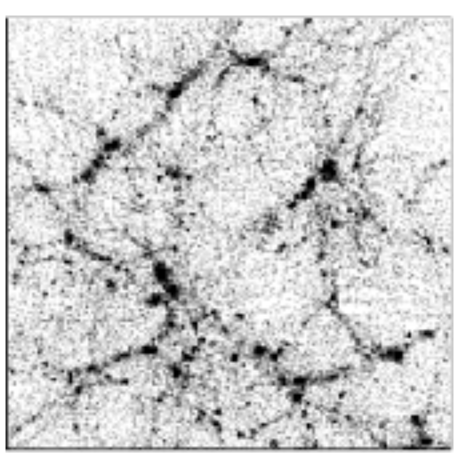
SDSS-III: BOSS

SIMBIG — 1. *generating training data of synthetic observations*



cosmological parameters

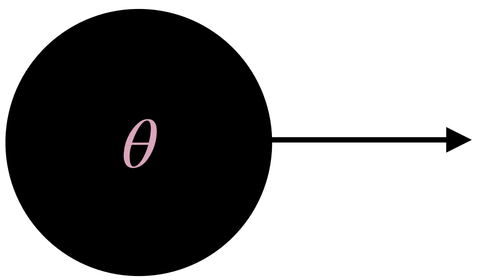
synthetic galaxy survey



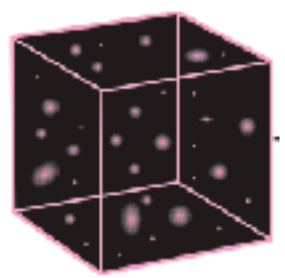
Quijote high-res
N-body simulations

SDSS-III: BOSS

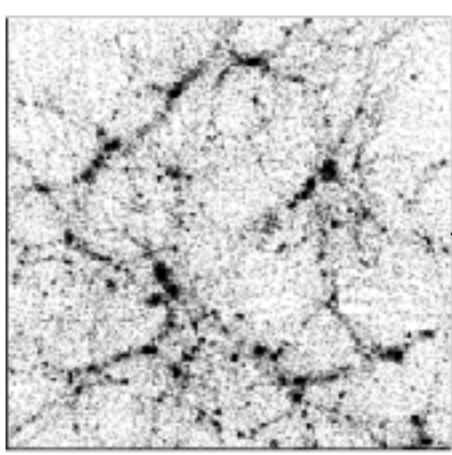
SimBIG — 1. *generating training data of synthetic observations*



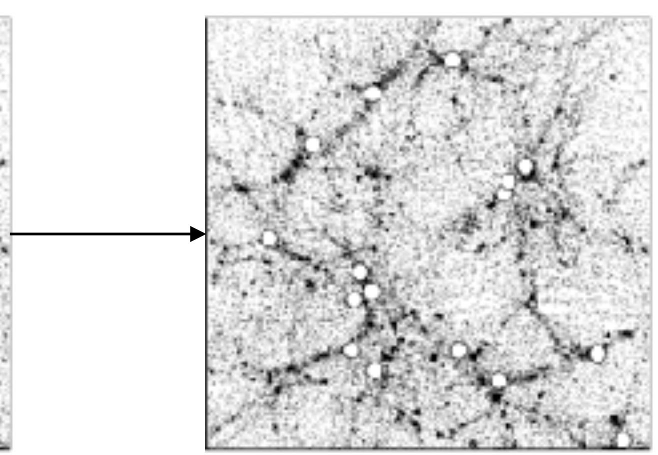
cosmological parameters



synthetic galaxy survey



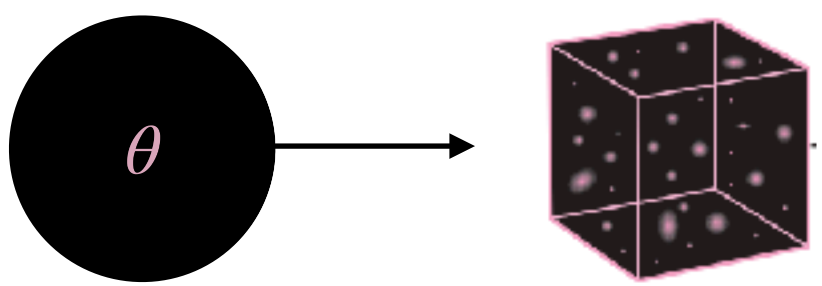
Quijote high-res
N-body simulations



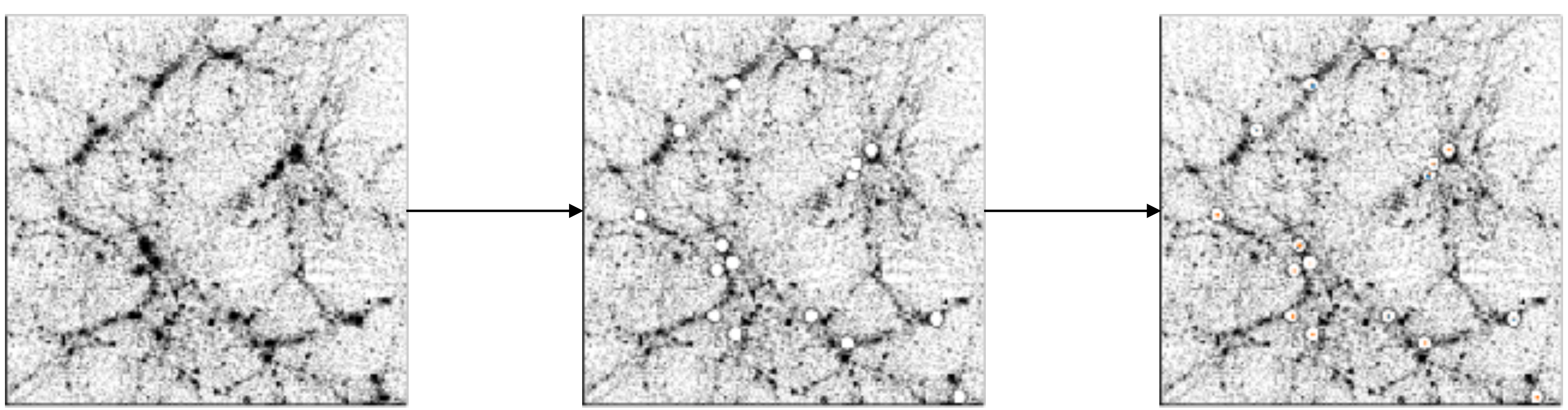
Rockstar phase-space
halo finder

SDSS-III: BOSS

SimBIG — 1. *generating training data of synthetic observations*



cosmological parameters *synthetic galaxy survey*



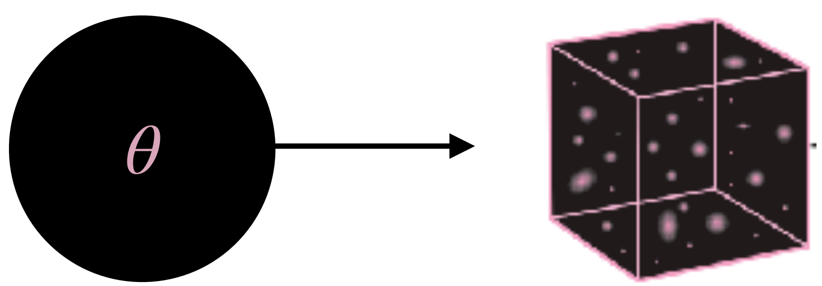
Quijote high-res N-body simulations

Rockstar phase-space halo finder

HOD model with *assembly, velocity, concentration biases*

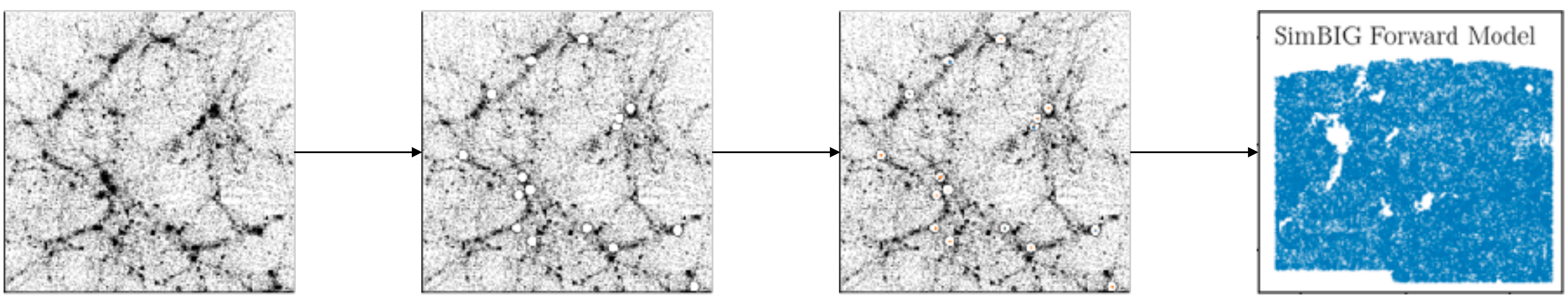
SDSS-III: BOSS

SimBIG — 1. *generating training data of synthetic observations*



cosmological parameters

synthetic galaxy survey



Quijote high-res N-body simulations

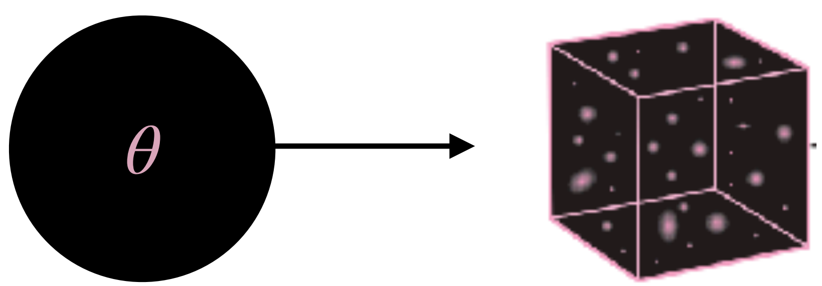
Rockstar phase-space halo finder

HOD model with *assembly, velocity, concentration biases*

survey realism: *redshift-space, geometry, mask, fiber collisions*

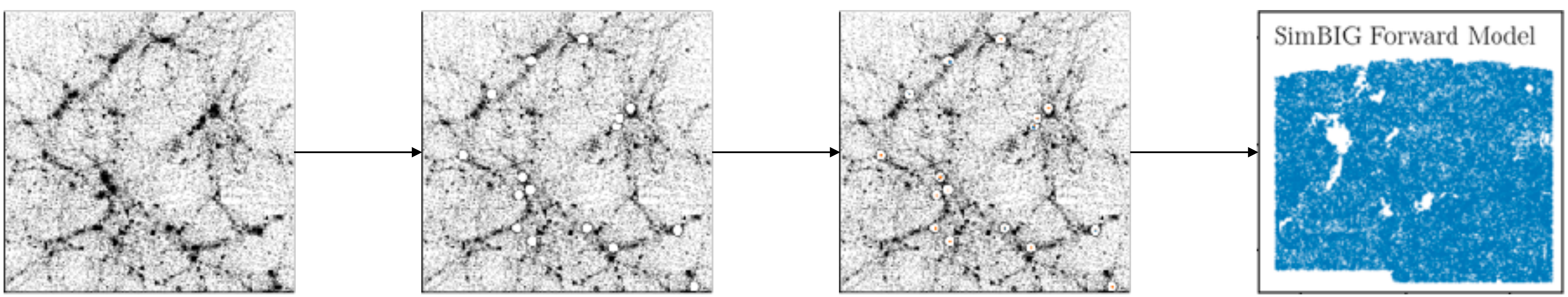
SDSS-III: BOSS

SimBIG — 1. *generating training data of synthetic observations*



cosmological parameters

synthetic galaxy survey



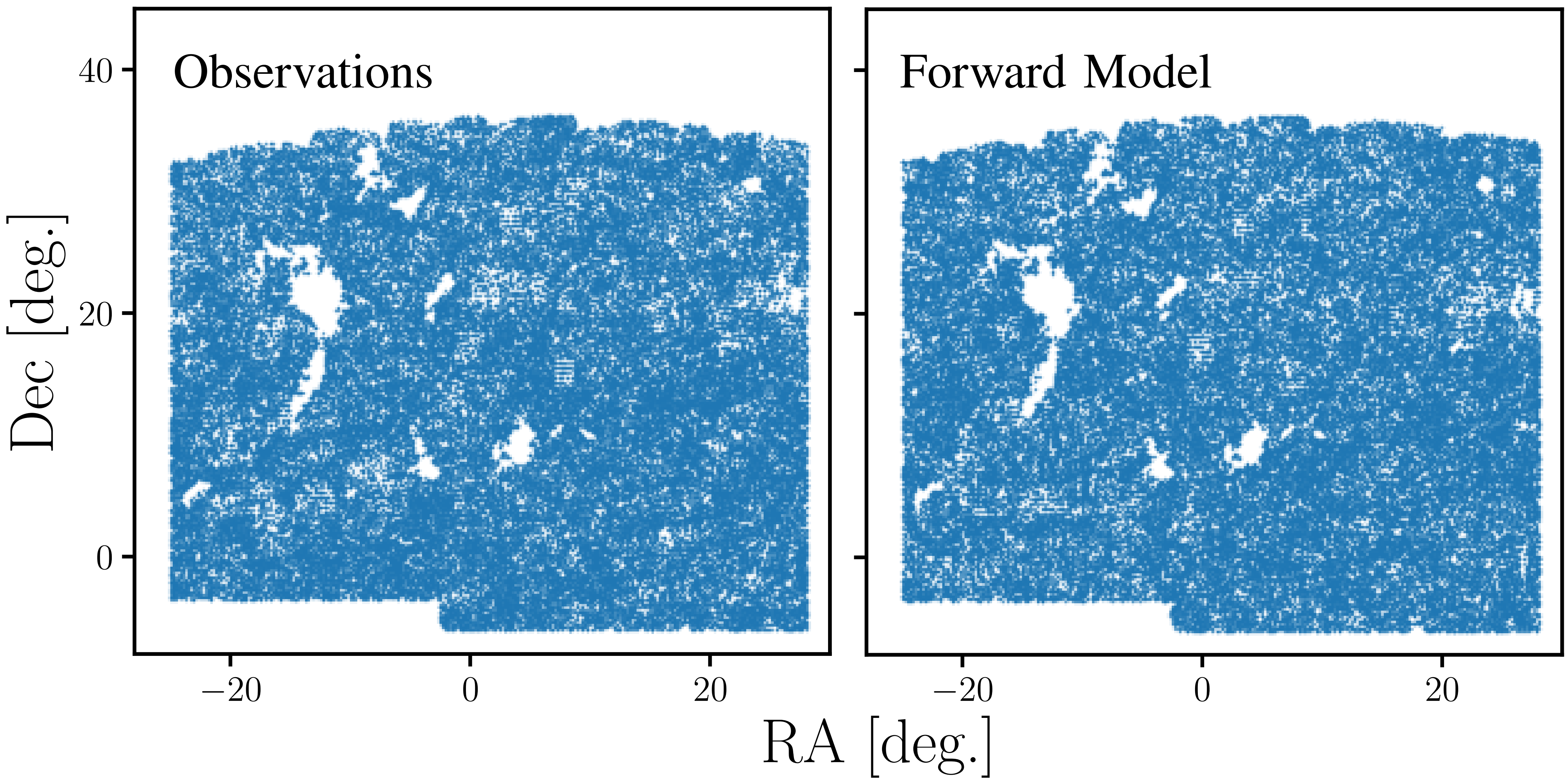
Quijote high-res N-body simulations

Rockstar phase-space halo finder

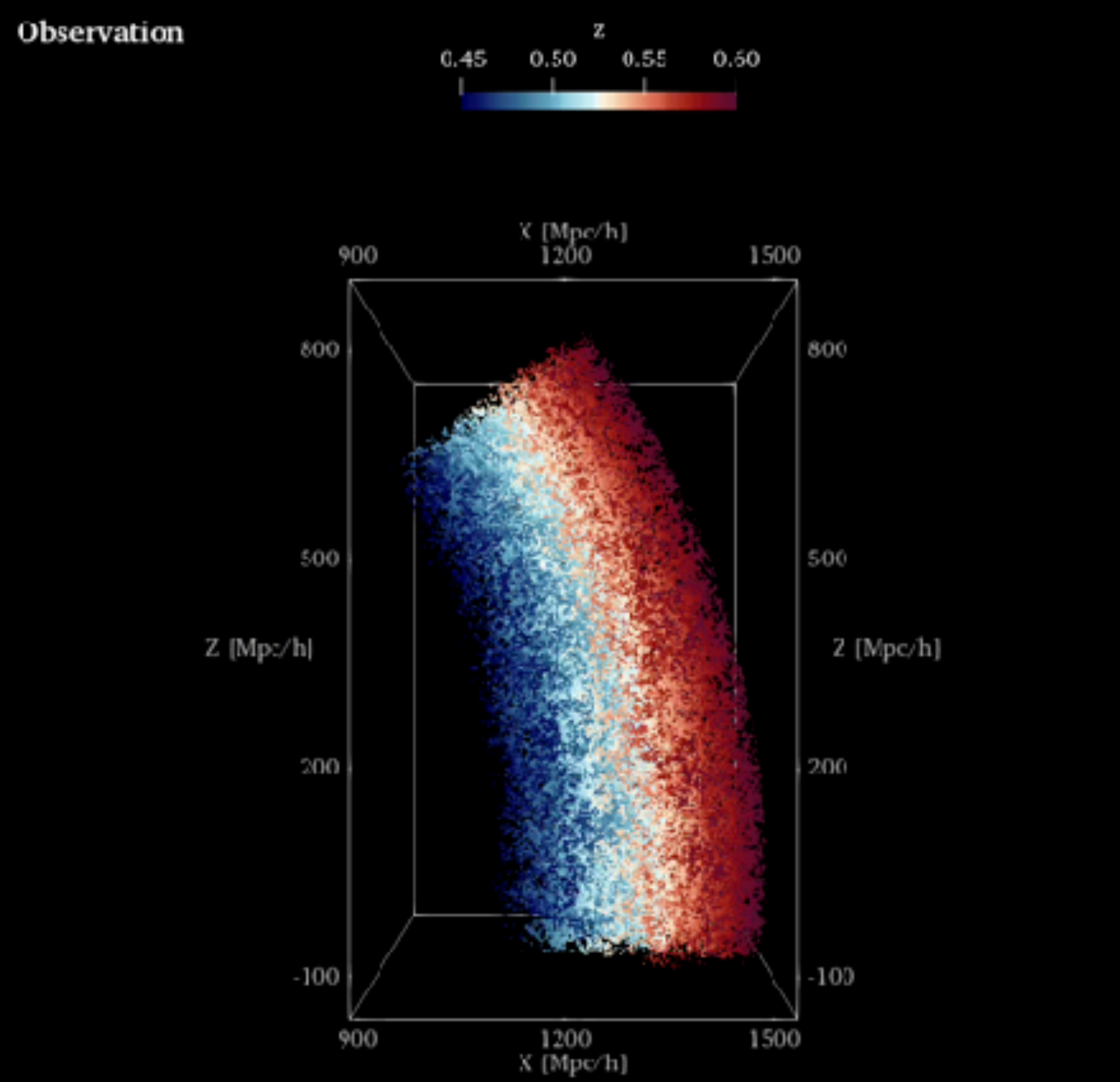
HOD model with *assembly, velocity, concentration biases*

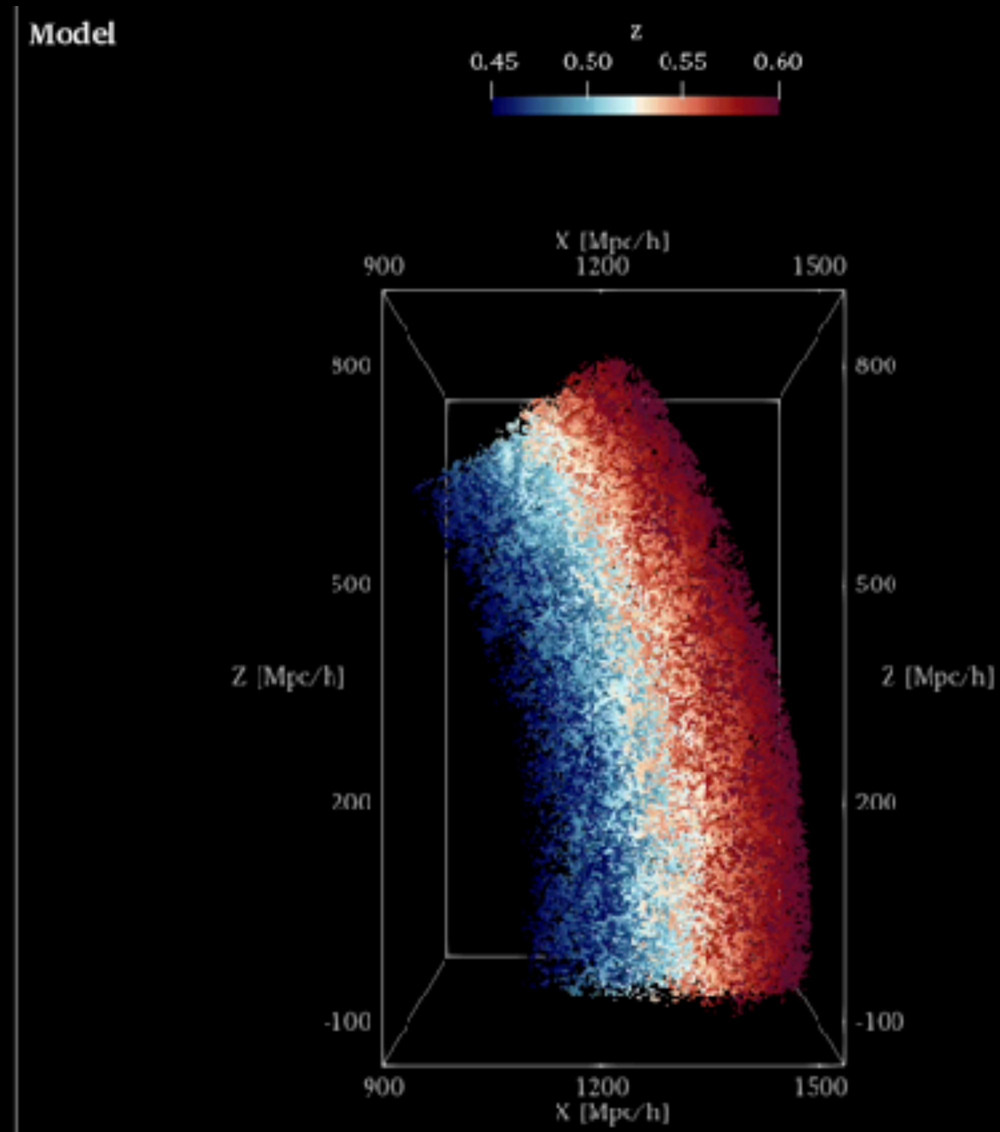
survey realism: *redshift-space, geometry, mask, fiber collisions*

20,000 training simulations spanning broad range of cosmologies and HOD parameters



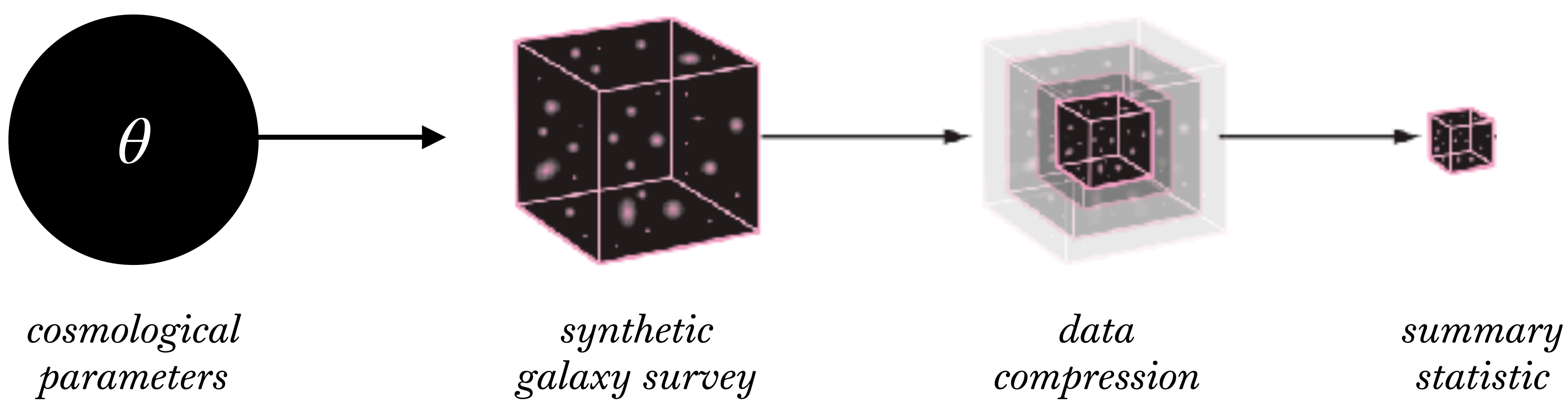
Hahn et al.(2023c, 2023d)



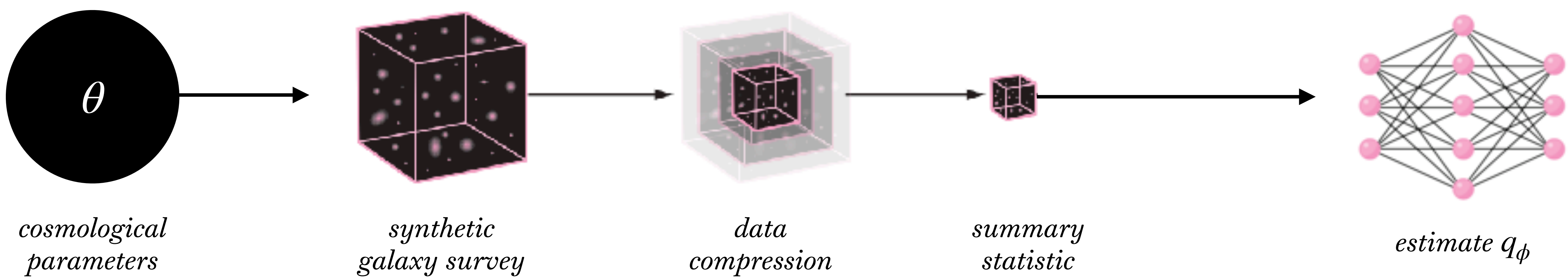


Hahn *et al.(2023c, 2023d)*
video: Bruno Régaldo-Saint Blancard

SimBIG — 1. *generating training data of synthetic observations*



SimBIG — 2. *estimating the neural posterior estimator* q_ϕ



normalizing flows: generative models that are easy to evaluate and flexibly expressive

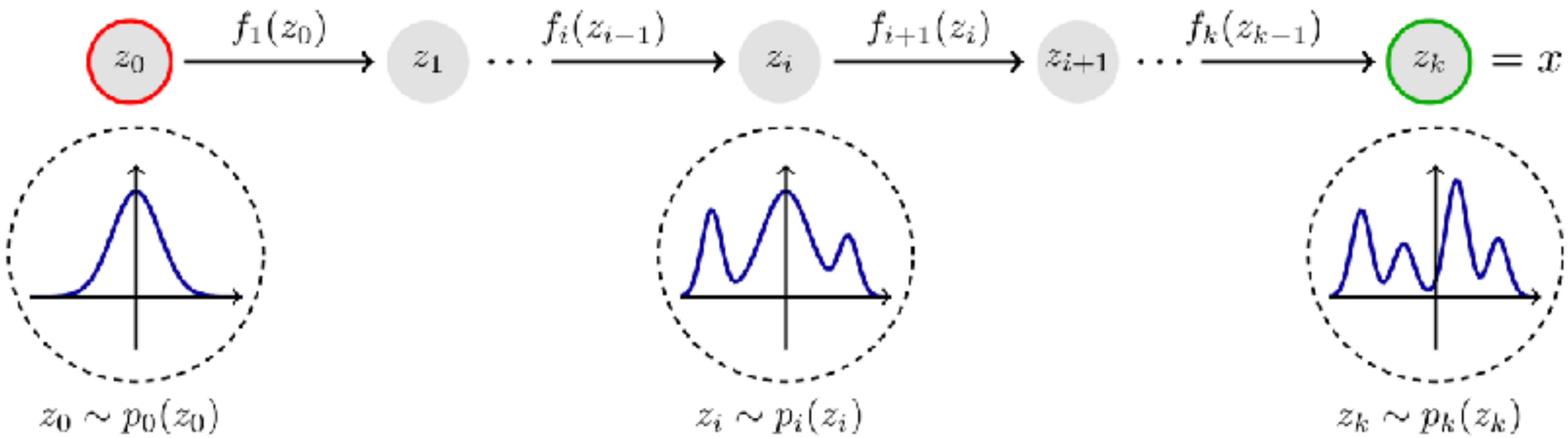
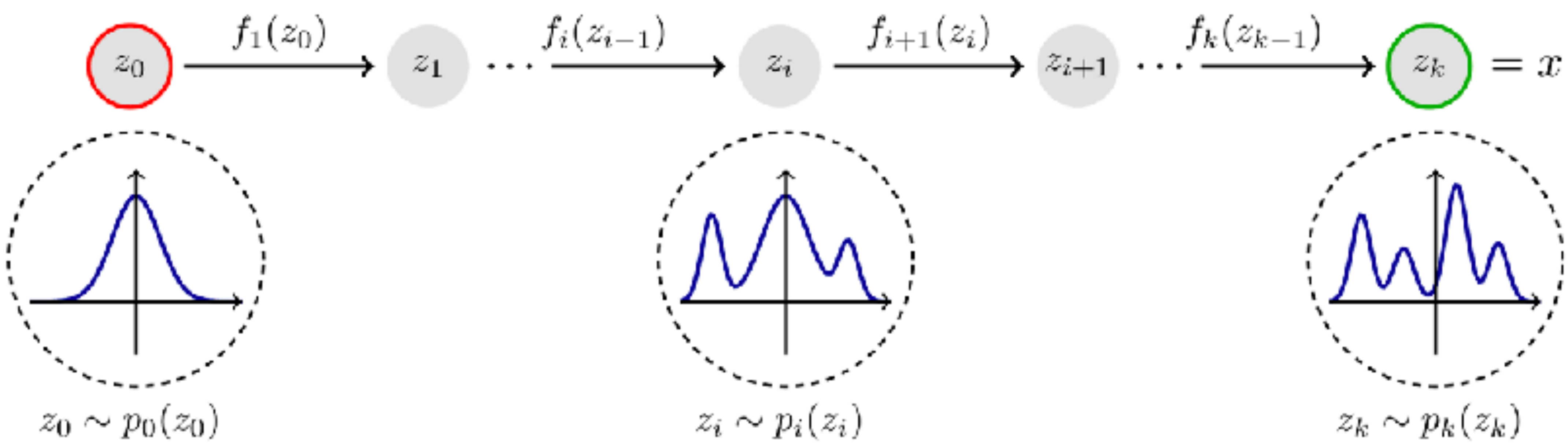


image: Janosh Riebesell

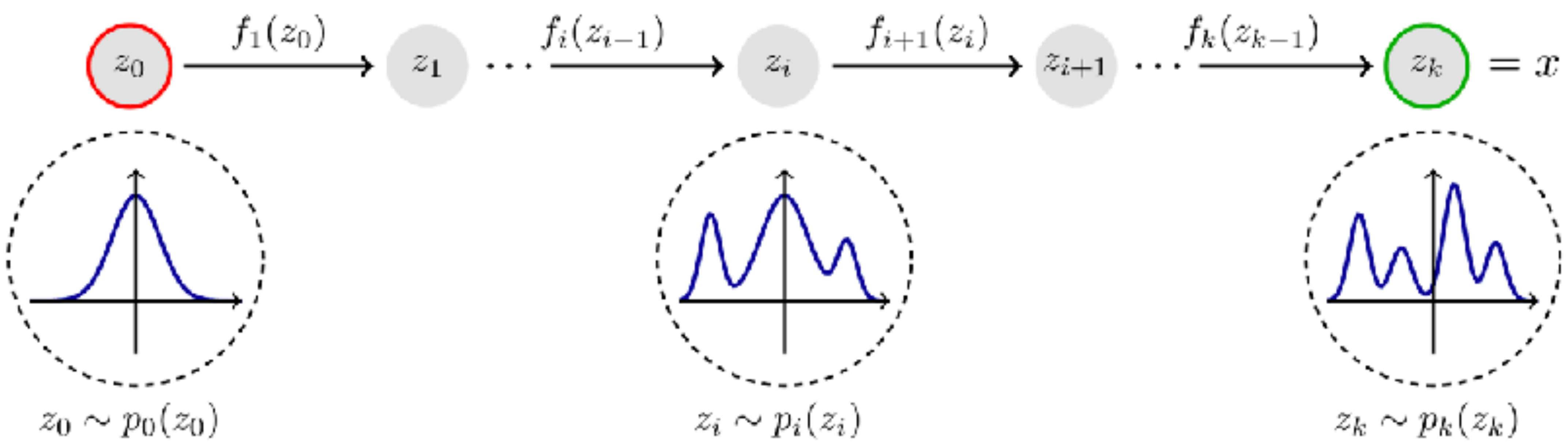
normalizing flows: generative models that are easy to evaluate and flexibly expressive



$z_i = f_i(z_{i-1})$ are invertible and differentiable transformations

$$p(z_i) = p(z_{i-1}) \left| \det \left(\frac{\partial f_i^{-1}}{\partial z_i} \right) \right|$$

normalizing flows: generative models that are easy to evaluate and flexibly expressive



$z_i = f_i(z_{i-1})$ are invertible and differentiable transformations

$f = f_1 \circ f_2 \ldots \circ f_{k-1} \circ f_k$ is also invertible and differentiable

normalizing flows: generative models that are easy to evaluate and flexibly expressive

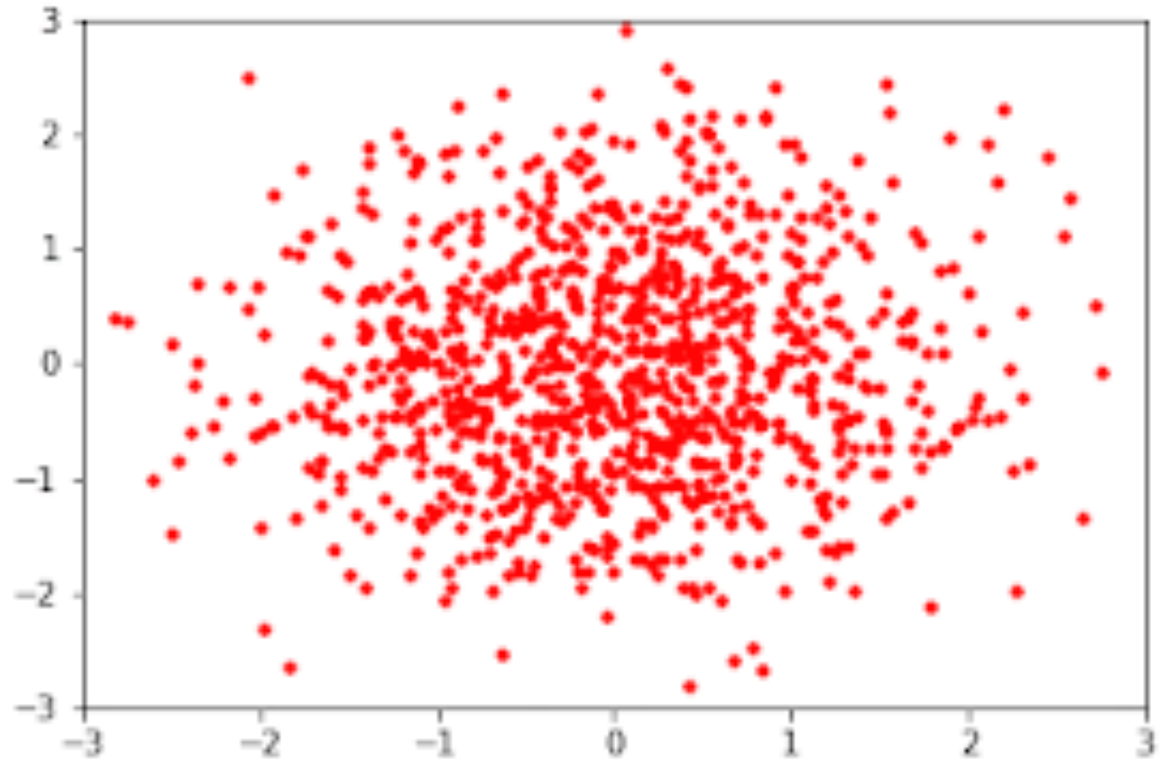
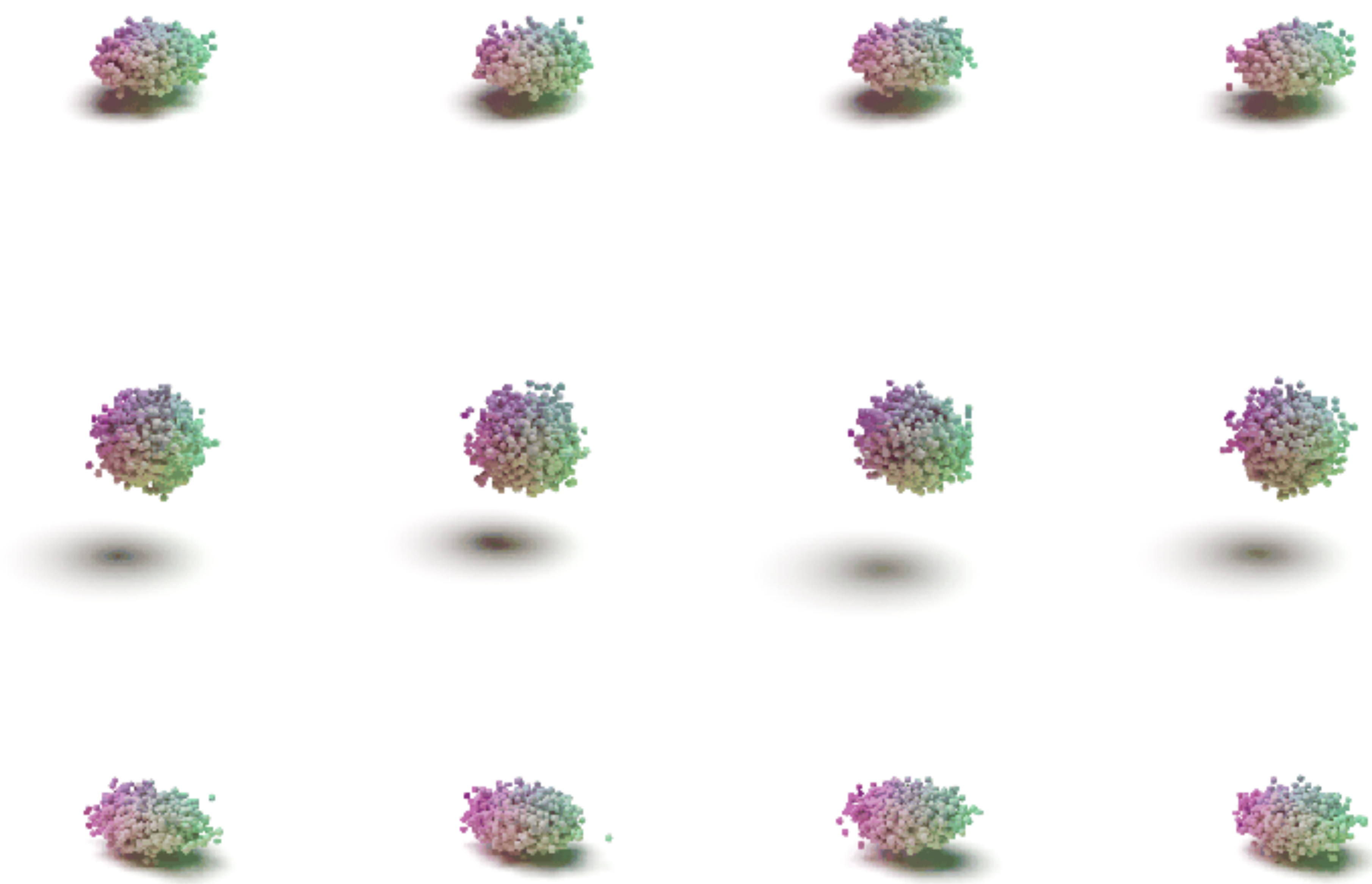


image: Eric Jang

normalizing flows: generative models that are easy to evaluate and flexibly expressive



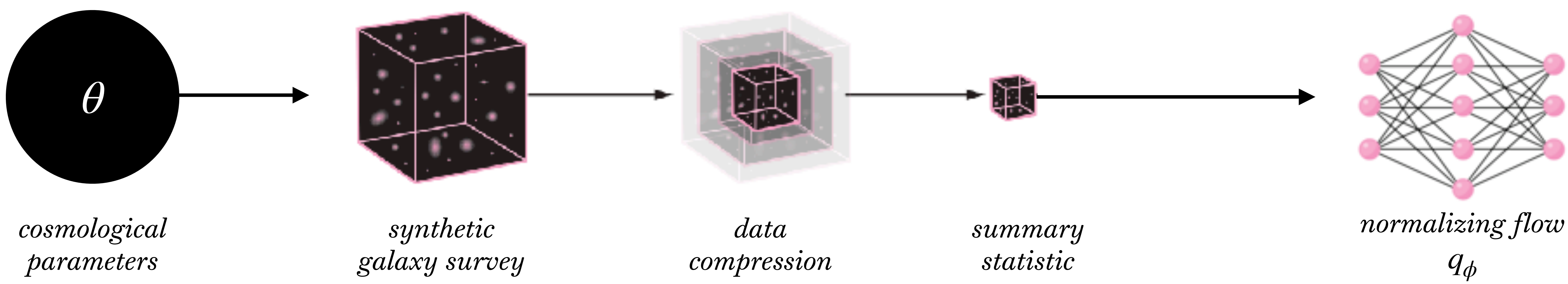
p(plane)

p(chair)

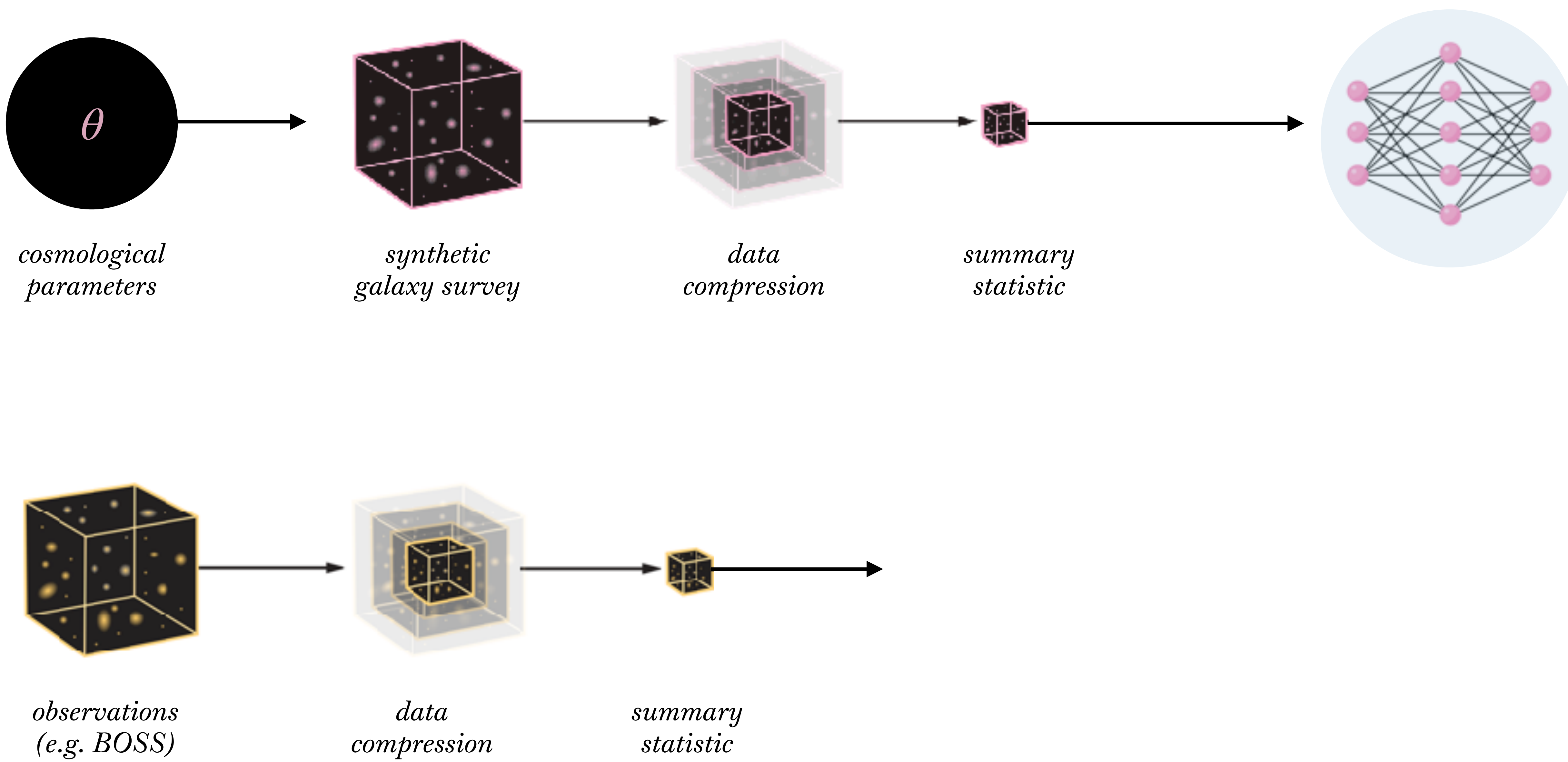
p(car)

Yang et al.(2019) PointFlow

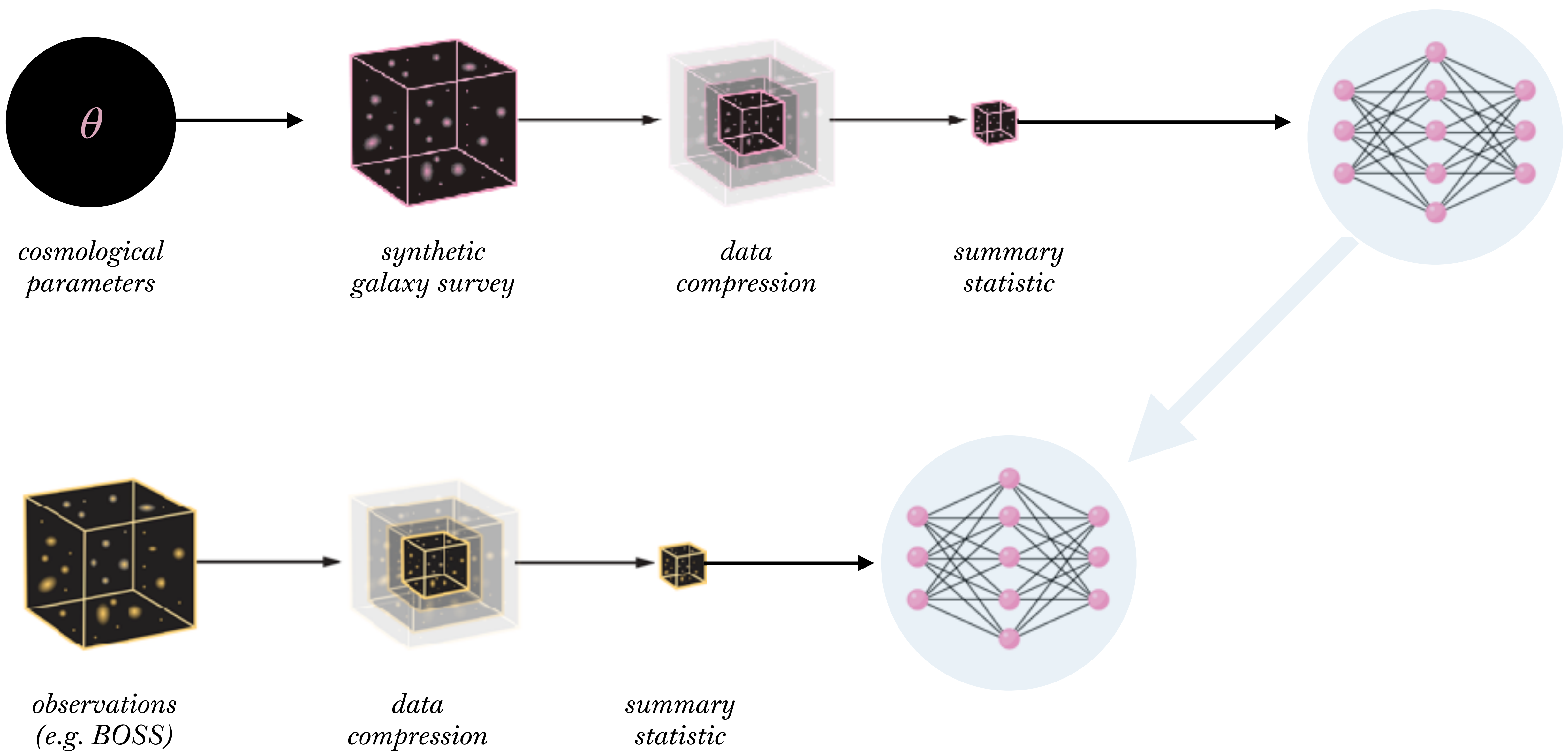
SIMBIG — 2. *training the normalizing flow*



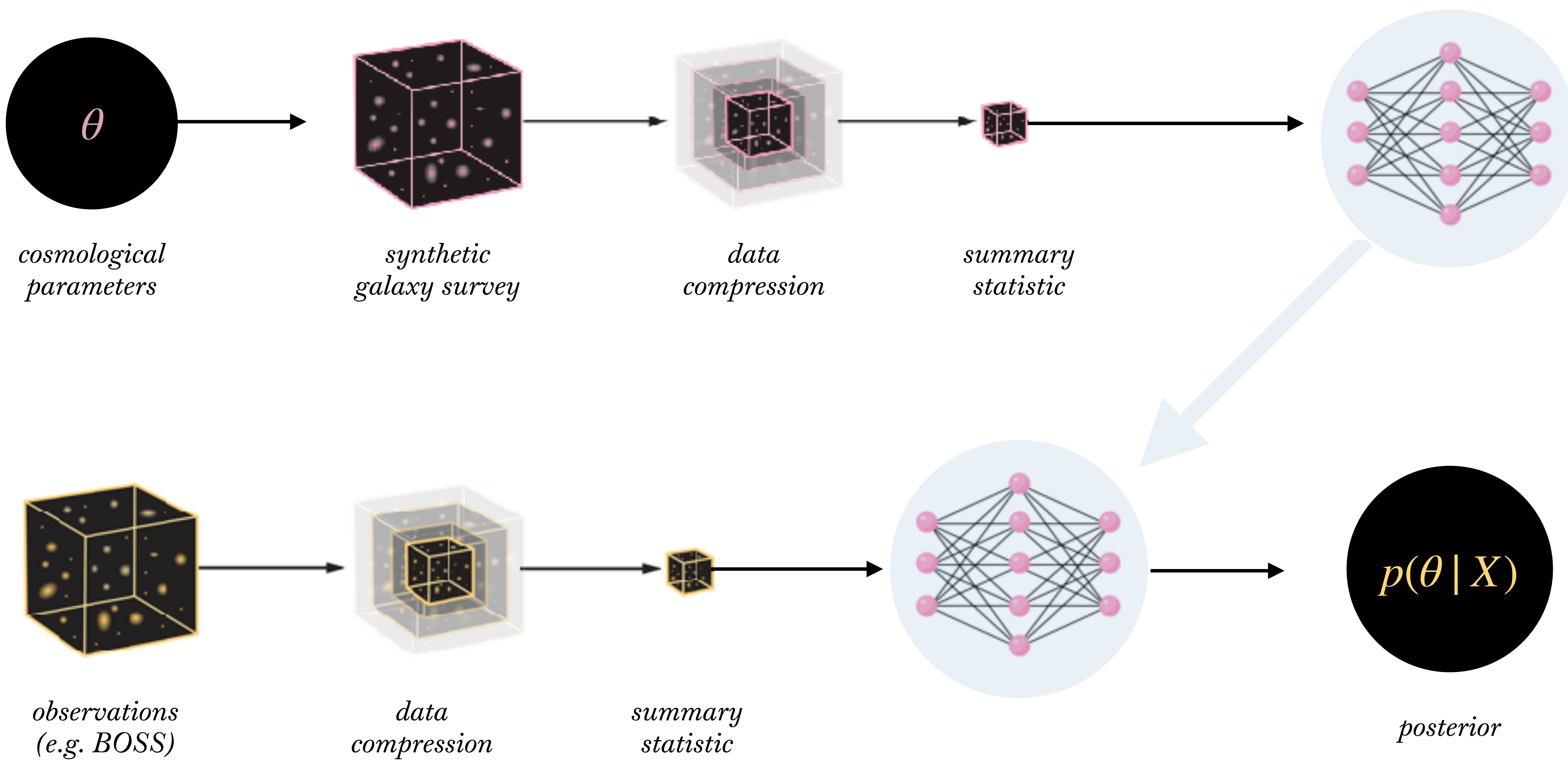
SimBIG — 3. inference using real observations



SIMBIG — 3. inference using real observations

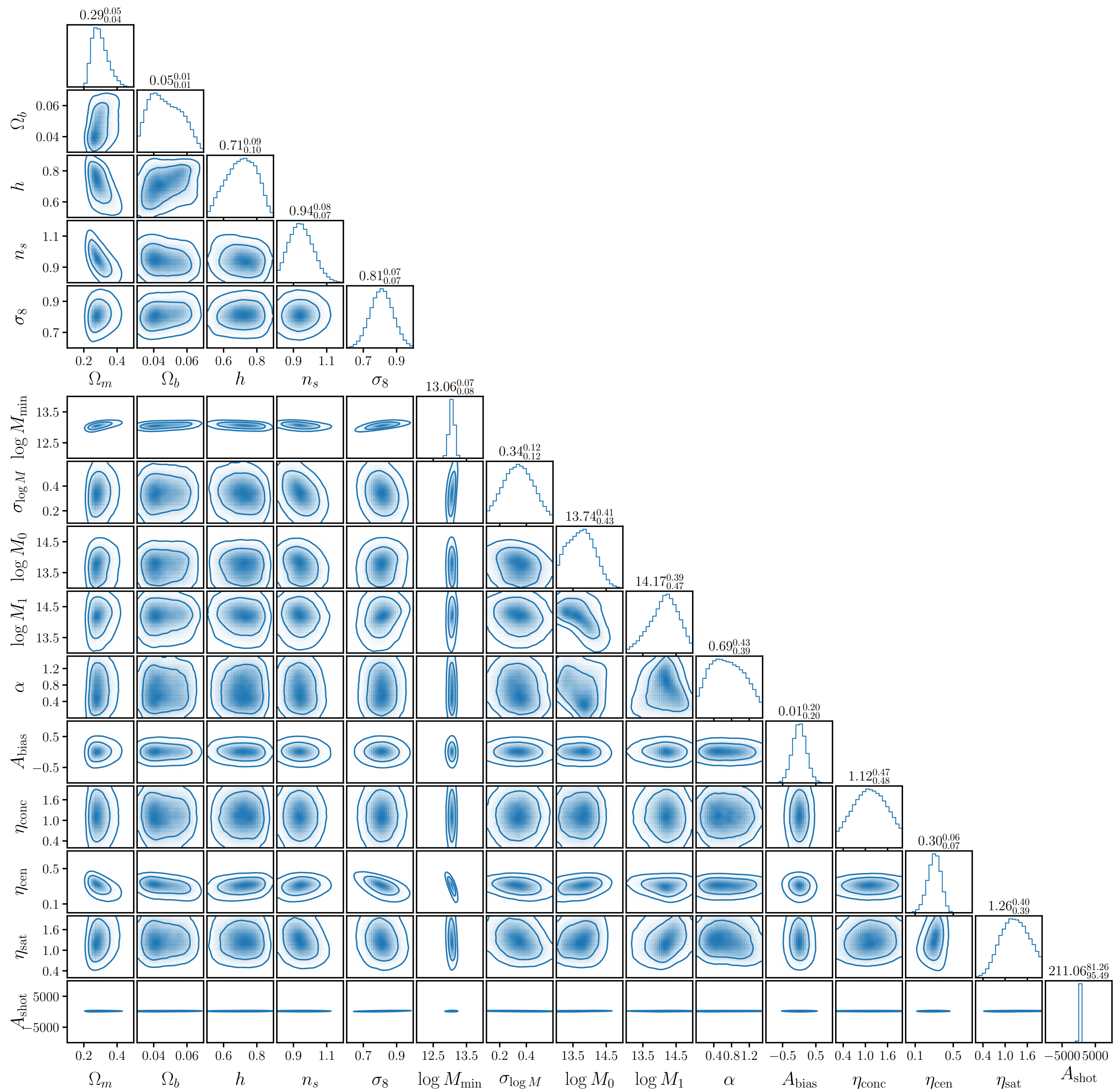


SimBIG — 3. inference using real observations



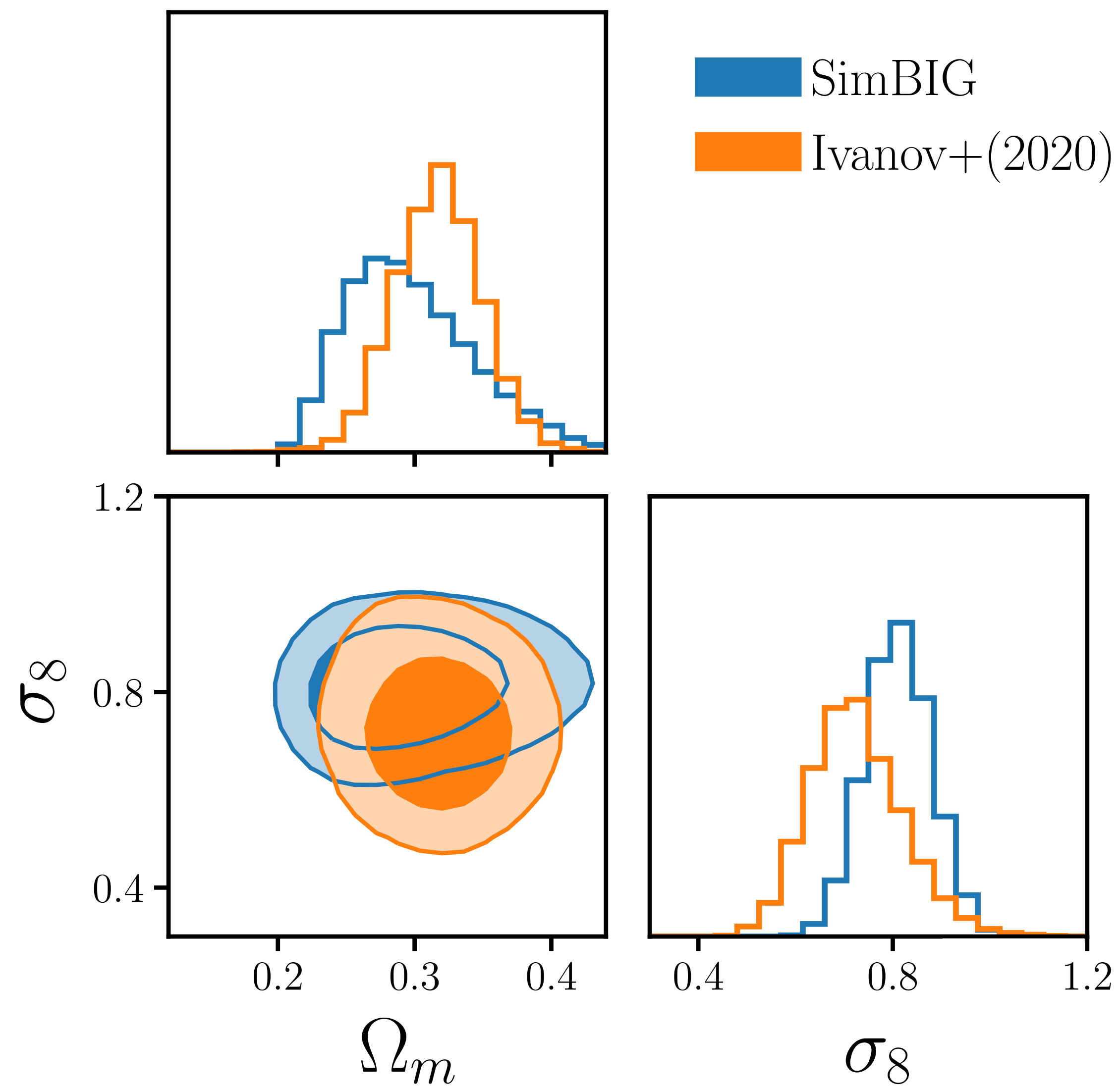
SIMBIG: non-linear galaxy power spectrum $P_\ell(k < 0.5\,h/\mathrm{Mpc})$

SimBIG: non-linear galaxy power spectrum $P_\ell(k < 0.5\, h/\mathrm{Mpc})$



Hahn et al.(2023c); Hahn et al.(2023d) PNAS

SimBIG: non-linear galaxy power spectrum $P_\ell(k < 0.5\,h/\mathrm{Mpc})$

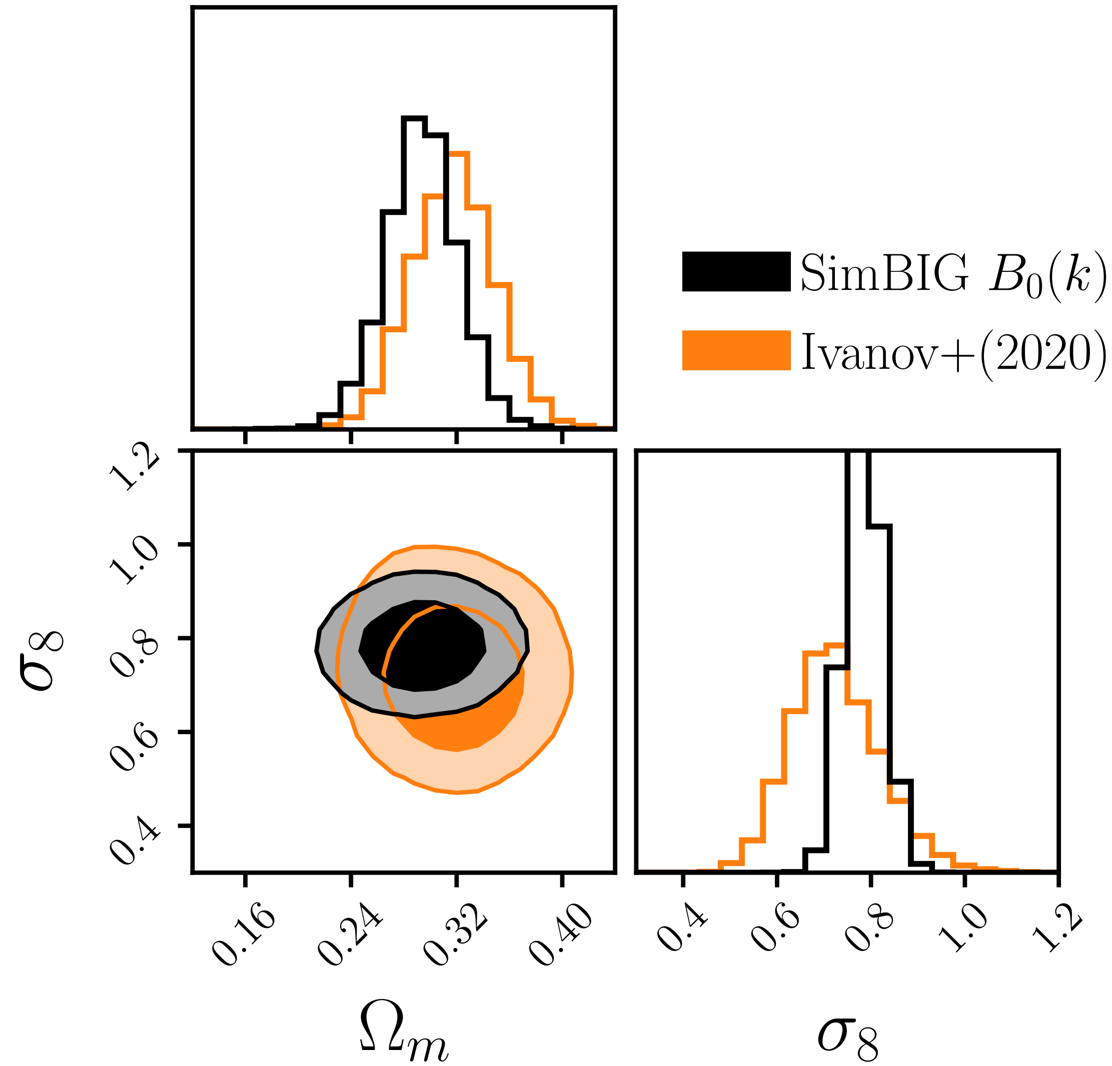


$1.4\times$ *tighter* σ_8 from non-linear scales

Hahn et al.(2023c); Hahn et al.(2023d) PNAS

SimBIG: non-linear galaxy bispectrum $B_0(k_1, k_2, k_3 < 0.5\, h/\mathrm{Mpc})$

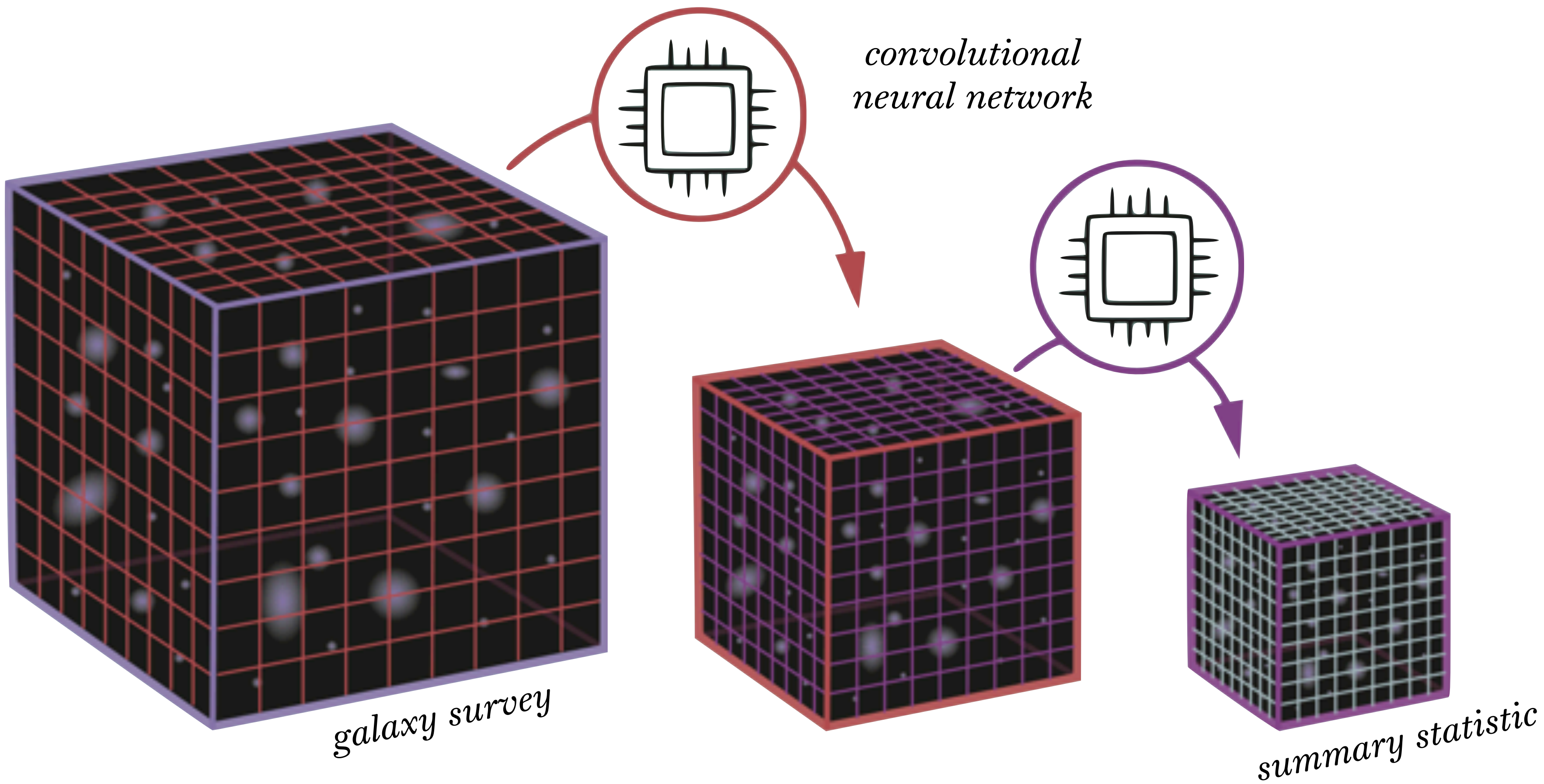
SIMBIG: non-linear galaxy bispectrum $B_0(k_1, k_2, k_3 < 0.5\,h/\mathrm{Mpc})$



1.2 and 2.4 $\times$ *tighter* Ω_m and σ_8 from **non-linear** + **higher-order** clustering

Hahn et al. (2023h)

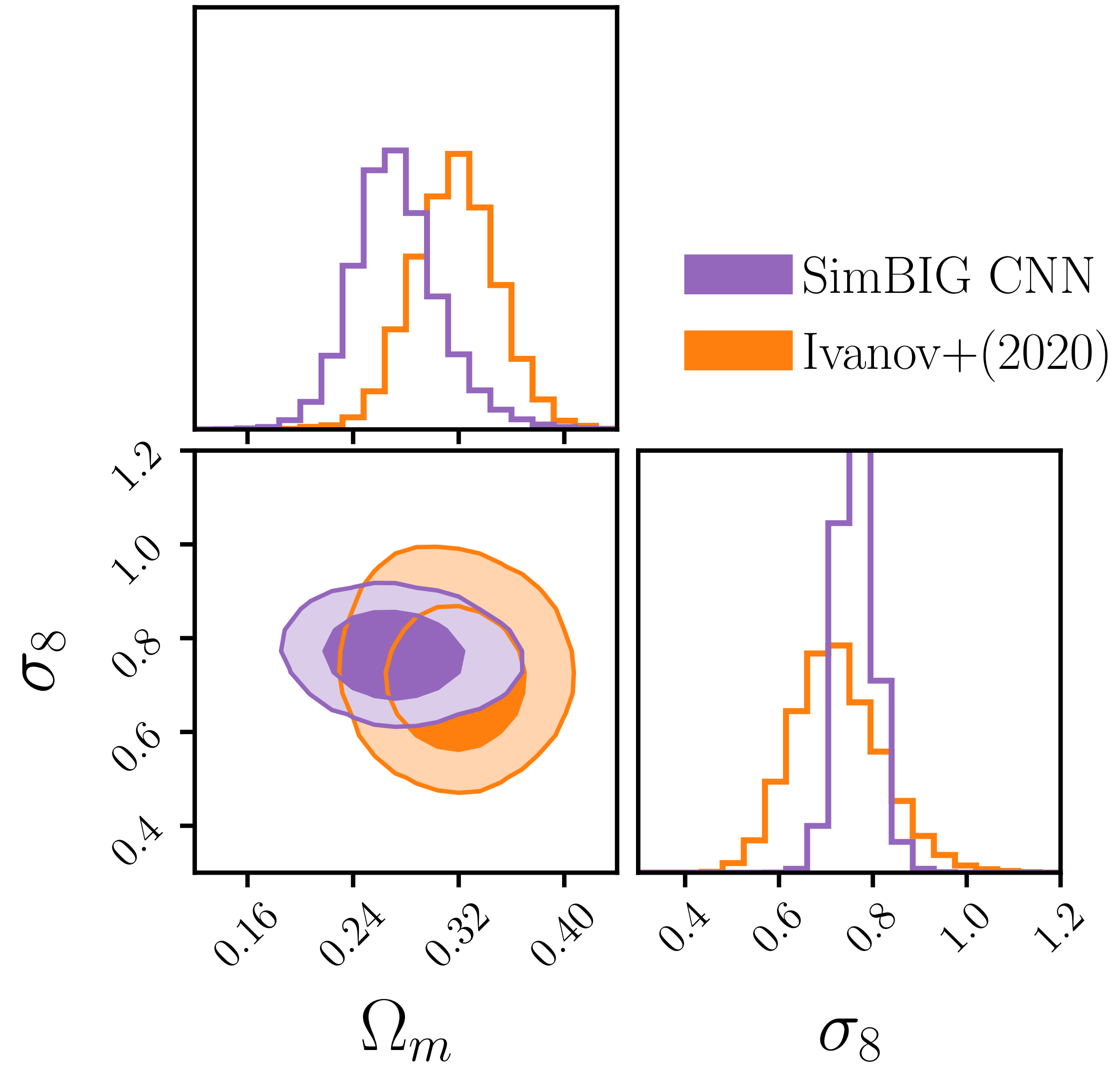
SimBIG: convolutional neural network field-level summary



extracting *all* relevant cosmological information in N-pt functions

Lemos, Parker, Hahn et al. (2023)

SimBIG: convolutional neural network field-level summary



extracting *all* relevant cosmological information in N-pt functions

Lemos, Parker, Hahn et al. (2023)

SimBIG

wavelet scattering transforms

Régaldo-Saint Blancard, Hahn et al. (2023)

Bruno Régaldo-Saint Blancard
CCM Flatiron

skew spectra

***Hou**, Moradinezhad Dizgah, Hahn et al. (2024)*

Jiamin Hou
Univ. of Florida

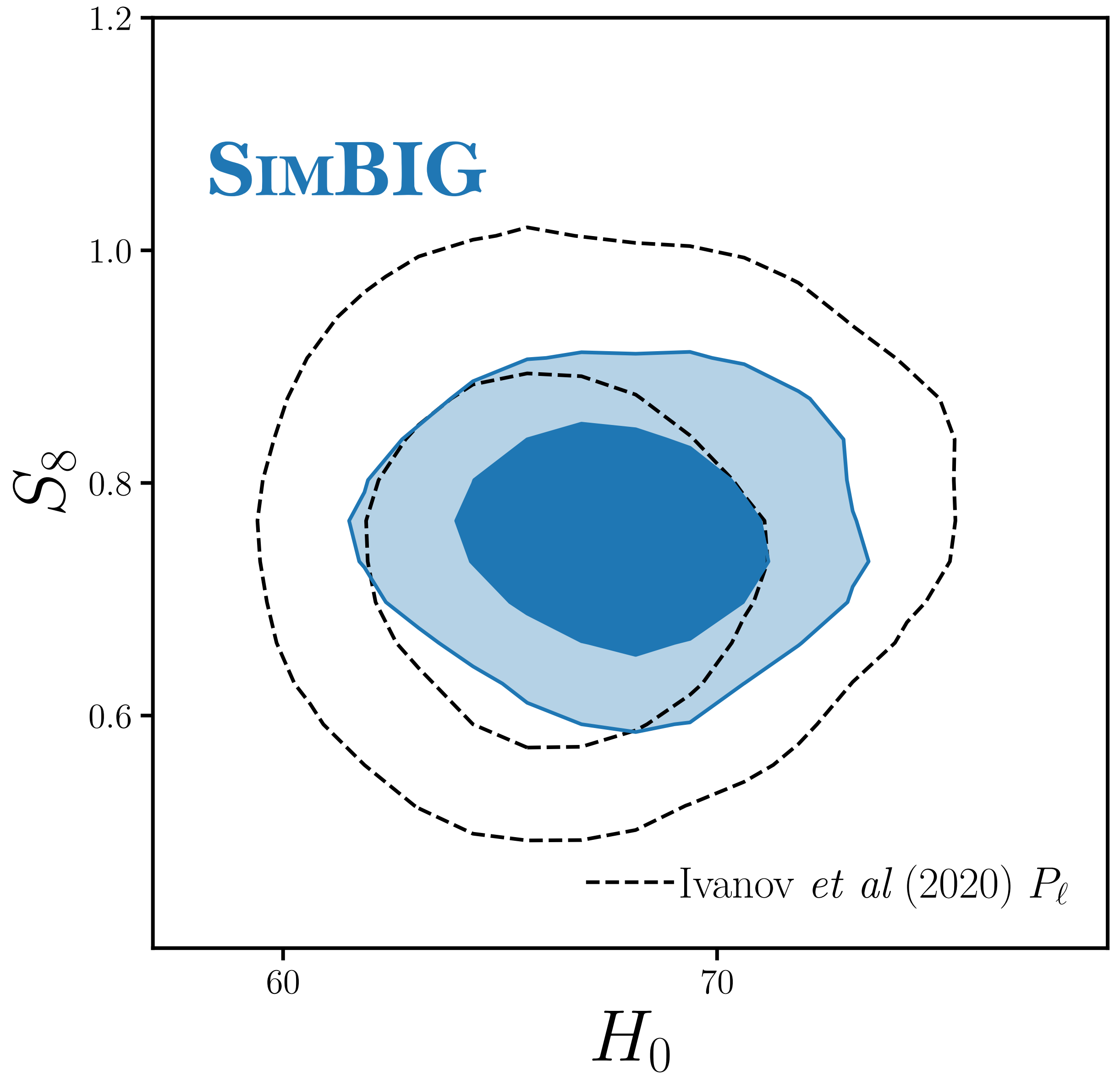
marked powerspectrum

Massara, Hahn et al. (2024)

Elena Massara
UWaterloo

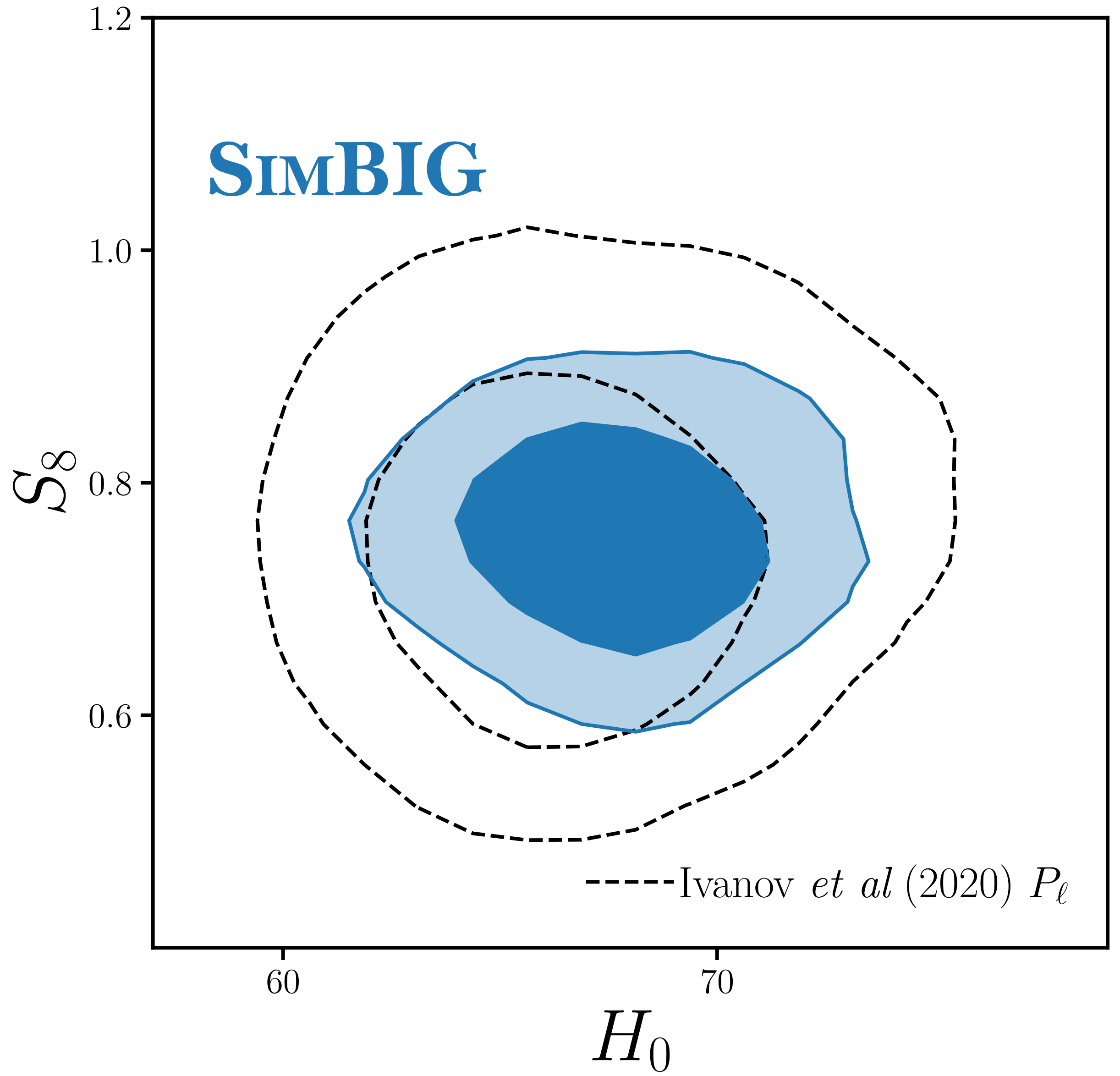
voids, clusters, (*your favorite summary statistic*)

SimBIG: ~1.9 and 1.5× tighter S_8 and H_0



Hahn et al. (2023i)

SimBIG: **~1.9 and 1.5× tighter** S_8 and H_0



S_8 improvement is equivalent to analyzing a *survey of ~4× larger volume*

Hahn et al. (2023i)

simulation-based inference* *in action*

see also many other cosmological SBI analyses: *Jeffrey et al.(2021), Fluri et al.(2022), Gatti et al.(2024), Moser et al.(2024), von Wietersheim-Kramsta et al. (2024)++*

**state-of-the-art SBI (e.g. neural posterior estimation)*

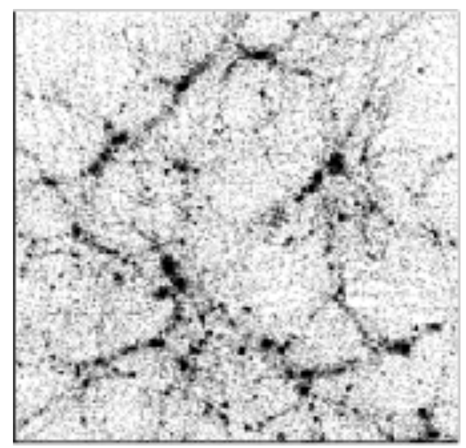
what is simulation-based inference?

opportunities for simulation-based inference?

challenges for simulation-based inference?

challenges for SBI: can forward models *scale* to the next-generation galaxy surveys?

challenges for SBI: can forward models *scale* to the next-generation galaxy surveys?



Quijote high-res
N-body simulations

$1\,\mathrm{Gpc}/h$ box with $\sim 10^{12} M_\odot$ halo mass limit

DARK ENERGY SPECTROSCOPIC INSTRUMENT

U.S. Department of Energy Office of Science

DESI will observe >40 million galaxies

15 million bright galaxies	$z < 0.6$
8 million Luminous Red Galaxies	$0.4 < z < 1.0$
16 million Emission Line Galaxies	$0.6 < z < 1.6$
3 million Quasars	$0.9 < z < 2.1$

see Hector's talk; DESI Year 1 papers

DARK ENERGY SPECTROSCOPIC INSTRUMENT
U.S. Department of Energy Office of Science

DESI will observe >40 million galaxies

	$V_{\rm eff}$
15 million bright galaxies	$1\ \mathrm{Gpc}^3$
8 million Luminous Red Galaxies	$12\ \mathrm{Gpc}^3$
16 million Emission Line Galaxies	$5\ \mathrm{Gpc}^3$
3 million Quasars	$1.5\ \mathrm{Gpc}^3$

see Hector's talk; DESI Year 1 papers

DESI will observe >40 million galaxies

	V_{eff}
15 million bright galaxies	1 Gpc^3
8 million Luminous Red Galaxies	12 Gpc^3
16 million Emission Line Galaxies	5 Gpc^3
3 million Quasars	1.5 Gpc^3

for year 1!

see Hector's talk; DESI Year 1 papers

DARK ENERGY SPECTROSCOPIC INSTRUMENT
U.S. Department of Energy Office of Science

DESI will observe >40 million galaxies

	$M_{h,\rm min}$
15 million bright galaxies	$< 10^{11} M_\odot$
8 million Luminous Red Galaxies	$\sim 10^{12} M_\odot$
16 million Emission Line Galaxies	$> 10^{11} M_\odot$
3 million Quasars	$\sim 10^{12} M_\odot$

Smith et al.(2024); Yuan et al.(2024); Garcia-Quintero

Prime Focus Spectrograph

SuMIRe

the SuMIRe Prime Focus Spectrograph (PFS) Cosmology Survey will observe on the **8.2m** Subaru telescope *next year*

5 million *emission line galaxy*

$0.6 < z < 2.4$

Yuka Yamada
Univ. of Tokyo
PFS Cosmology target selection

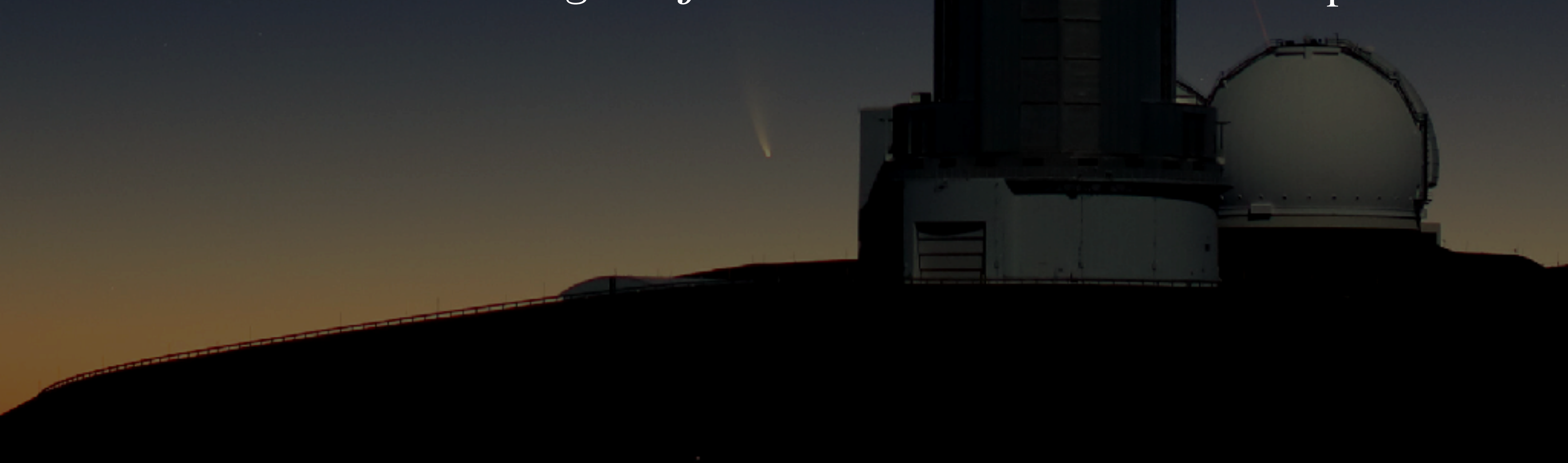
Prime Focus Spectrograph

SuMIRe

the SuMIRe Prime Focus Spectrograph (PFS) Cosmology Survey will observe on the **8.2m** Subaru telescope *next year*

$$\underline{V_{\rm eff}}$$

5 million *emission line galaxy* $\sim 10\,{\rm Gpc}^3$



Prime Focus Spectrograph

SuMIRe

the SuMIRe Prime Focus Spectrograph (PFS) Cosmology Survey will observe on the **8.2m** Subaru telescope *next year*

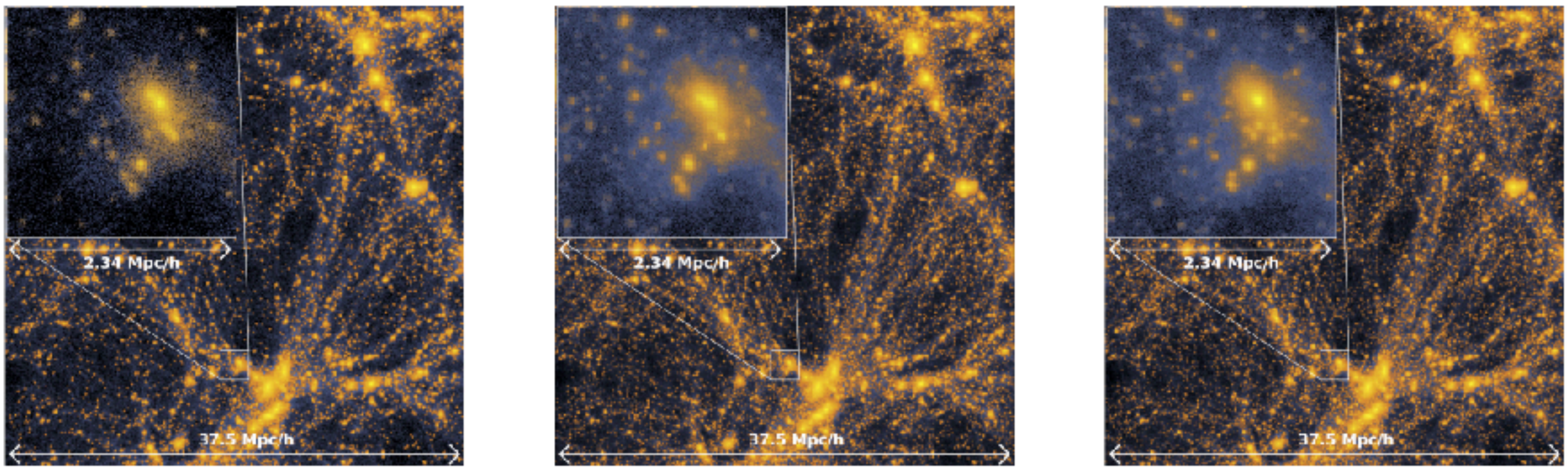
	$M_{h,\mathrm{min}}$
5 million *emission line galaxy*	$> 10^{11} M_{\odot}$

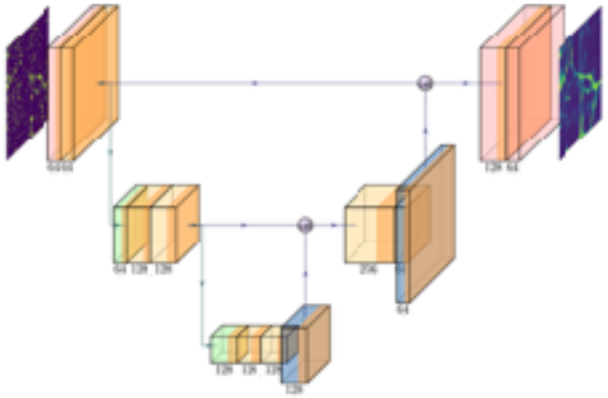
see Martin's talk for Euclid's similar situation

SimBIGGER — we need larger volumes, higher resolution

SimBIGGER — we need larger volumes, higher resolution

emulation

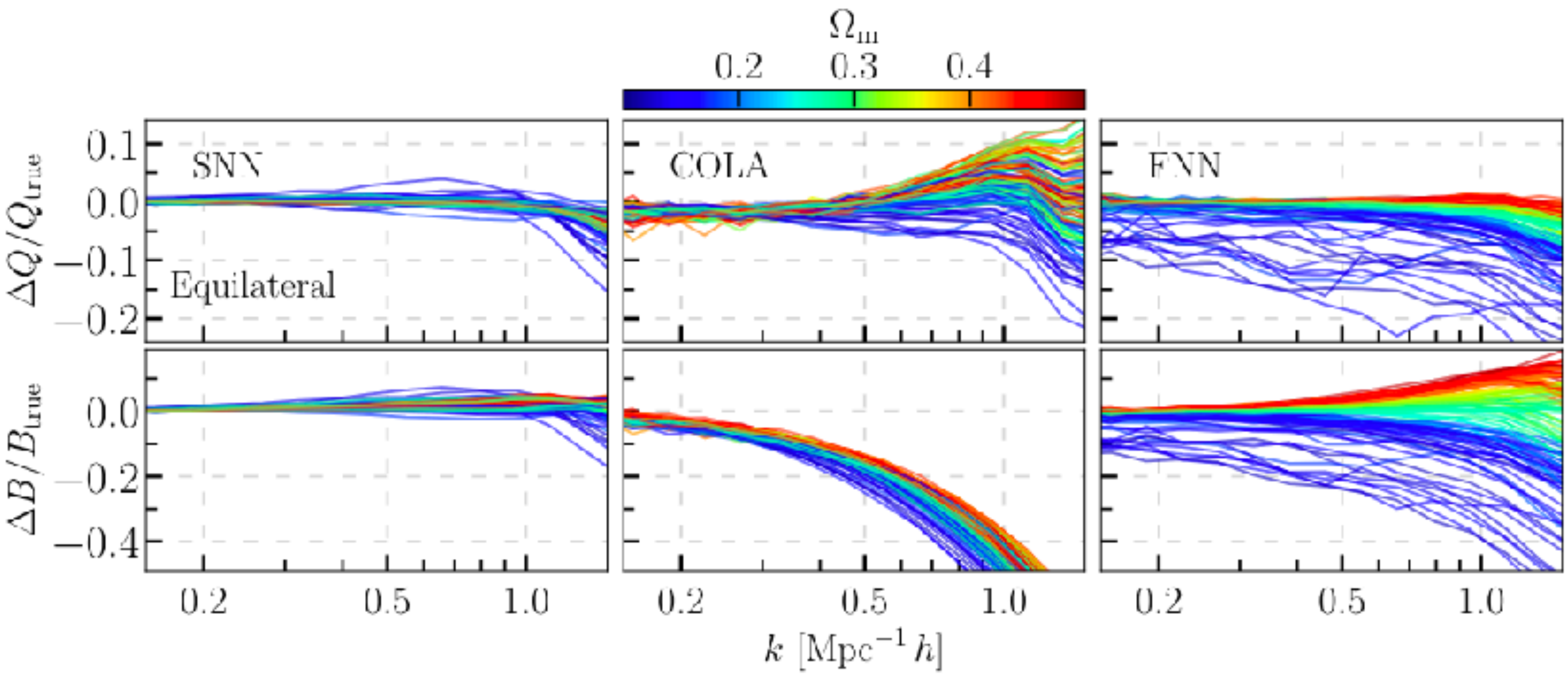




e.g. Alves de Oliveira et al.(2020), Li et al. (2021), Schaurecker et al.(2022),
***Jamieson** et al.(2022), Giusarma et al.(2023), Zhang et al.(2023), **Ariel's** emulator*

SimBIGGER — we need larger volumes, higher resolution

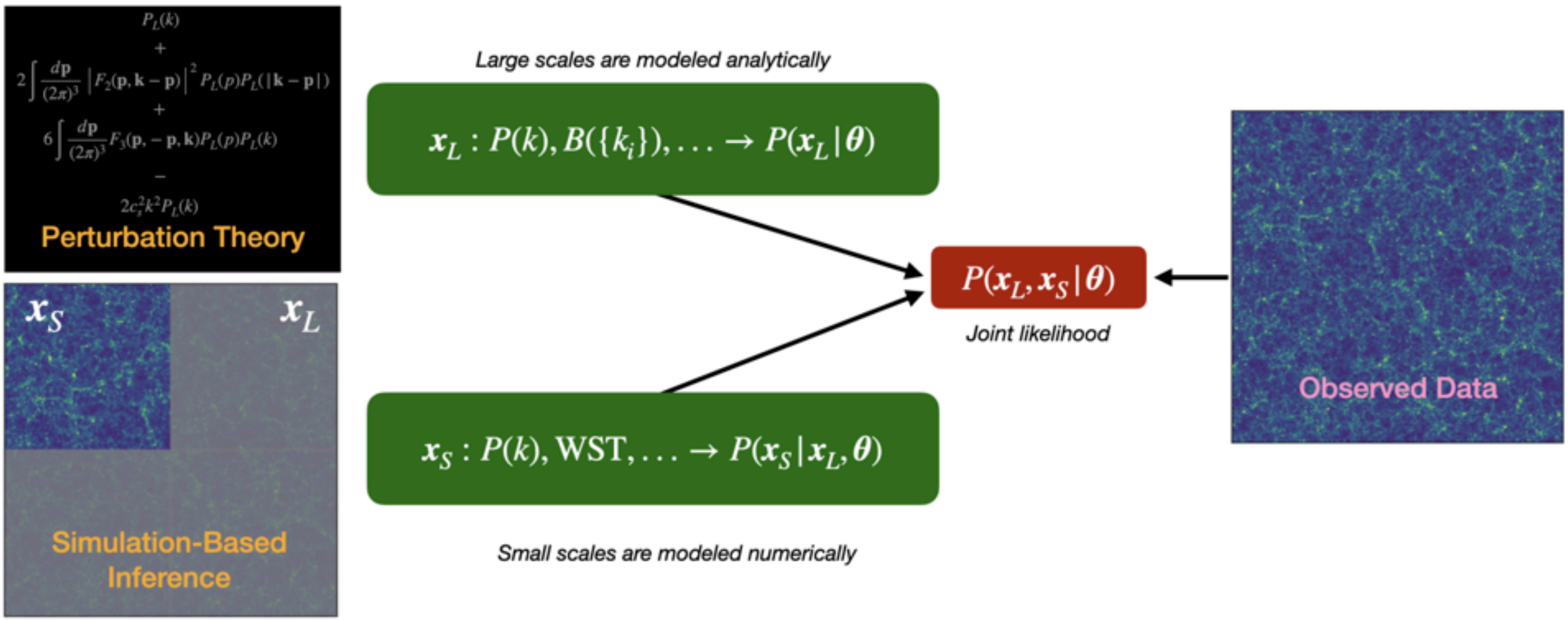
emulation



e.g. ***Jamieson** et al.(2022)*

SimBIGGER — we need larger volumes, higher resolution

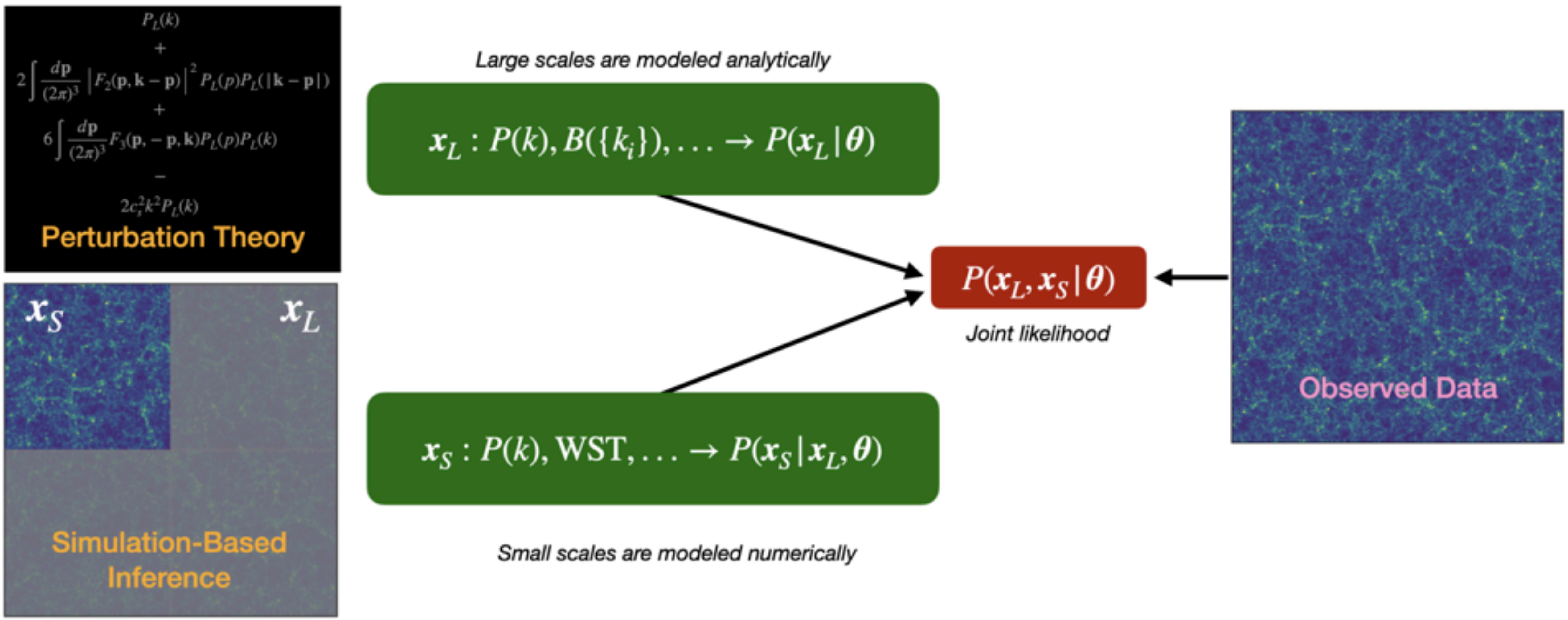
hybrid SBI



Modi & Philcox (2023)

SimBIGGER — we need larger volumes, higher resolution

hybrid SBI



$$p(\mathbf{X} \mid \theta) = p(\mathbf{X}_L \mid \theta)\; p(\mathbf{X}_S \mid \mathbf{X}_L, \theta)$$

Modi & Philcox (2023)

challenges for SBI: how can we trust SBI results?

challenges for SBI: how can we trust SBI results?

posterior validation

$q_\phi(\theta \mid \mathbf{X})$ is guaranteed to converge to $p(\theta \mid \mathbf{X})$ if

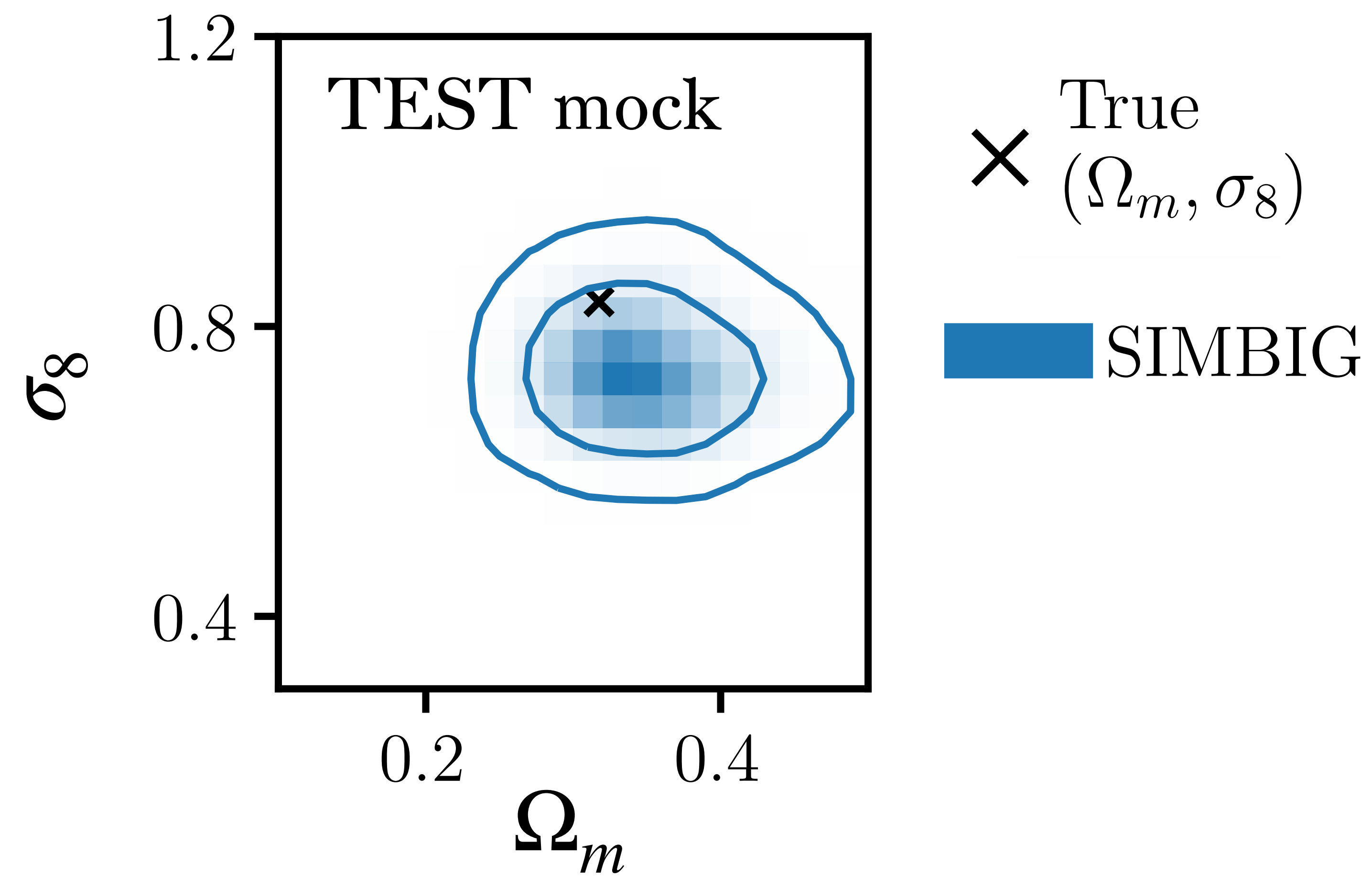
q_ϕ is flexibly expressive

$N \to \infty$ samples from $p(\mathbf{X}, \theta)$

successful optimization

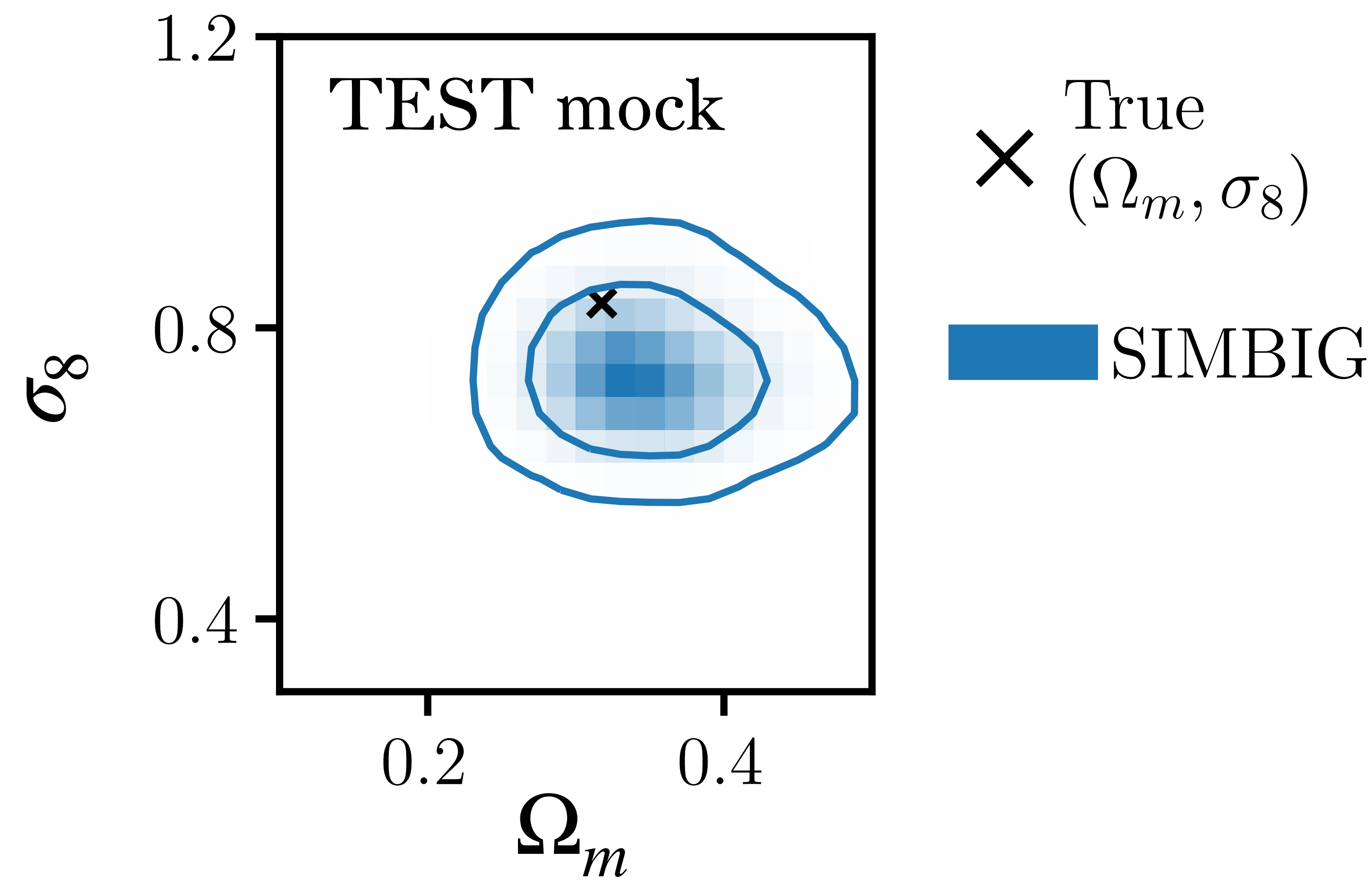
challenges for SBI: how can we trust SBI results?

posterior validation



challenges for SBI: how can we trust SBI results?

posterior validation

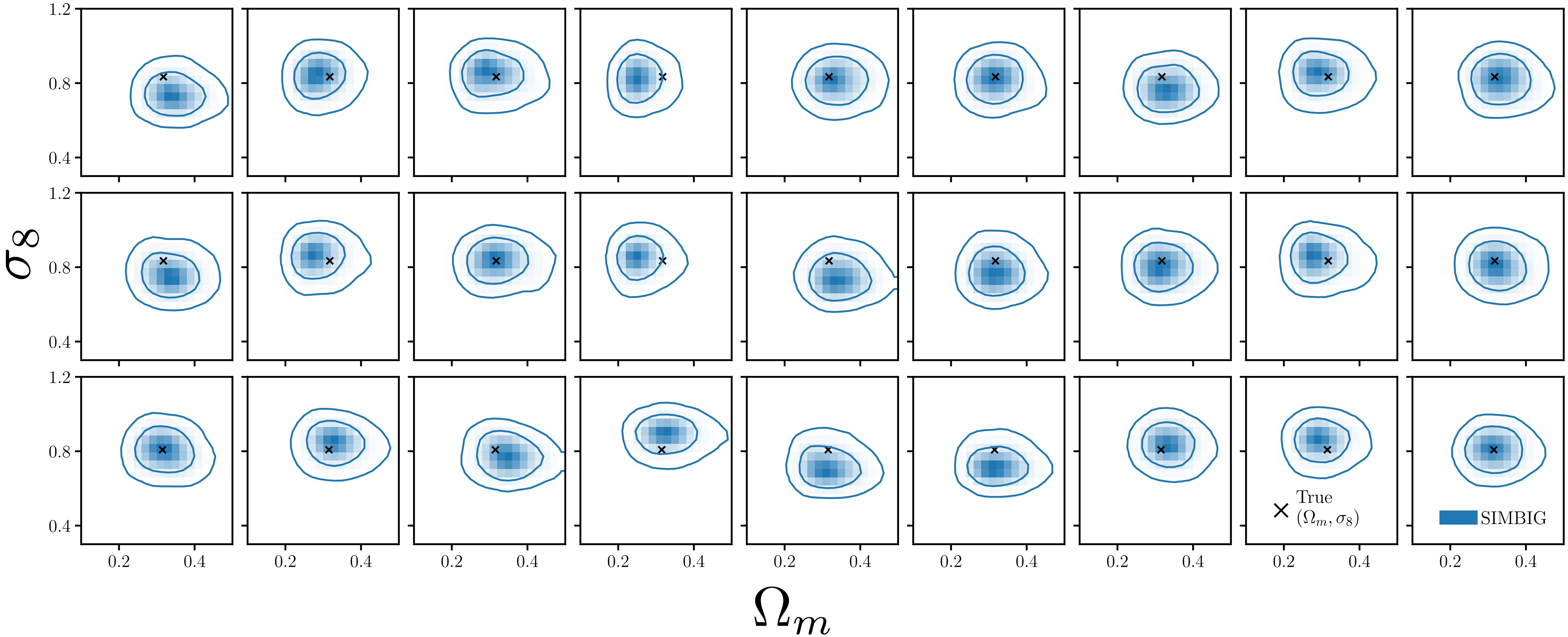


we need to validate *both* accuracy and precision

challenges for SBI: how can we trust SBI results?

posterior validation

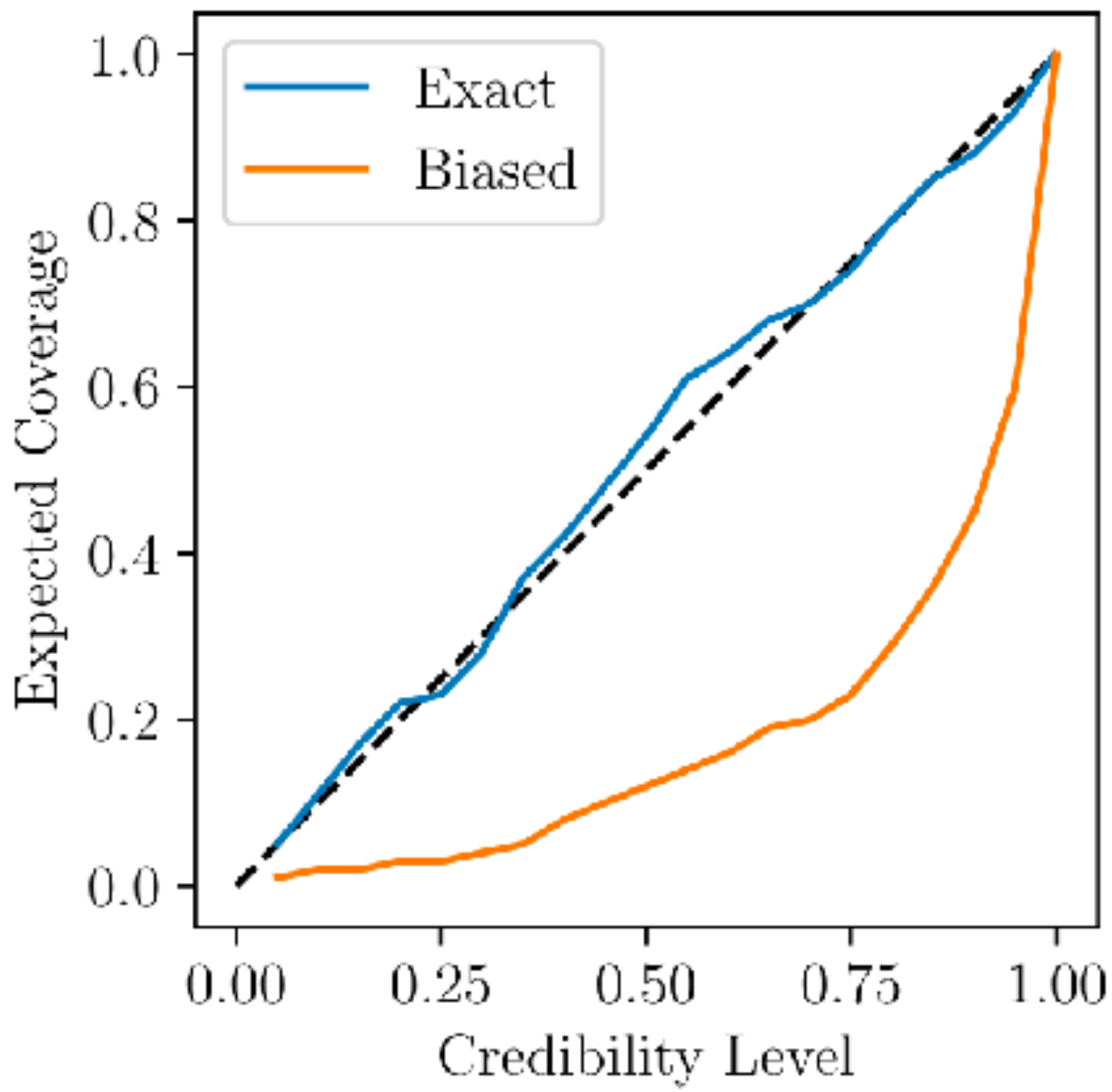
TEST mocks



a single validation mock is not sufficient!

challenges for SBI: how can we trust SBI results?

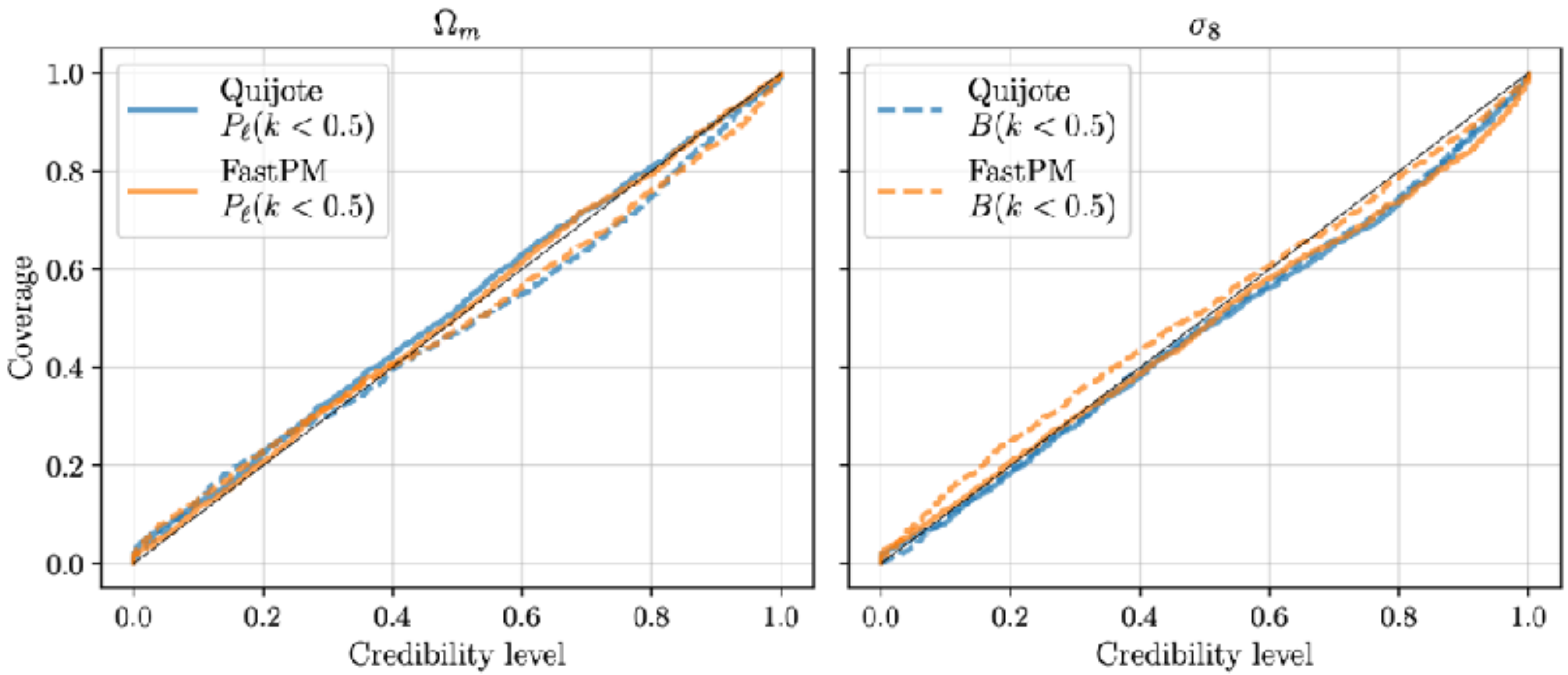
posterior validation



coverage tests — e.g. Lemos et al.(2023); see also Talts et al. (2020)

challenges for SBI: how can we trust SBI results?

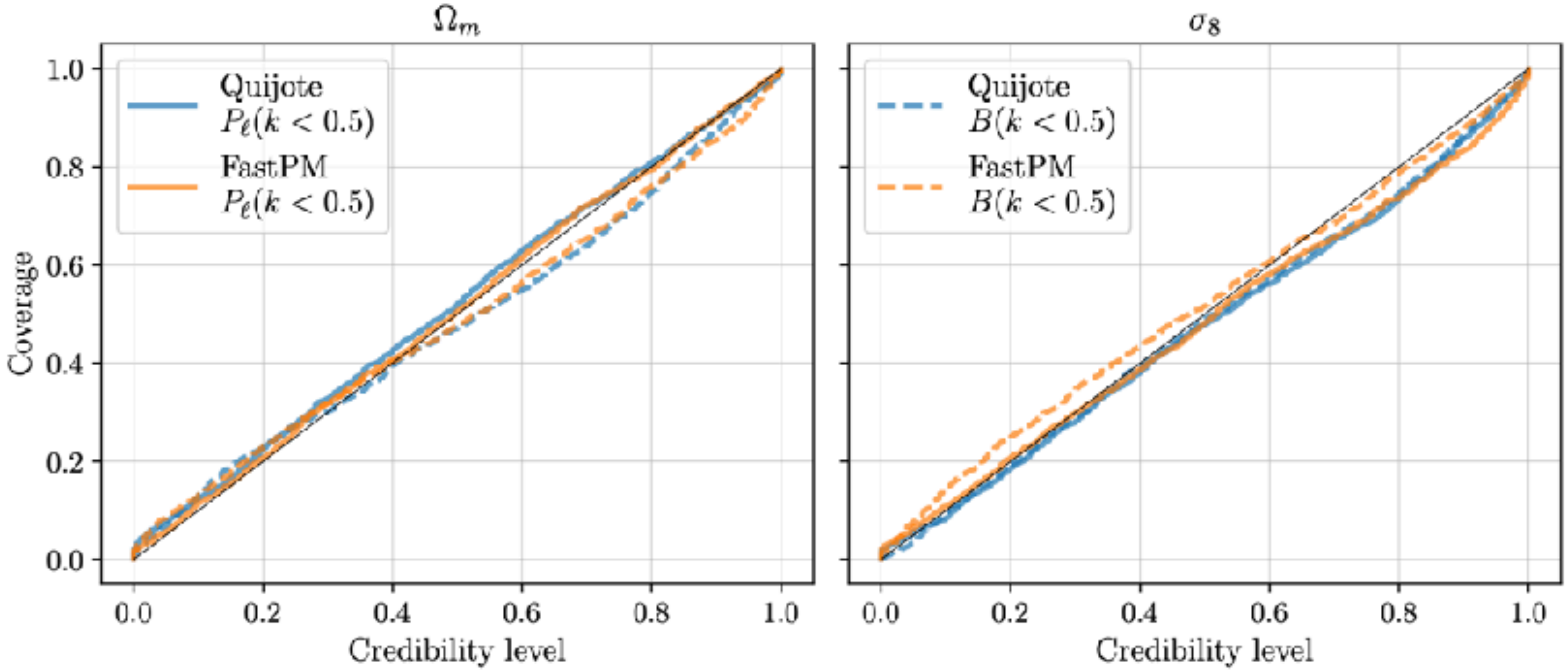
posterior validation



Modi…Hahn et al. (2023)

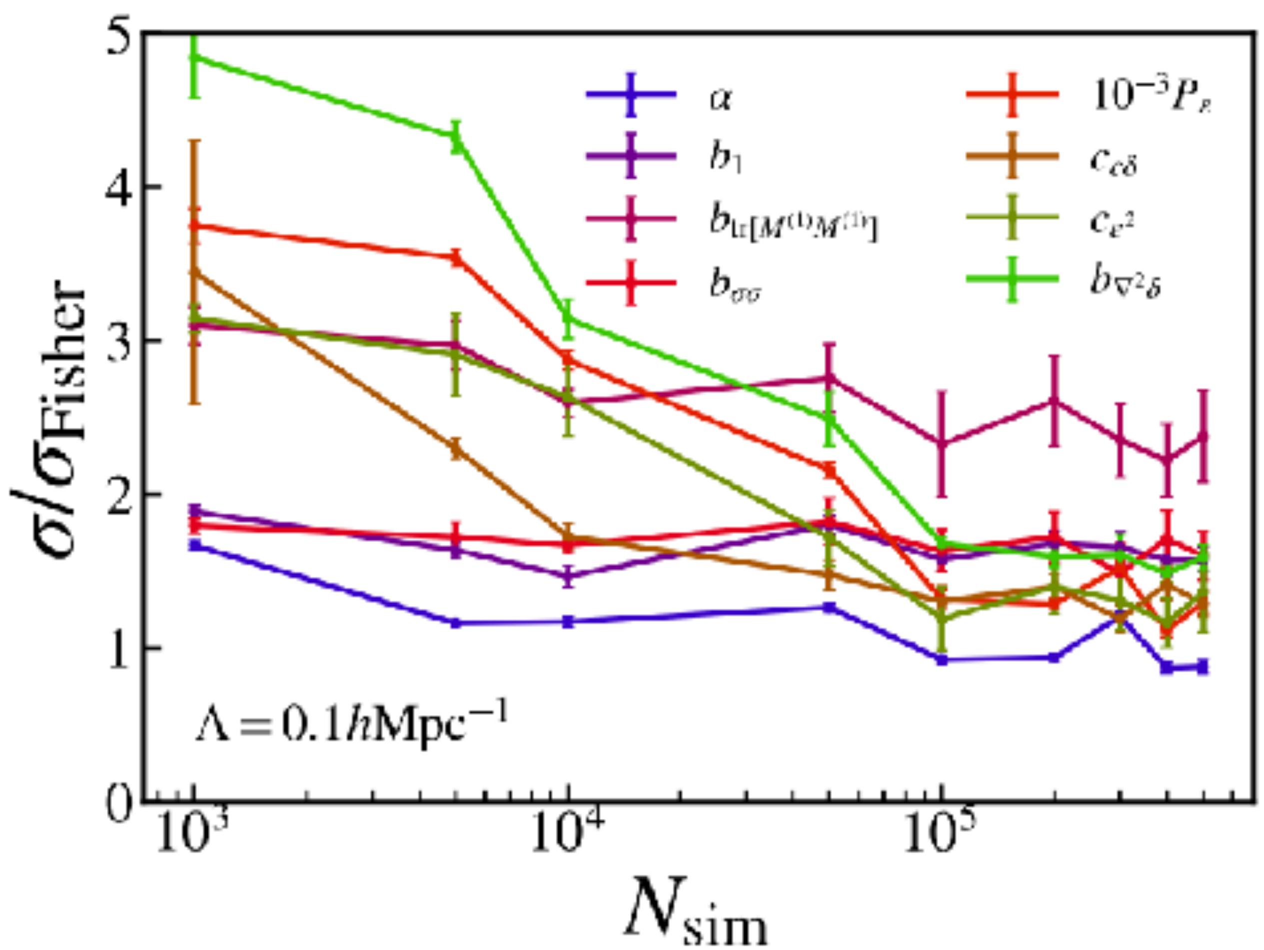
challenges for SBI: how can we trust SBI results?

posterior validation



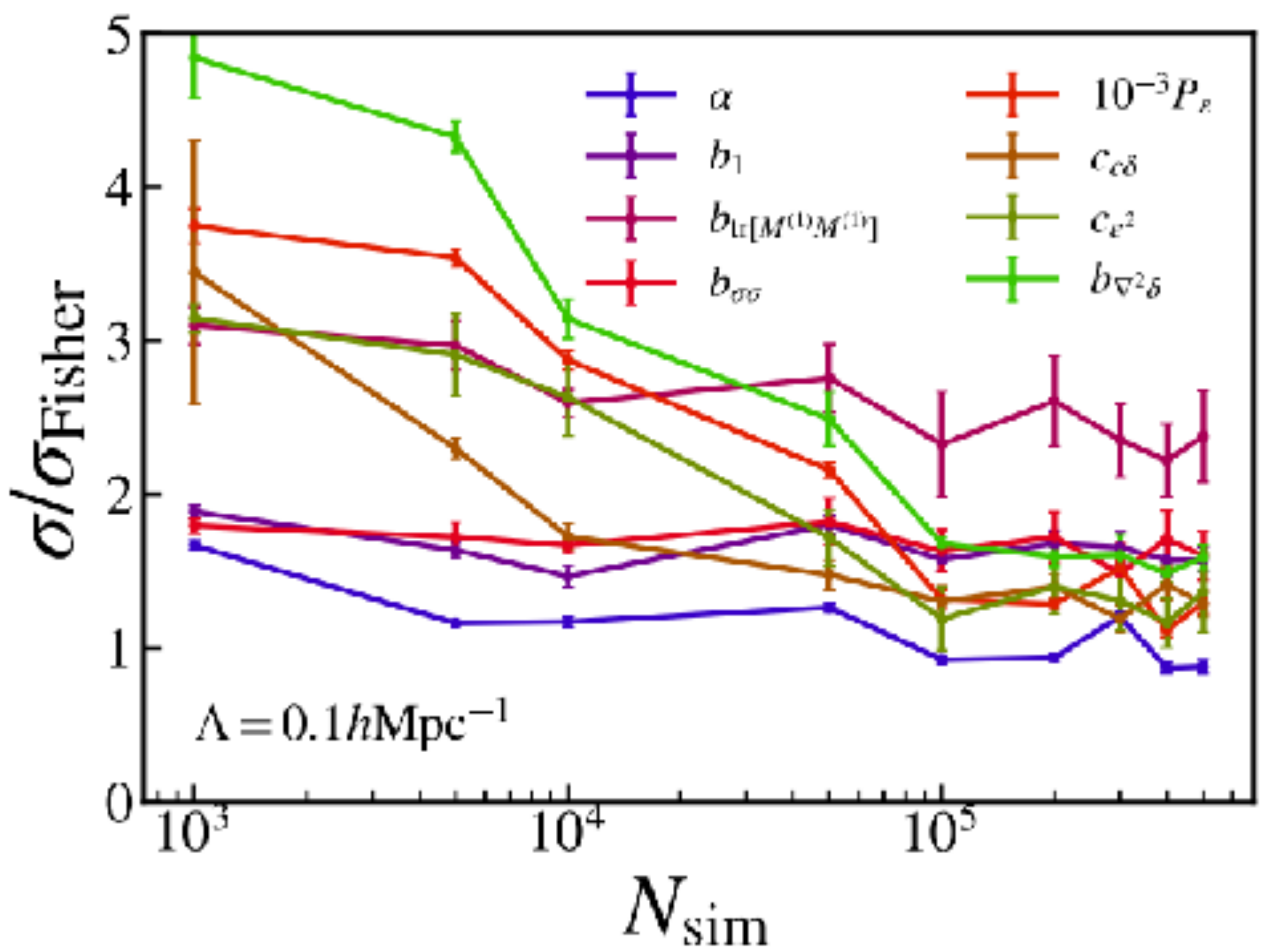
SBI for galaxy clustering is possible with *just* ~2,000 simulations

caution: “good” coverage doesn’t guarantee “optimal”



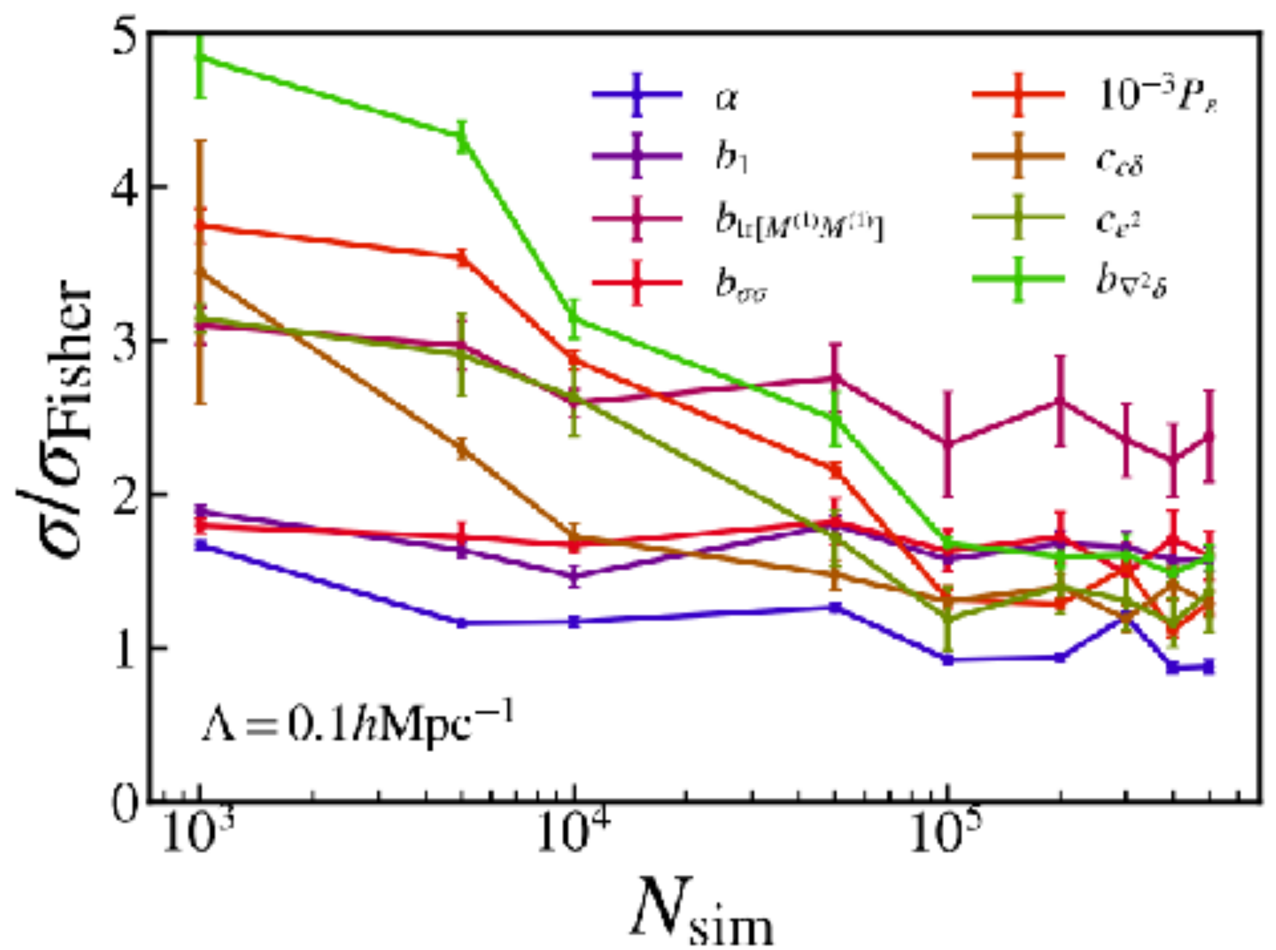
Tucci & Schmidt (2024)

caution: “good” coverage doesn’t guarantee “optimal”



“Il meglio è l'inimico del bene”
“perfect is the enemy of good”

caution: “good” coverage doesn’t guarantee “optimal”

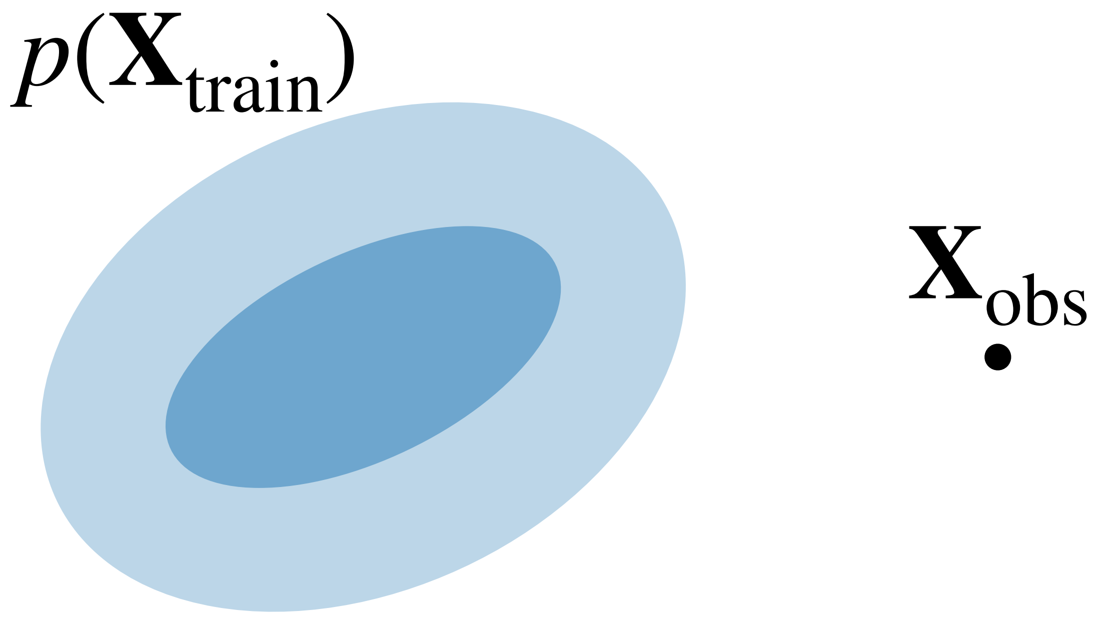


“Il meglio è l'inimico del bene”
“perfect is the enemy of good”
Friday discussion: ***Field-level vs Summaries***

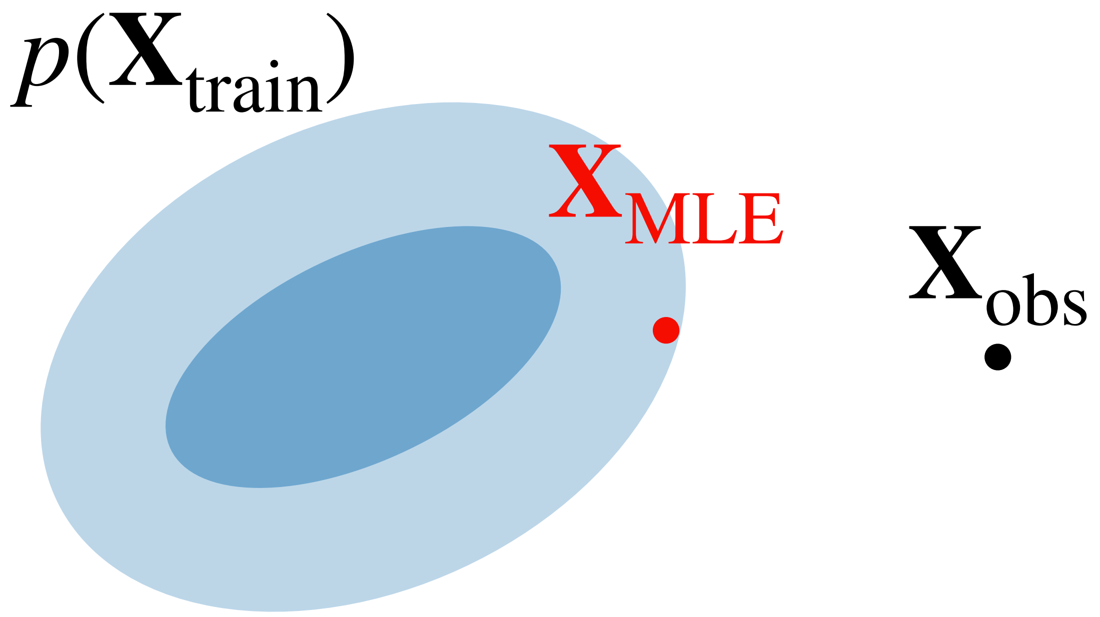
challenges for SBI: how can we trust SBI results?

model misspecification

model misspecification is a *concern for everyone* not just SBI



model misspecification is a *concern for everyone* not just SBI



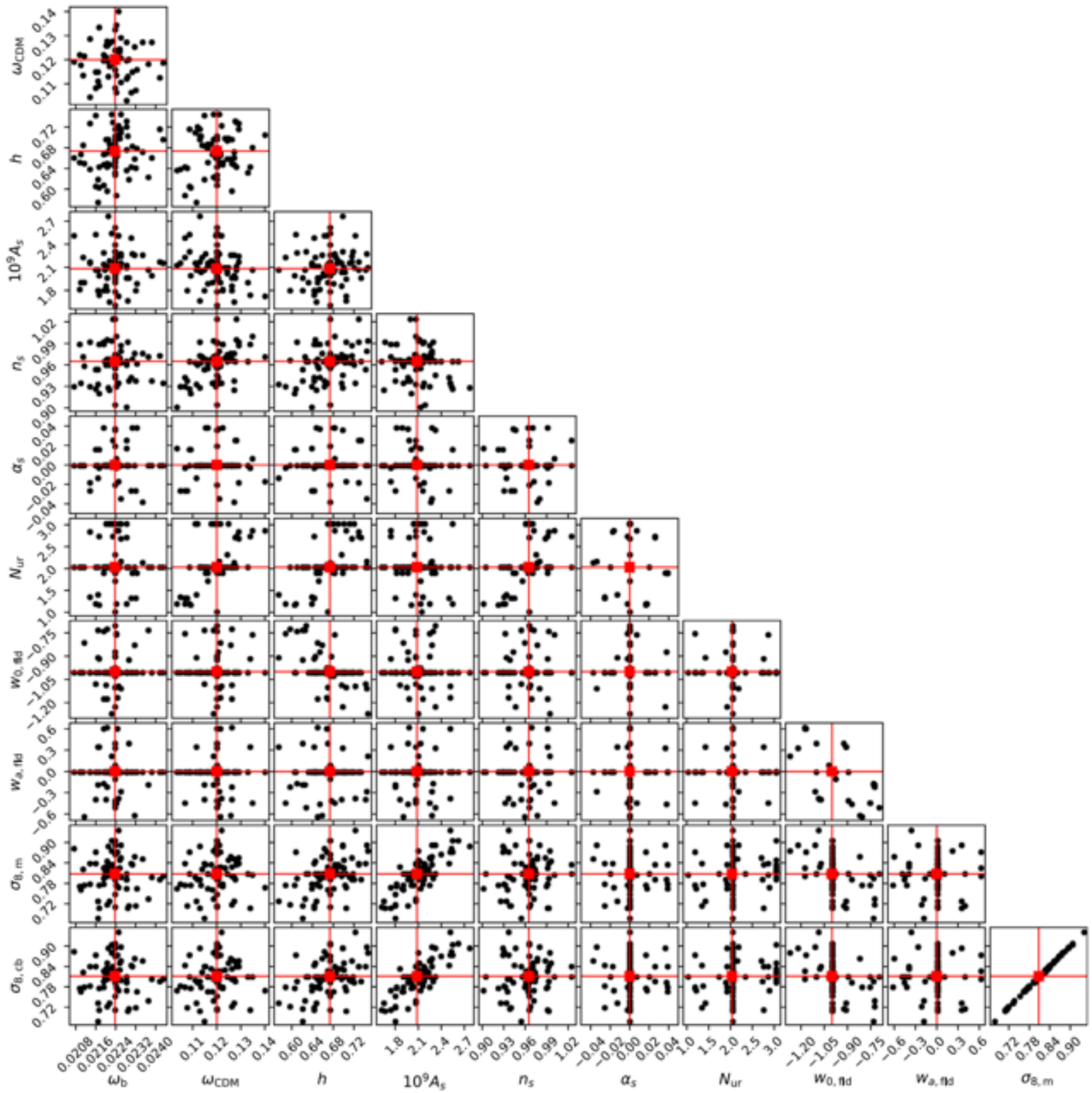
model misspecification concerns for “standard” analyses

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

**lets say the true likelihood is Gaussian*

model misspecification concerns for “standard” analyses

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

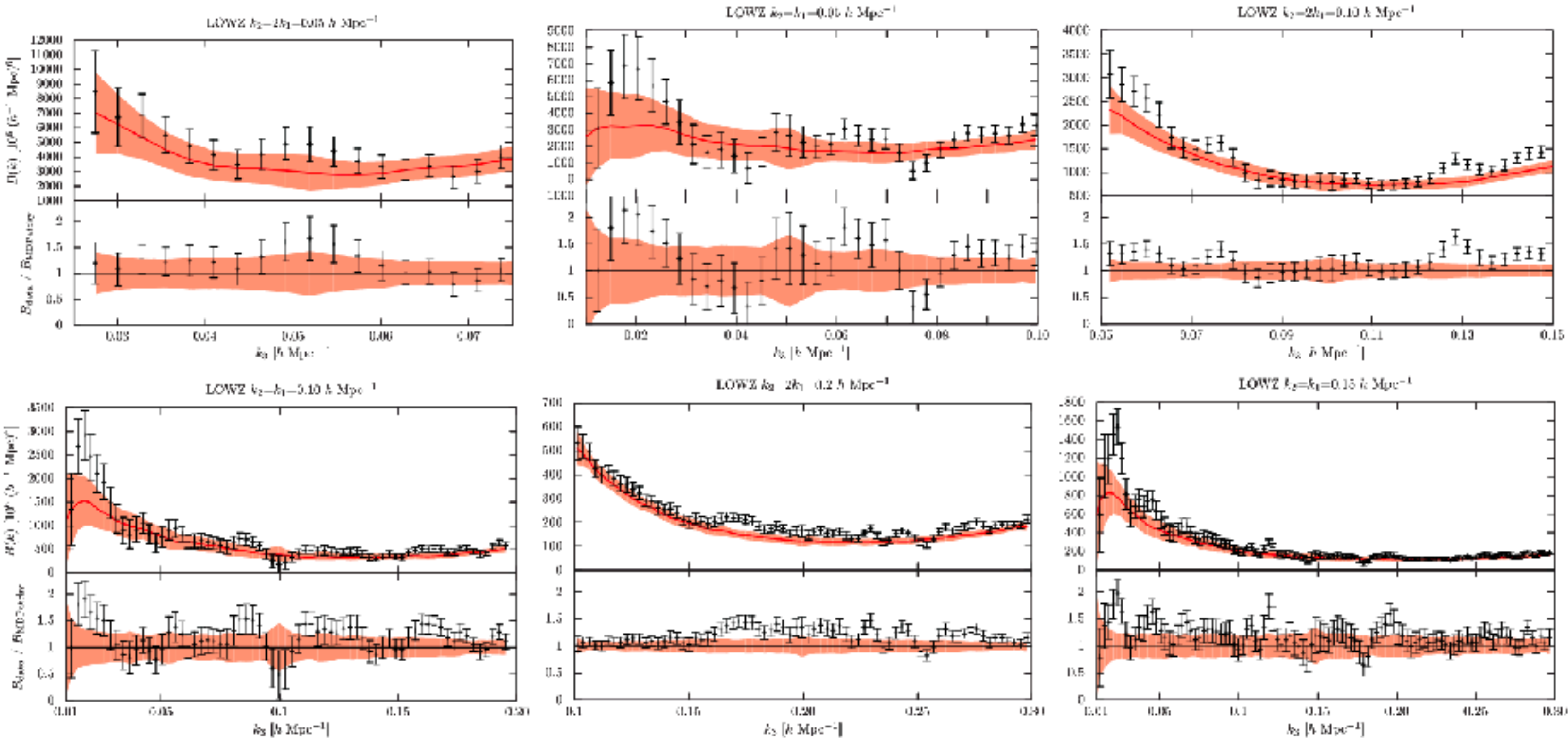


Maksimova et al.(2021); ***Hector's comment***

model misspecification concerns for “standard” analyses

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

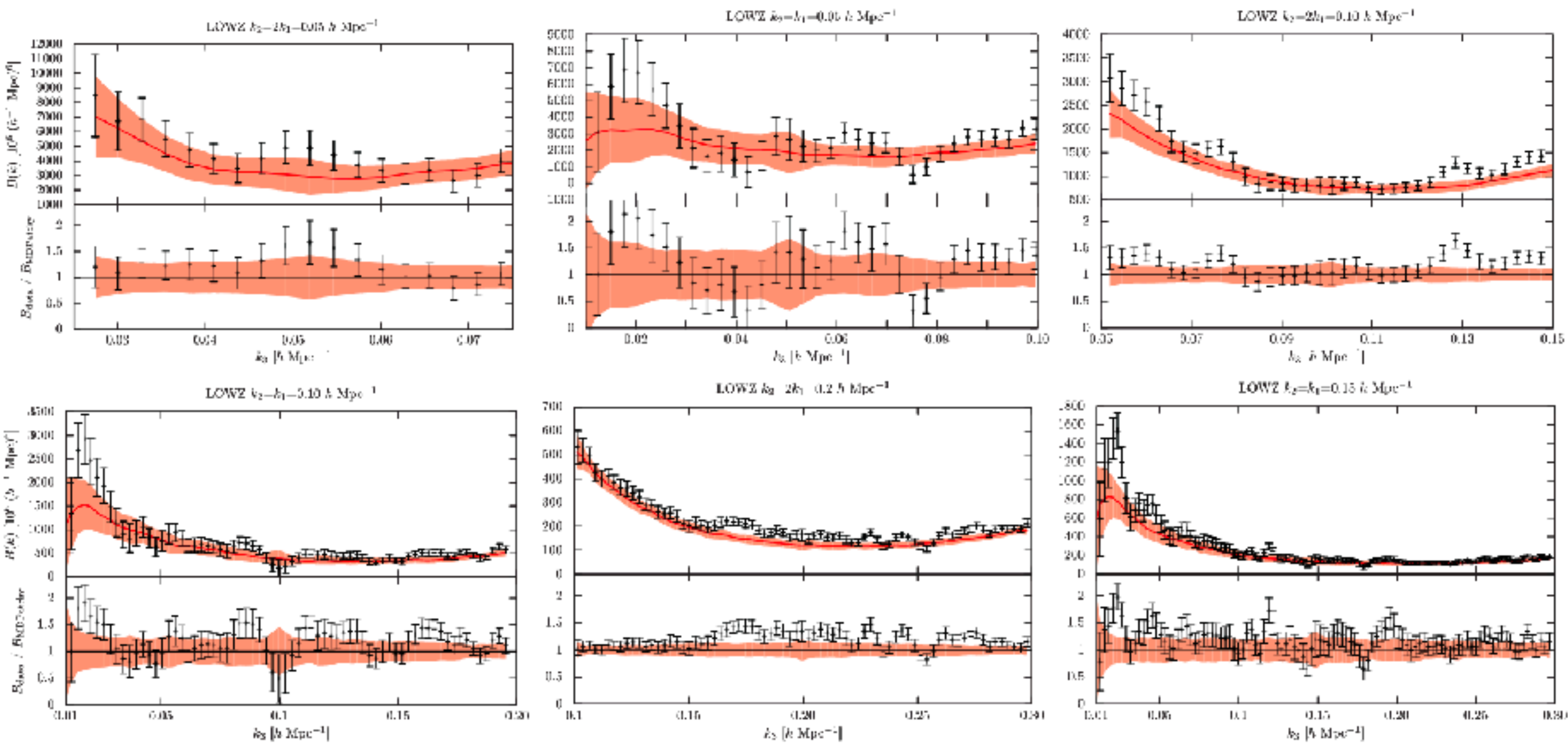
model misspecification concerns for “standard” analyses



100 PATCHY mocks **BOSS**

bispectrum; Kitaura et al. (2016)

model misspecification concerns for “standard” analyses — can we use approximate mocks *designed for 2pt analyses* for beyond 2pt?



model misspecification concerns for “standard” analyses

$$\mathbf{X}' \sim F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$$

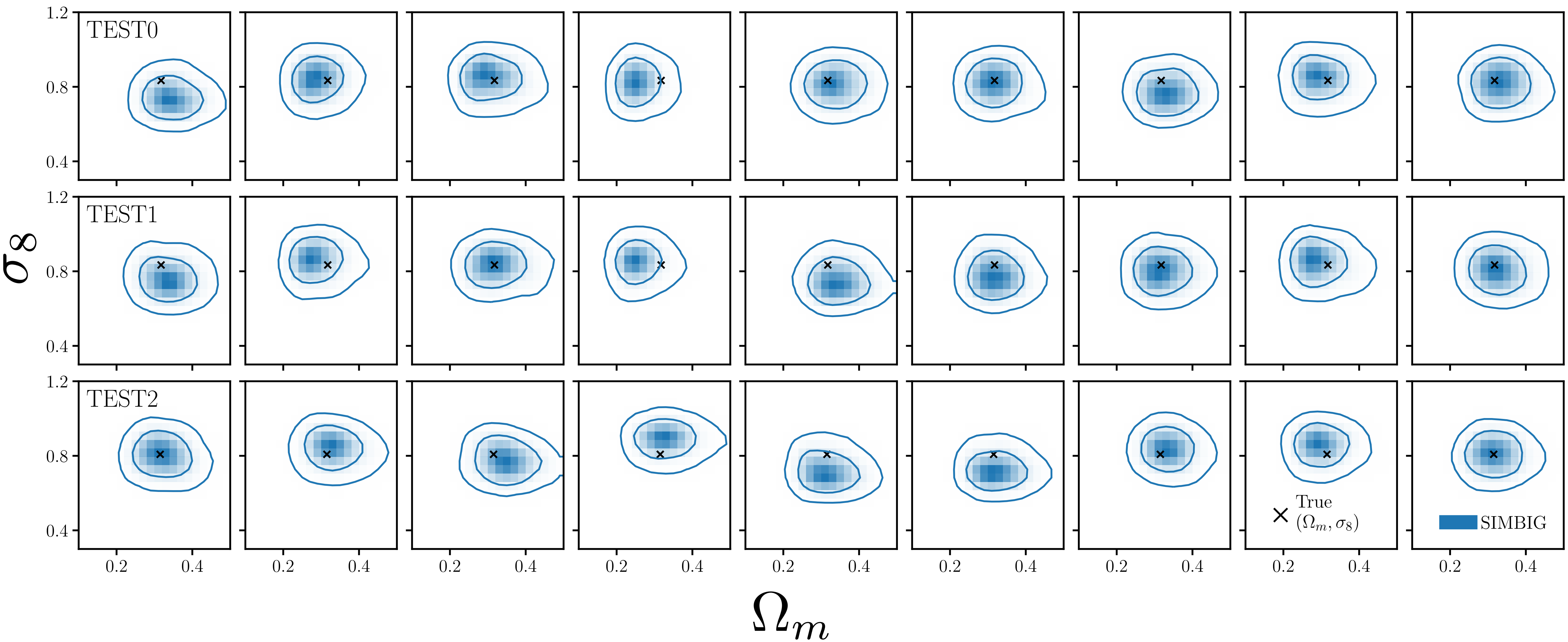
would you trust an SBI analysis with forward model $F(\theta') = \mathcal{N}(m(\theta'), \mathbf{C})$*?*

tackling model misspecification with *cross-validation*

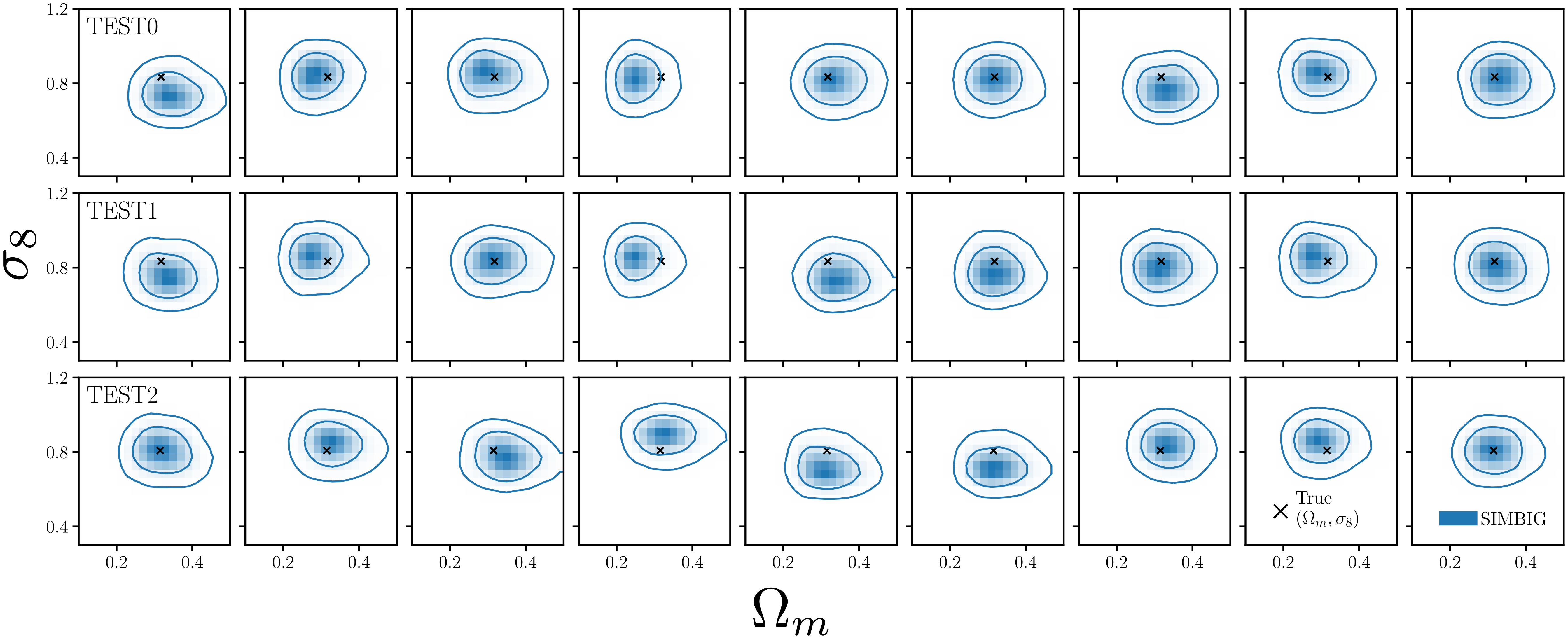
tackling model misspecification with *cross-validation* — SimBIG: 2,000 simulations with three set of different forward models

	N-body	*halo finder*	*HOD*	N_{sim}
TEST0	Quijote	Rockstar	fiducial	500
TEST1	Quijote	FoF	*Zheng+(2007)*	500
TEST2	AbacusSummit	CompaSO	fiducial	1000

tackling model misspecification with *cross-validation* — SIMBIG: 2,000 simulations with three set of different forward models

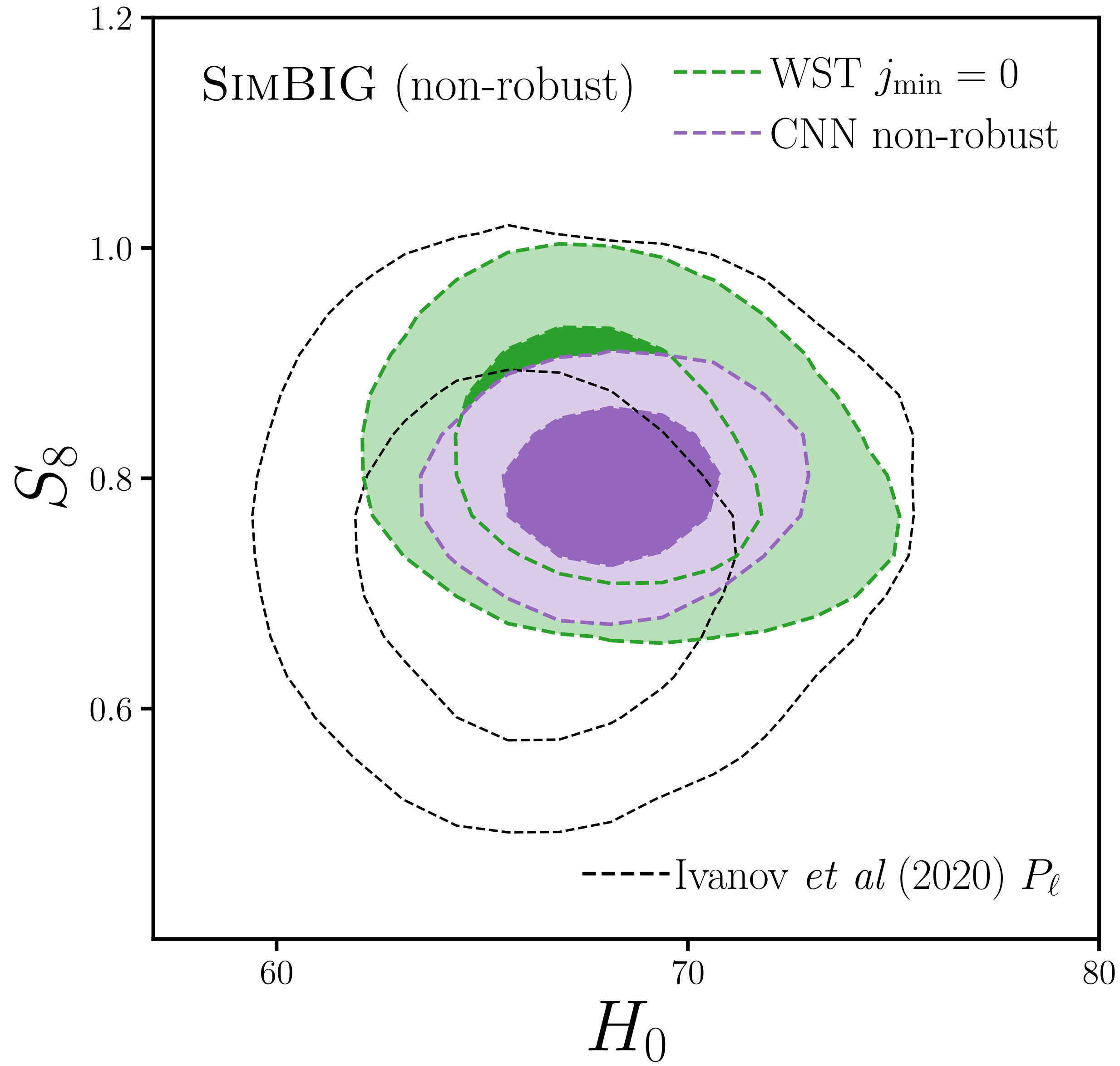


tackling model misspecification with *cross-validation* — SimBIG: 2,000 simulations with three set of different forward models



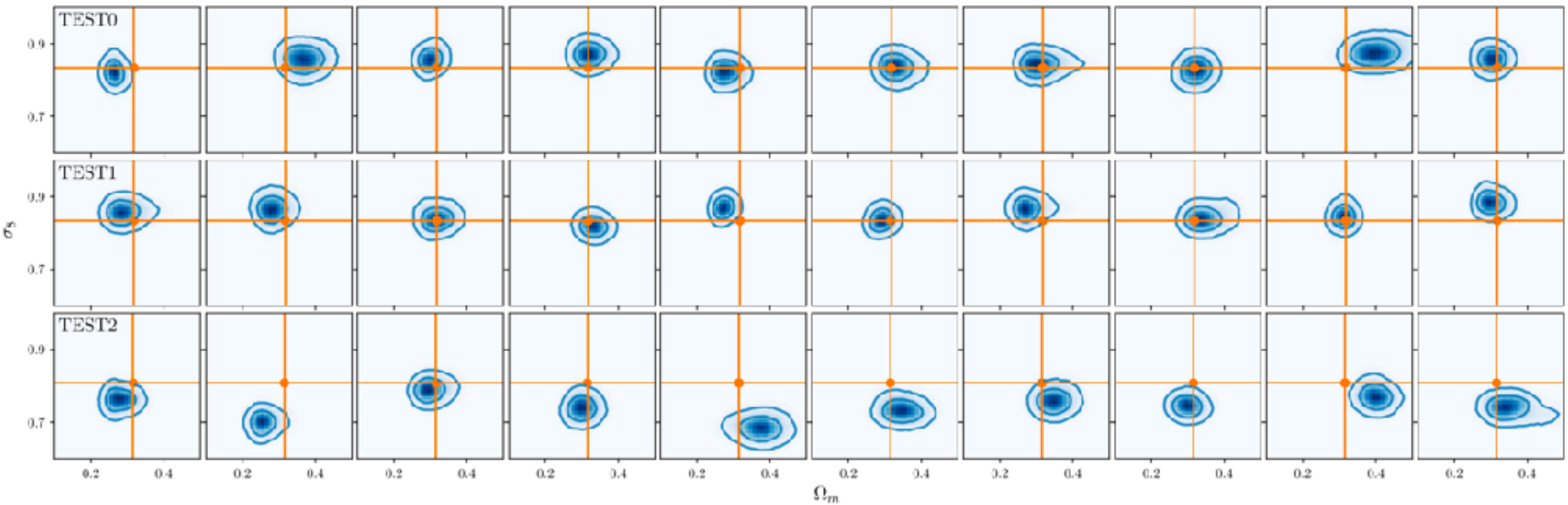
cosmological analyses are only as good as their validation

cautionary tale on "optimal observables" — consistency ≠ robustness

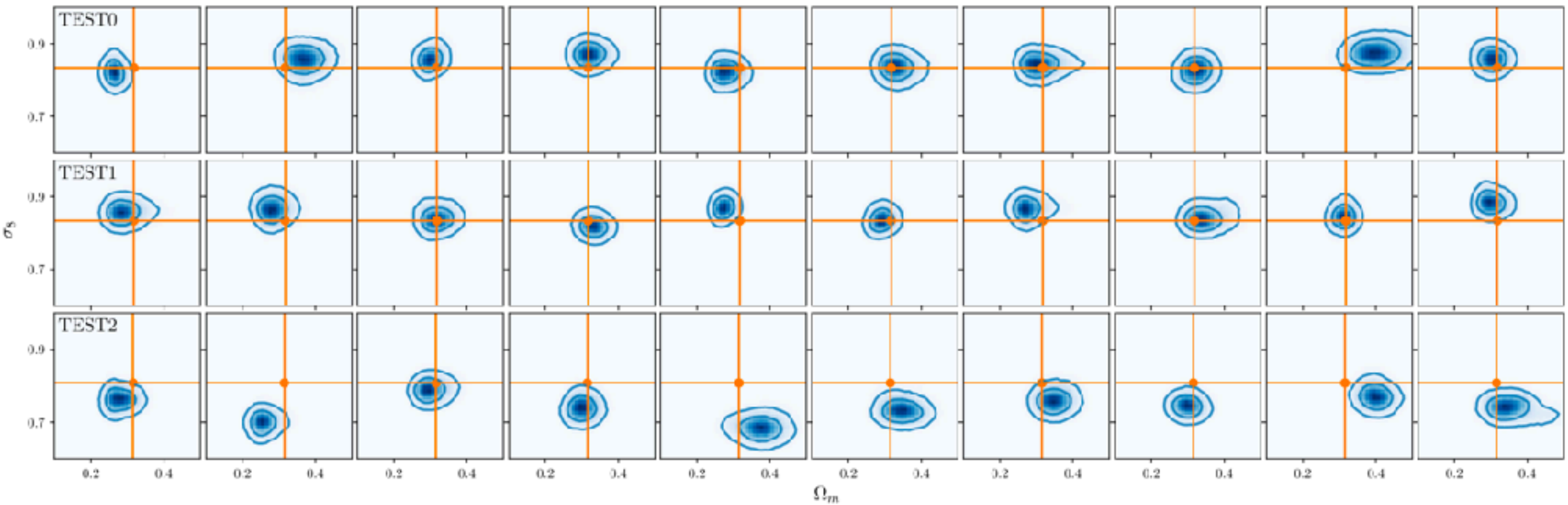


NON-ROBUST *SimBIG posteriors*

cautionary tale on "optimal observables" — consistency ≠ robustness



wavelet validation (Régaldo-Saint Blanchard, Hahn et al. 2024) *cautionary tale* on “optimal observables” — consistency ≠ robustness



our goal should be optimal **robust** observables

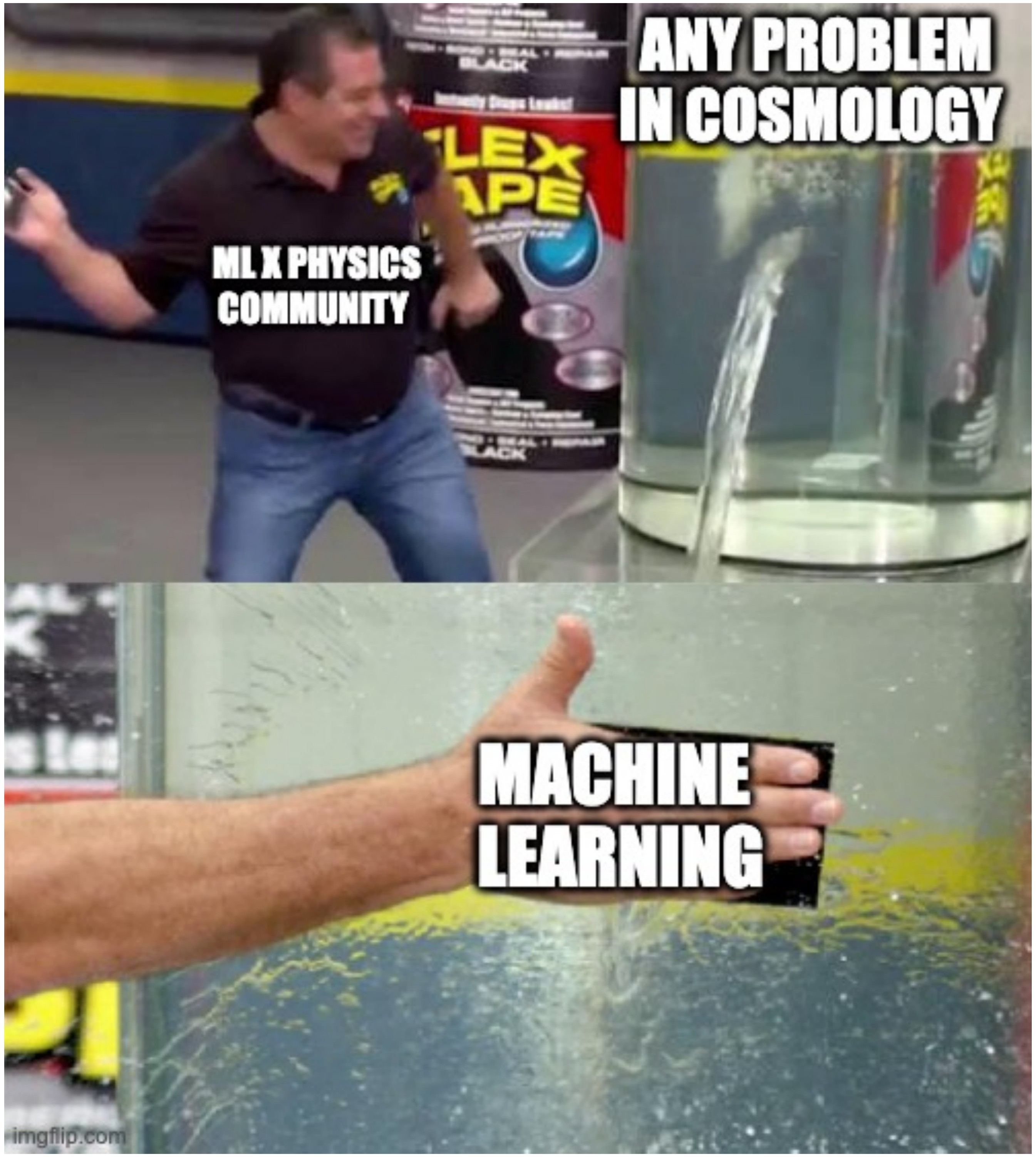
wavelet validation (Régaldo-Saint Blanchard, Hahn et al. 2024)

challenges for SBI: how can we trust SBI results?

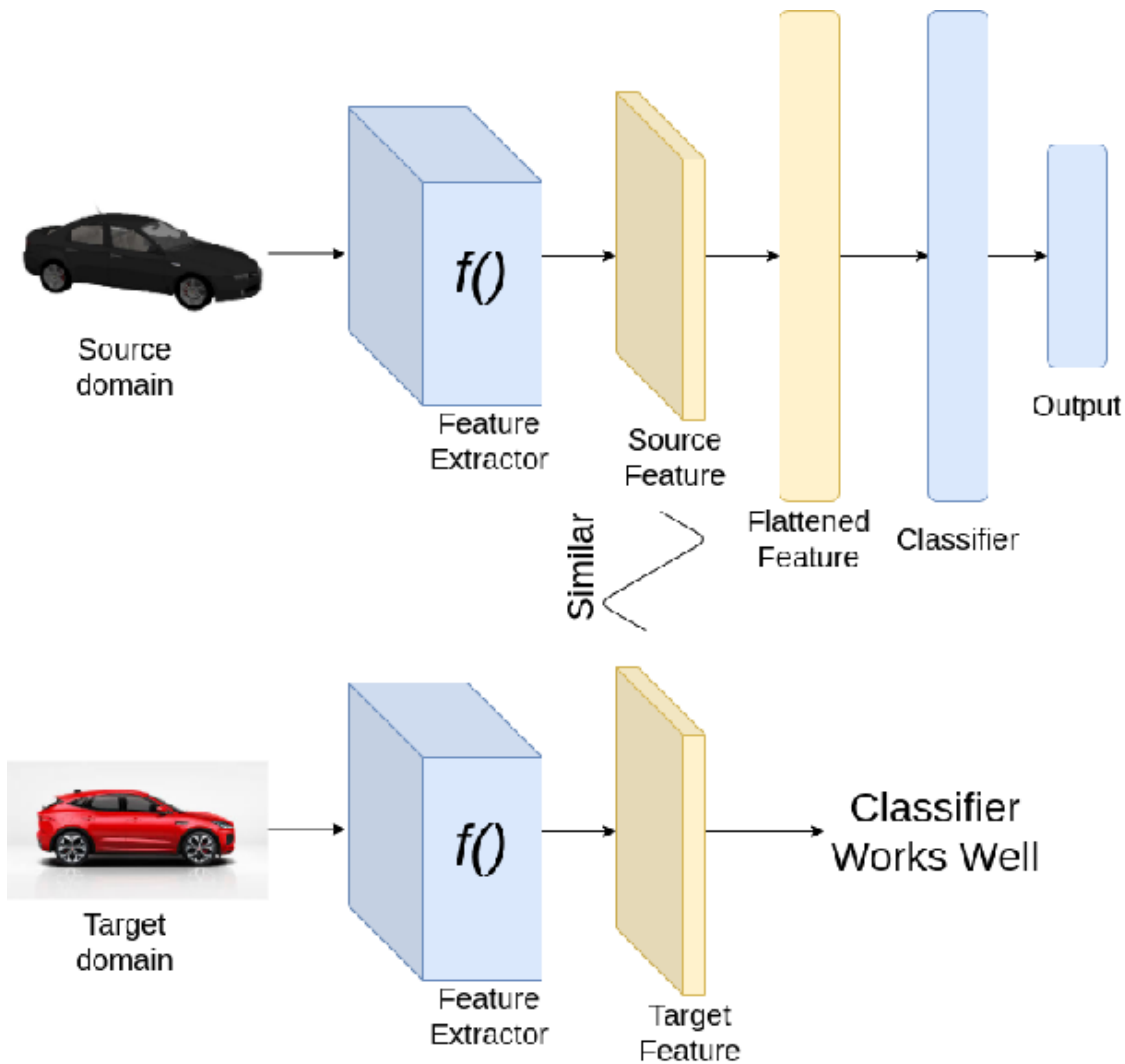
model misspecification

challenges for SBI: how can we trust SBI results?

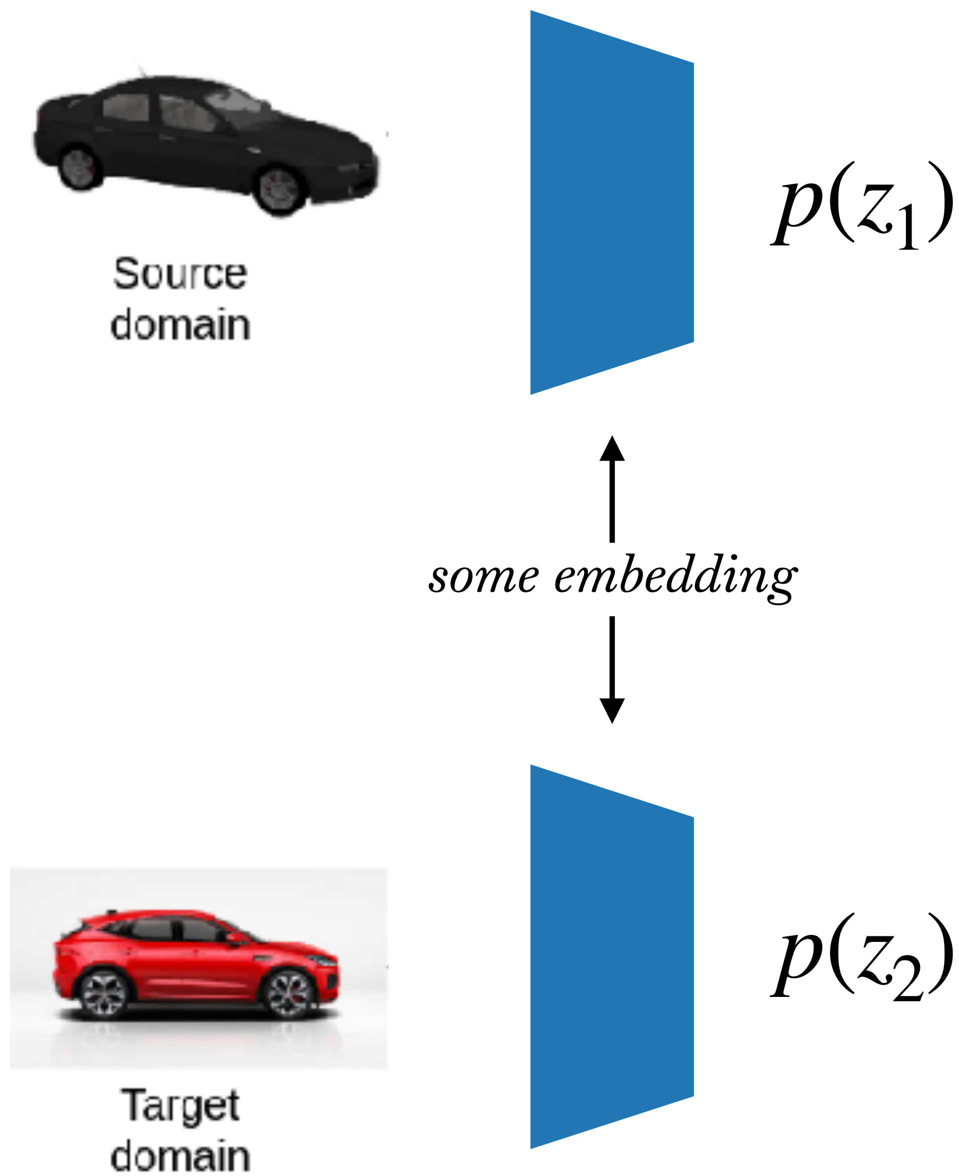
model misspecification



can **domain adaptation** help us with model misspecification?



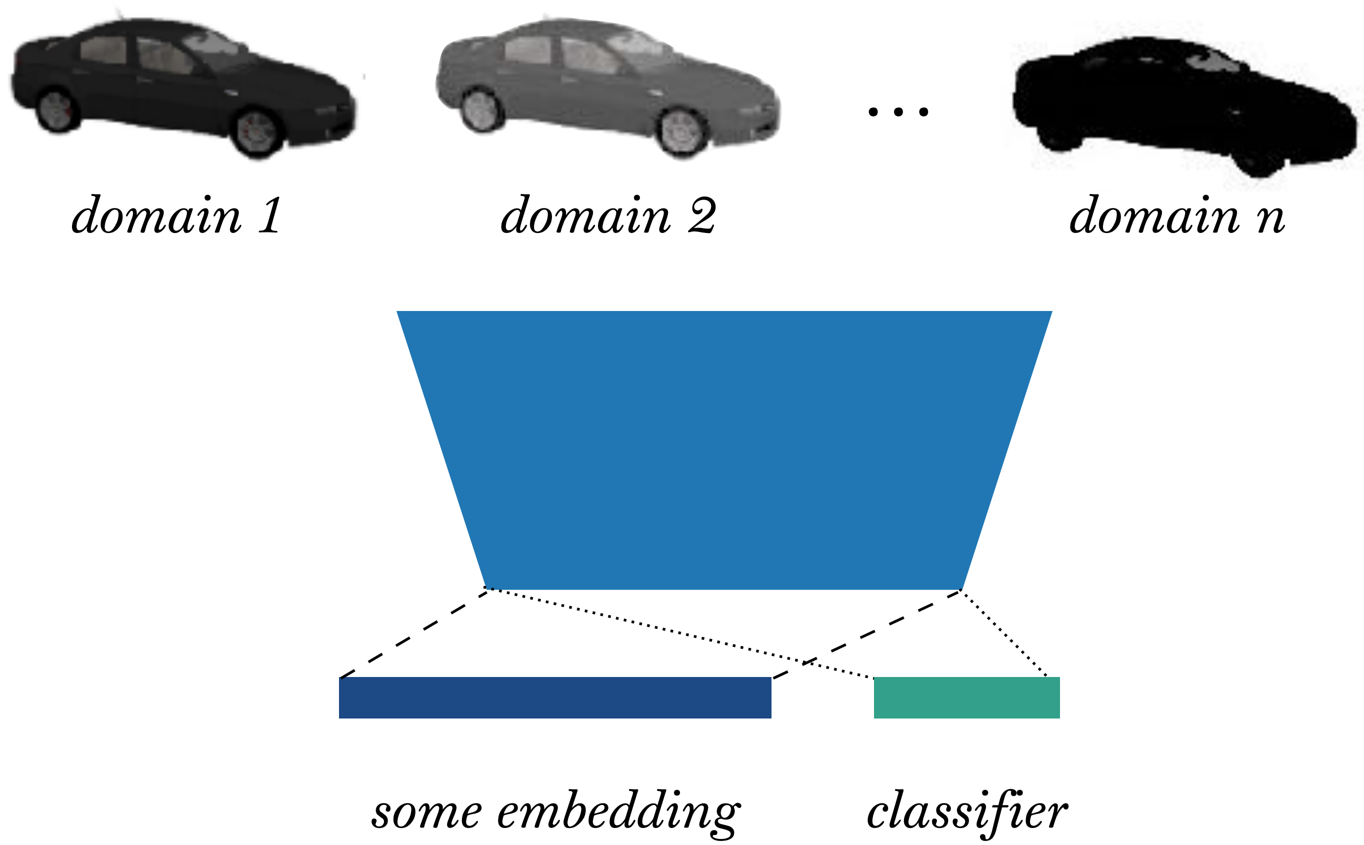
can **domain adaptation** help us with model misspecification?



loss penalizes difference between $p(z_1)$ and $p(z_2)$ (*e.g. MMD)*

e.g. Roncoli et al.(2024)

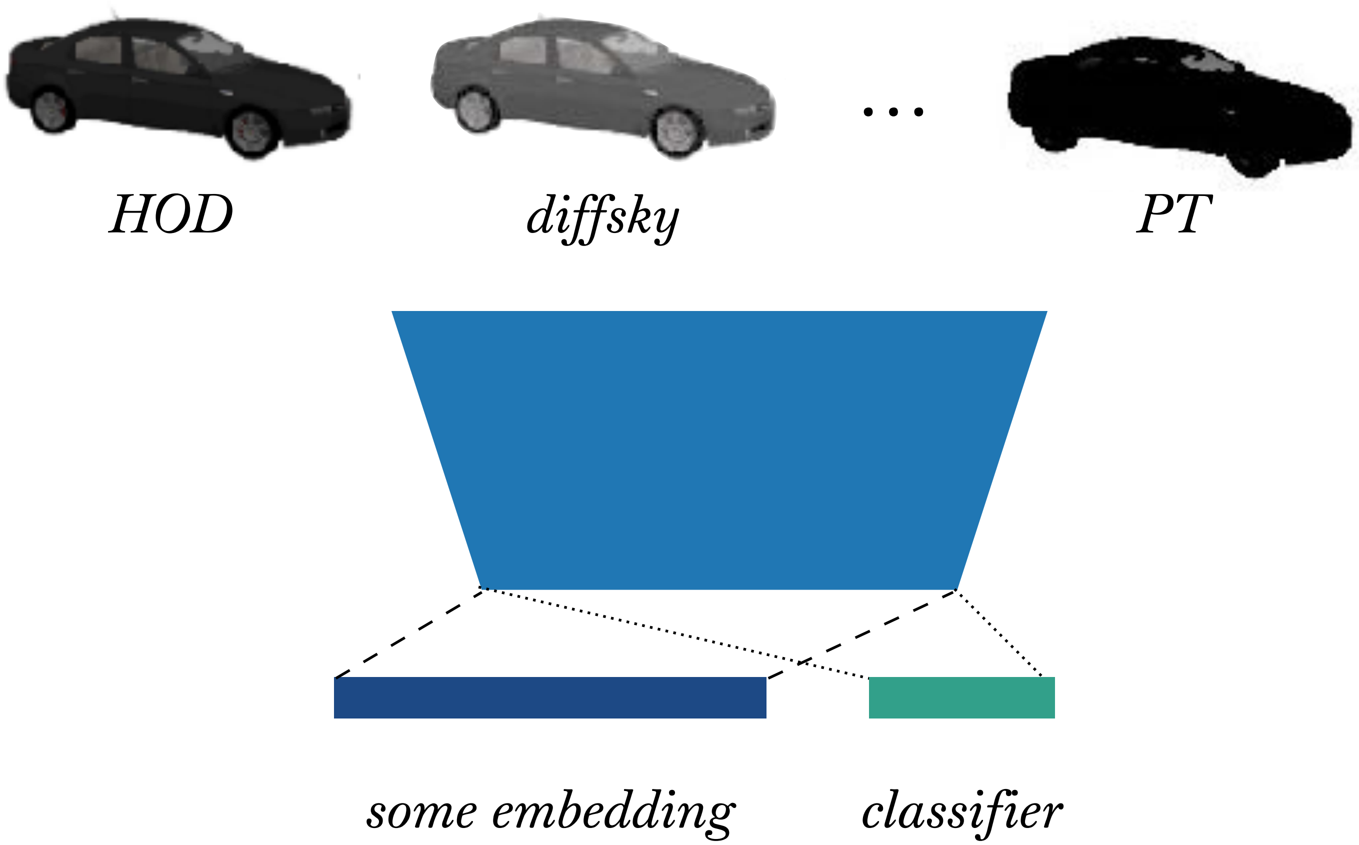
can **domain adaptation** help us with model misspecification?



loss = MSE + classifier penalty

Domain Adversarial Neural Networks (e.g. Ćiprijanović et al. 2020)

can **domain adaptation** help us with model misspecification?



loss = MSE + classifier penalty

Domain Adversarial Neural Networks (e.g. Ćiprijanović et al. 2020) we are all SBI! — just with different assumptions

state-of-the-art SBI provide opportunities to extract information in higher-order and non-linear galaxy clustering — *e.g.* SimBIG: $1.9\times$ improvement in S_8

still many challenges:

scaling up to next-generation surveys (*emulation, hybrid SBI?*)

posterior validation — accuracy and precision (*coverage plots?*)

model misspecification (*cross-validation, domain adaptation?)*

ChangHoon Hahn

changhoon.hahn@princeton.edu

changhoonhahn.github.io