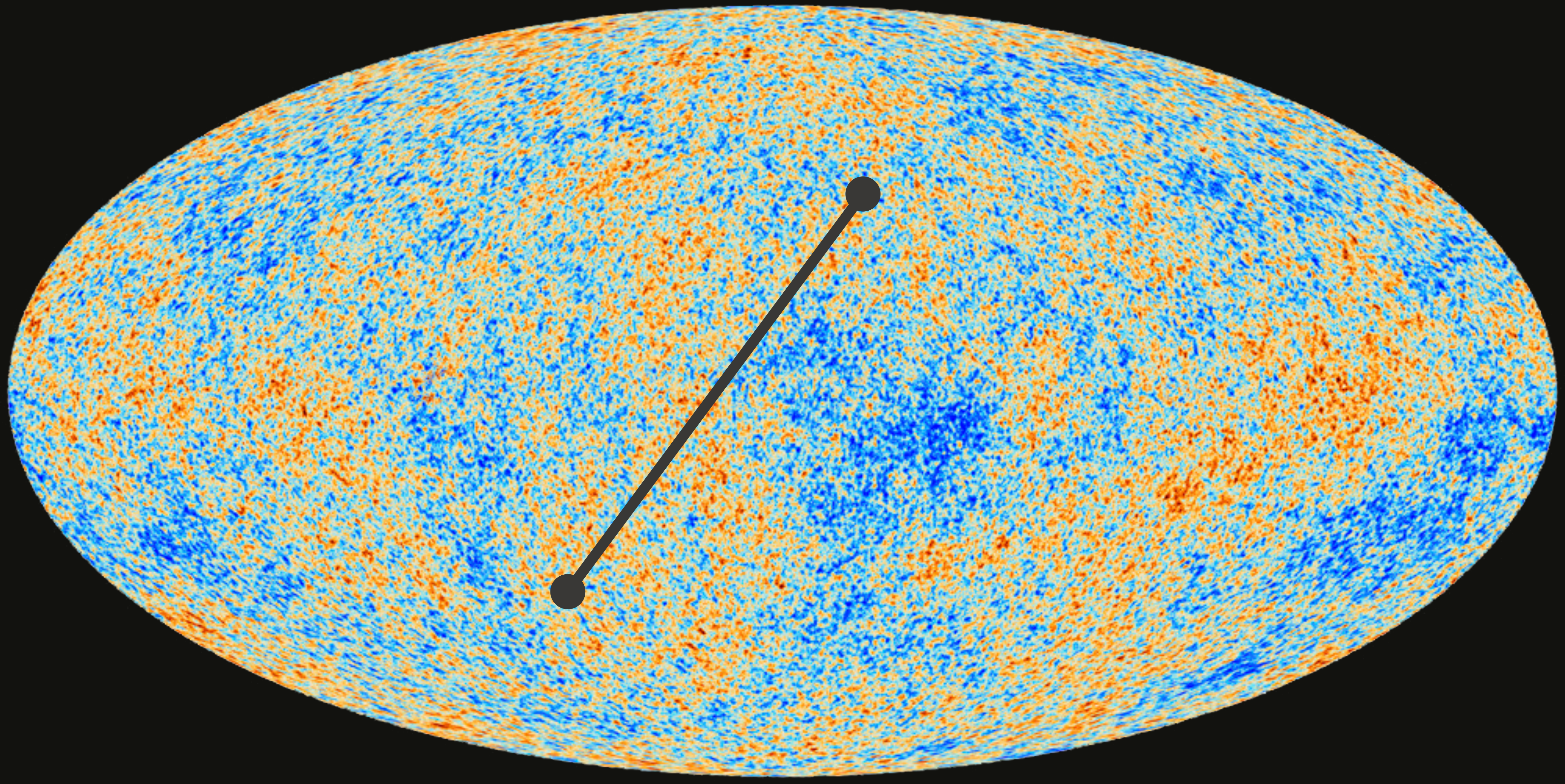
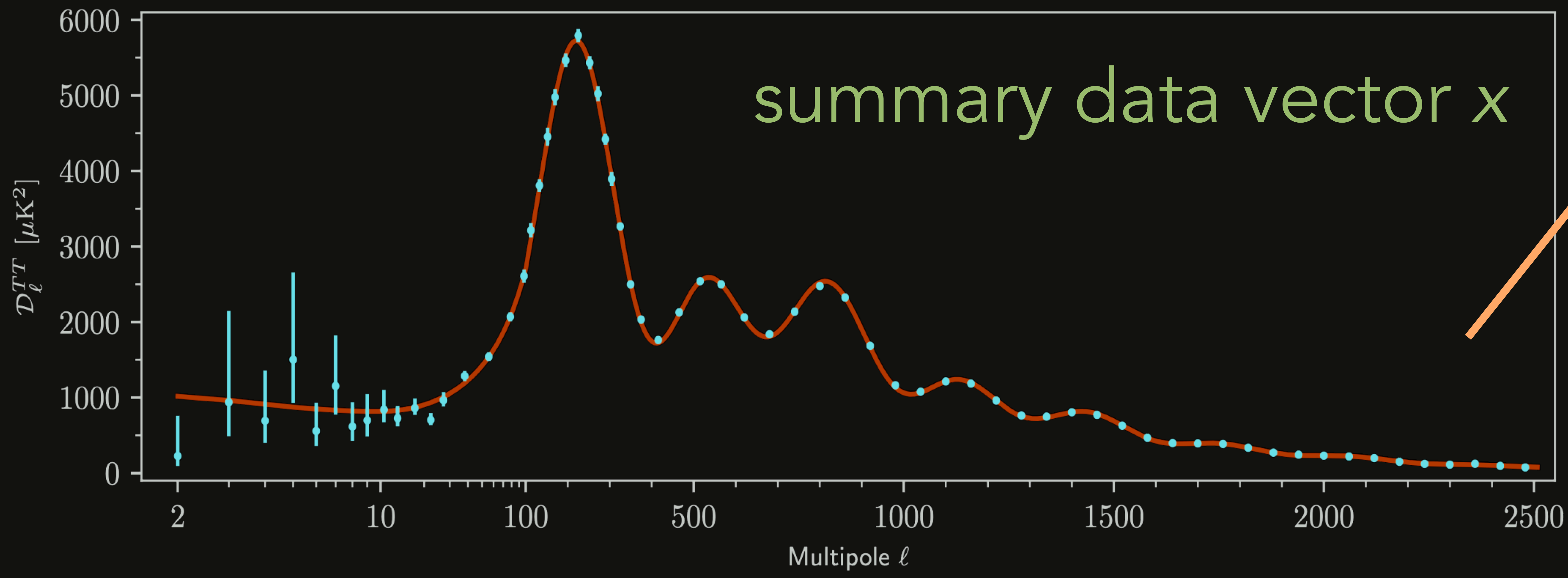
Implicit Likelihood Inference for Late-Time Cosmology

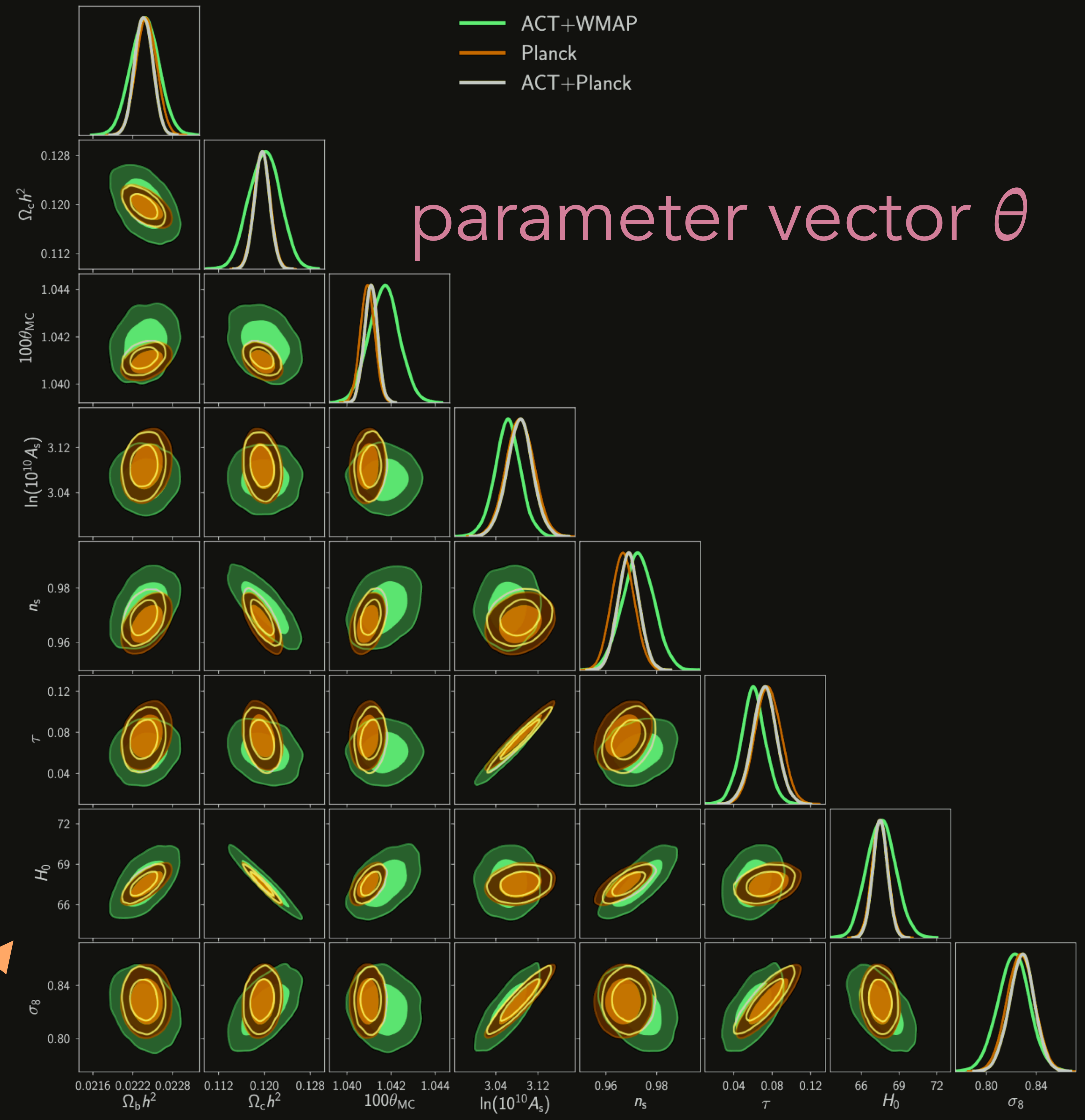
Leander Thiele (Center for Data-Driven Discovery, Kavli IPMU)

NPML, 10/28/2025

Likelihood-based cosmology





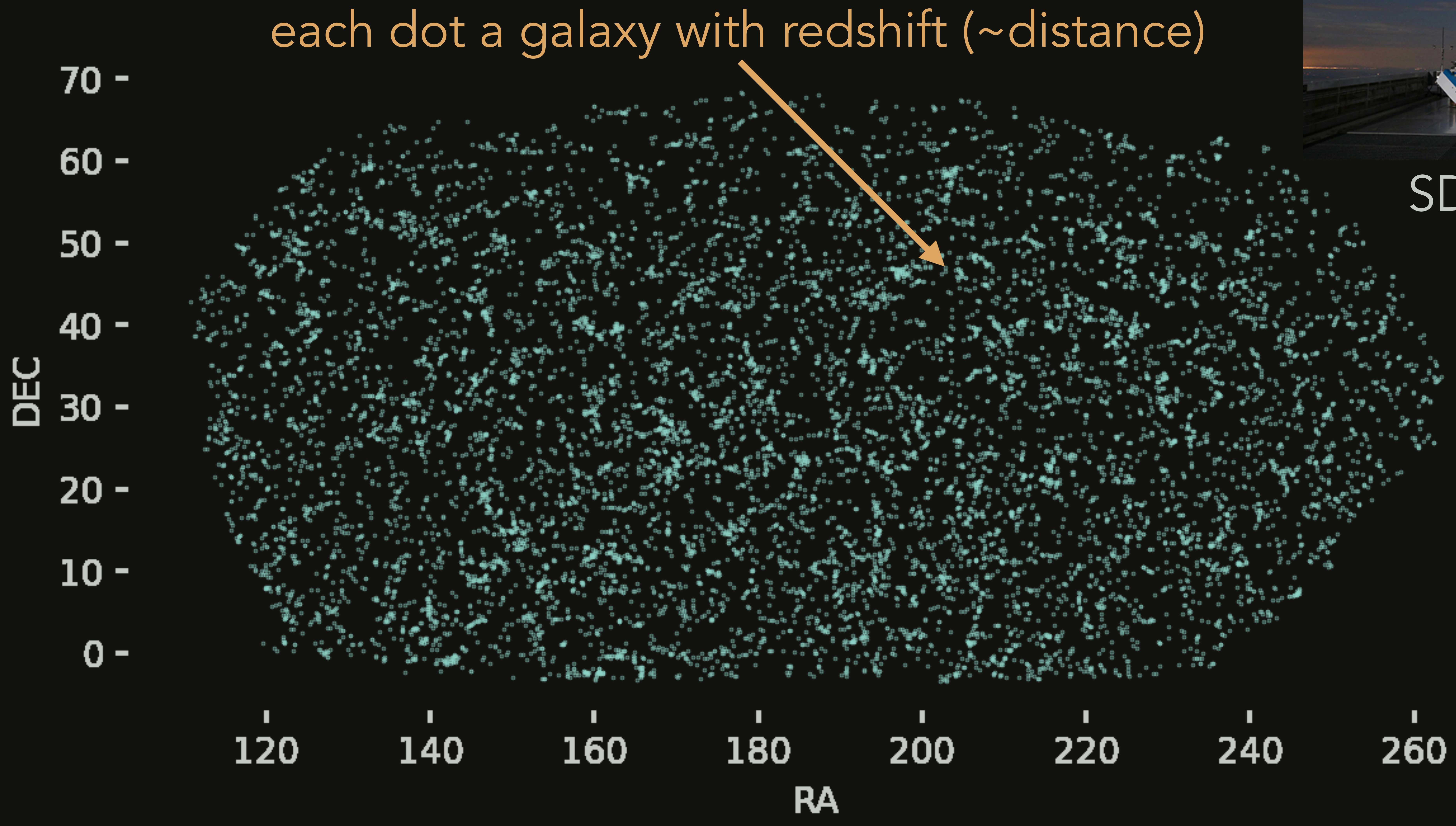


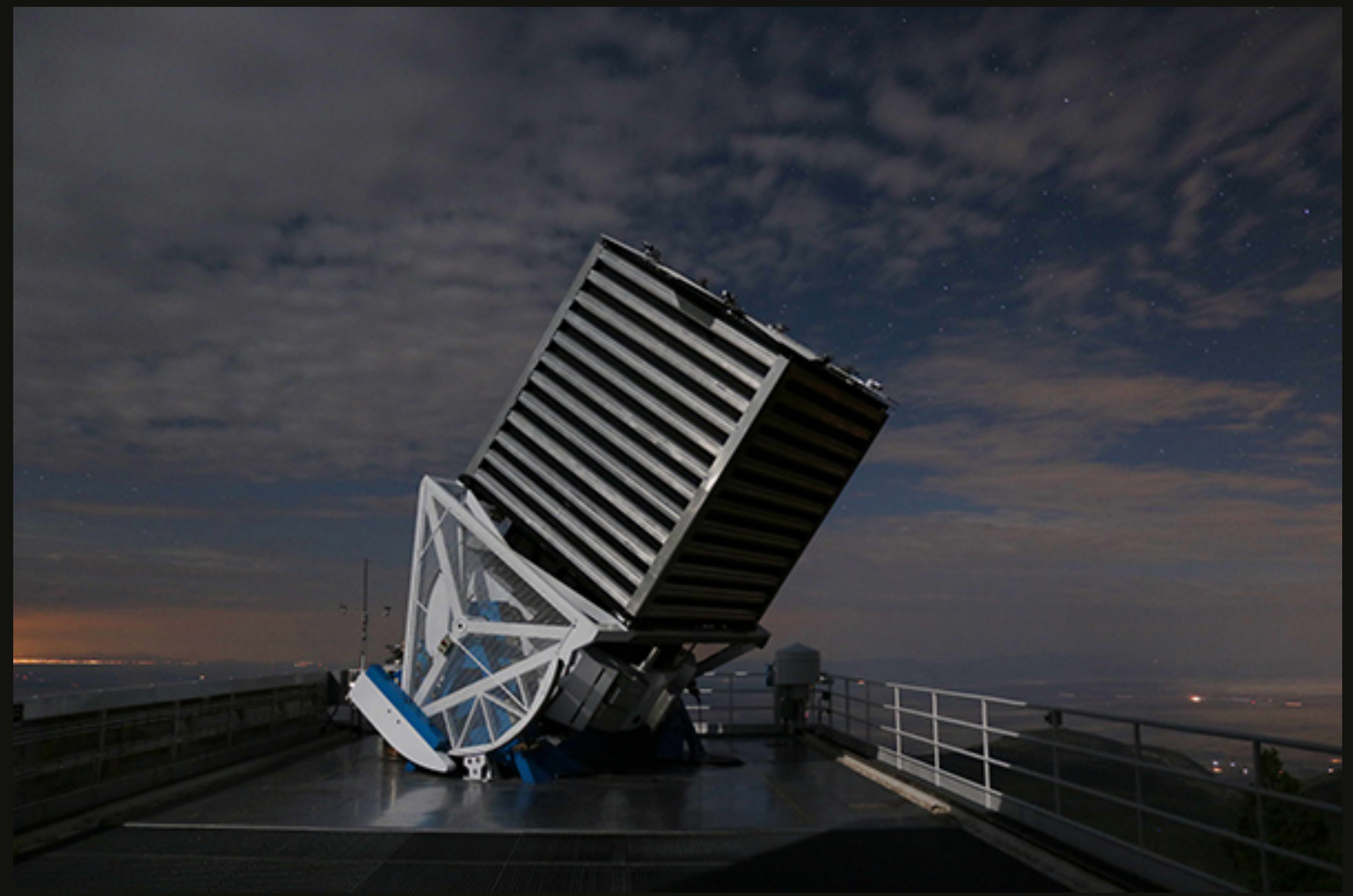
parameter vector θ

Explicit Likelihood:

$$-2\log p(x\,|\,\theta) \propto (x-\mu(\theta))^T\Sigma^{-1}(x-\mu(\theta))$$

A 3-D Map of the Universe





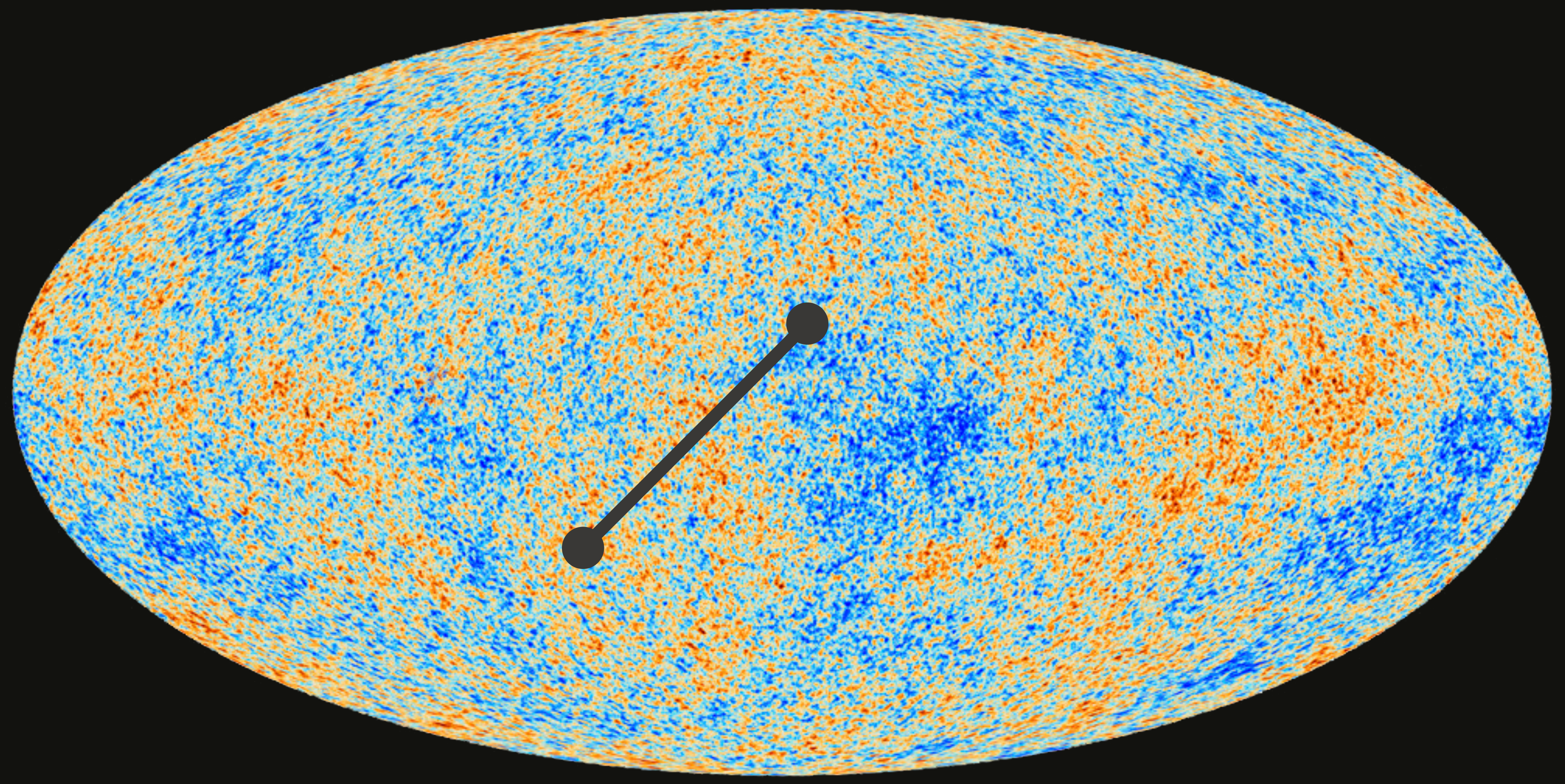
SDSS/BOSS survey

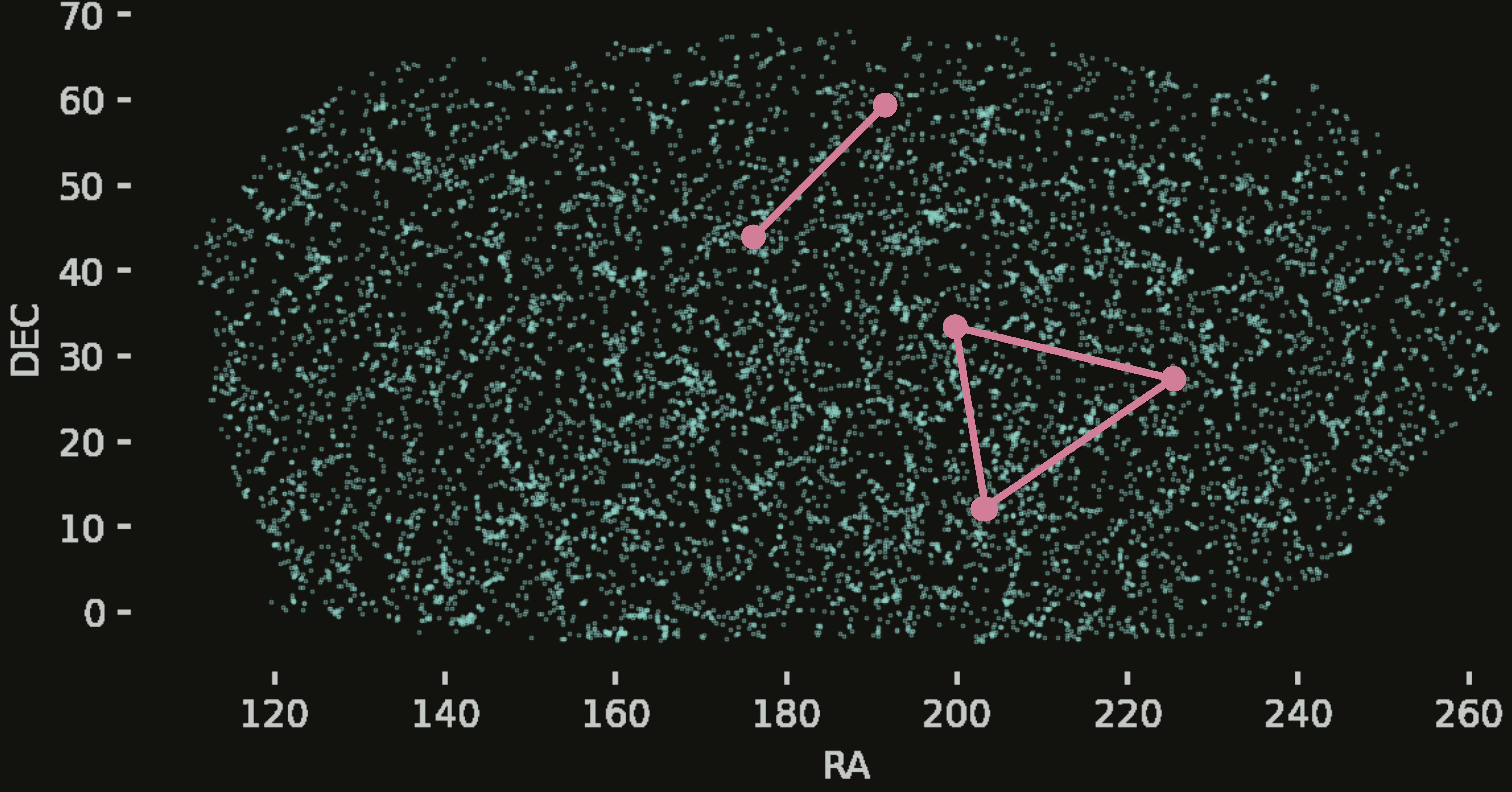
How to summarize this map?

1) pairs of galaxies (power spectrum)

2) triangles of galaxies (bispectrum)

3) …





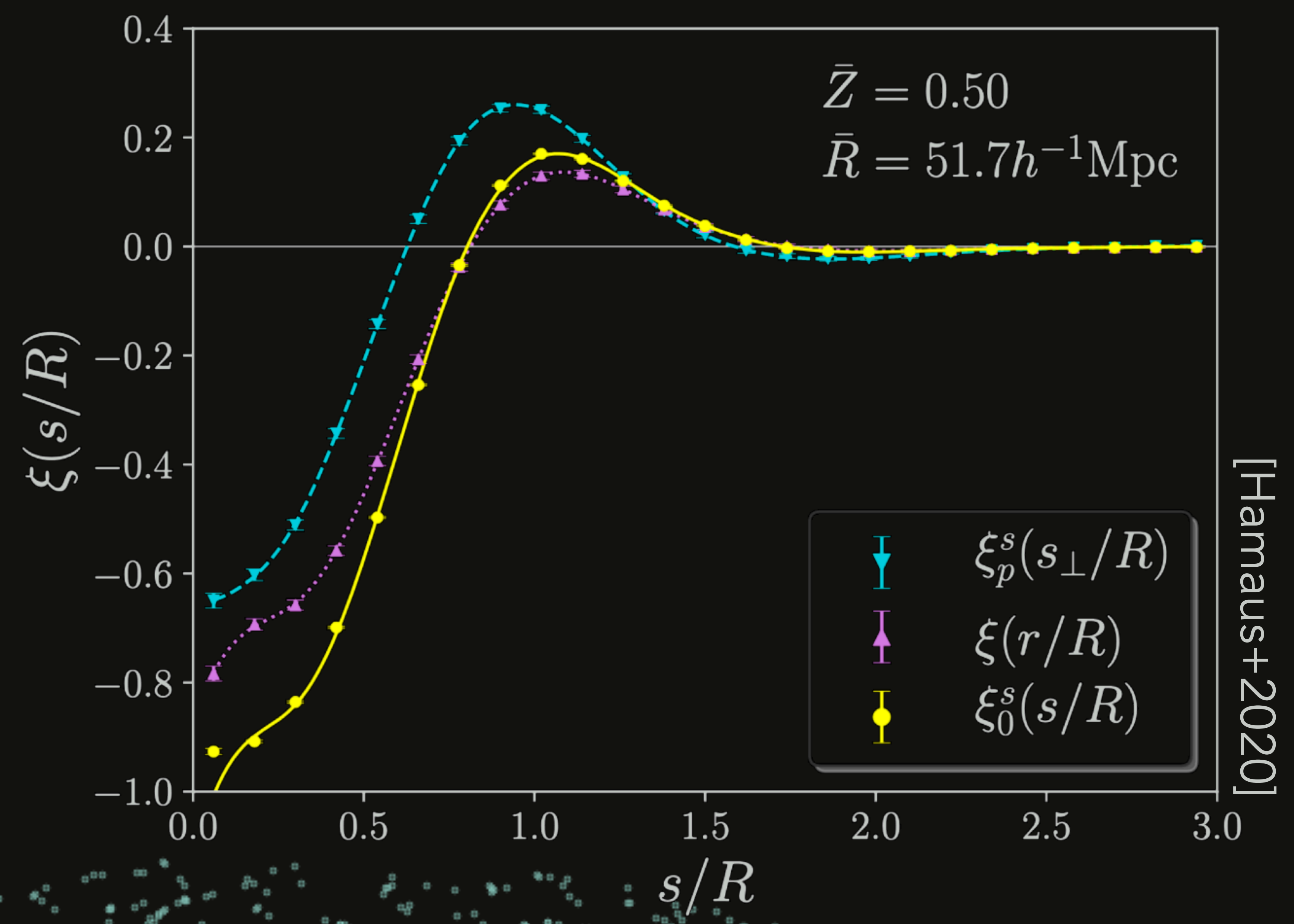
How to summarize this map?

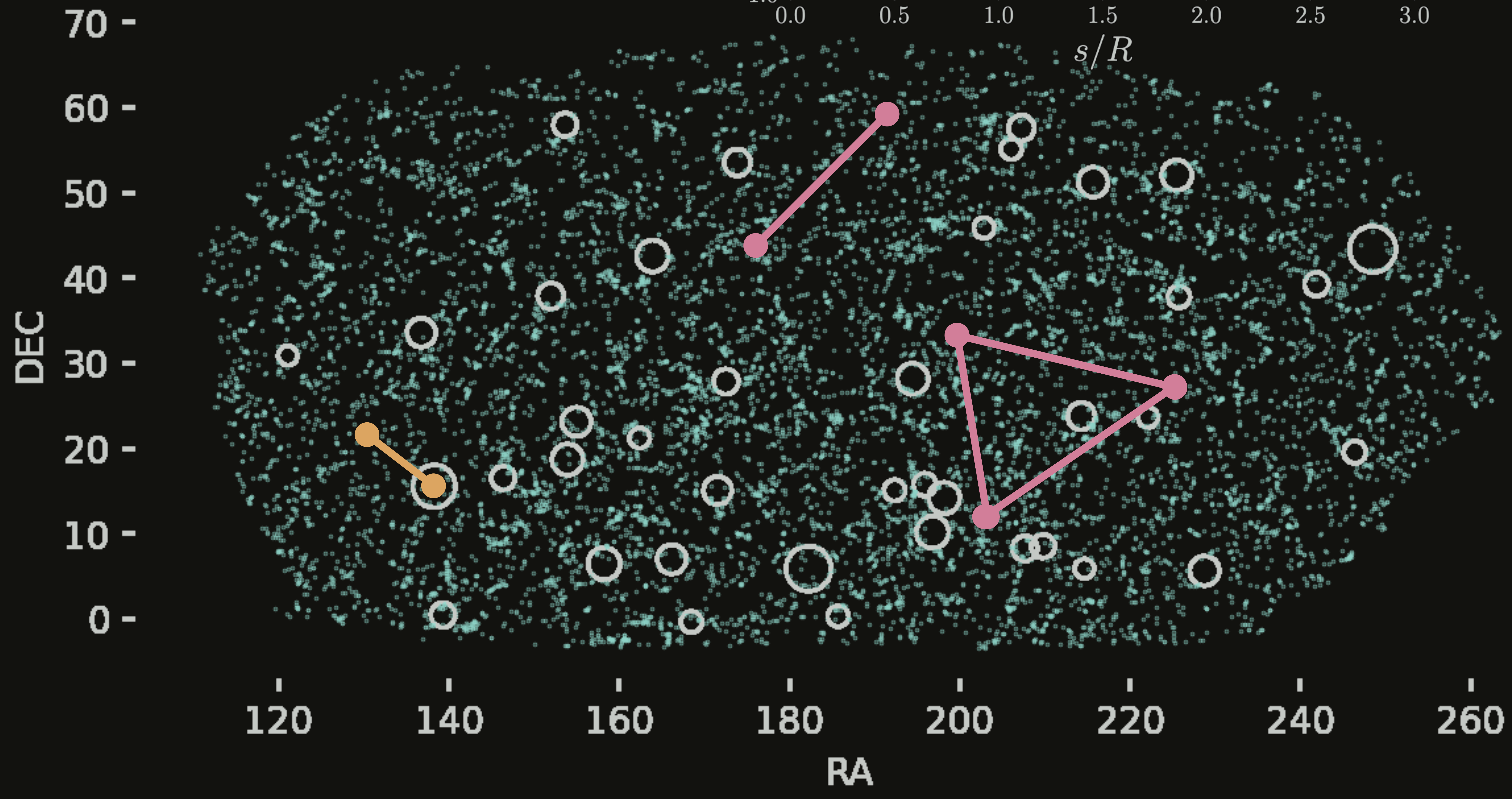
1) pairs of galaxies (power spectrum)

2) triangles of galaxies (bispectrum)

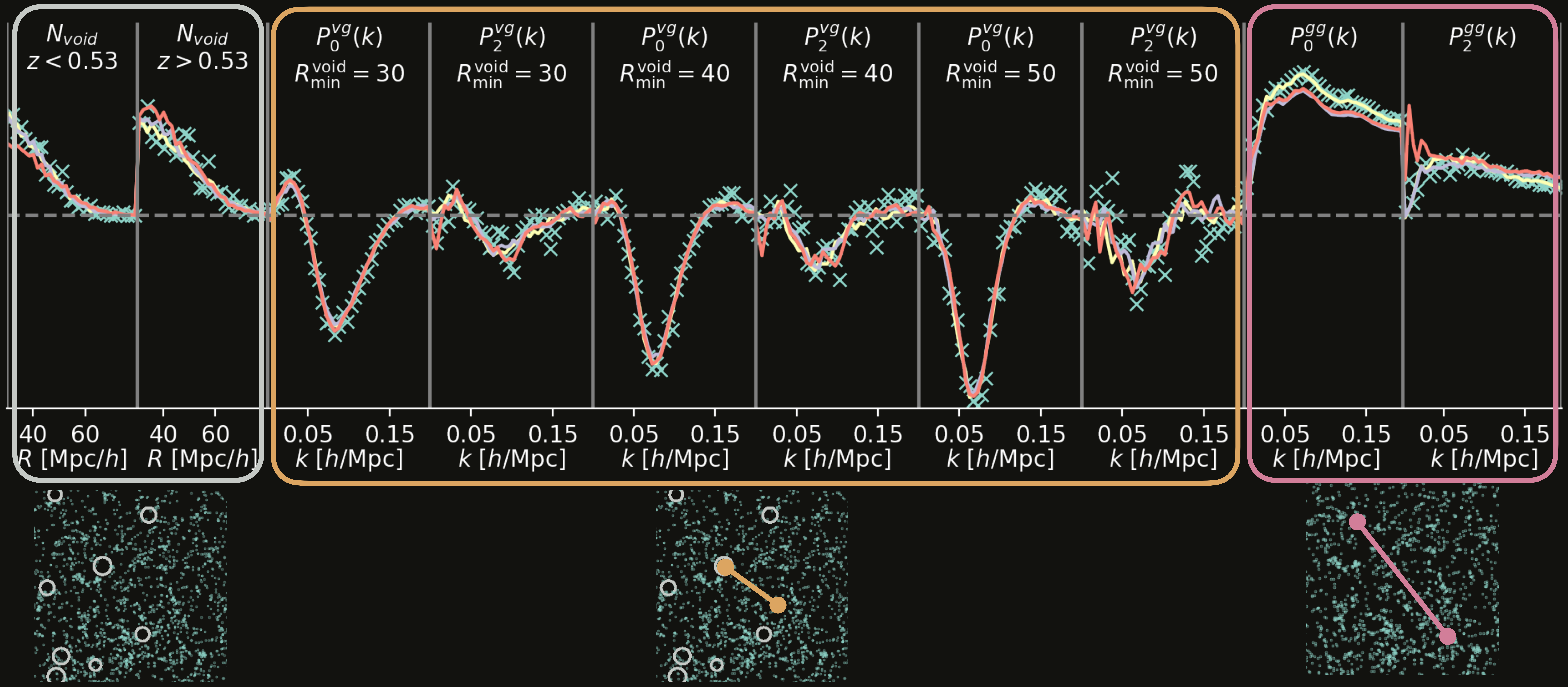
3) …

4) "empty regions": *cosmic voids*
   - size distribution
   - void-galaxy pairs
   - …





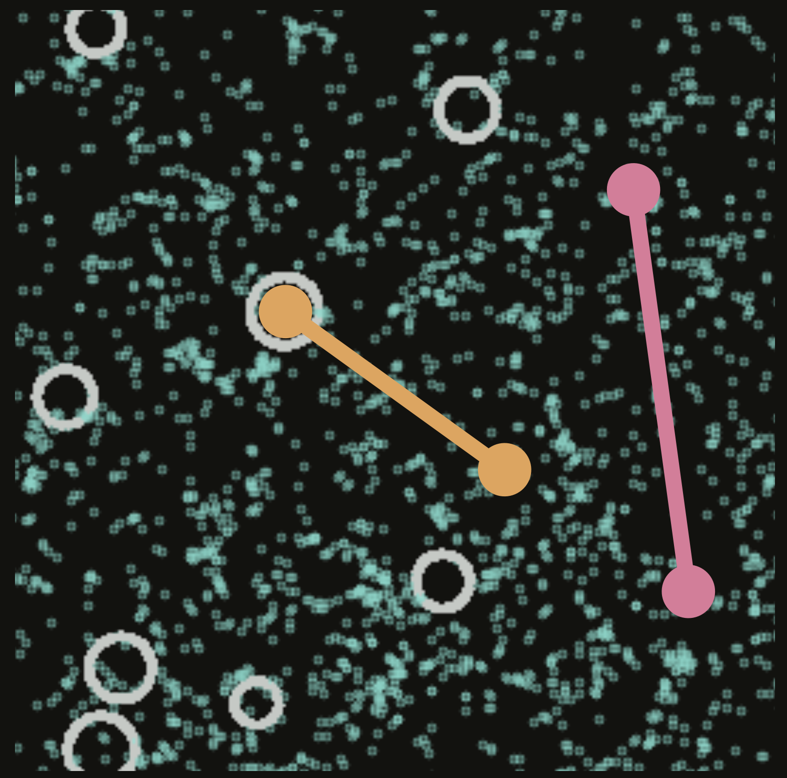
Data Vector

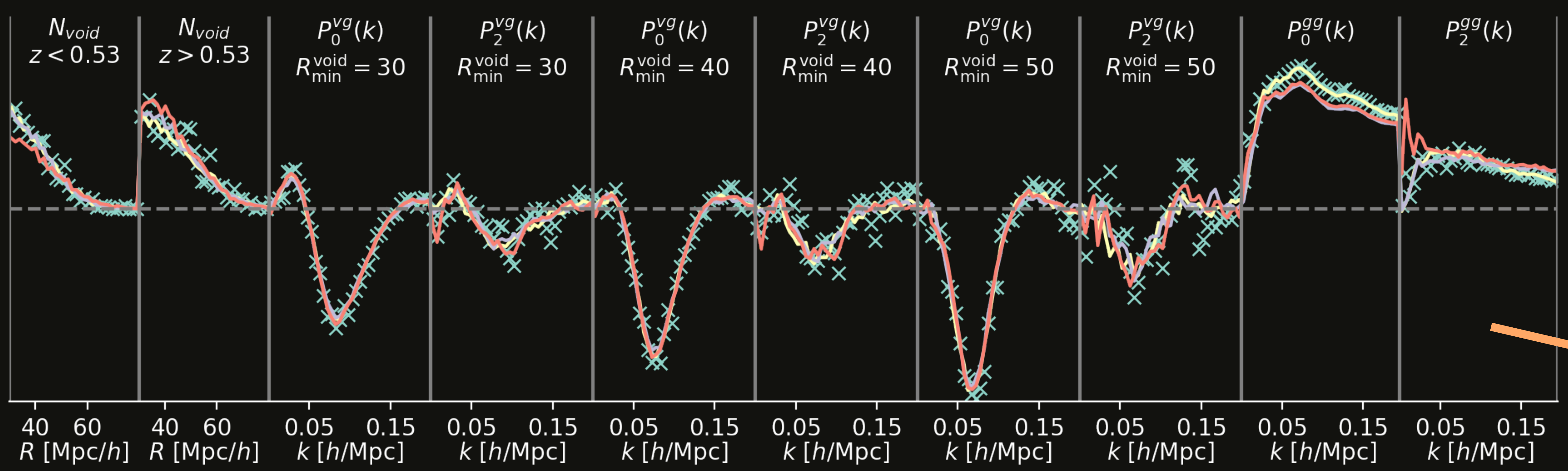


N_{void}
$z < 0.53$
N_{void}
$z > 0.53$
$P_0^{vg}(k)$
$R_{min}^{void} = 30$
$P_2^{vg}(k)$
$R_{min}^{void} = 30$
$P_0^{vg}(k)$
$R_{min}^{void} = 40$
$P_2^{vg}(k)$
$R_{min}^{void} = 40$
$P_0^{vg}(k)$
$R_{min}^{void} = 50$
$P_2^{vg}(k)$
$R_{min}^{void} = 50$
$P_0^{gg}(k)$
$P_2^{gg}(k)$
40 60
R [Mpc/h]
40 60
R [Mpc/h]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]
0.05 0.15
k [h/Mpc]

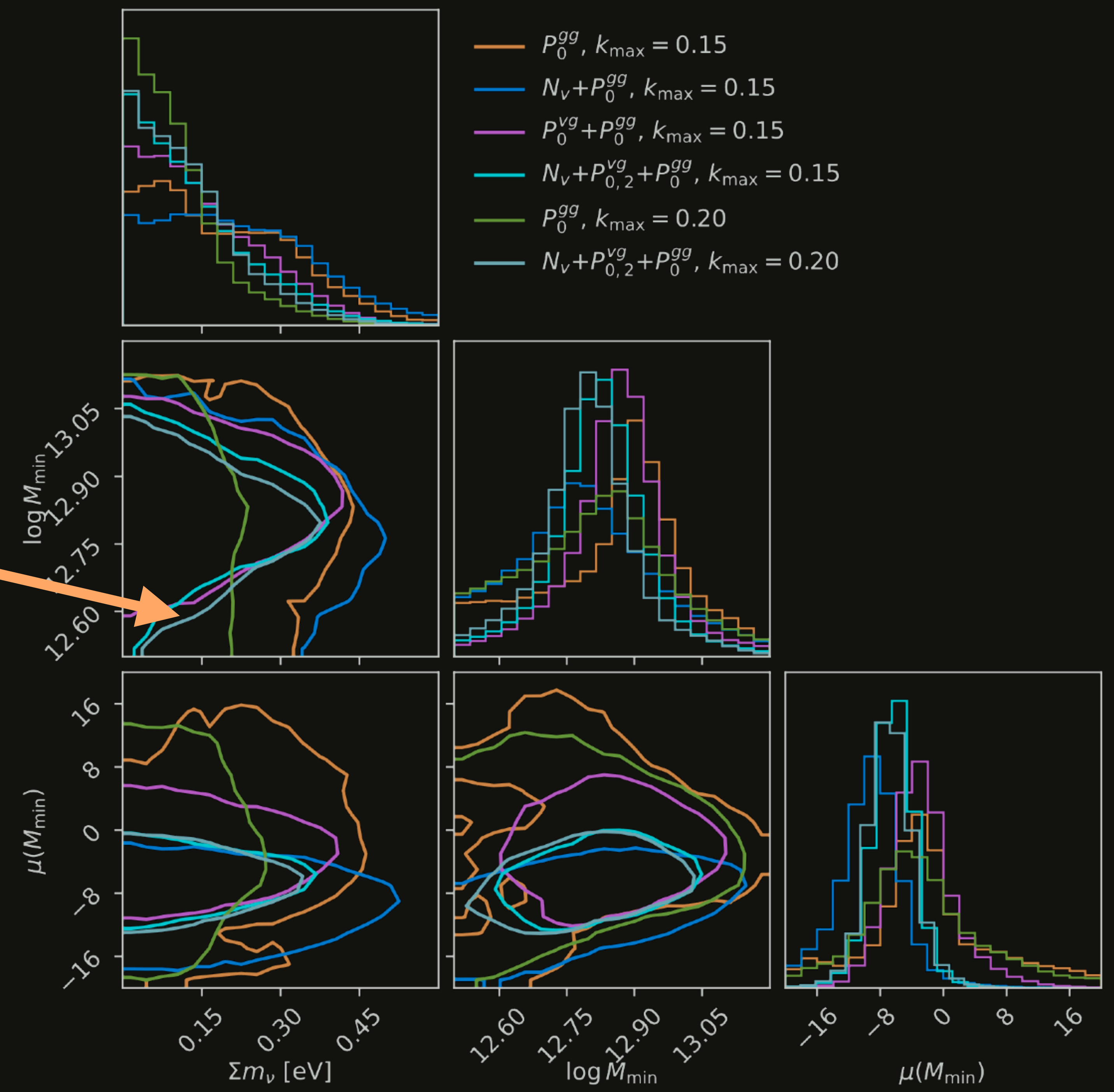
Inference?

- Want to constrain cosmology with BOSS data:
  - galaxy auto power spectrum
  - void size function (histogram of void sizes)
  - void galaxy cross power spectrum (void profile)



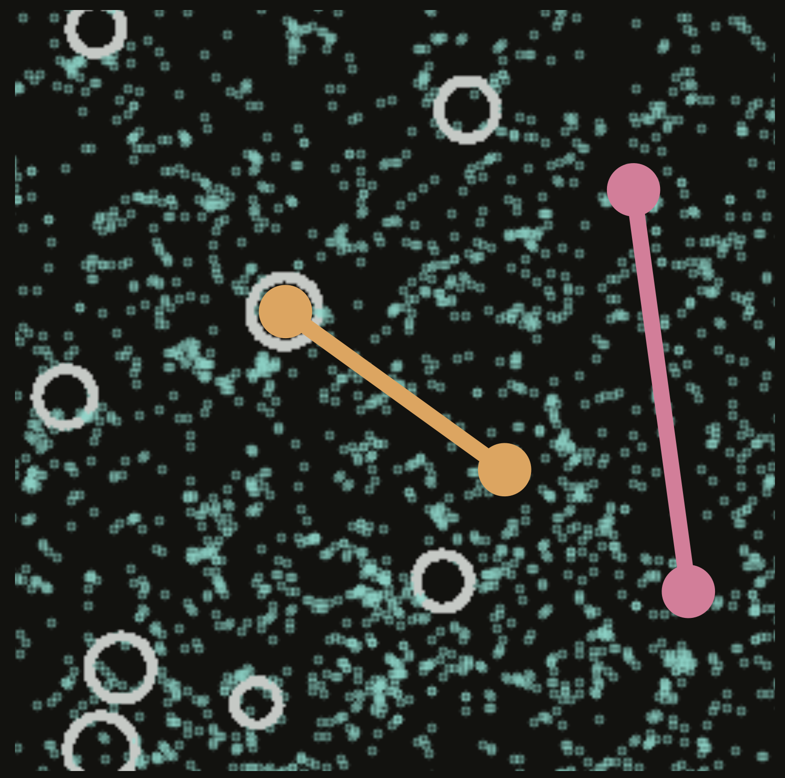


?

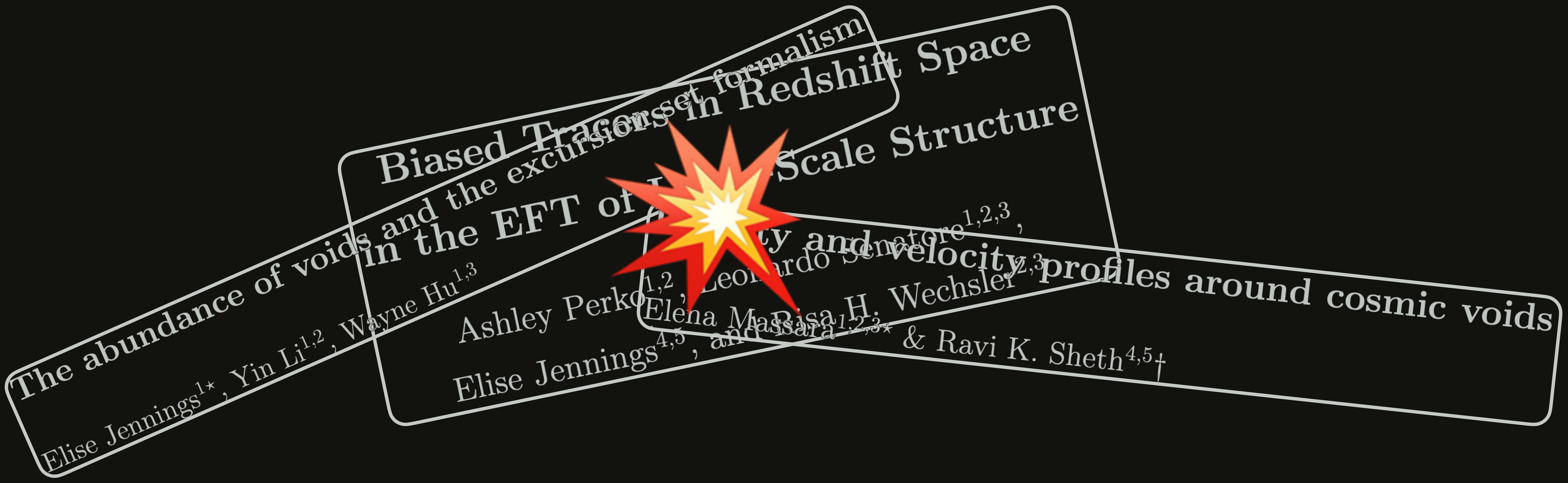


Inference?

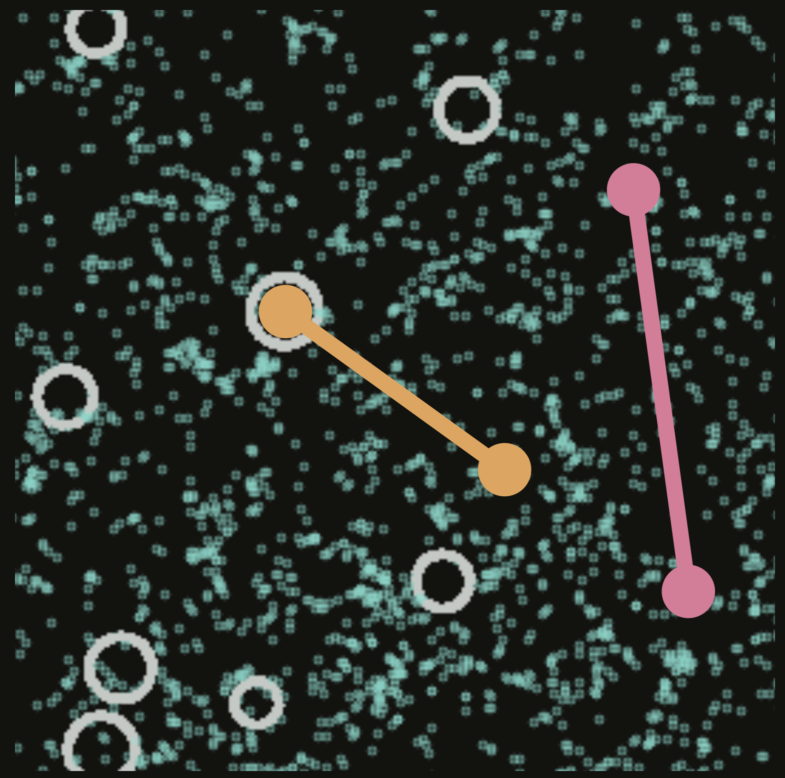
- Want to constrain cosmology with BOSS data:
  - galaxy auto power spectrum
  - void size function (histogram of void sizes)
  - void galaxy cross power spectrum (void profile)



- Joint modeling of these statistics difficult with analytic methods



Inference?



- Want to constrain cosmology with BOSS data:
    - galaxy auto power spectrum
    - void size function (histogram of void sizes)
    - void galaxy cross power spectrum (void profile)
- Joint modeling of these statistics difficult with analytic methods
- Likelihood unknown

$$-2\log p(x \mid \theta) \; 💥 \; (x-\mu(\theta))^T\Sigma^{-1}(x-\mu(\theta))$$

Outline

Why should nature have conspired to put all information into the (quasi-)linear modes we happen to be able to describe with perturbation theory?

Accessing the information in non-linear regime typically requires simulation-based methods.

Outline

Implicit Likelihood Inference: A tool to solve inverse problems which are implicitly defined through simulations, typically using deep neural networks

1. Motivation & Overview of Methods

2. Example application to BOSS voids

3. The “elephant in the room”: How to make it reliable and computationally feasible

A word about words

Simulation-based Inference

Explicit-likelihood inference

Implicit-likelihood inference
("likelihood-free inference")

summary statistics

Field-level forward modeling
(e.g., galaxy clustering: F.Schmidt,
weak lensing: N.Porqueres)

Field-level likelihood
(e.g., weak lensing: Dai&Seljak)

Bayesian inference

$$\underbrace{\boxed{P(\text{parameters} \mid \text{data})}}_{\text{posterior}} = \frac{\overbrace{\boxed{P(\text{data} \mid \text{parameters})}}^{\text{likelihood}} \; \overbrace{\boxed{P(\text{parameters})}}^{\text{prior}}}{\underbrace{\boxed{P(\text{data})}}_{\text{evidence}}}$$

Bayesian inference

More concretely:

θ=interesting parameters, η=nuisance parameters, ζ=initial conditions,

x=data, m=model:

$$P(x \mid \theta) = \int D\eta\, D\zeta\, \delta[\, x - m(\theta, \eta, \zeta)\,]$$

$$\underset{\text{posterior}}{\boxed{P(\text{parameters}\,|\,\text{data})}} = \frac{\overset{\text{likelihood}}{\boxed{P(\text{data}\,|\,\text{parameters})}}\quad \overset{\text{prior}}{\boxed{P(\text{parameters})}}}{\underset{\text{evidence}}{\boxed{P(\text{data})}}}$$

Bayesian inference

More concretely:

θ=interesting parameters, η=nuisance parameters, ζ=initial conditions,

x=data, m=model:

$$P(x \mid \theta) = \int D\eta\, D\zeta\, \delta[\, x - m(\theta, \eta, \zeta)\,]$$

Traditional case:

$$P(x \mid \theta) = \int D\eta\ \mathrm{Gaussian}[\, x - \mu(\theta, \eta), \Sigma\,]$$

[do remaining low-dimensional η-integral with Monte Carlo]

Bayesian inference

More concretely:

θ=interesting parameters, η=nuisance parameters, ζ=initial conditions,

x=data, m=model:

$$P(x \mid \theta) = \int D\eta\, D\zeta\, \delta[\, x - m(\theta, \eta, \zeta)\,]$$

Traditional case:

$$P(x \mid \theta) = \int D\eta\ \mathrm{Gaussian}[\, x - \mu(\theta, \eta), \Sigma\,]$$

But what do we do if the Gaussian approximation doesn't hold?

→ assume we have simulator that evaluates m(θ, η, ζ) accurately

Neural Implicit Likelihood Inference (ILI)

More concretely:

θ=interesting parameters, η=nuisance parameters, ζ=initial conditions,

x=data, m=model:

$$P(x \mid \theta) = \int D\eta\, D\zeta\, \delta[\, x - m(\theta, \eta, \zeta)\,]$$

But what do we do if the Gaussian approximation doesn't hold?

→ assume we have simulator that evaluates m(θ, η, ζ) accurately

neural likelihood estimation (NLE) neural posterior estimation (NPE)

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})\, P(\text{parameters})}{P(\text{data})}$$

neural ratio estimation (NRE)

Neural Implicit Likelihood Inference (ILI)

neural likelihood estimation (NLE) neural posterior estimation (NPE)

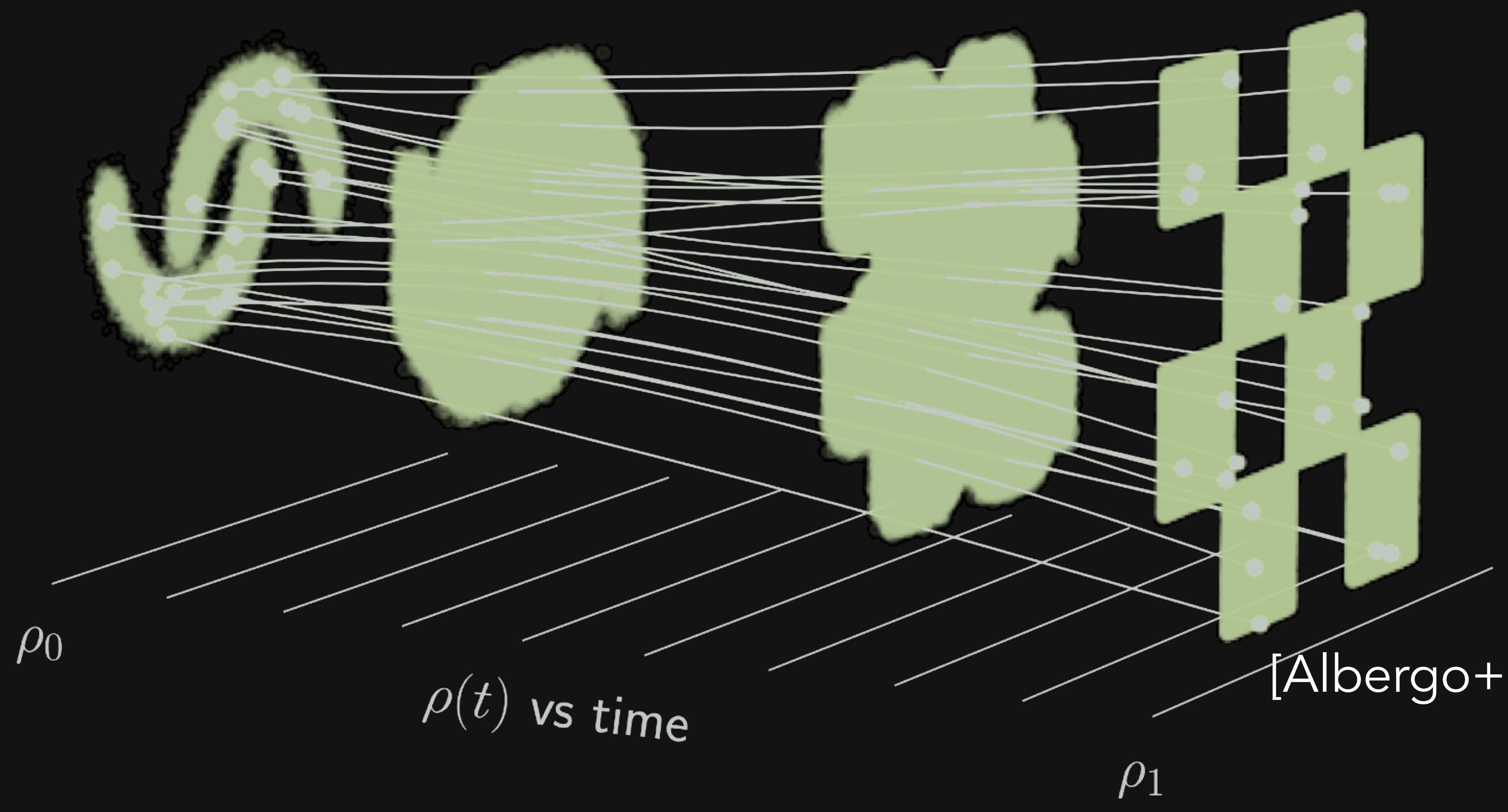
$$P(\text{parameters}\,|\,\text{data}) = \frac{P(\text{data}\,|\,\text{parameters}) \quad P(\text{parameters})}{P(\text{data})}$$

neural ratio estimation (NRE)

neural likelihood estimation (NLE):
conditioned flow q(x|θ)

neural posterior estimation (NPE):
conditioned flow q(θ|x)

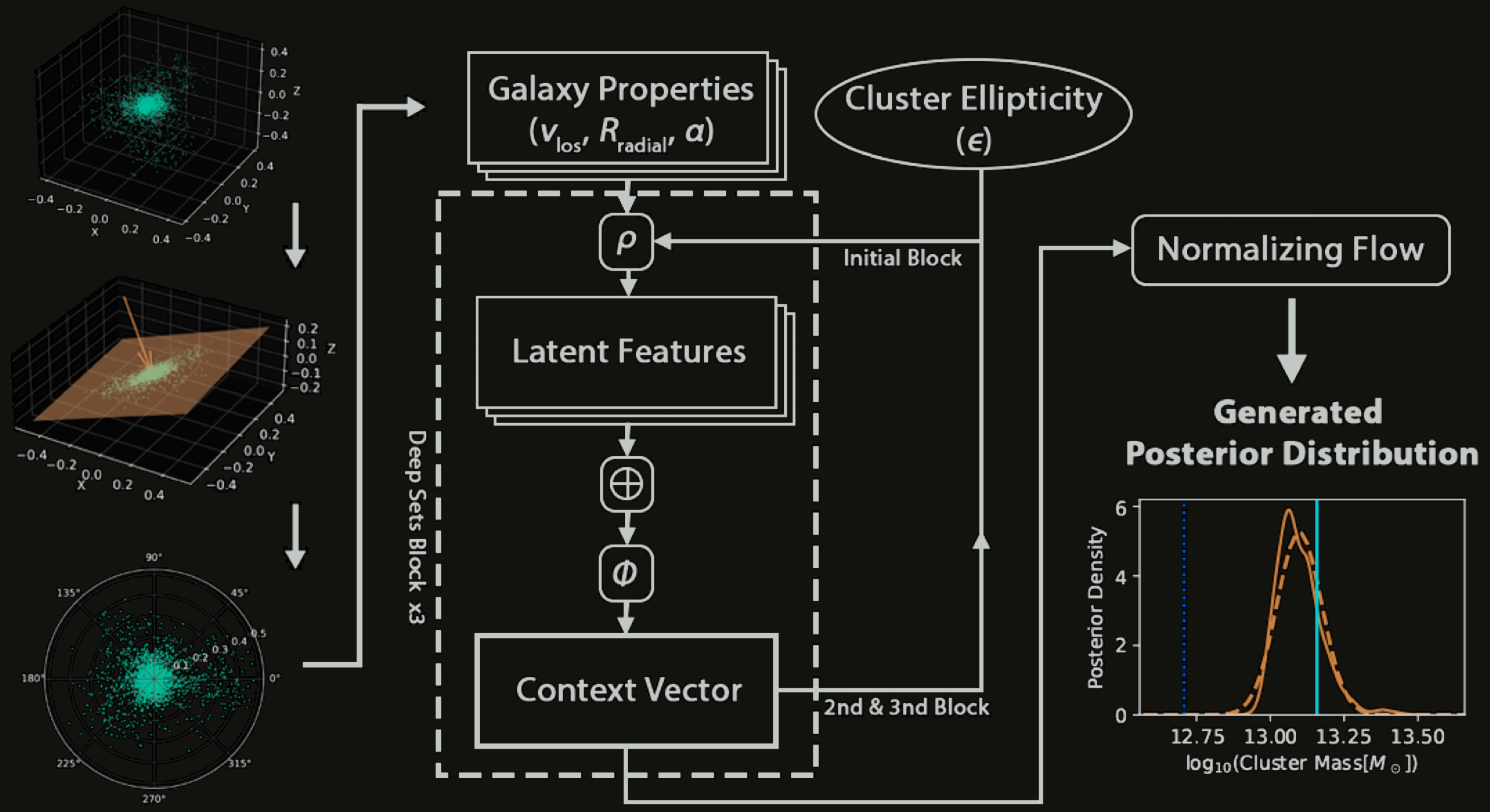
neural ratio estimation (NRE):
classifier between (x, θ) ~p(x, θ) and ~p(x)p(θ)



[Albergo+2023]

Practicality: data size & structure

Usually, data vectors too high dimensional or live in awkward spaces
→ We deal with this by constructing a useful latent representation through *embedding networks*



[B.Y.Wang & LT 2025]

Neural Implicit Likelihood Inference (ILI)

- *in the limit*, any likelihood learnable
- any simulate-able effect can be incorporated
- no formal difference between nuisance parameters and initial conditions
- primary choice at the moment:
  - NPE: empirically good performance, need to deal with flow, not so convenient for i.i.d. data
  - NRE: classification → super flexible, empirically more tuning required

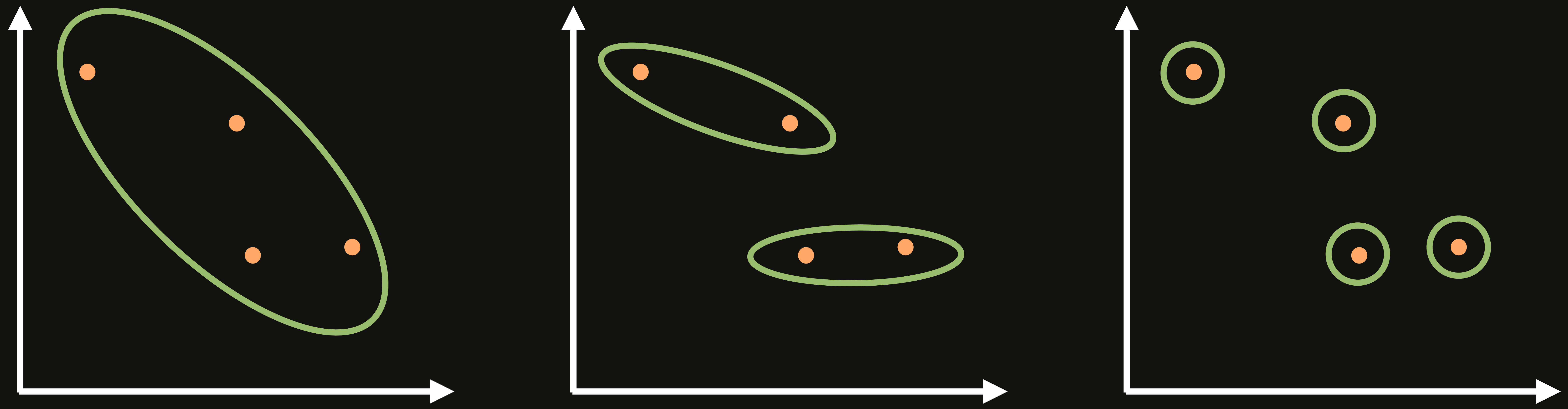
Curse of dimensionality

Let's say: 200,000 simulations, 10 model parameters, 10 data points
→ this is optimistic!

We're trying to learn a density in this 10+10 dimensional space

$(200{,}000)^{1/(10+10)} = 1.8$, let's say 2

What is the implicit prior?
Maybe theory can help?



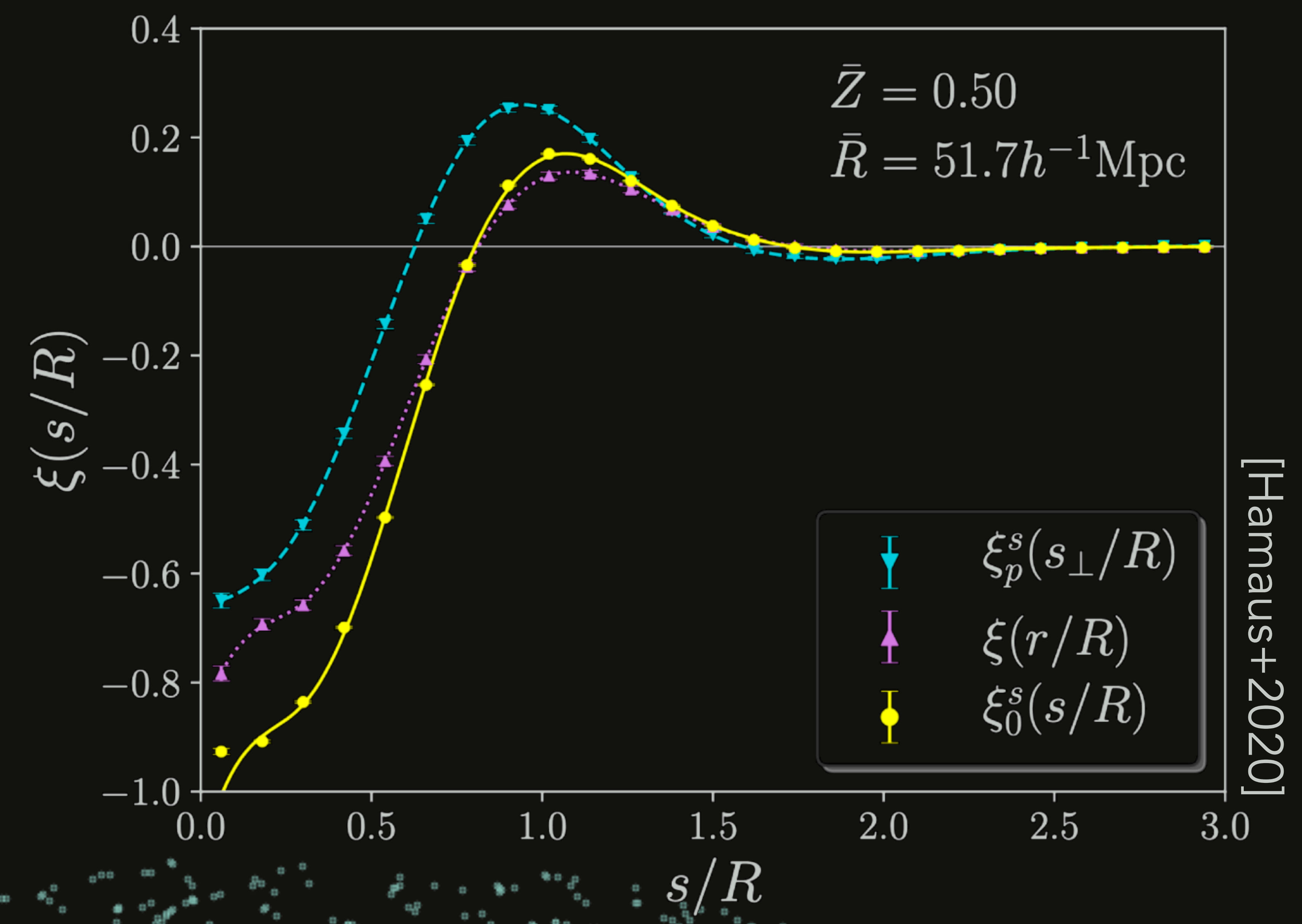
How to summarize this map?

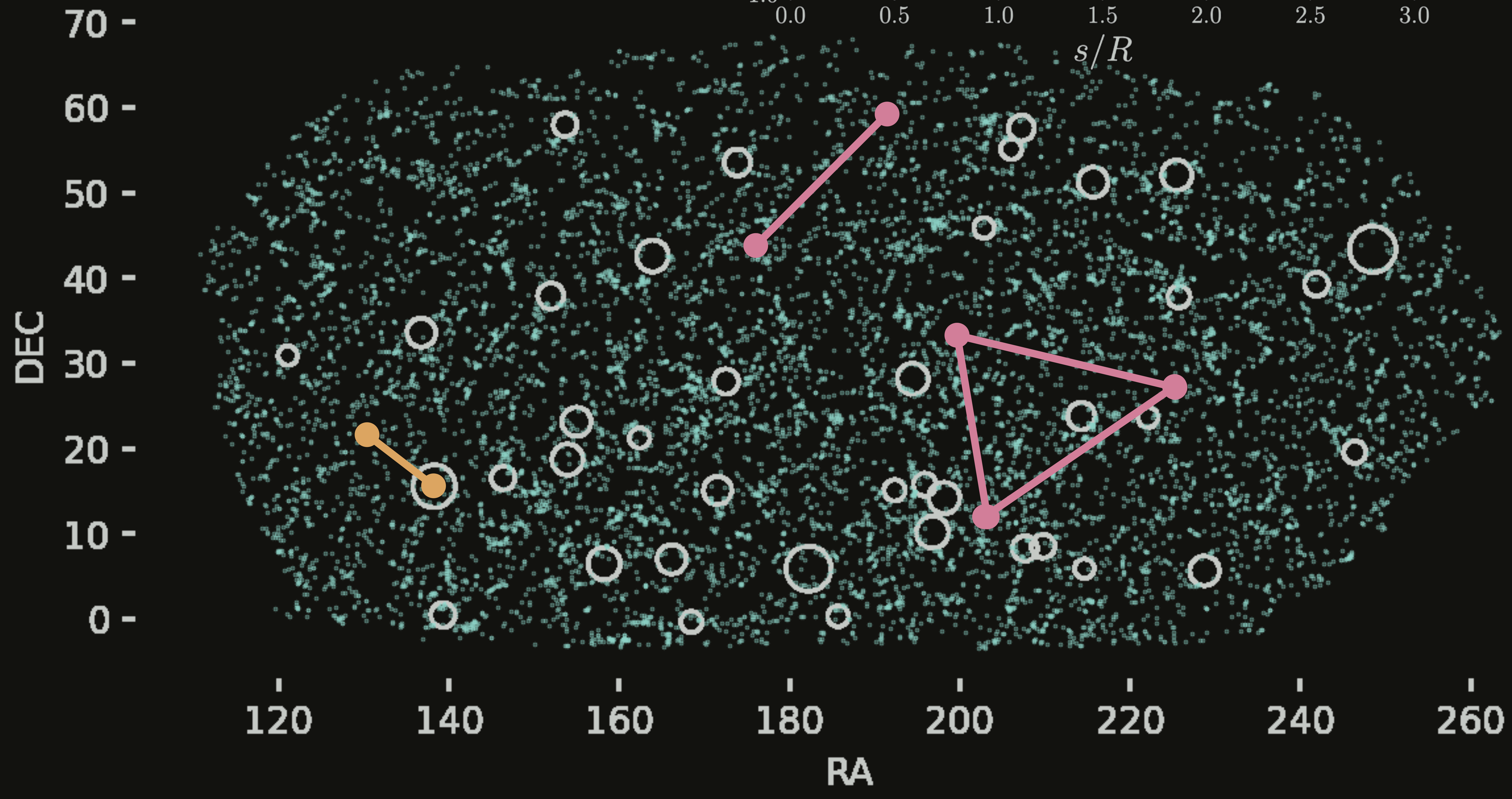
1) pairs of galaxies (power spectrum)

2) triangles of galaxies (bispectrum)

3) …

4) “empty regions”: *cosmic voids*
    - size distribution
    - void-galaxy pairs
    - …

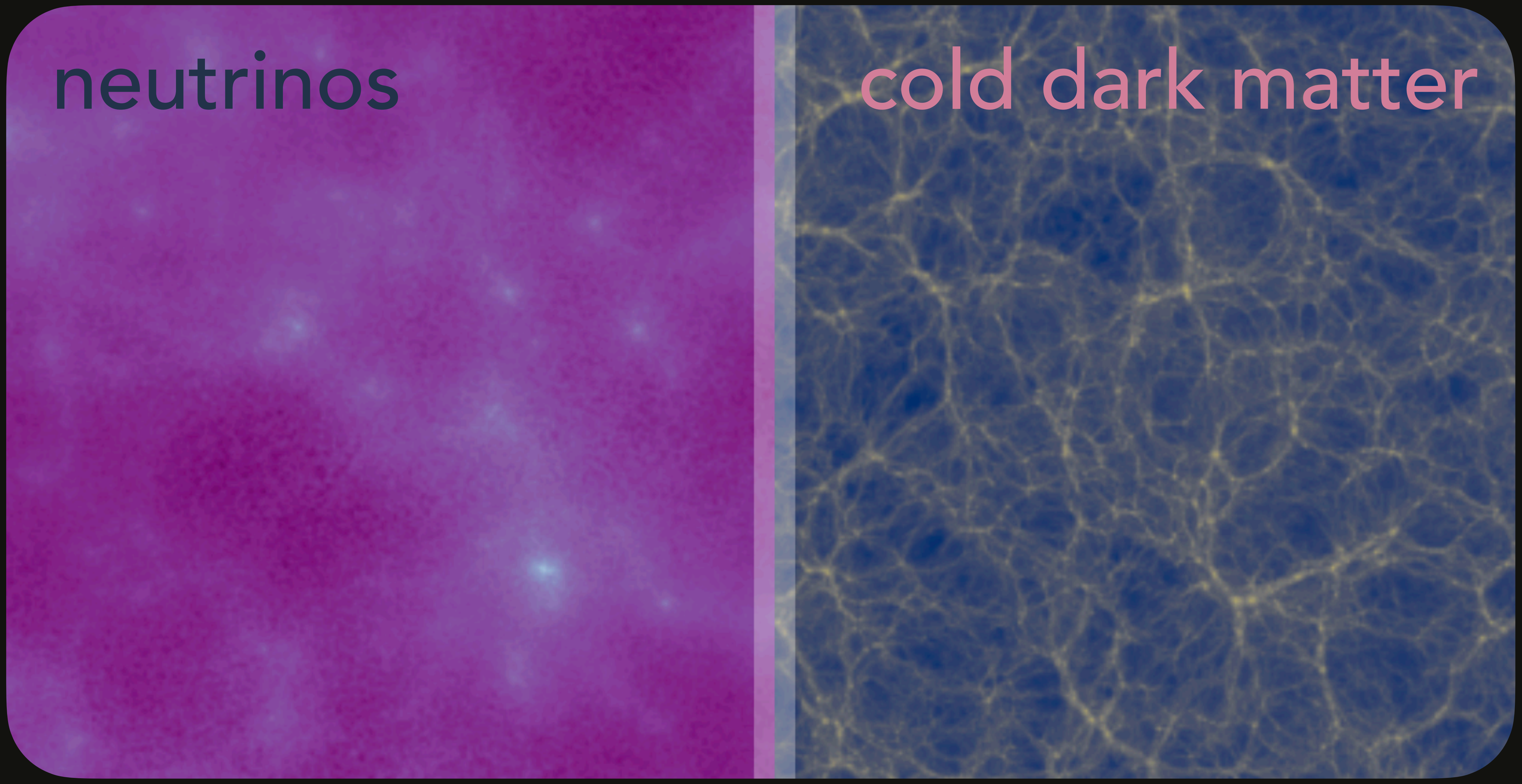


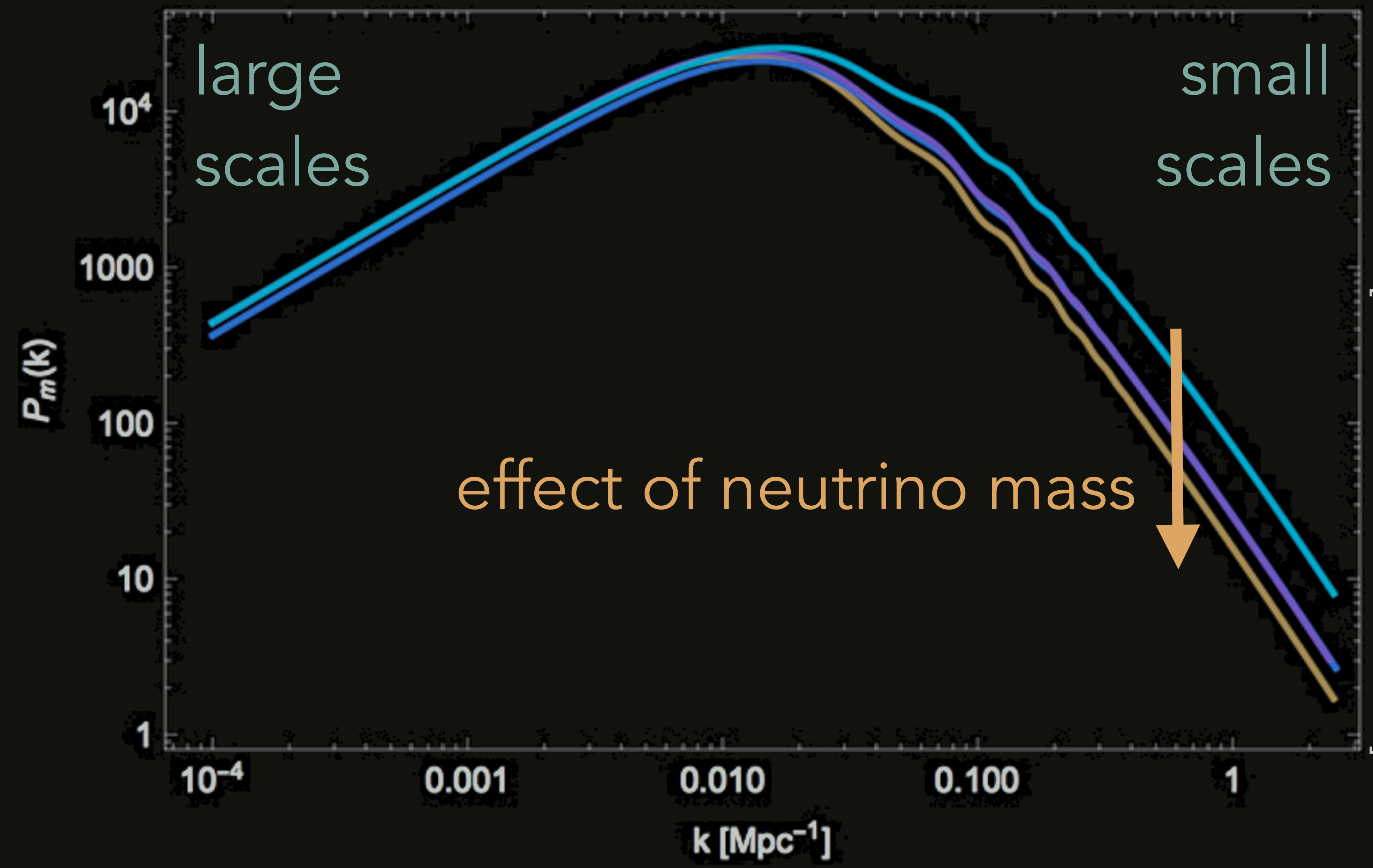


What can voids do for us?

- upweight underdensities → complementary to correlation functions
  - corrections to general relativity
  - dark energy
  - neutrino mass

[FLAMINGO]

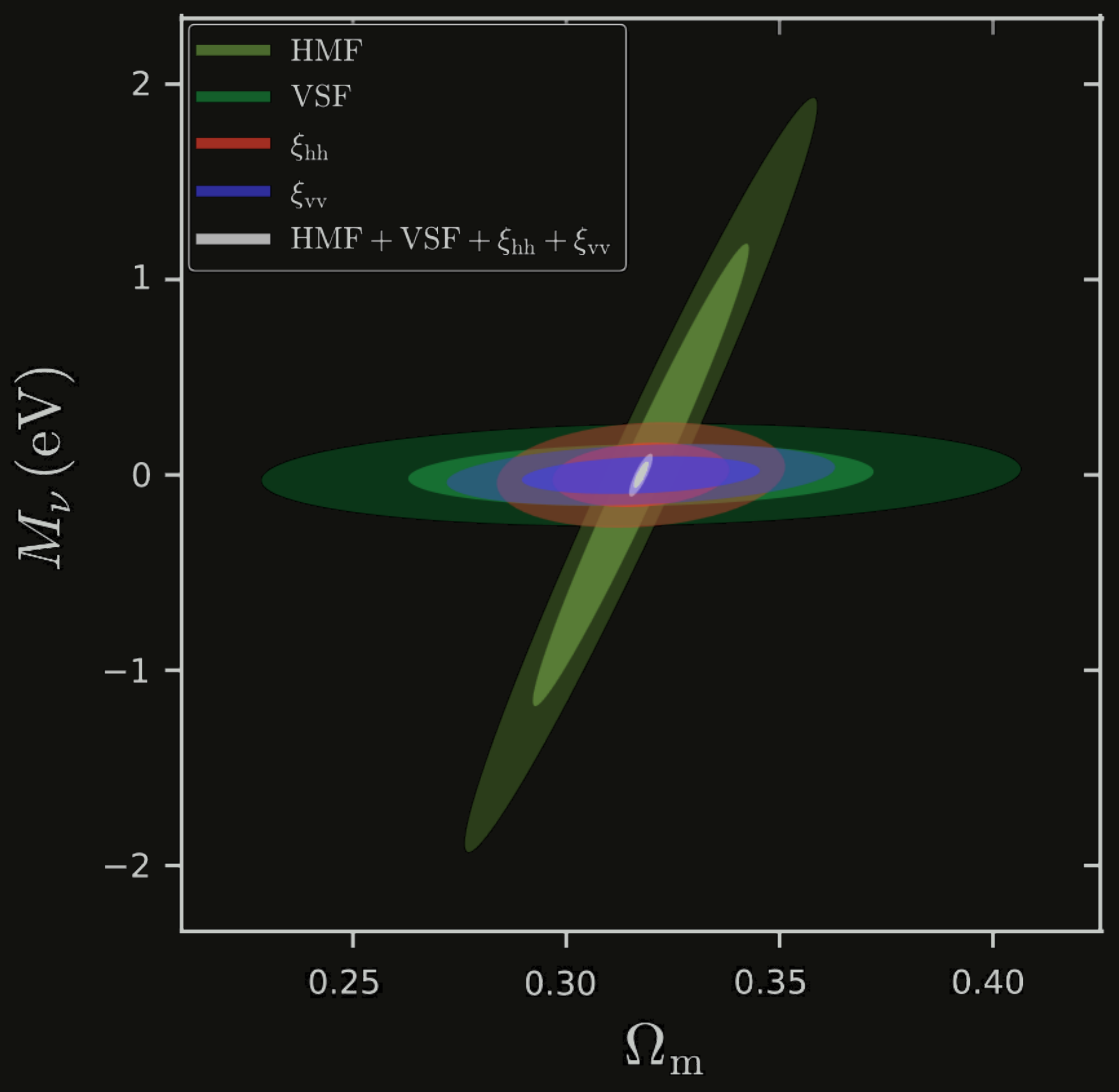




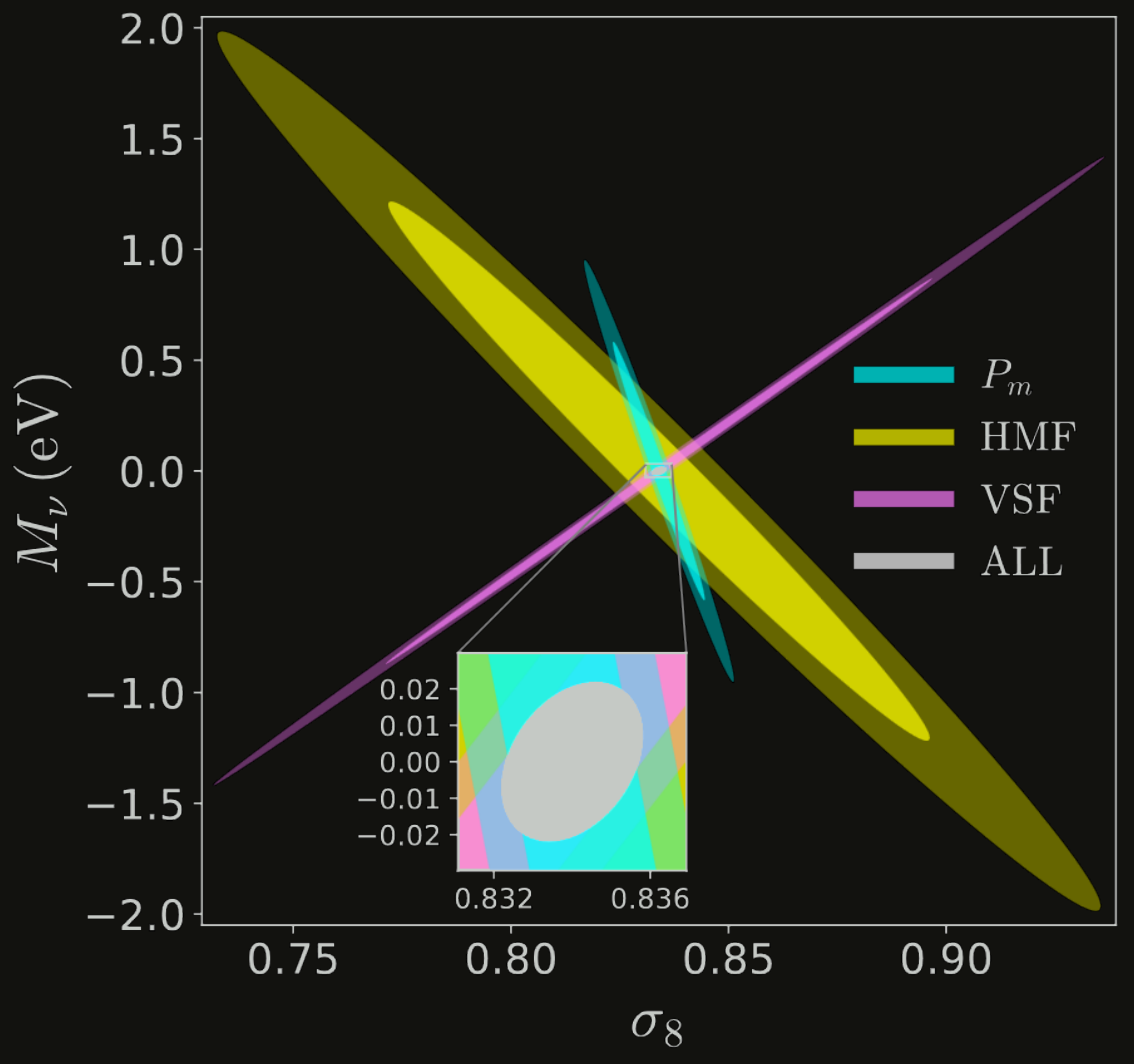
[Gerbino&Lattanzi2018]

Voids & Neutrino Mass: Forecasts

QUIJOTE halos (1 Gpc/h)

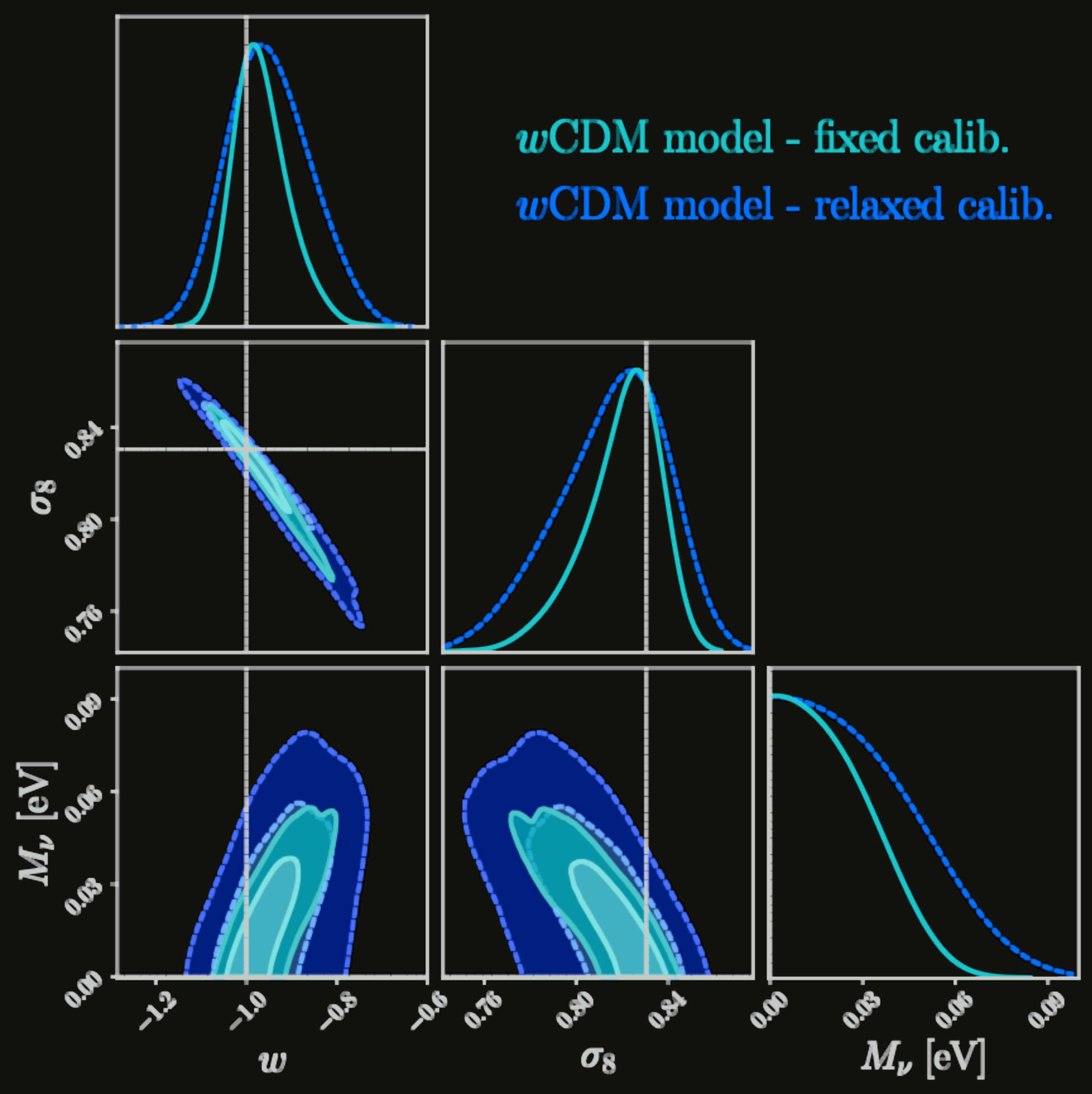


[Kreisch+2021]



[Bayer+2021]

"Euclid" spec-z galaxies

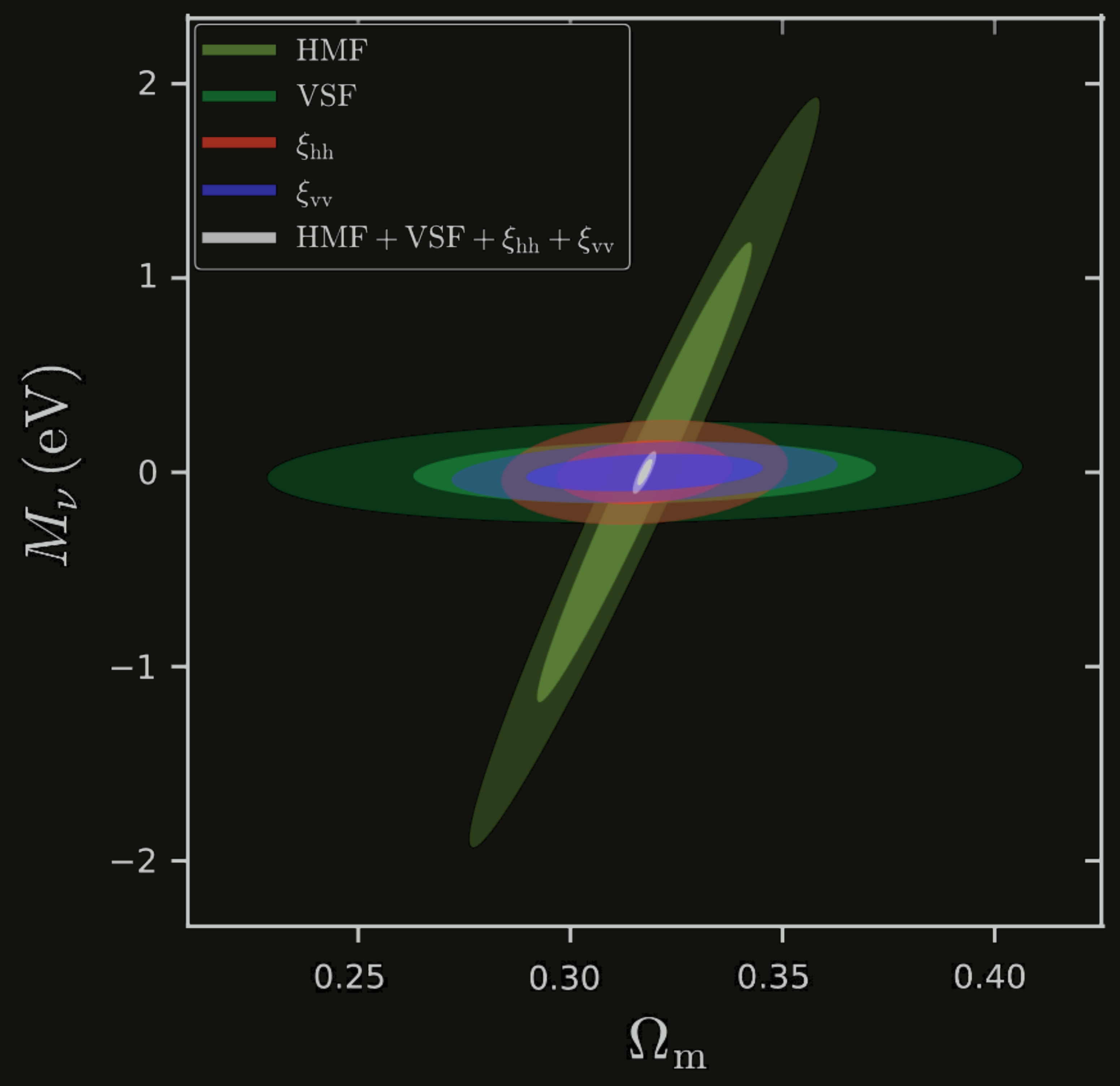


[Contarini+2021]

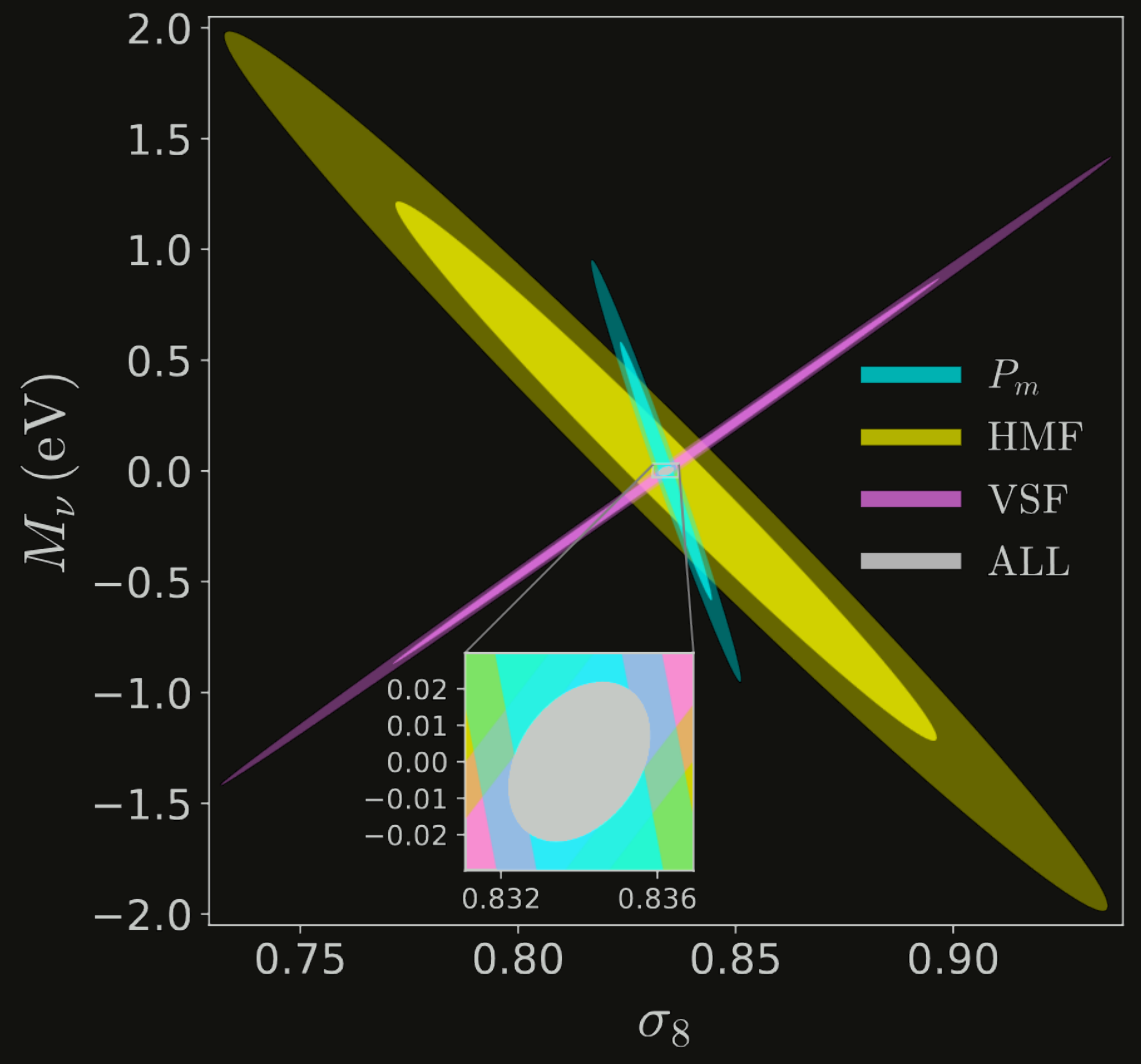
Voids & Neutrino Mass: Forecasts

QUIJOTE halos (1 Gpc/h)

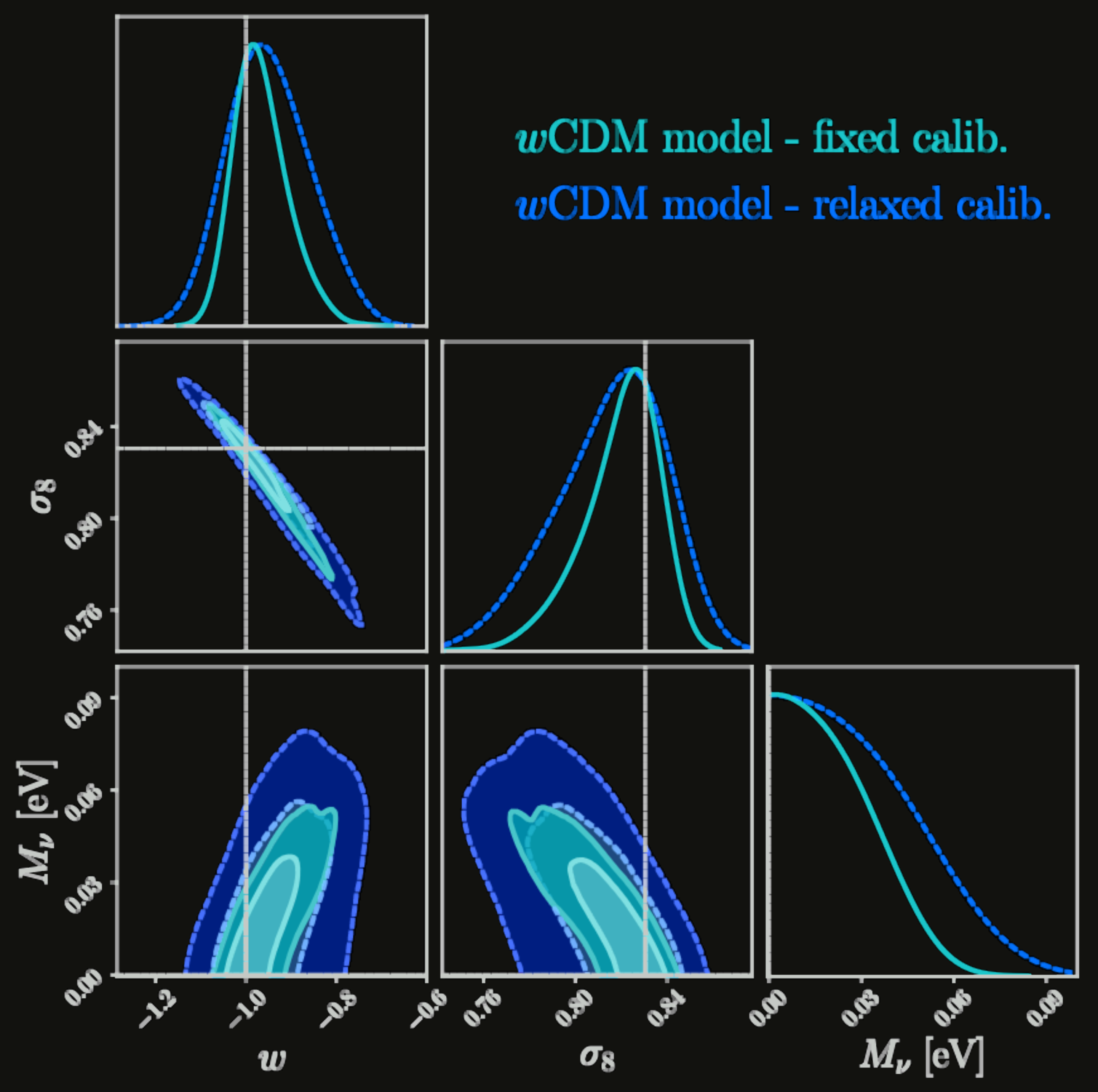
"Euclid" spec-z galaxies



[Kreisch+2021]



[Bayer+2021]

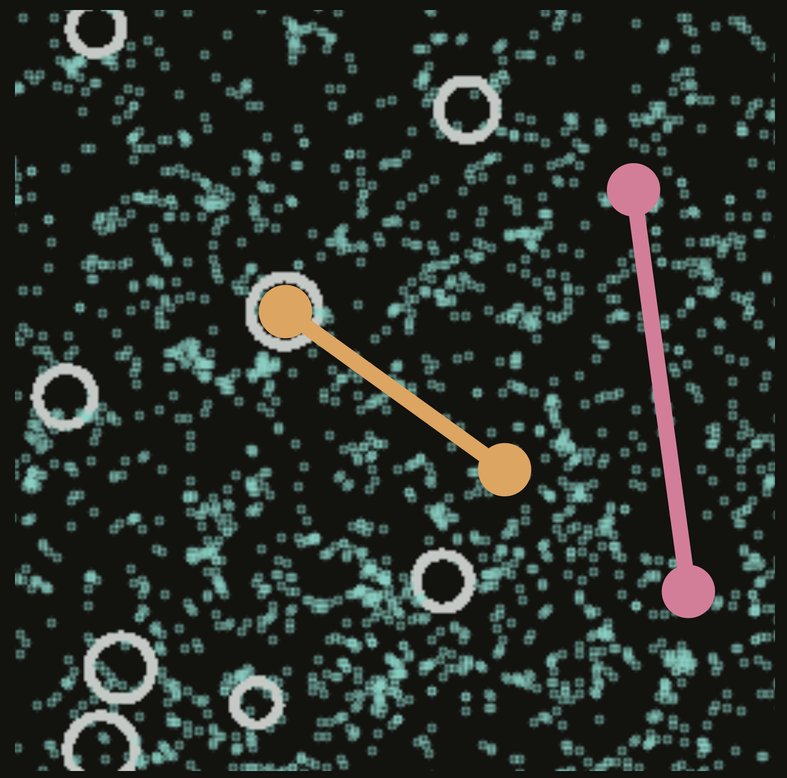


[Contarini+2021]

→ take this to real data from BOSS,
including complications from galaxy formation & survey systematics

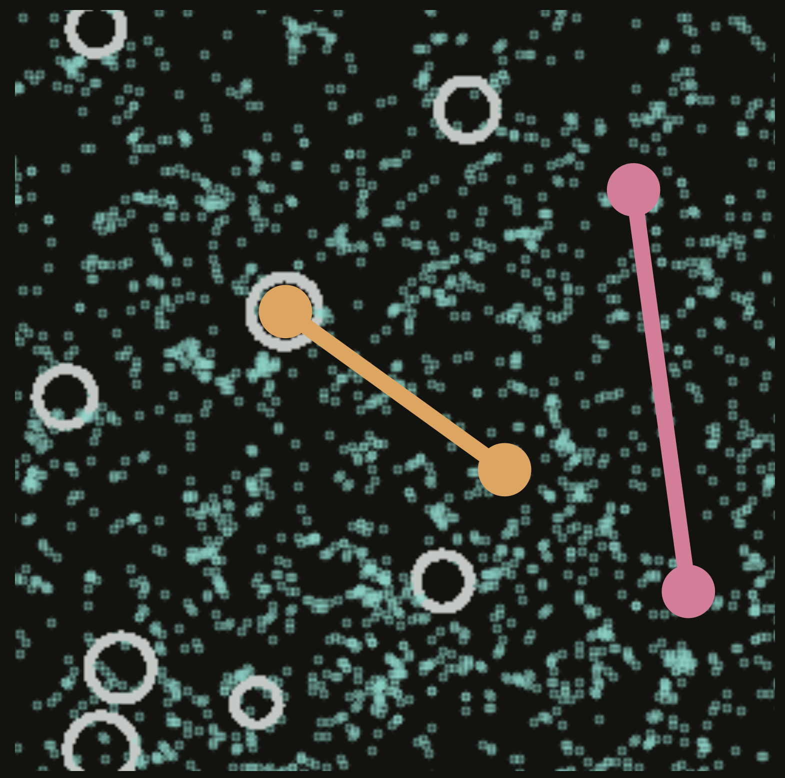
Simulation-based inference

- Want to constrain neutrino mass sum, $\sum m_\nu$, with BOSS data:
  - galaxy auto power spectrum
  - void size function (histogram of void sizes)
  - void galaxy cross power spectrum (void profile)

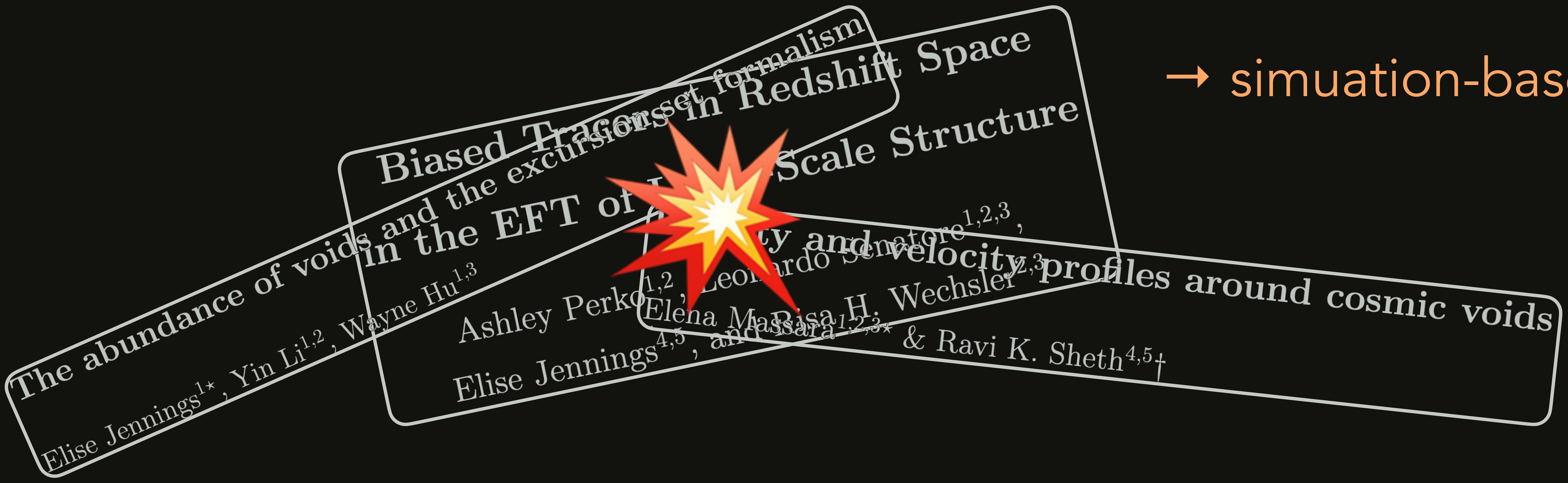


Simulation-based inference

- Want to constrain neutrino mass sum, $\sum m_\nu$, with BOSS data:
  - galaxy auto power spectrum
  - void size function (histogram of void sizes)
  - void galaxy cross power spectrum (void profile)



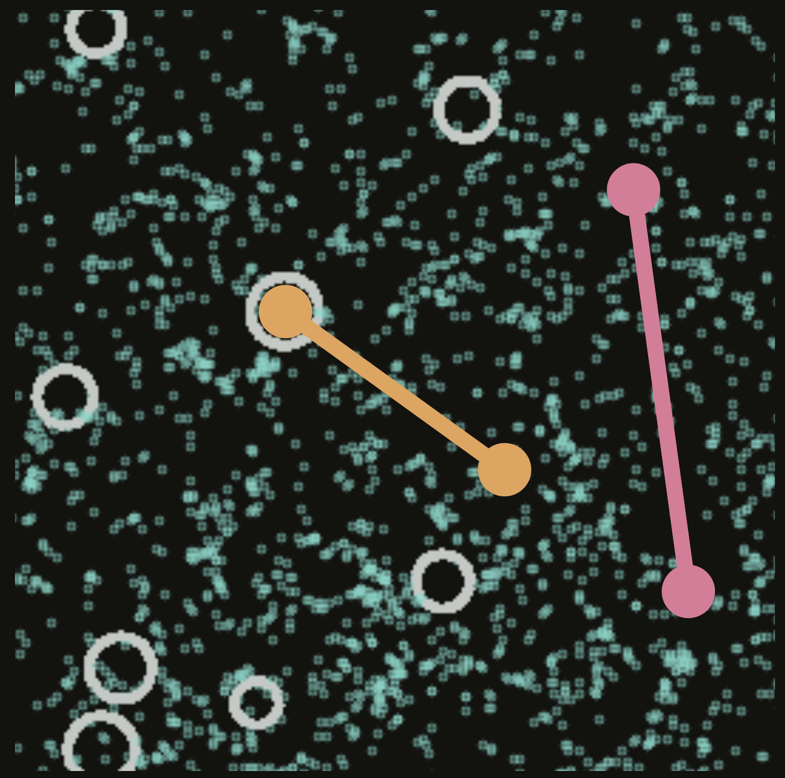
- Joint modeling of these statistics difficult with analytic methods



→ simuation-based

Simulation-based inference

- Want to constrain neutrino mass sum, $\sum m_\nu$, with BOSS data:
  - galaxy auto power spectrum
  - void size function (histogram of void sizes)
  - void galaxy cross power spectrum (void profile)



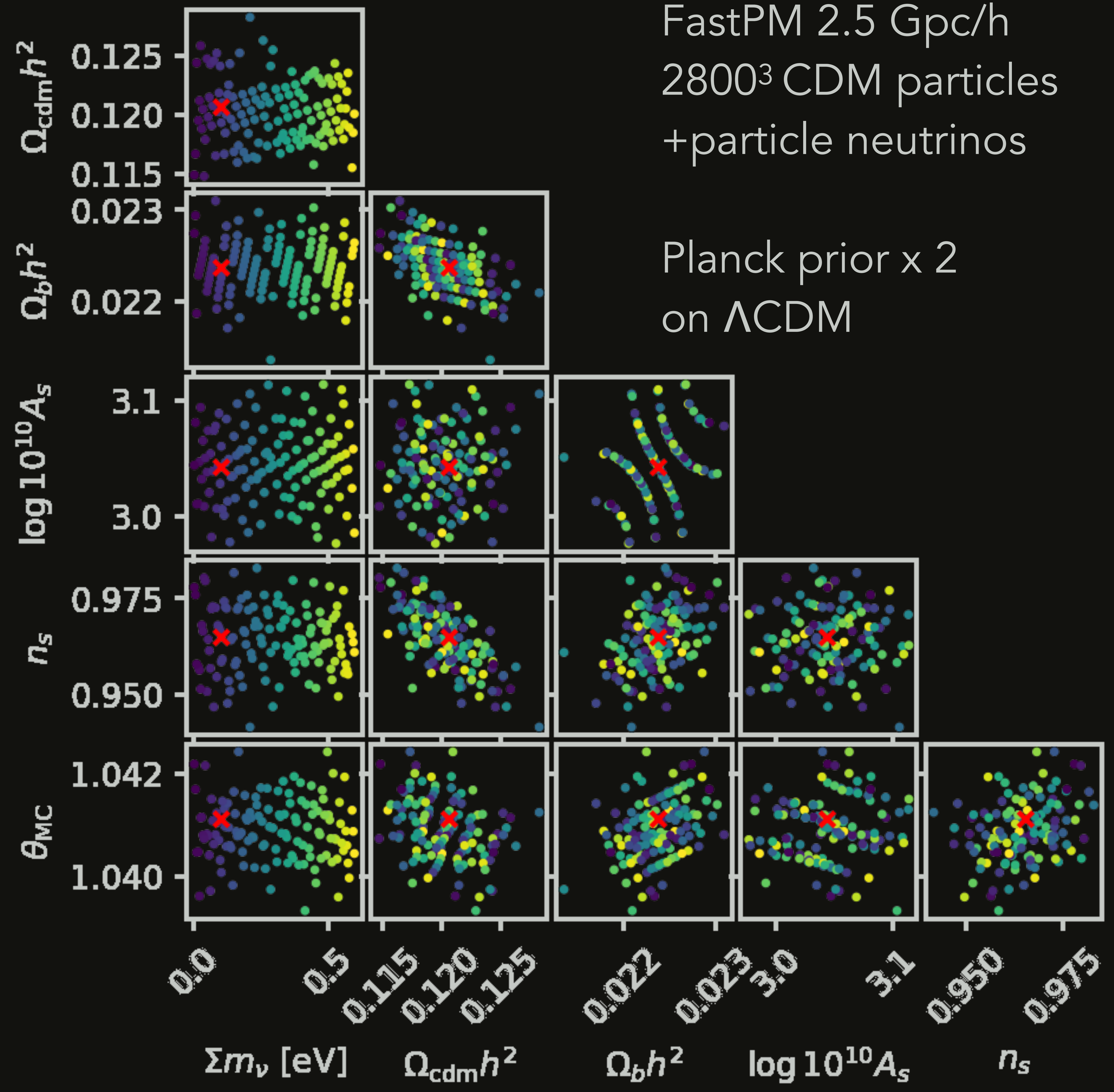
- Joint modeling of these statistics difficult with analytic methods

- Likelihood unknown

$$-2\log p(x\,|\,\theta) \quad (x-\mu(\theta))^T\Sigma^{-1}(x-\mu(\theta))$$

→ implicit-likelihood

Simulations



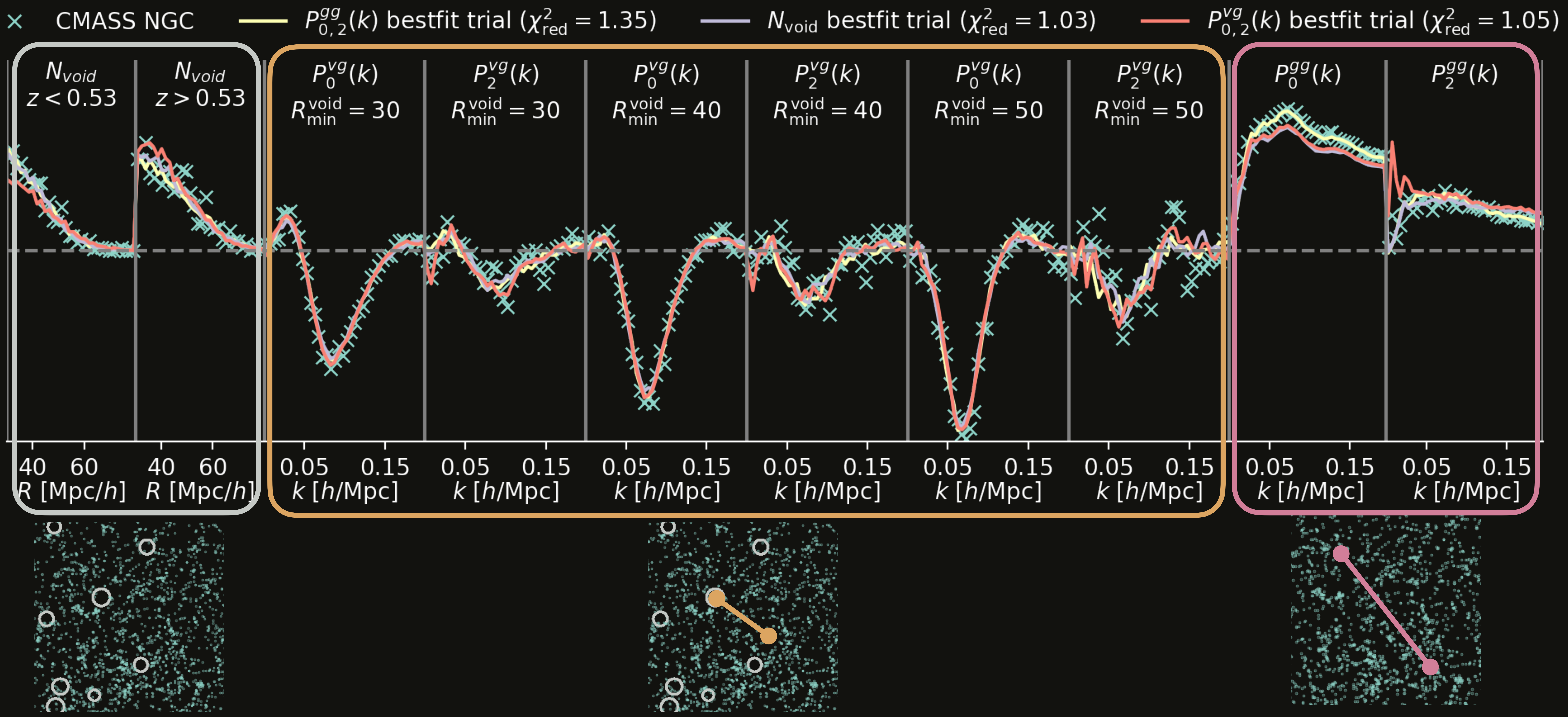
FastPM 2.5 Gpc/h
2800^3 CDM particles
+particle neutrinos

Planck prior x 2
on ΛCDM

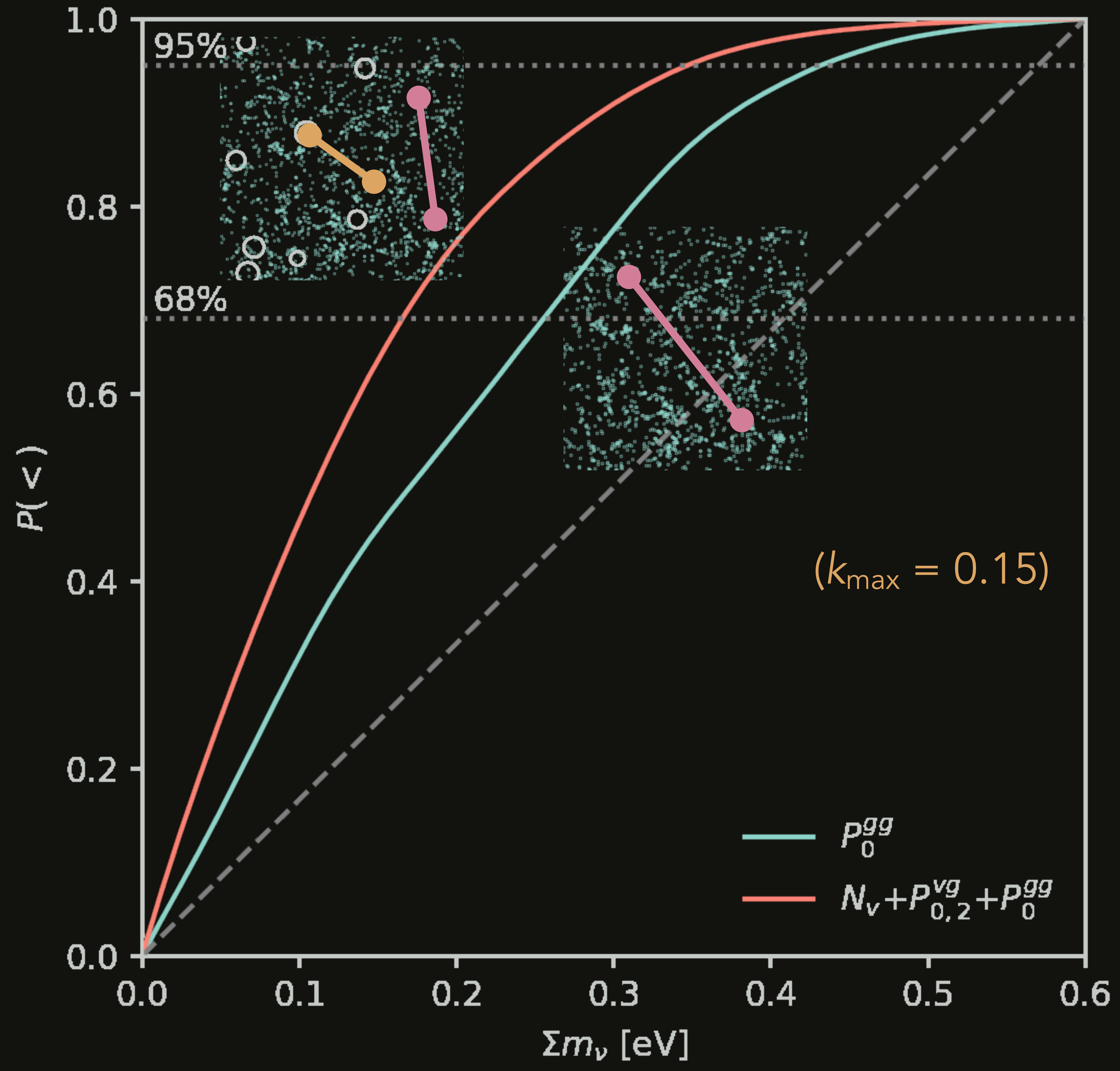
- populate gravity-only simulations with galaxies using HOD
- project on lightcone and add survey realism

Data Vector

Use MOPED compression to reduce dimensionality.



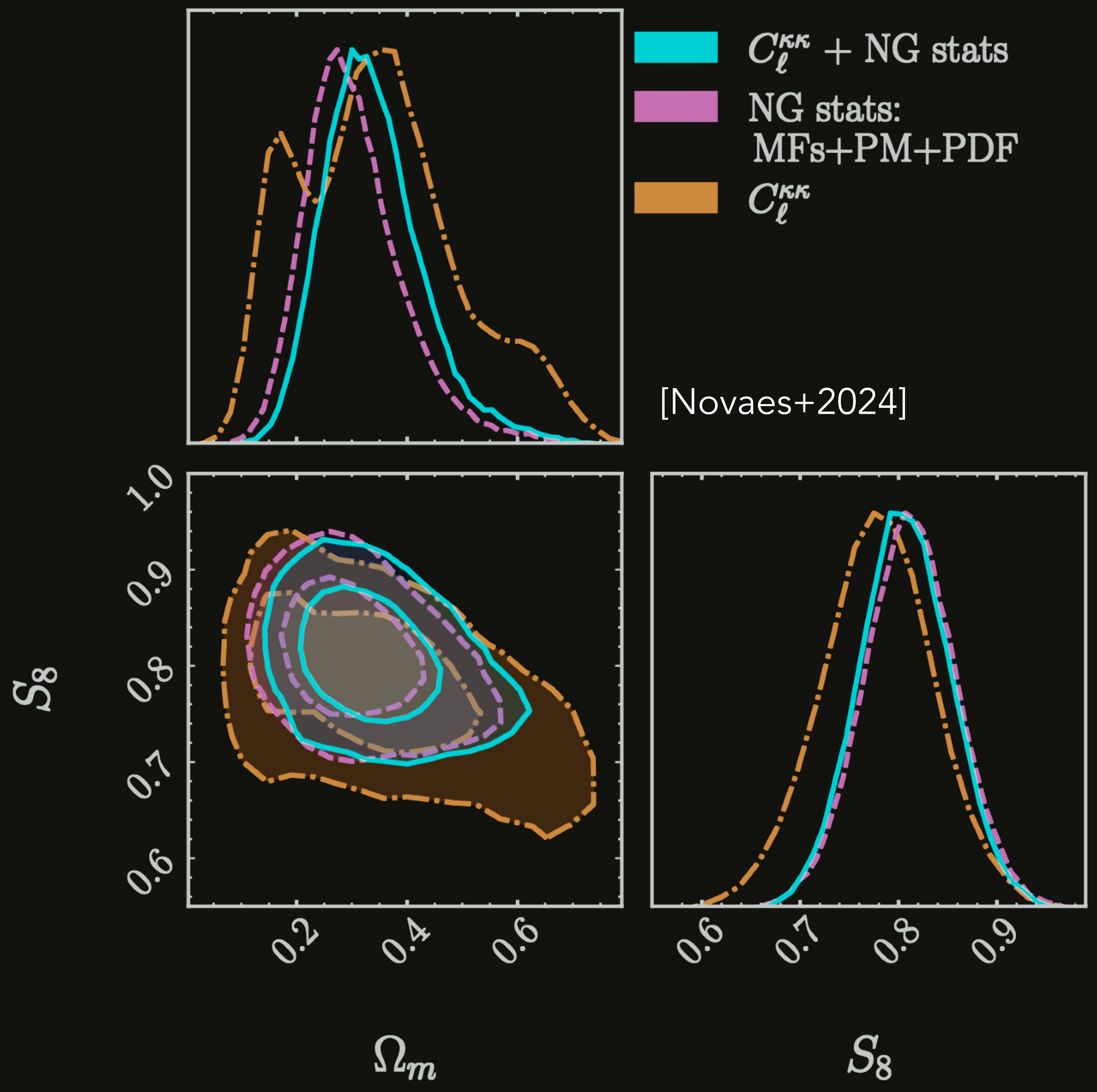
Main posterior



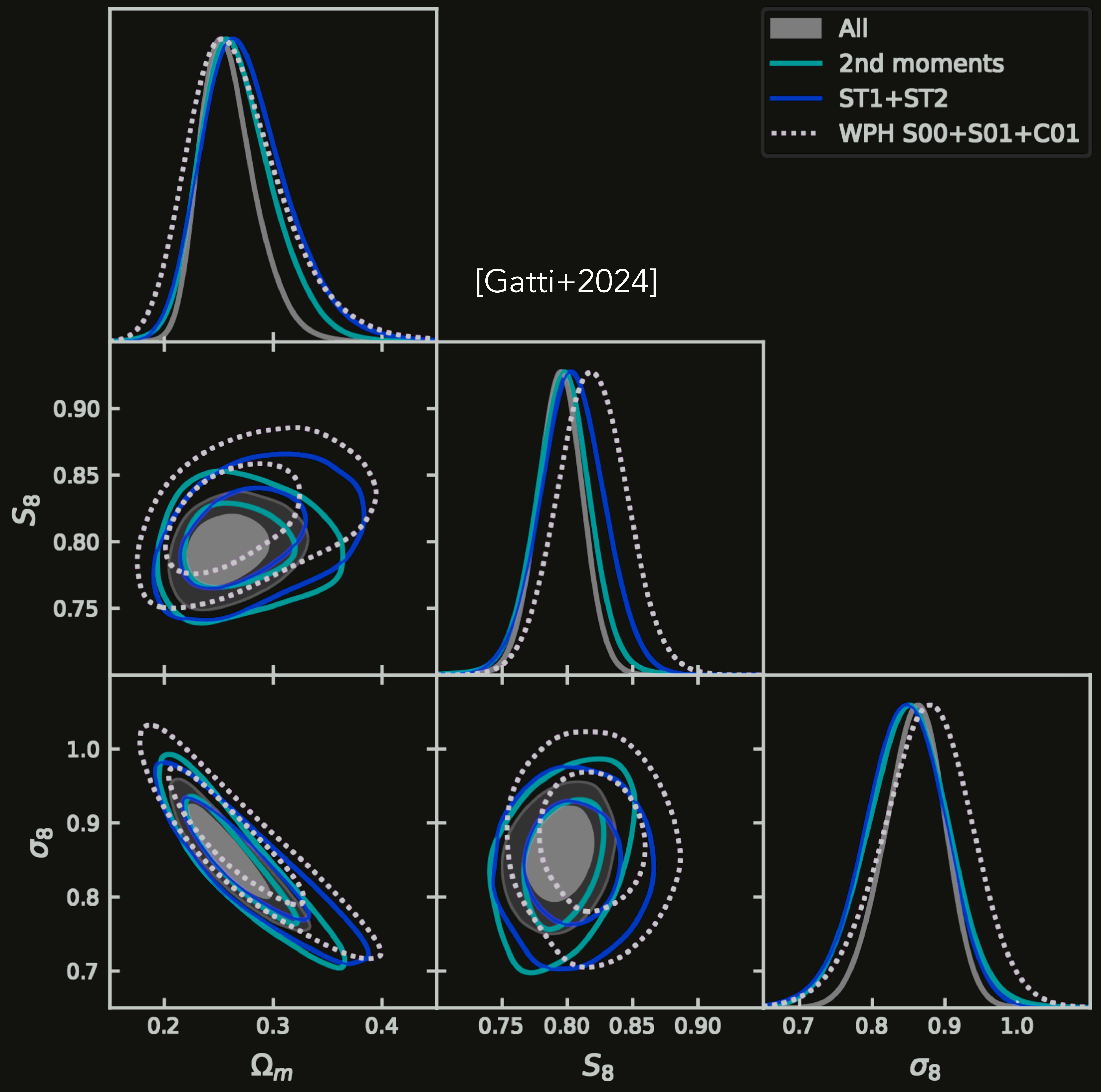
With conservative scale cut of k_{max}=0.15 hMpc^{-1}, voids tighten upper bound on neutrino mass.

ILI applied in weak lensing

HSC Y1



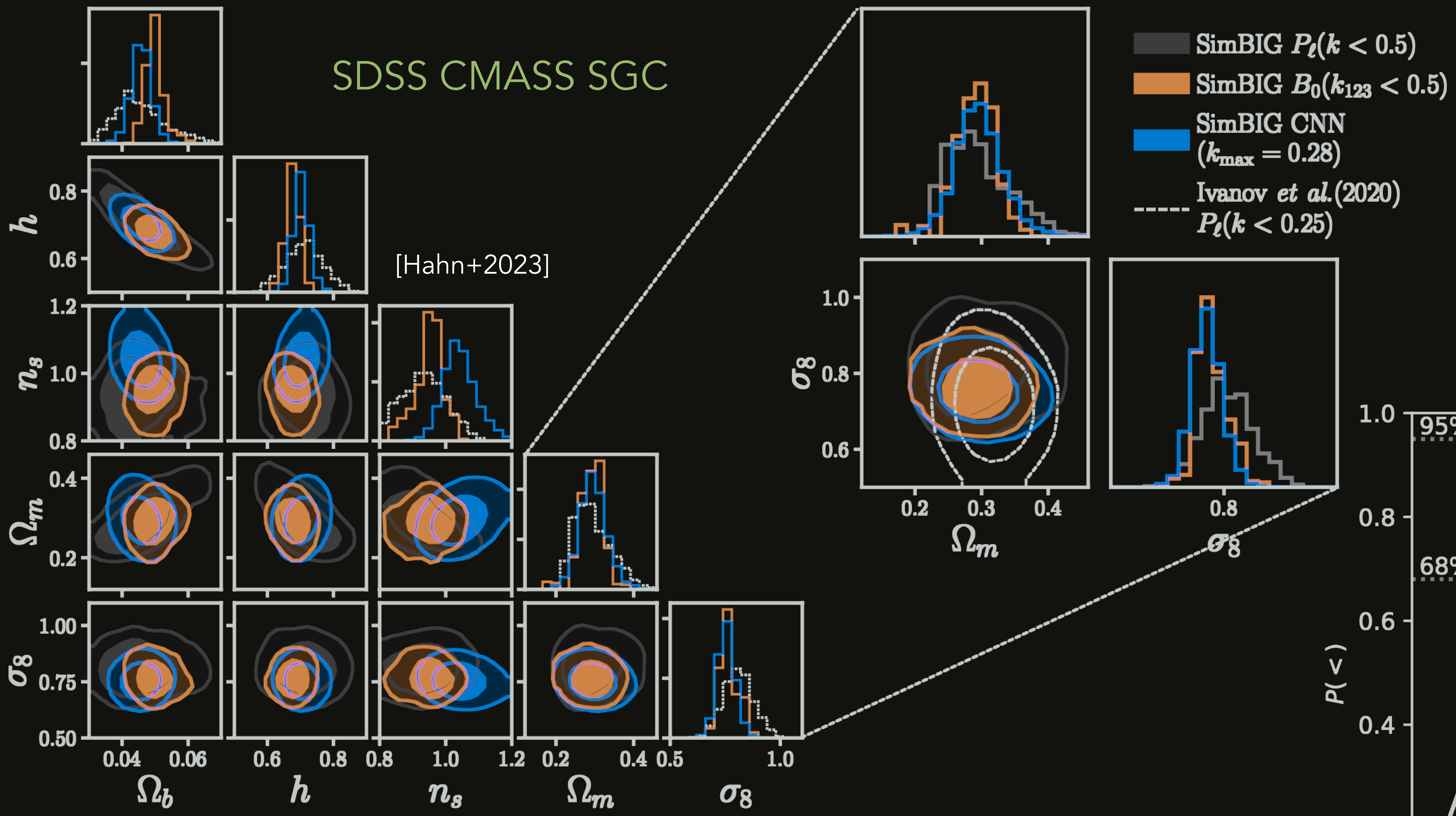
DES Y3

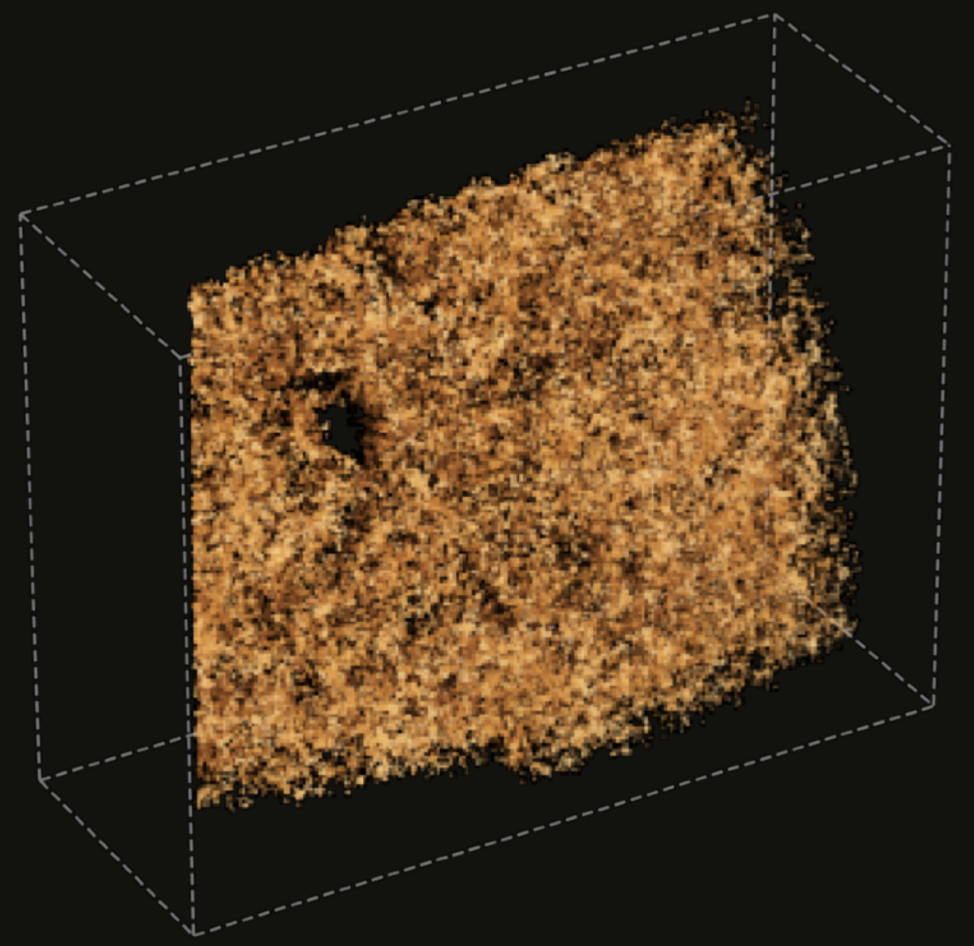


ILI applied in galaxy clustering

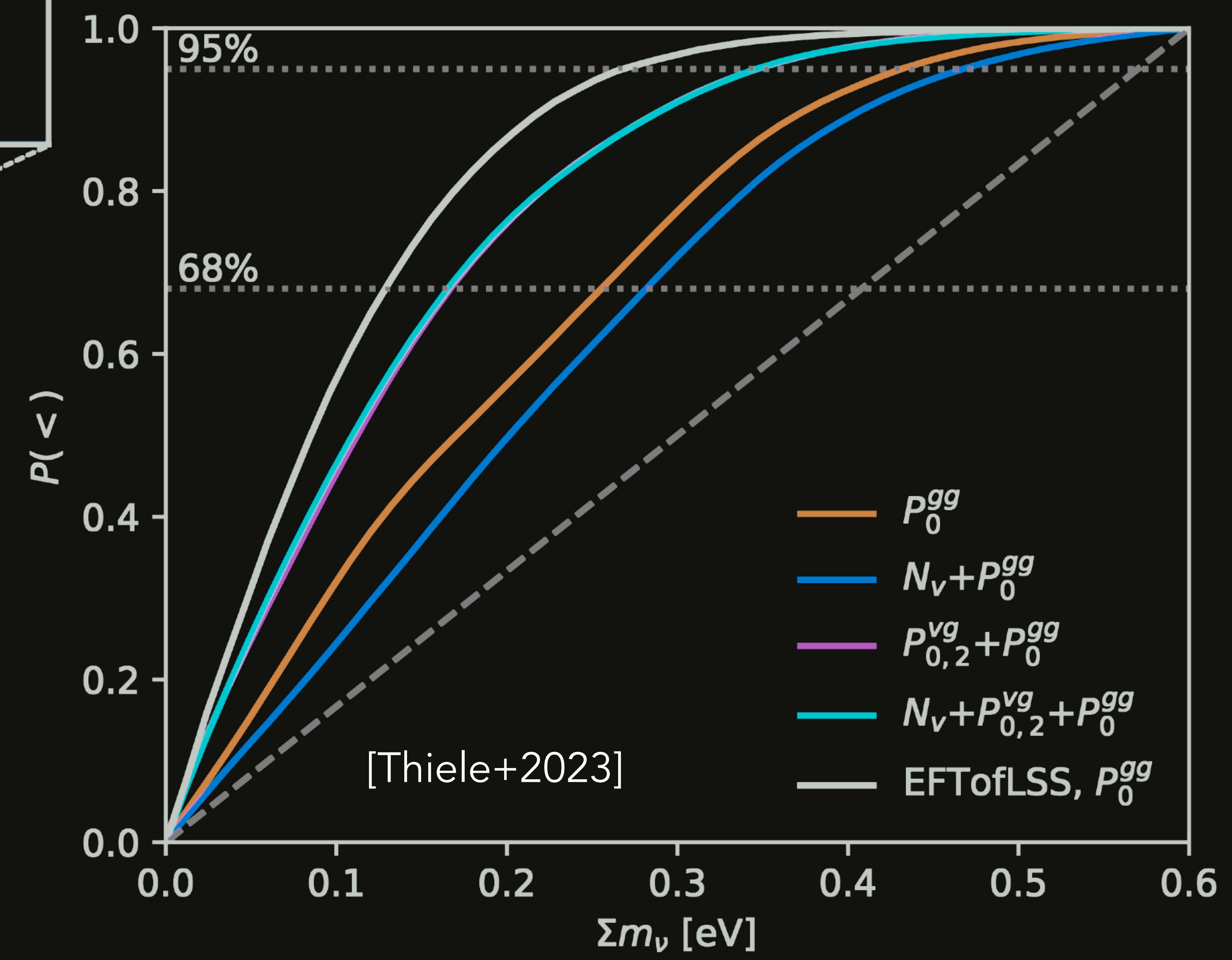
SDSS CMASS SGC

[Hahn+2023]





SDSS CMASS NGC voids



So far, more or less “toy examples”.

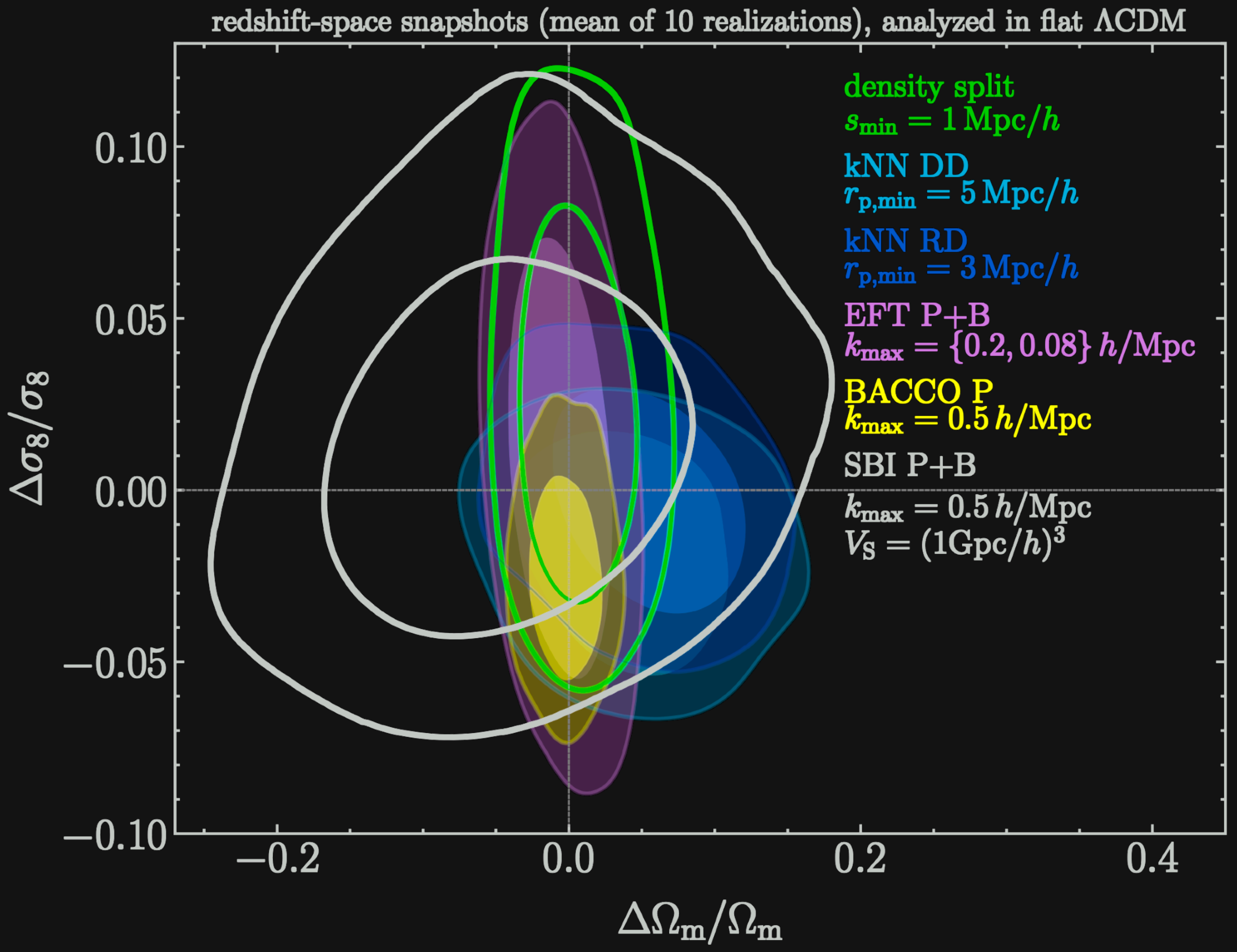
Not the same level of trust as traditional analyses.

A Parameter-Masked Mock Data Challenge for Beyond-Two-Point Galaxy Clustering Statistics*

The Beyond-2pt Collaboration

Elisabeth Krause,[1] Yosuke Kobayashi,[1,2] Andrés N. Salcedo,[1] Mikhail M. Ivanov,[3] Tom Abel,[4,5,6] Kazuyuki Akitsu,[7] Raul E. Angulo,[8,9] Giovanni Cabass,[10] Sofia Contarini,[11,12,13] Carolina Cuesta-Lazaro,[14,15,16] ChangHoon Hahn,[17] Nico Hamaus,[18,19] Donghui Jeong,[20,21] Chirag Modi,[22,23] Nhat-Minh Nguyen,[24,25] Takahiro Nishimichi,[2,26,27] Enrique Paillas,[28,29] Marcos Pellejero Ibañez,[30] Oliver H. E. Philcox,[31,32] Alice Pisani,[33,22,34,17] Fabian Schmidt,[35] Satoshi Tanaka,[26] Giovanni Verza,[36,22] Sihan Yuan,[4,6] Matteo Zennaro,[37]

building trust



Issues away from the limit

- *in the limit*, any likelihood learnable
- what is the limit?
  - infinite model expressivity (usually ok in cosmology)
  - ability to find good global optimum (usually ok)
  - infinite training set size / fast & accurate simulation codes

$$L = \int_{p(x,\theta)} \mathscr{L} \approx \sum_{\text{training set}} \mathscr{L}_{\text{approx}}$$

Implicit Likelihood Inference in Crisis?

A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful

Joeri Hermans*
Unaffiliated

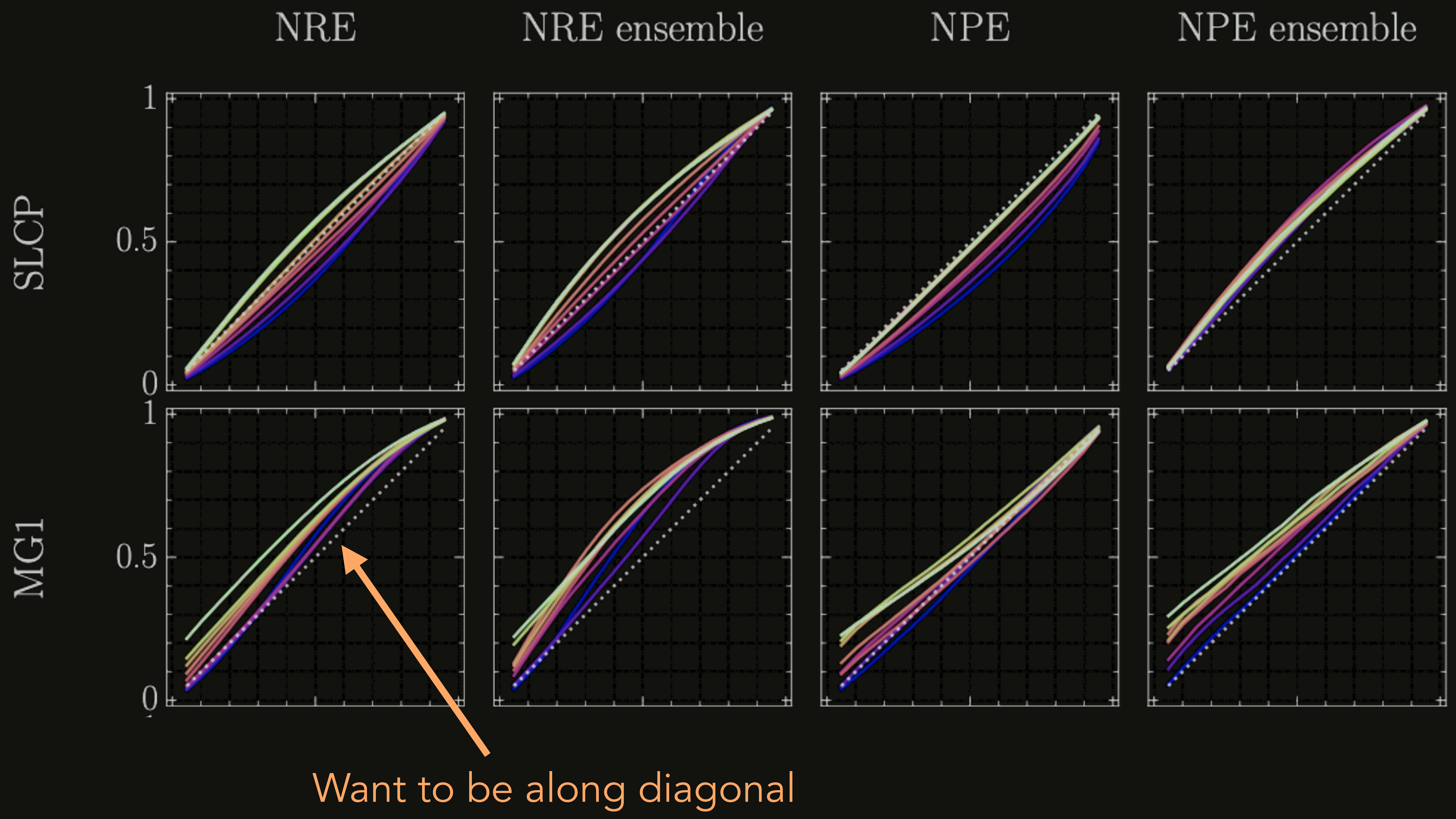
Arnaud Delaunoy*
University of Liège

François Rozet
University of Liège

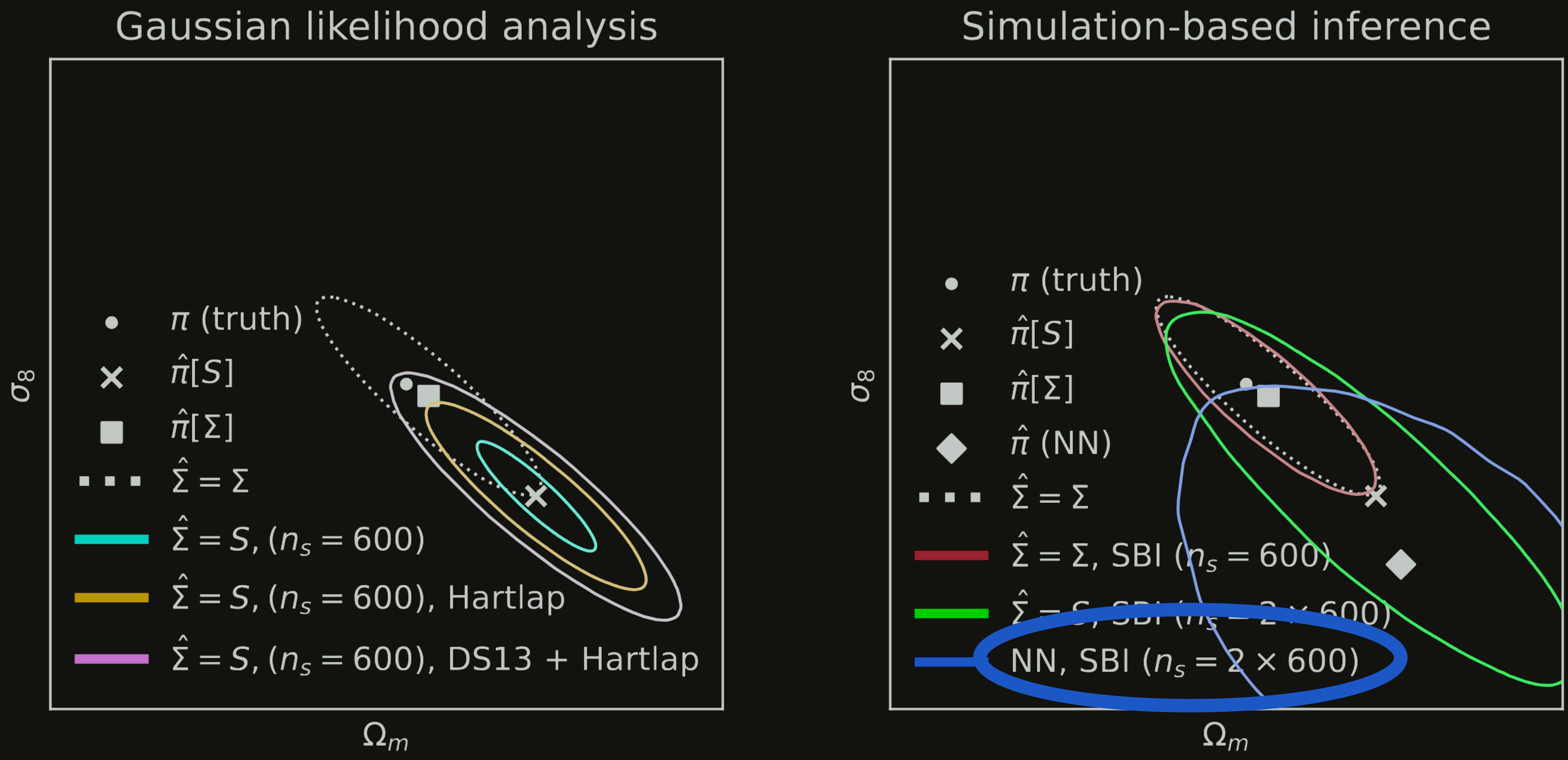
Antoine Wehenkel
University of Liège

Volodimir Begy
University of Vienna

Gilles Louppe
University of Liège

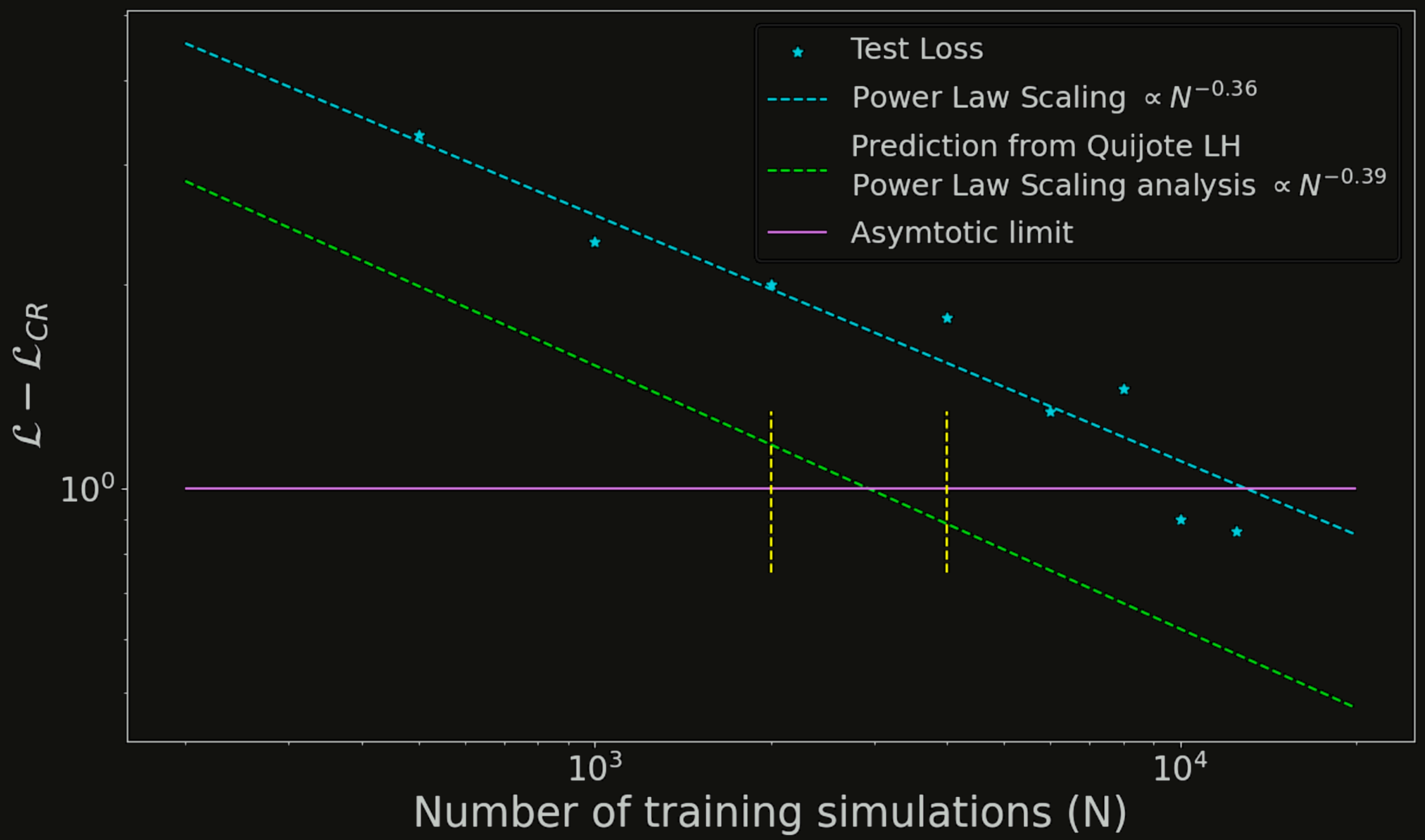


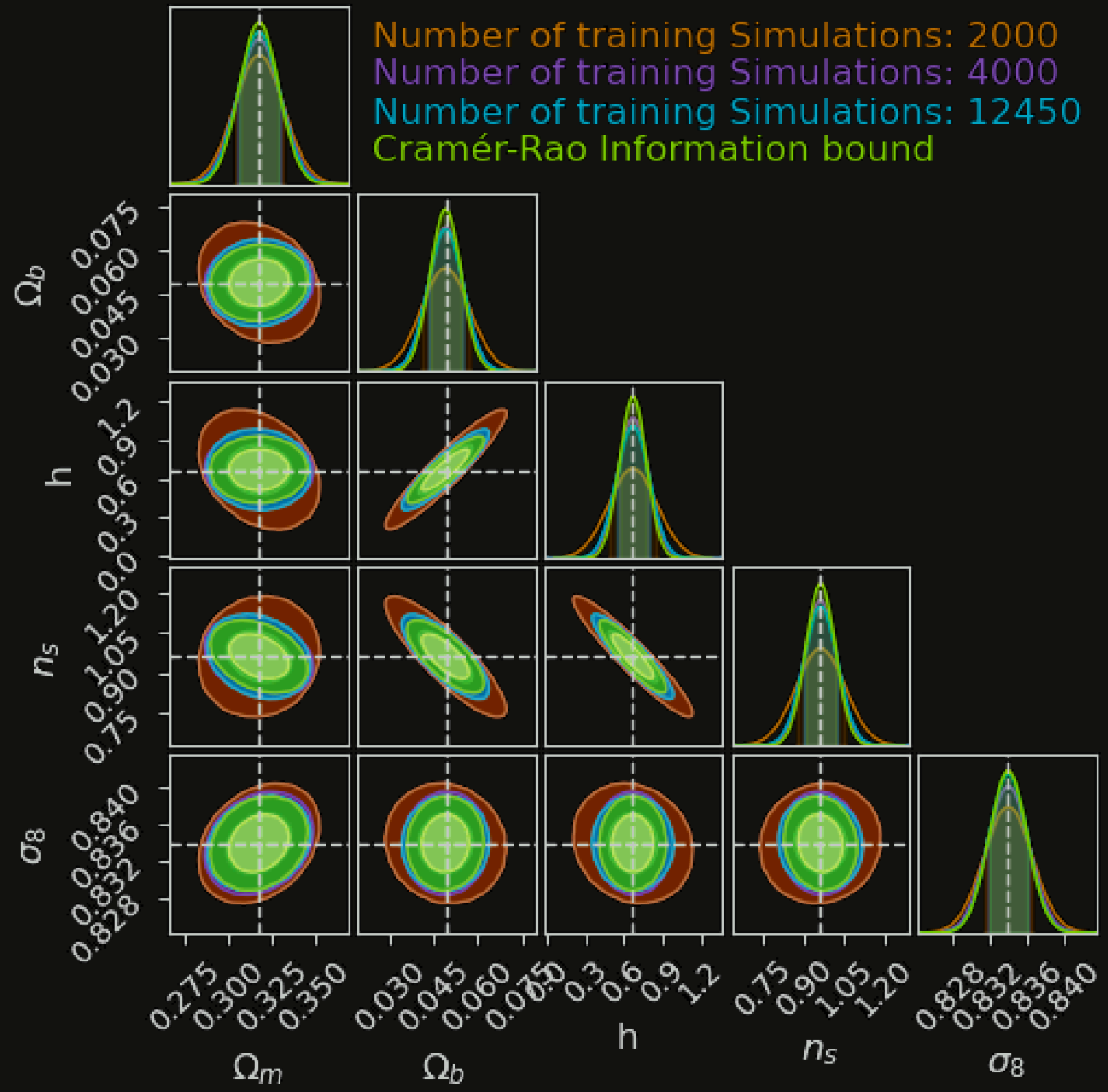
Implicit Likelihood Inference in Crisis?



[Homer, Friedrich, Gruen 2024]

Implicit Likelihood Inference in Crisis?





[Bairagi, Wandelt, Villaescusa-Navarro 2025]

So far, more or less “toy examples”.

Not the same level of trust as traditional analyses.

So far, more or less “toy examples”.

Not the same level of trust as traditional analyses. **For good reason!**

Stage-IV deluge of data will make the problem much more challenging…

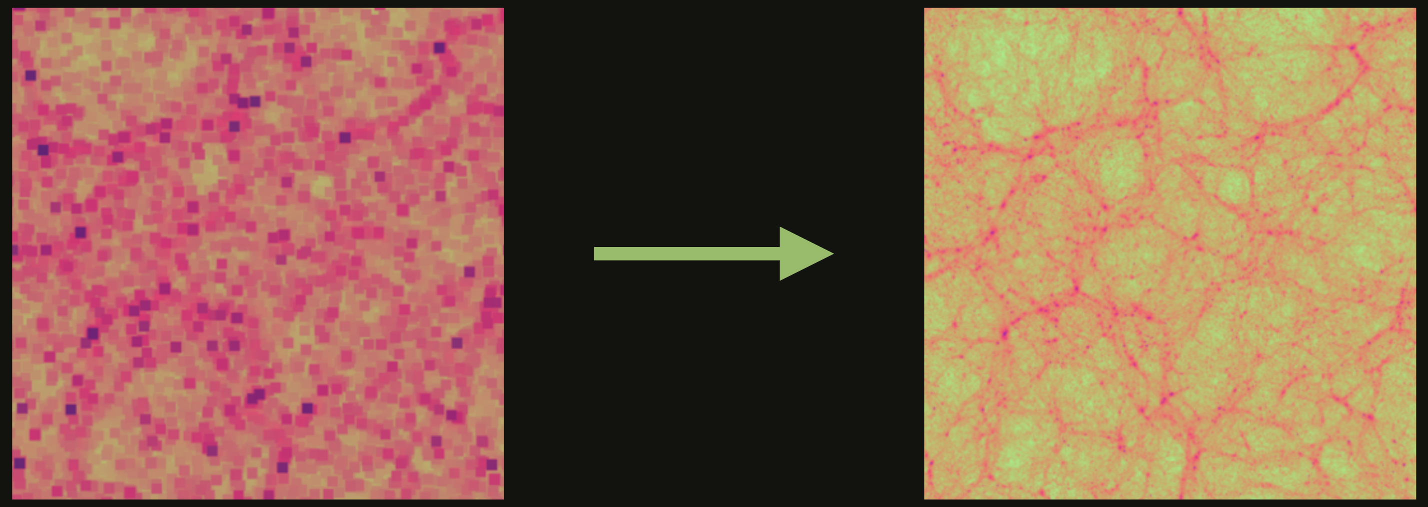
Now is the time to become clever:

- **— run simulations for cheaper**
- **— run them where it counts**
- **— combine simulations of different qualities**

Some promising approaches developed already!

Scaling up

- (ML) accelerated simulations

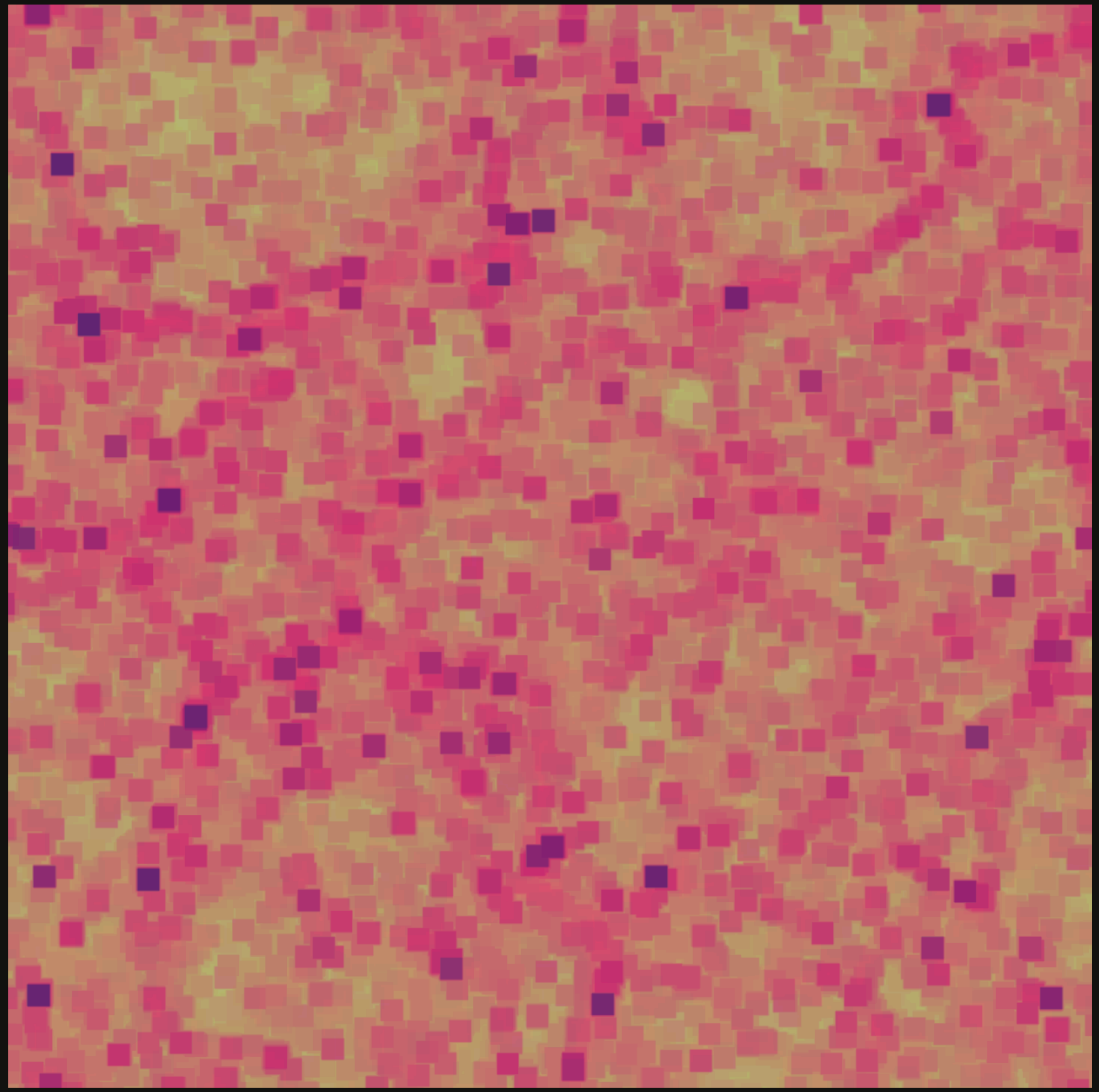


- "painting" into simulations [gas, galaxies, …]

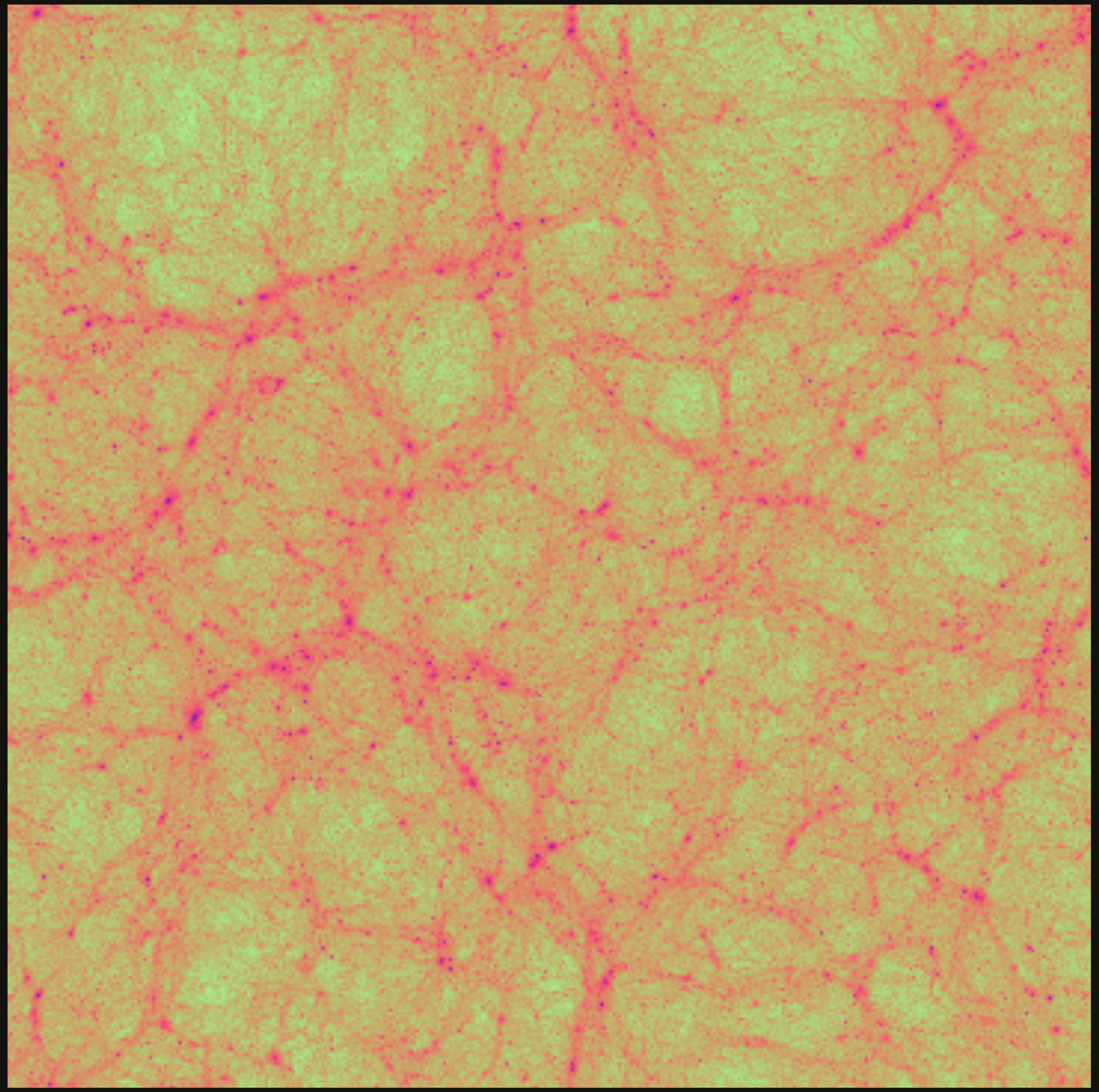
NB: This is most of the other activities in AI/ML for cosmology

I choose to present them in the SBI context, but there are certainly other motivations

Learning corrections

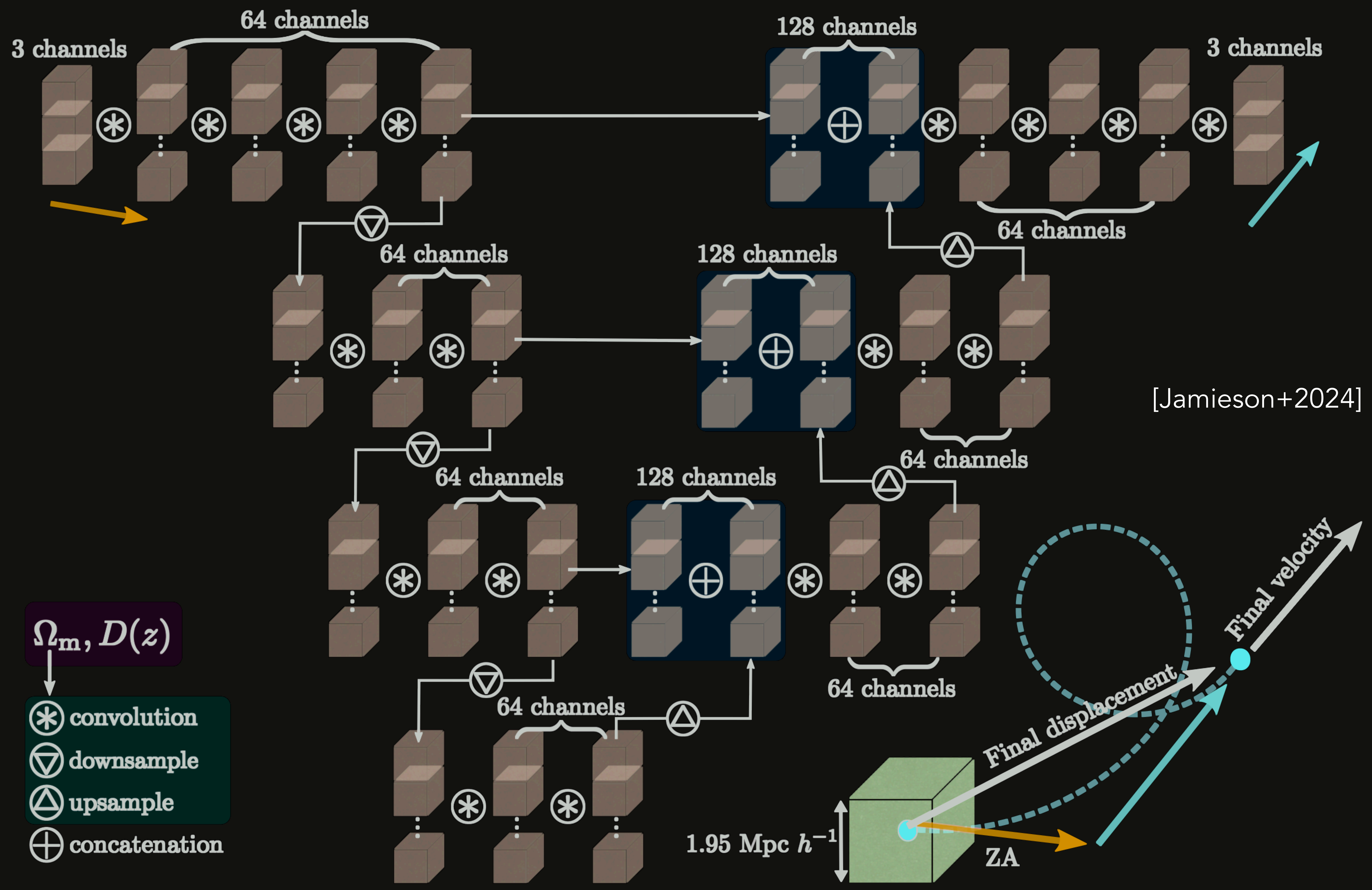


HODs, SAMs, …,
machine learning



Learning corrections

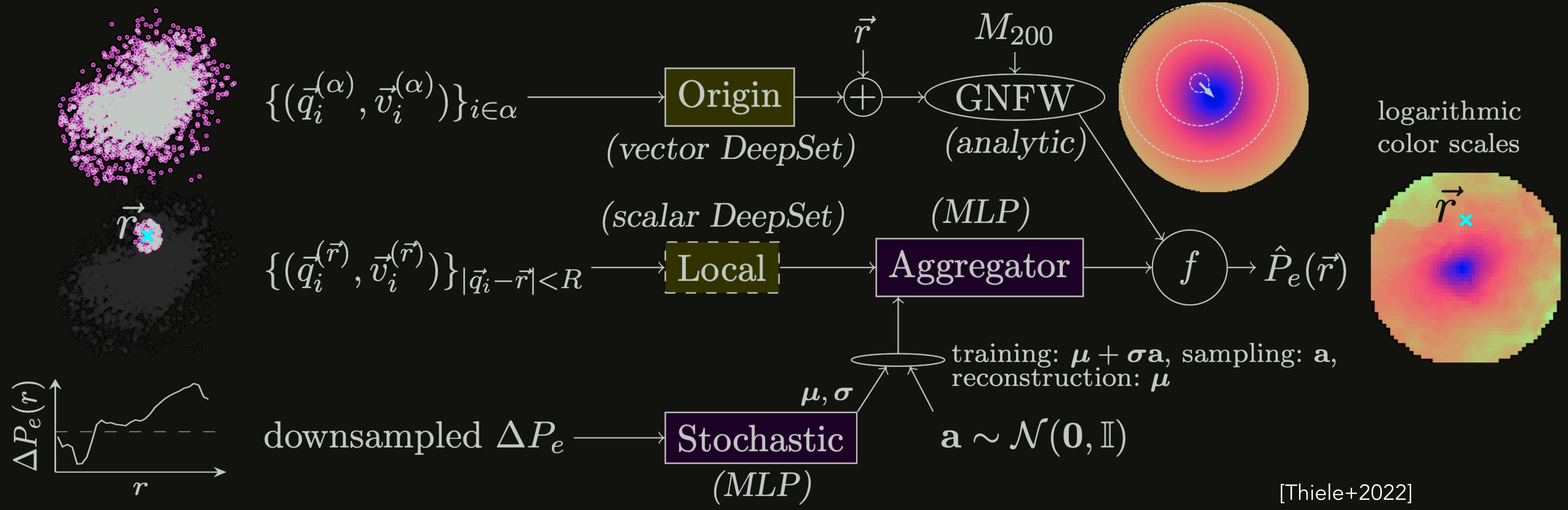
enhancing accuracy of gravity solvers



[Jamieson+2024]

Learning corrections

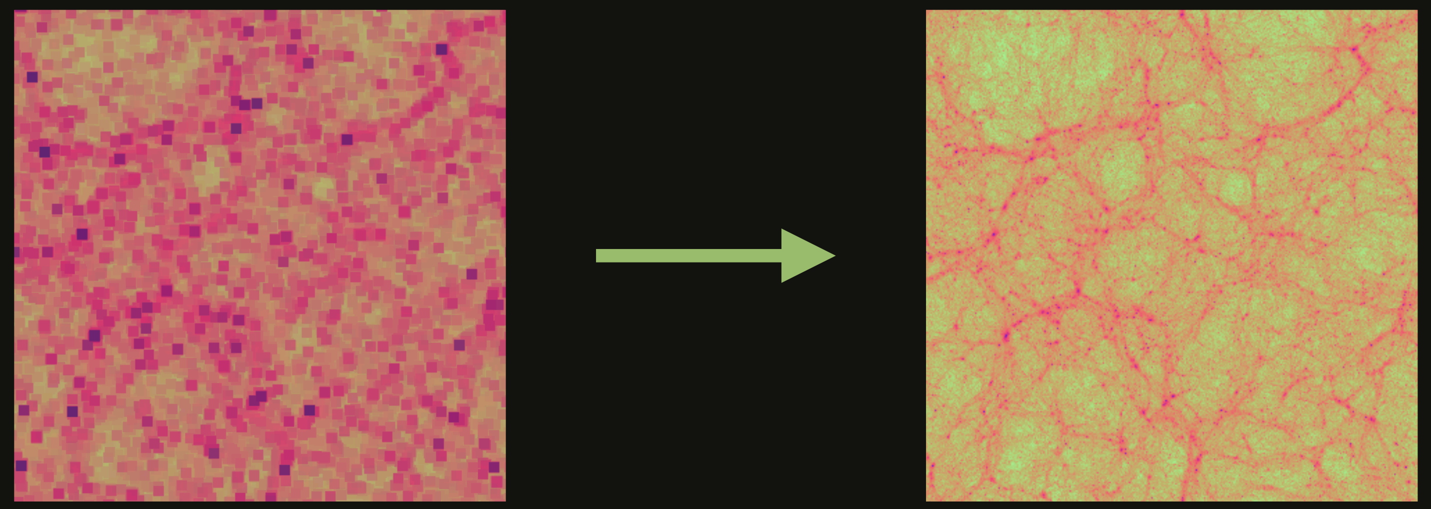
painting tSZ on gravity-only halos



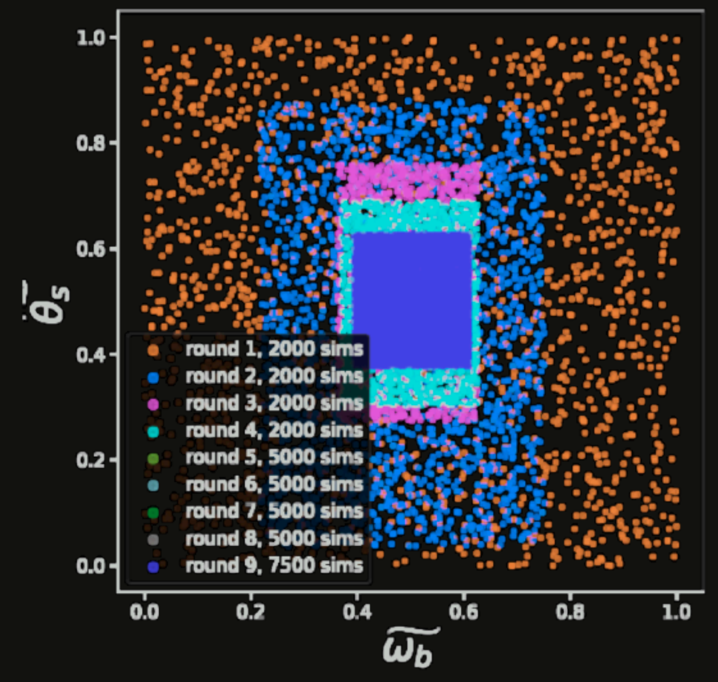
[Thiele+2022]

Scaling up

- (ML) accelerated simulations

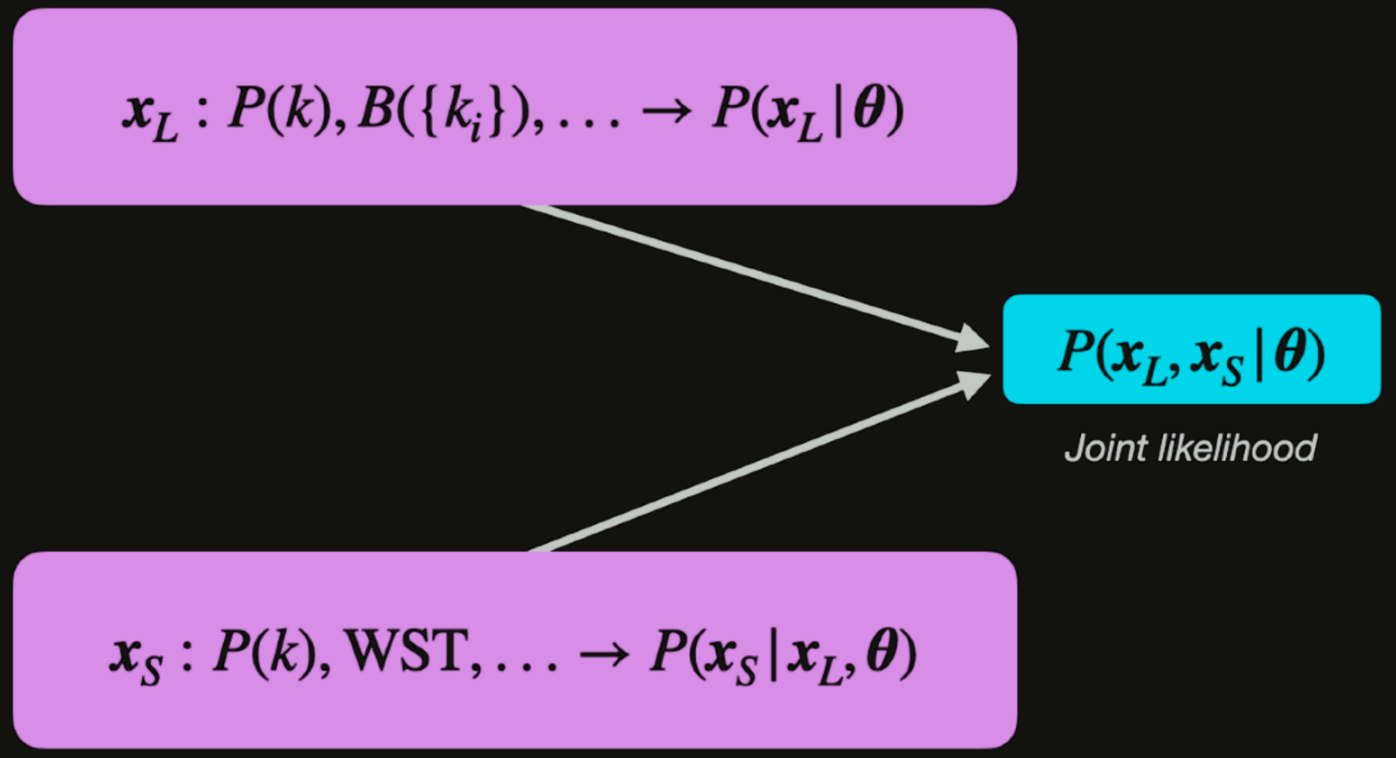


- “painting” into simulations [gas, galaxies, …]



- sequential inference

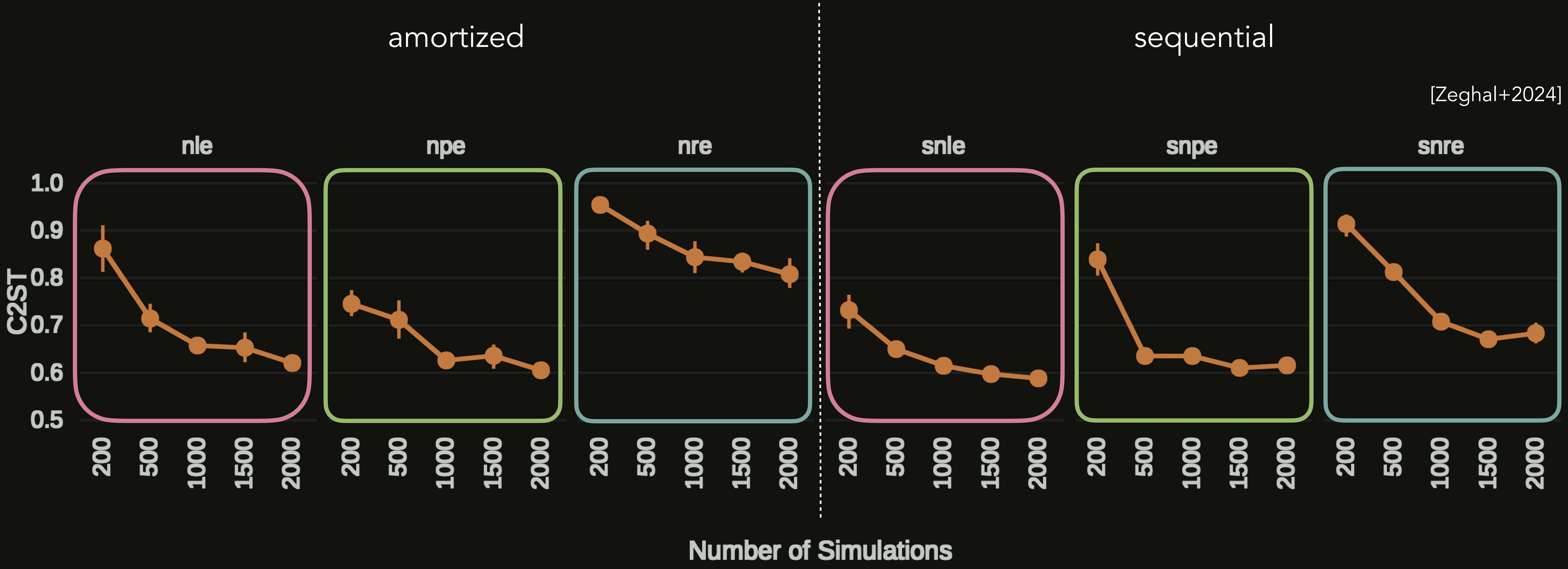
- hybrid analytic & SBI



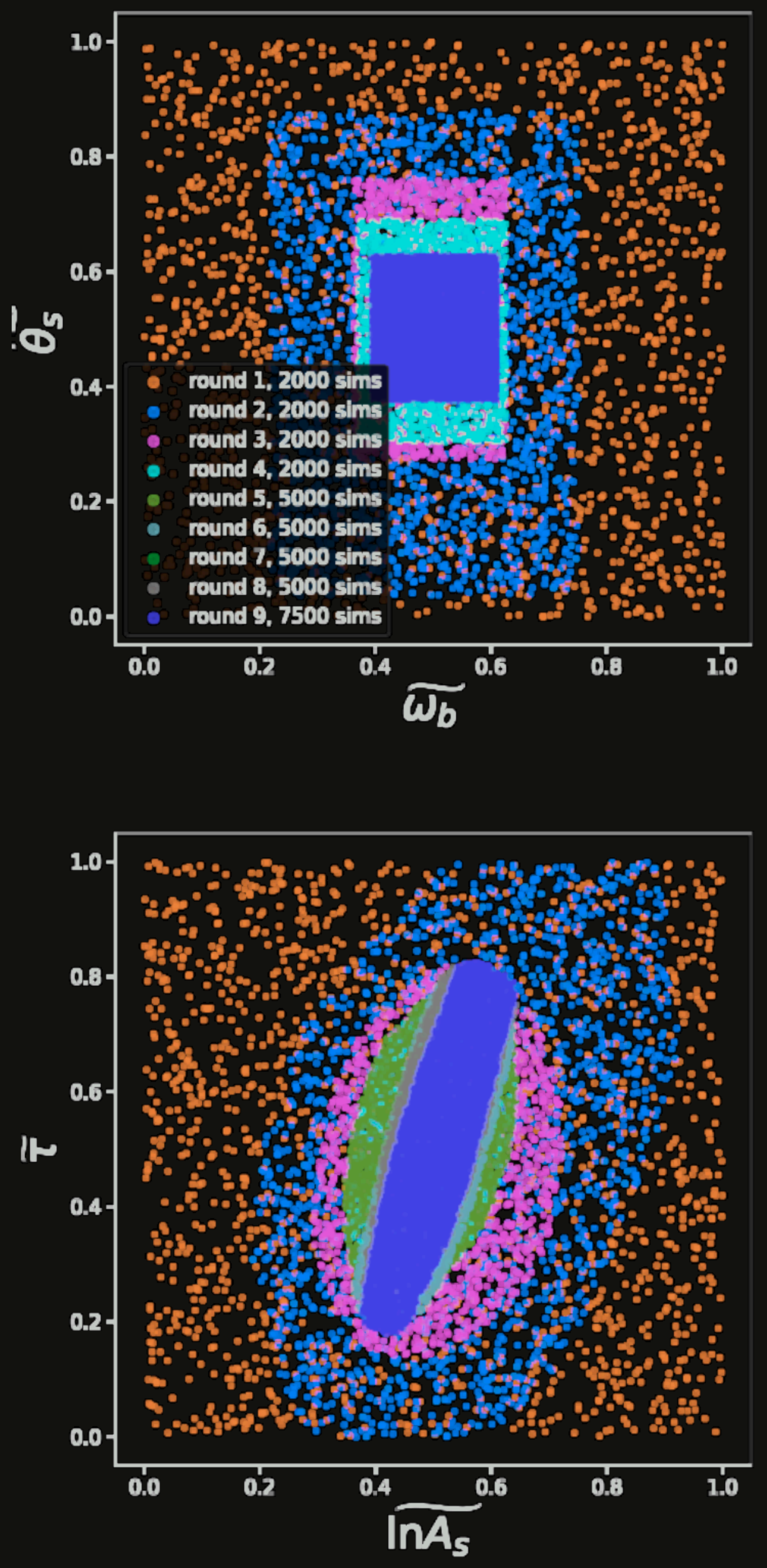
- multi-fidelity

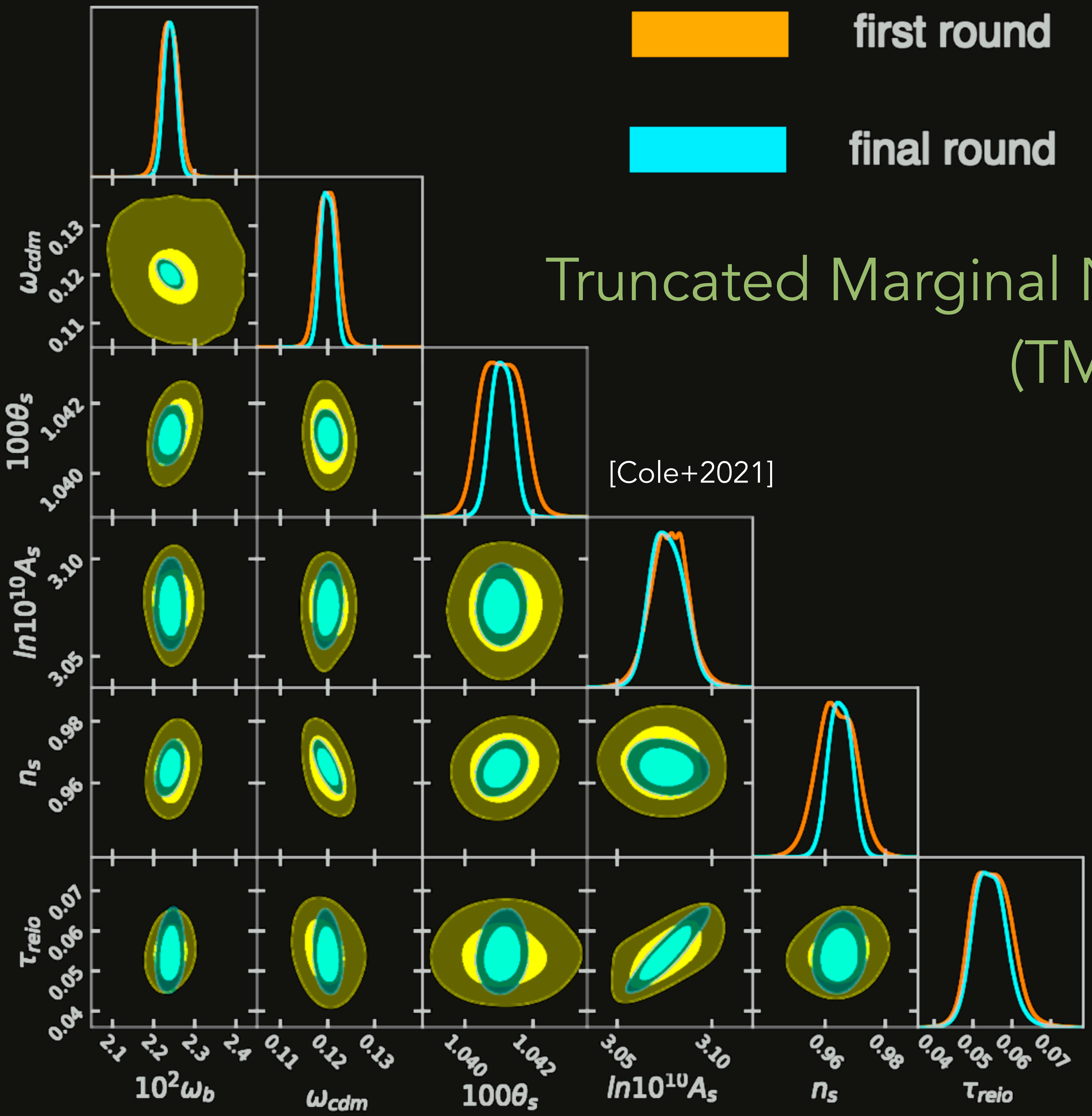
Sequential inference

representation learning only gets us so far…



Sequential inference





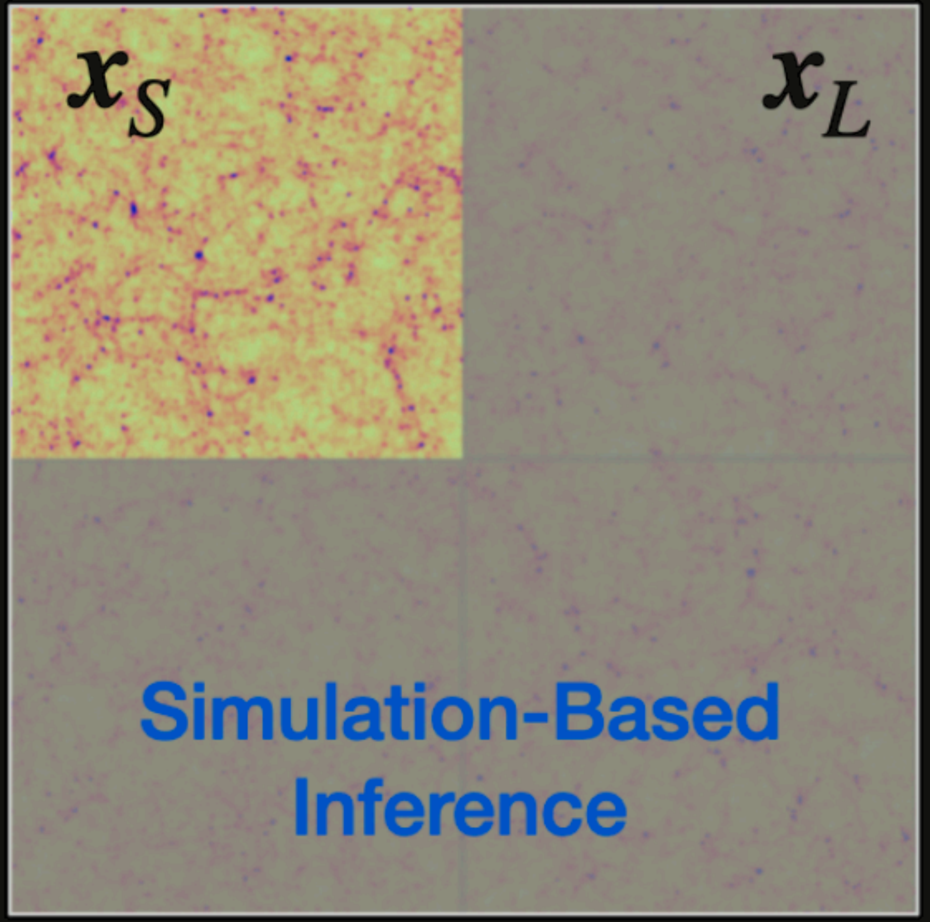
Truncated Marginal Neural Ratio Estimation (TMNRE)

[Cole+2021]

Hybrid ILI

$$P_L(k) + 2\int \frac{d\mathbf{p}}{(2\pi)^3}\left|F_2(\mathbf{p},\mathbf{k}-\mathbf{p})\right|^2 P_L(p)P_L(|\mathbf{k}-\mathbf{p}|) + 6\int \frac{d\mathbf{p}}{(2\pi)^3}F_3(\mathbf{p},-\mathbf{p},\mathbf{k})P_L(p)P_L(k) - 2c_s^2k^2P_L(k)$$

Perturbation Theory



Large scales are modeled analytically

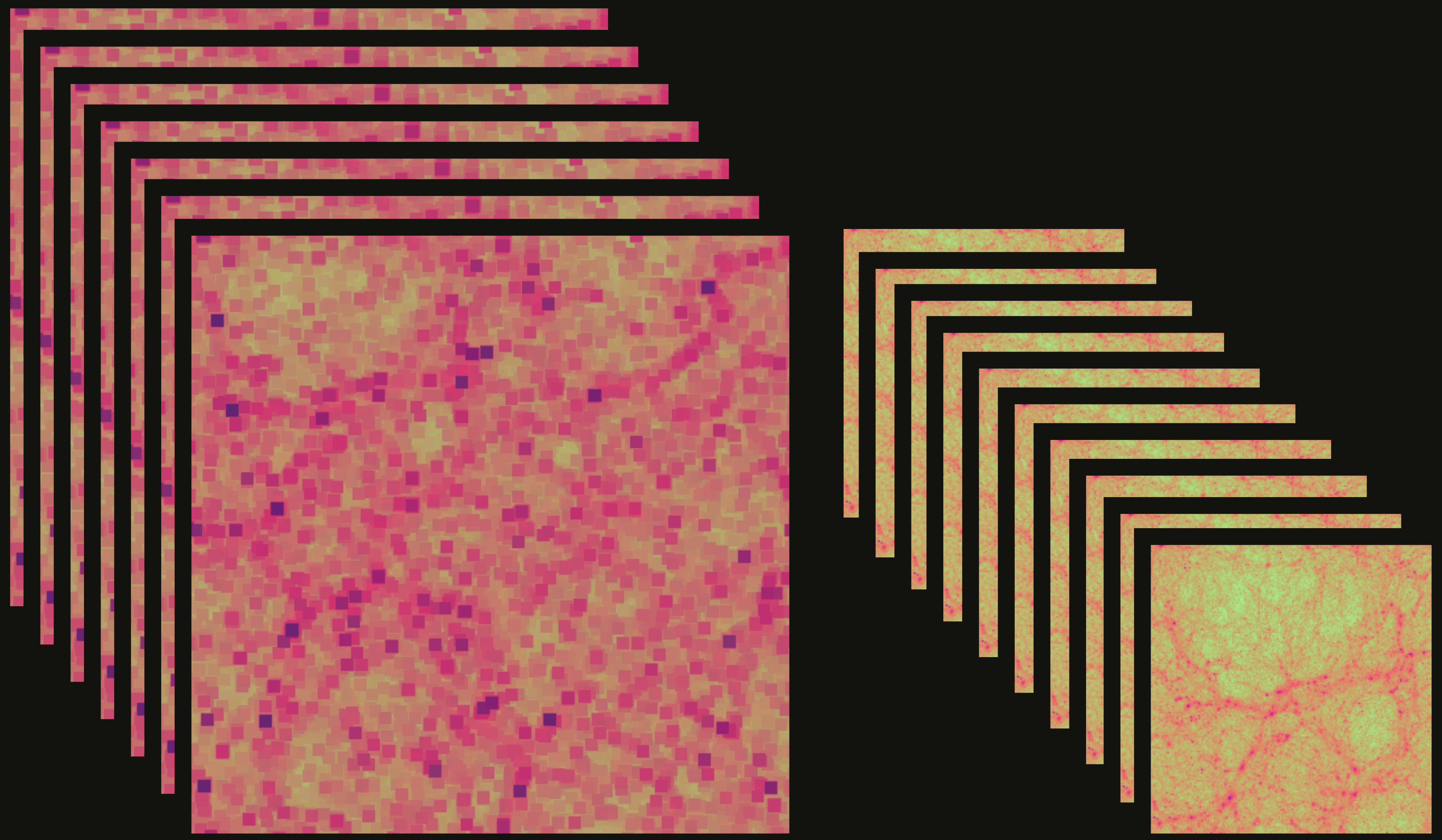
$$x_L : P(k), B(\{k_i\}), \ldots \rightarrow P(x_L|\theta)$$

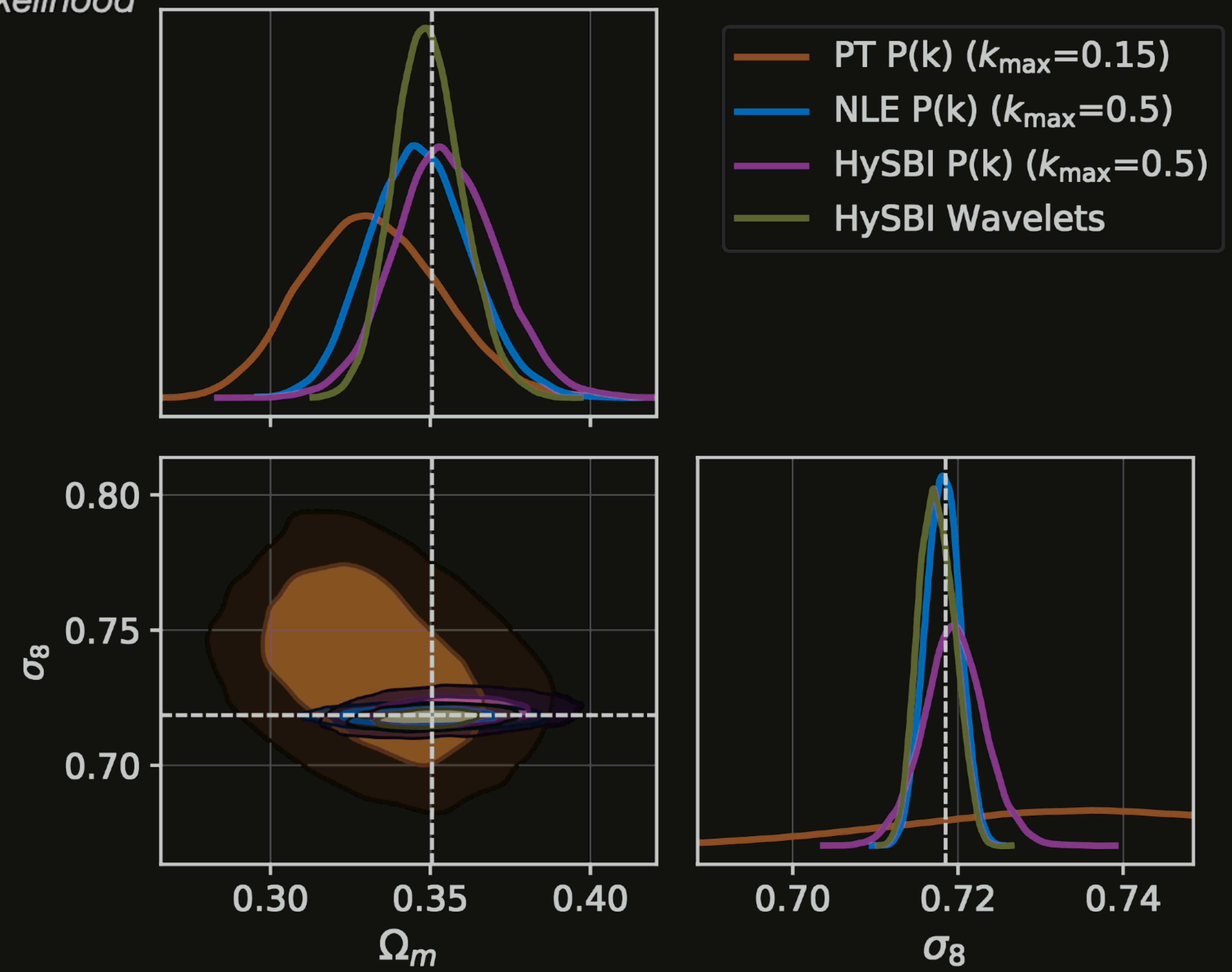
$$x_S : P(k), \mathrm{WST}, \ldots \rightarrow P(x_S|x_L,\theta)$$

Small scales are modeled numerically

$$P(x_L, x_S|\theta)$$

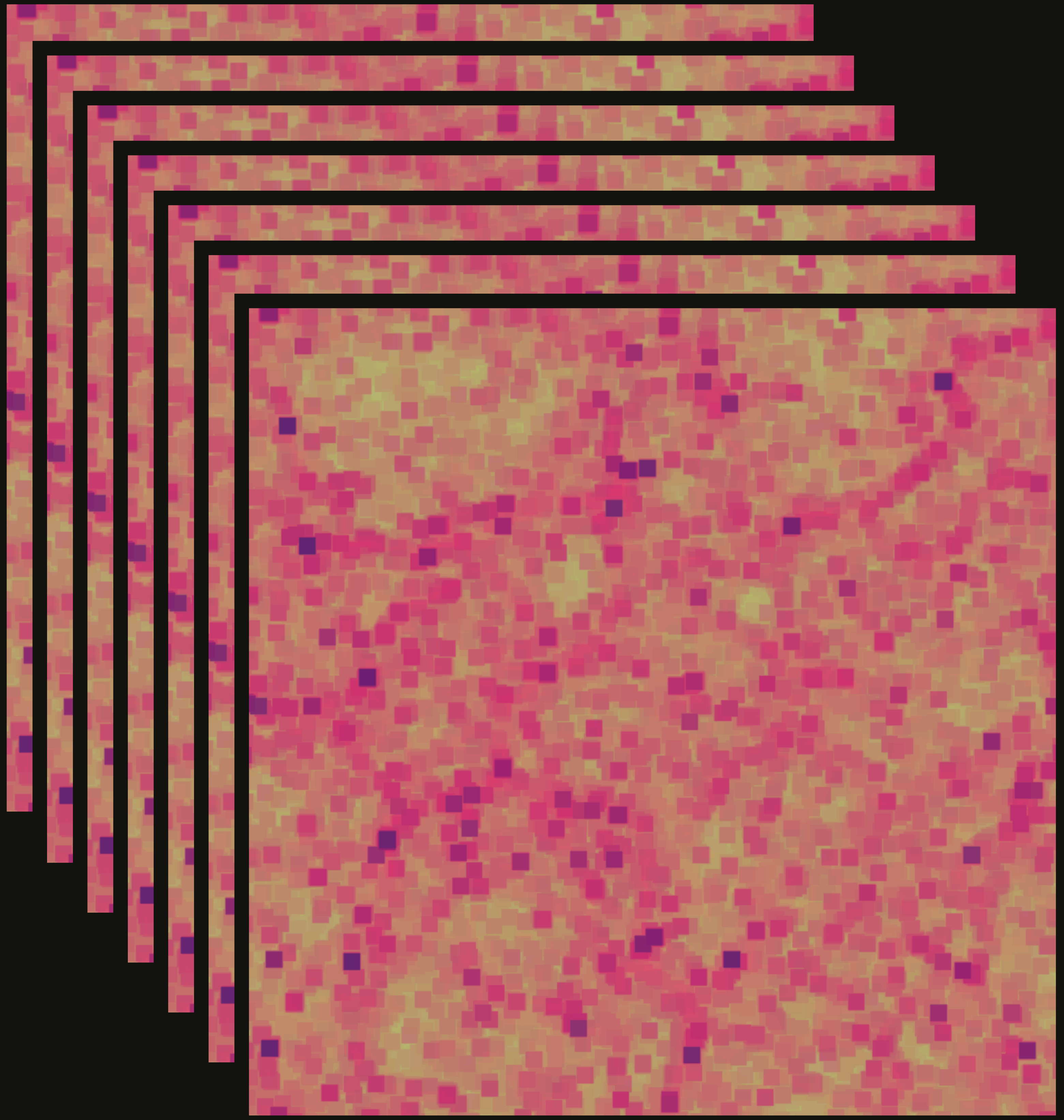
Joint likelihood

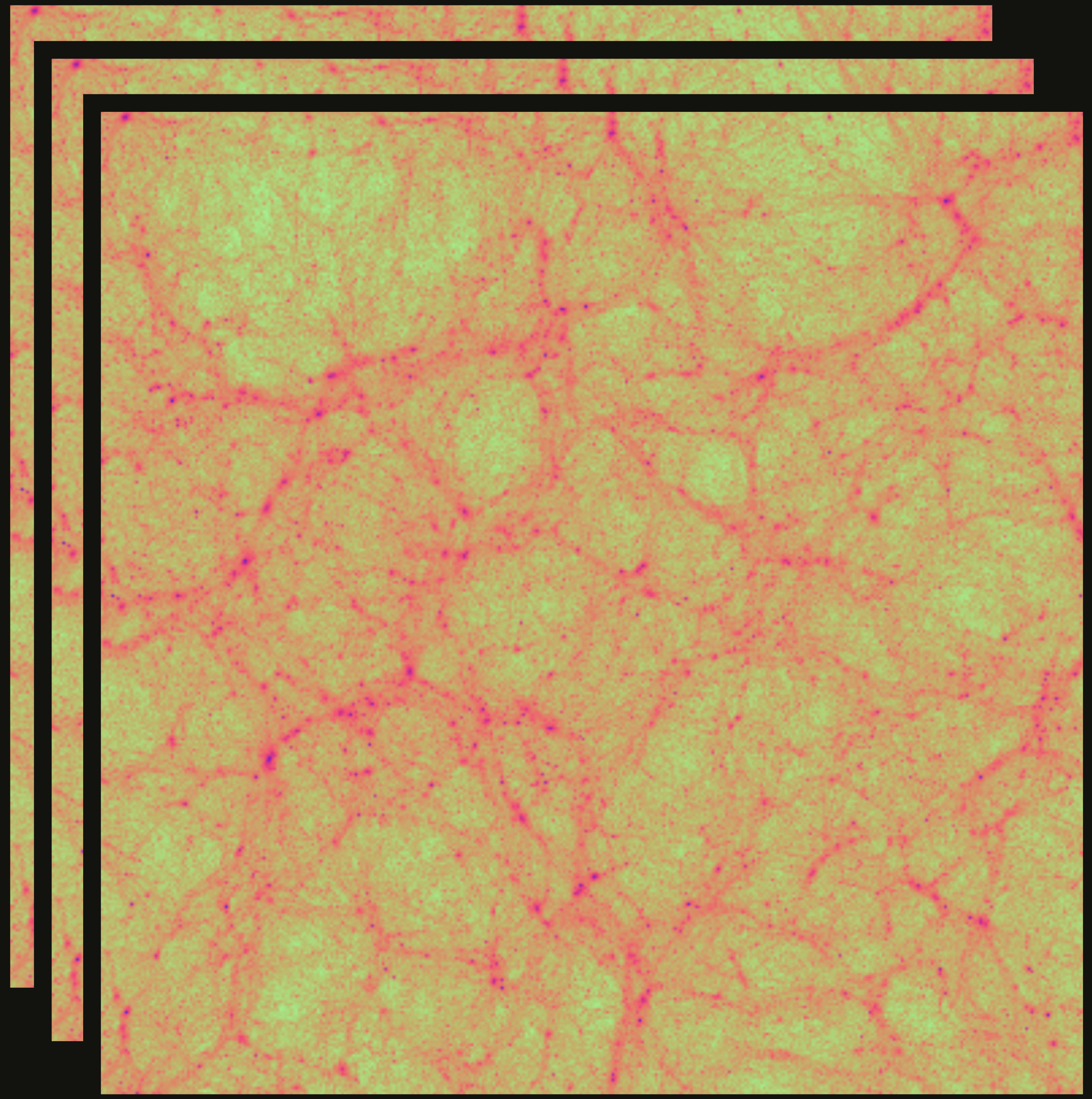




[Modi & Philcox 2023]

Multi-fidelity





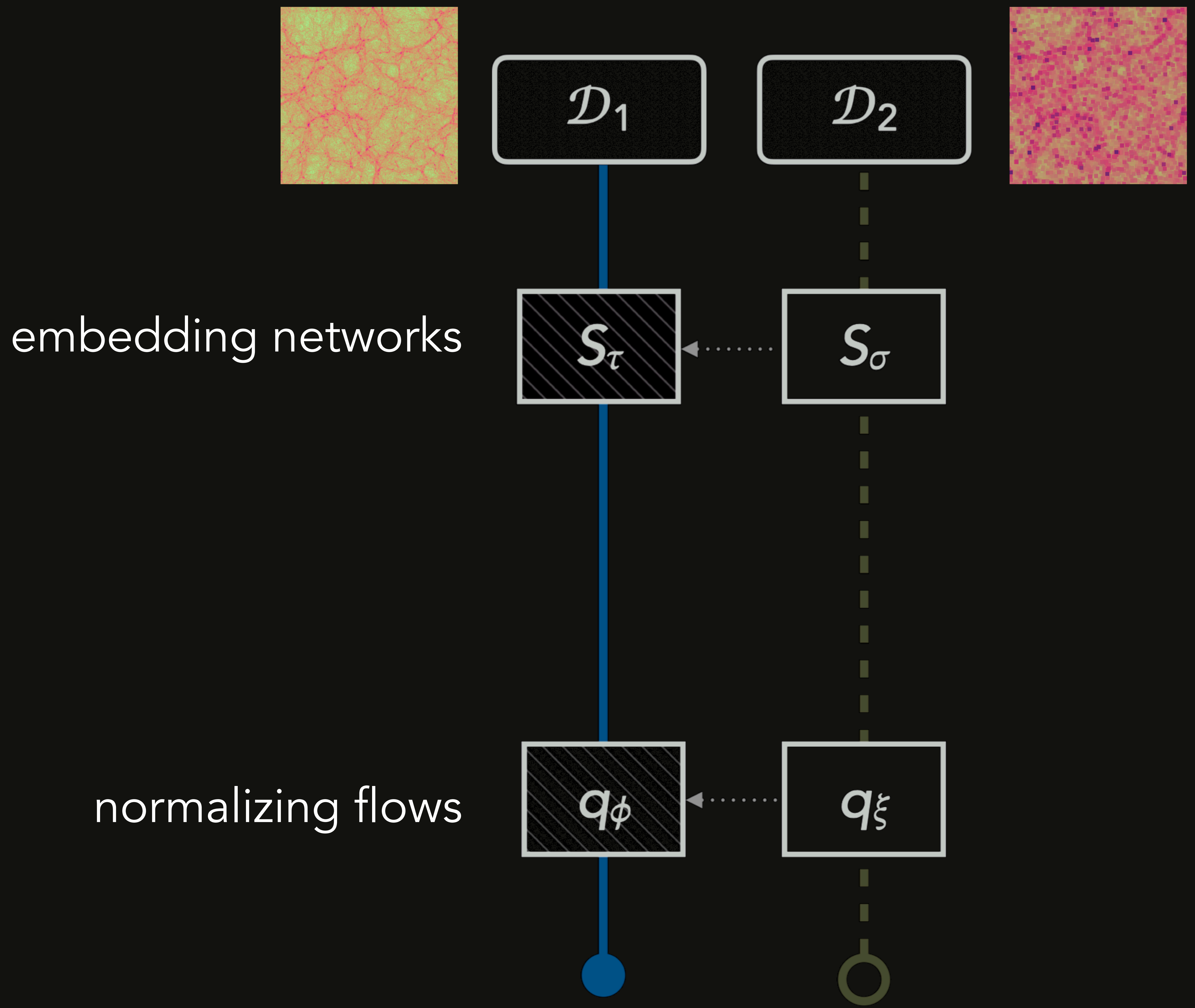
tiny low-fidelity set(s)
e.g., hydrodynamic

large low-fidelity set(s)
e.g., linear theory + 2LPT + particle-mesh + tree + HOD + SAM

[LT, A.Bayer, N.Takeishi 2025]

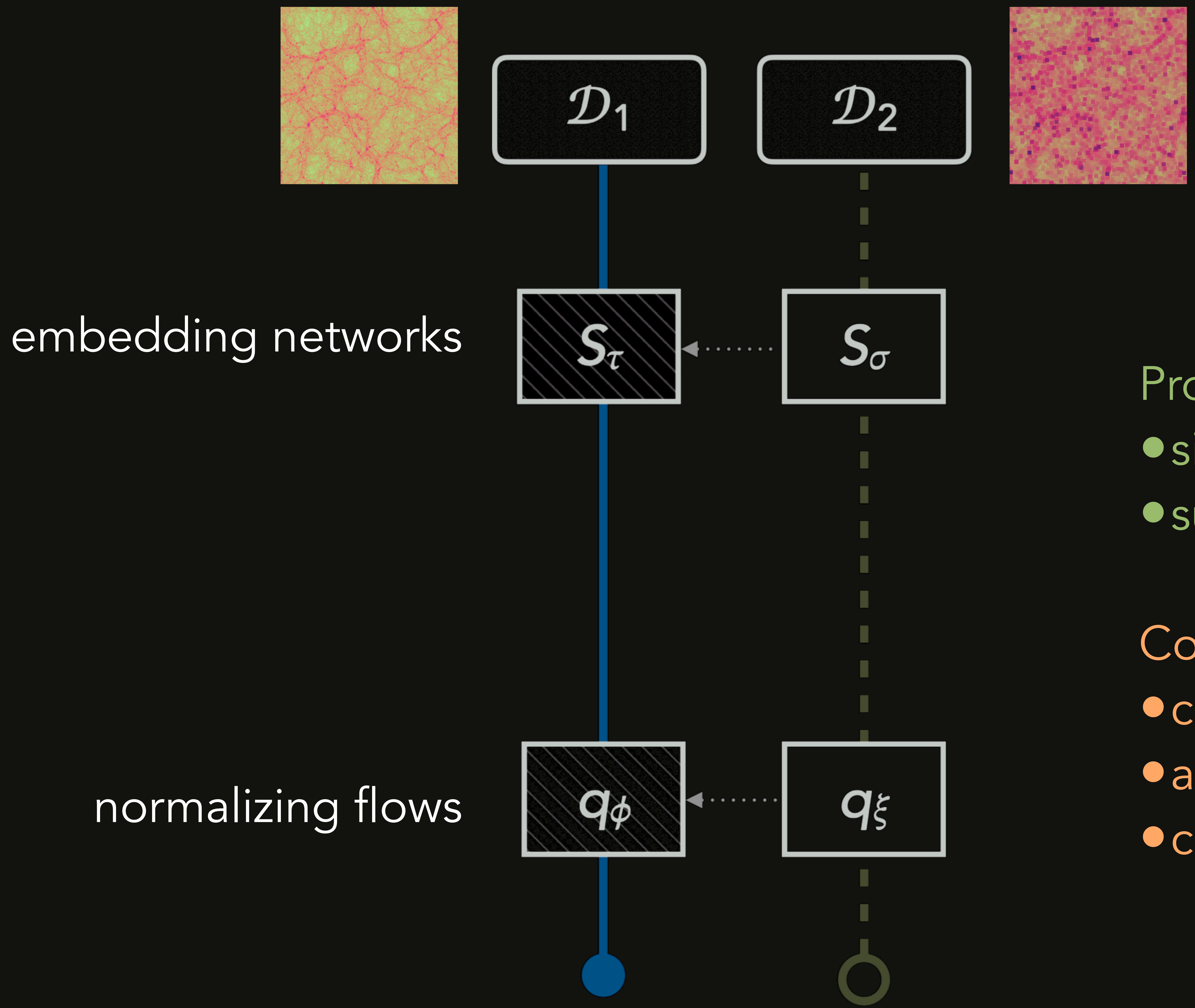
Multi-fidelity

simplest idea: transfer learning via weight initialization



Multi-fidelity

simplest idea: transfer learning via weight initialization



Pros:

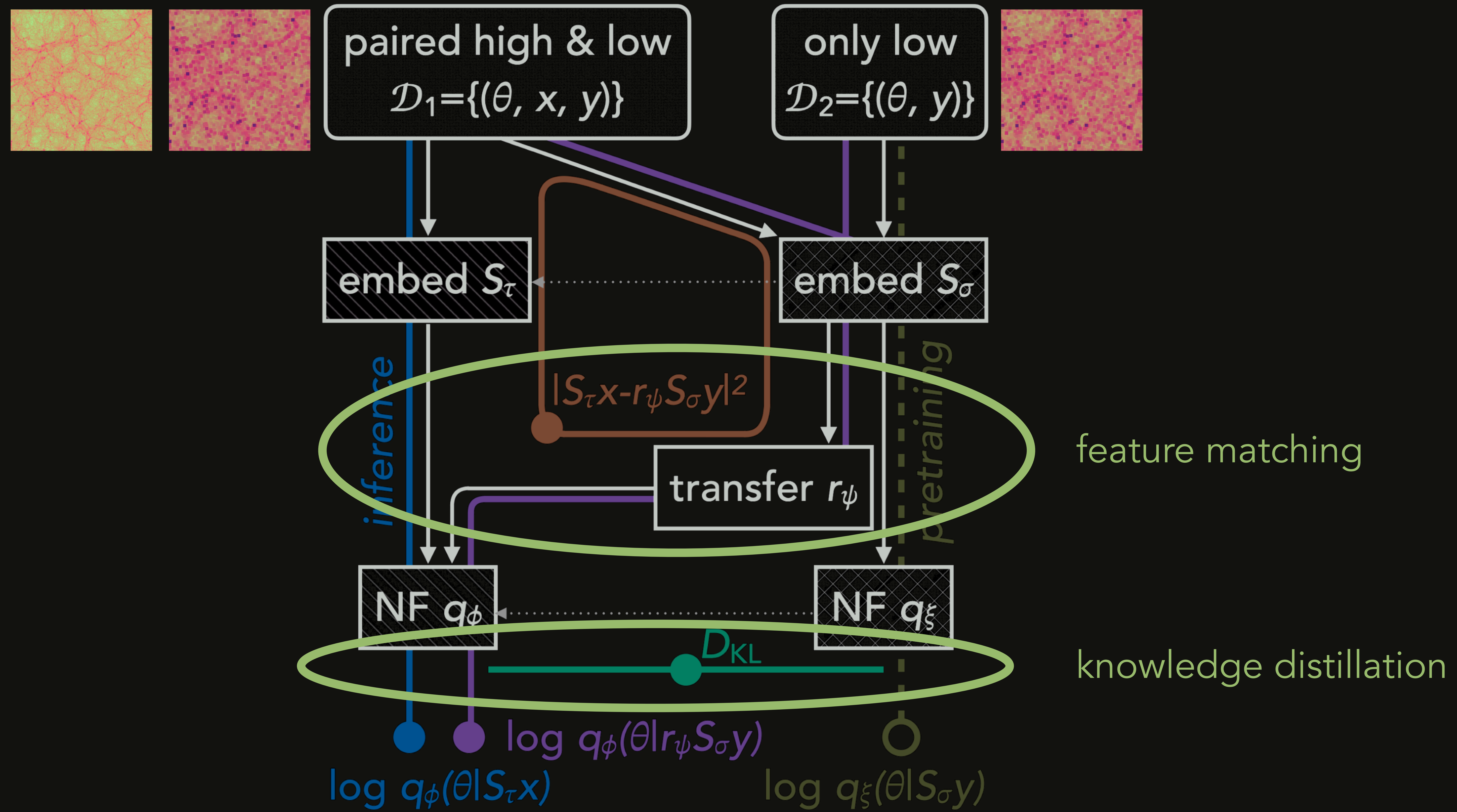
- simple
- surprisingly effective

Cons:

- clearly suboptimal
- architecture constrained
- can be unstable

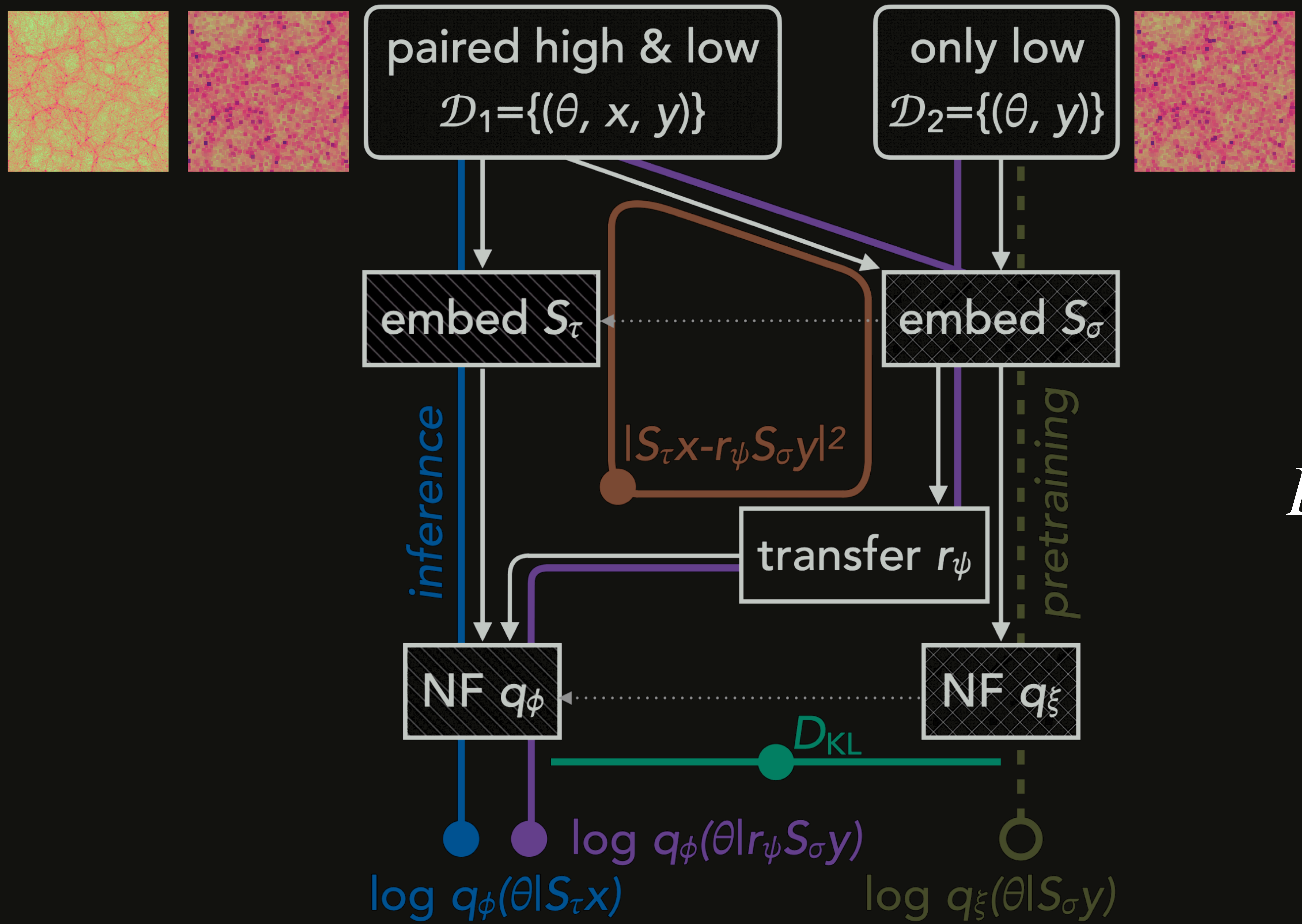
Multi-fidelity

improvement: feature matching & knowledge distillation



Multi-fidelity

improvement: feature matching & knowledge distillation



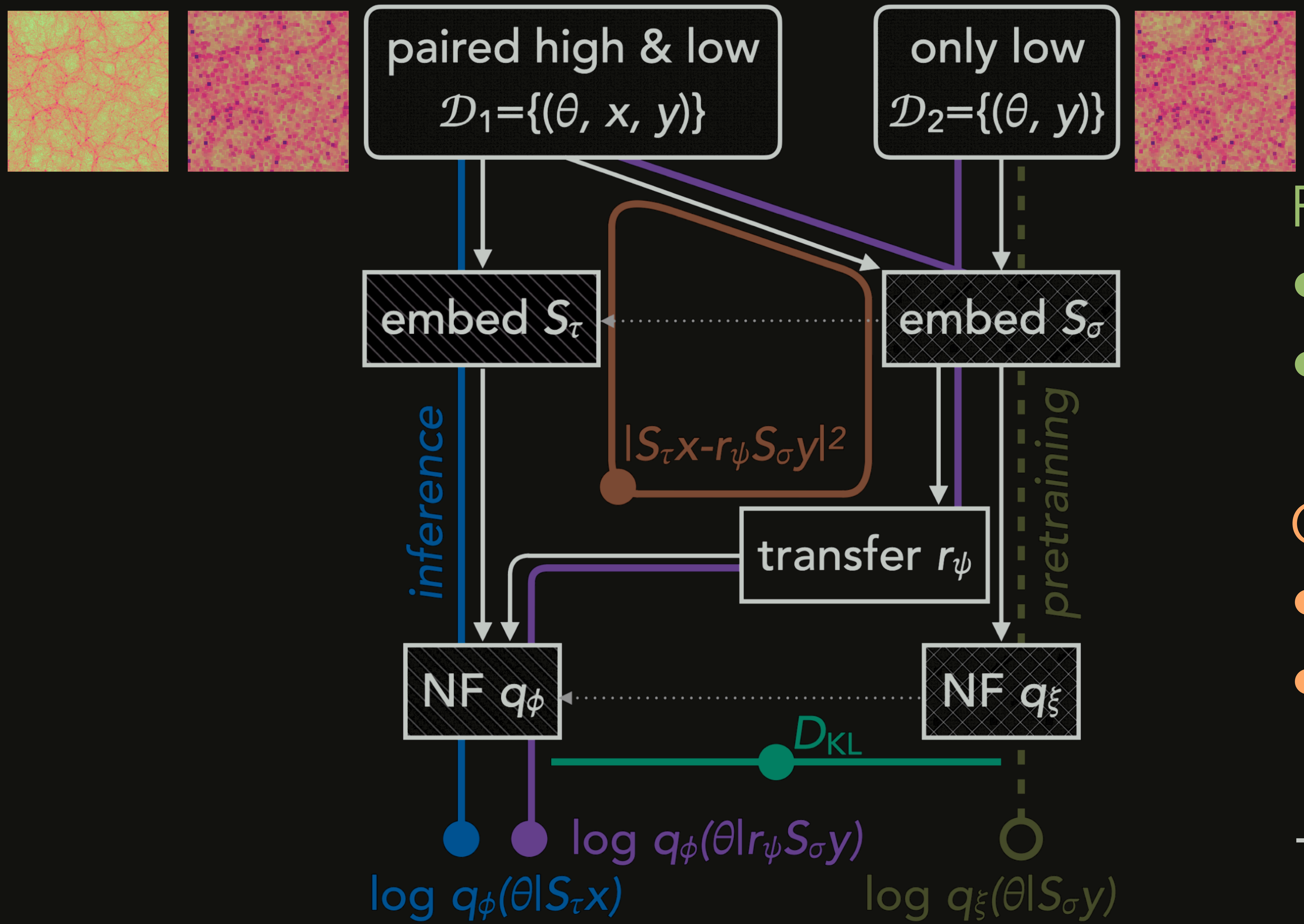
typical variance reduction at cost of introducing a bias

$$L = \int_{p(x,\theta)} \mathscr{L}$$

$$\approx \sum_{\text{training set}} \mathscr{L} + \sum \lambda_a R_a$$

Multi-fidelity

improvement: feature matching & knowledge distillation



Pros:

- very flexible (structure, levels)
- "soft": often better

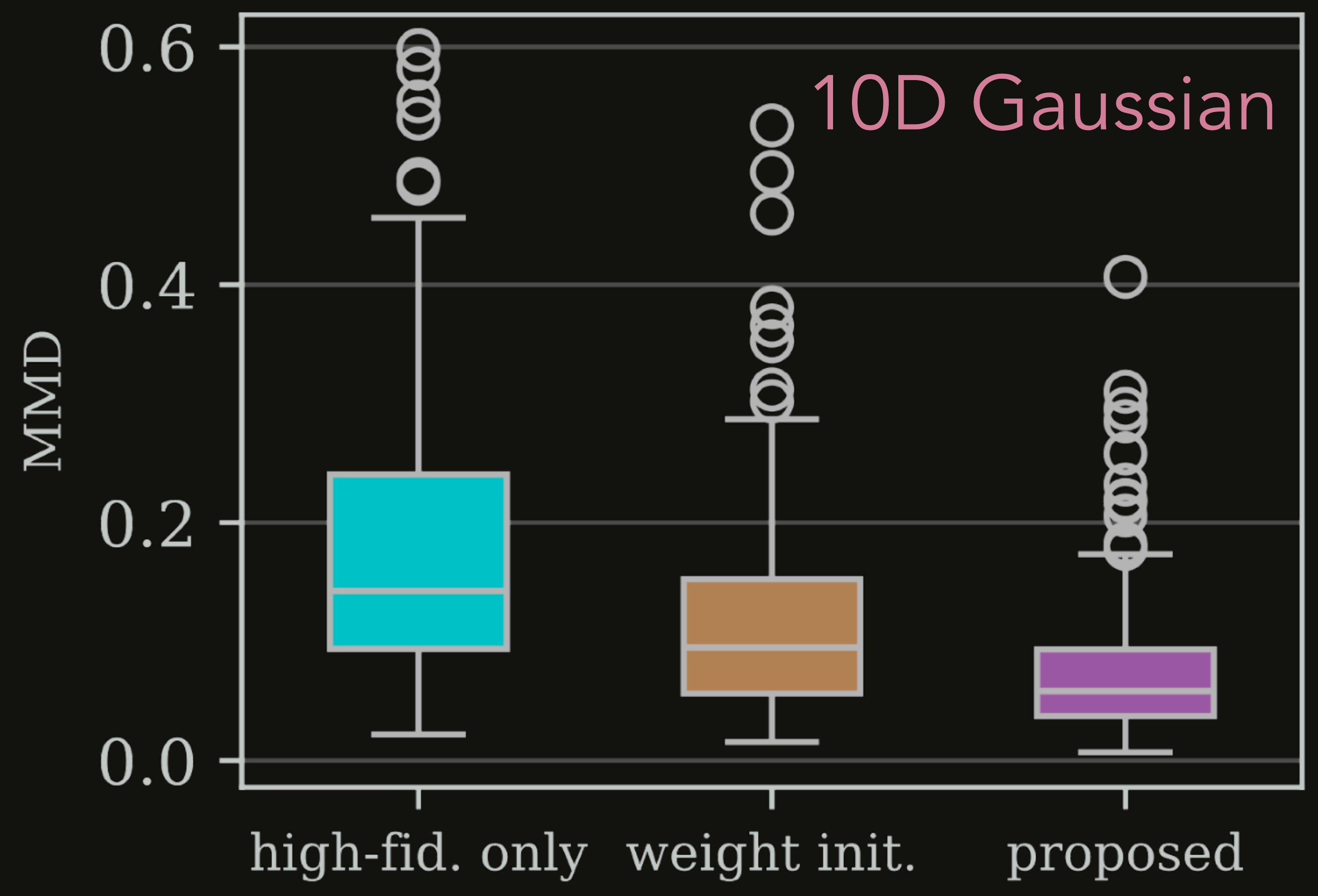
Cons:

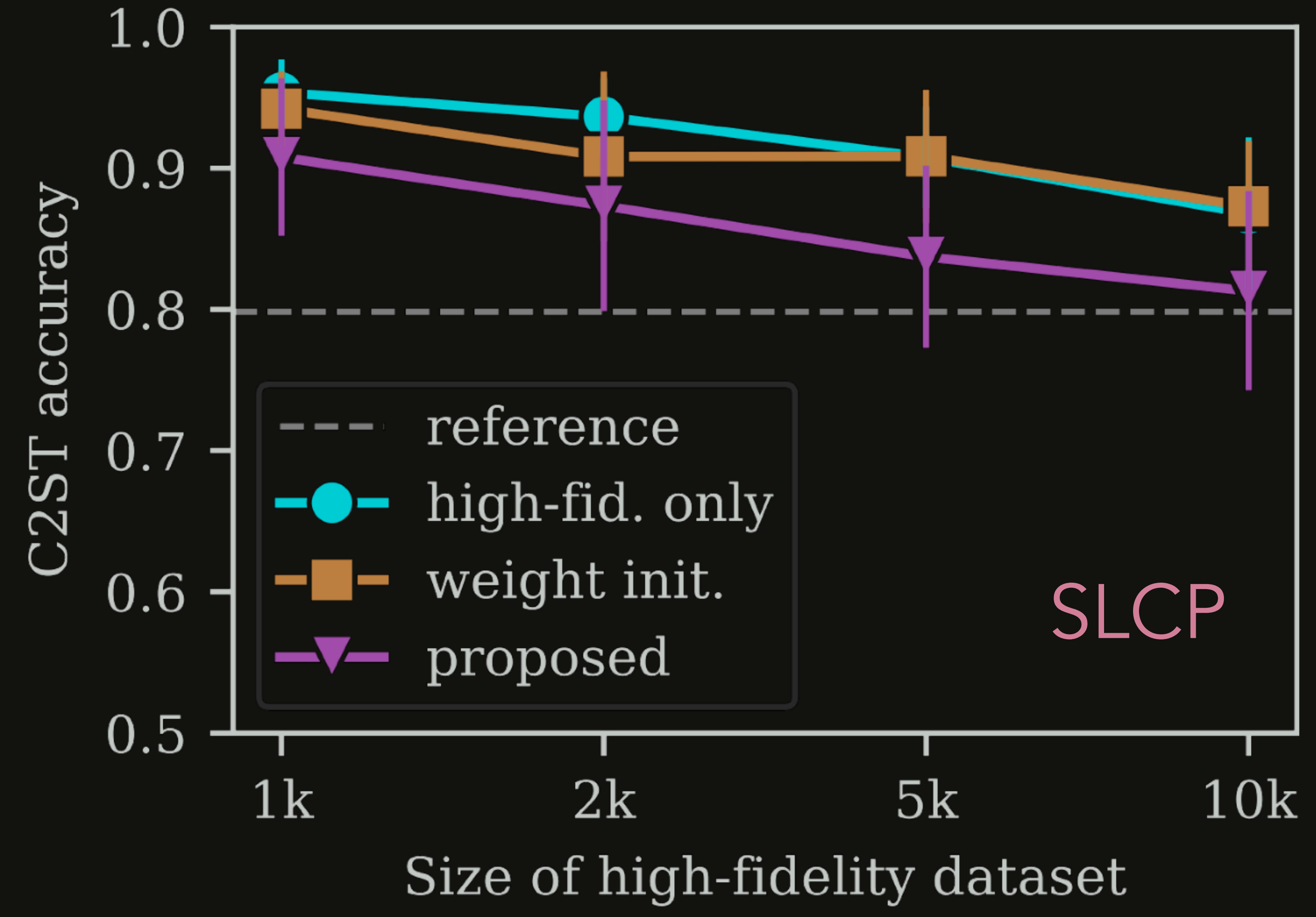
- multiple hyperparameters
- heuristics

→ improvement work in progress

Multi-fidelity: results

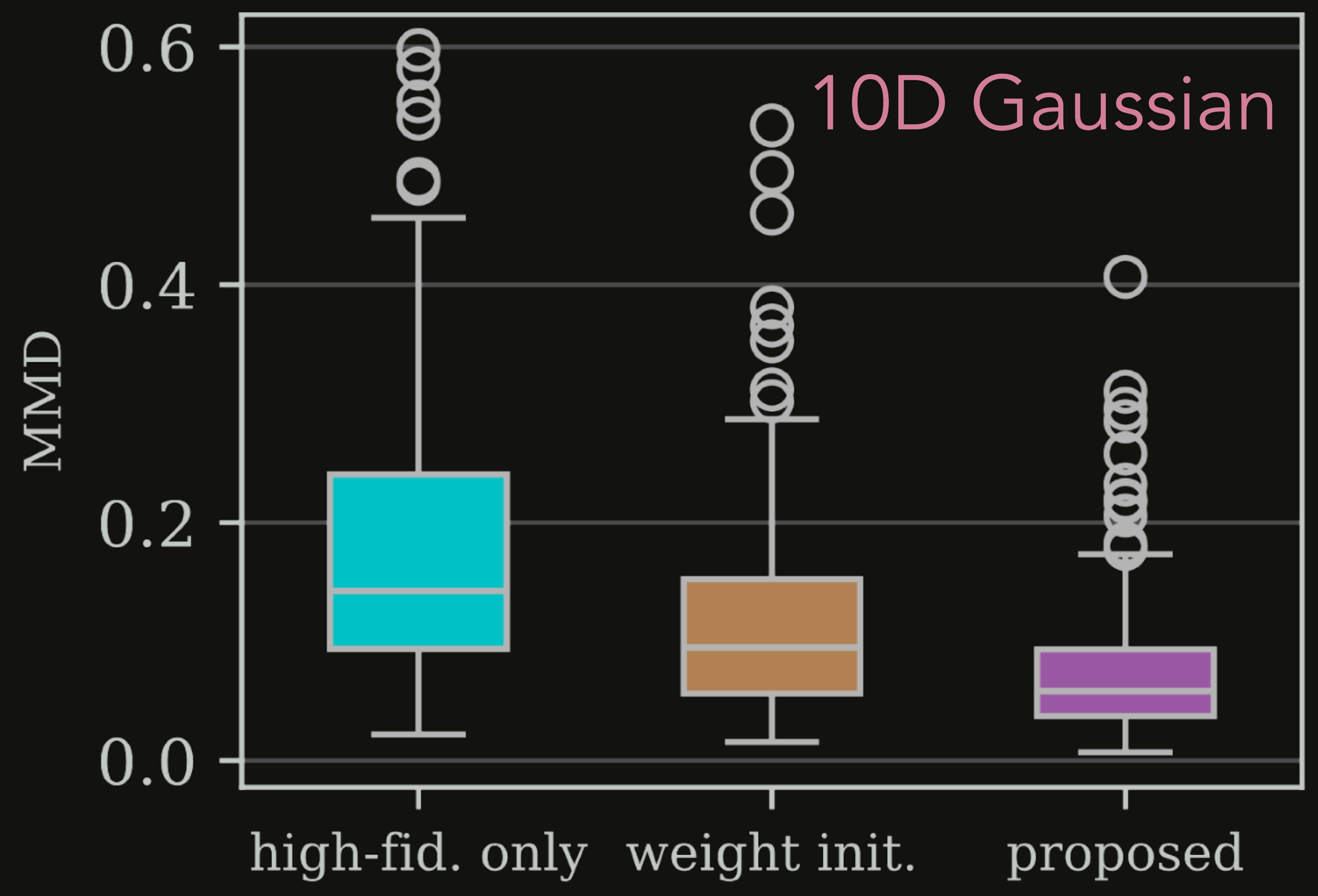
synthetic examples
(lower is better for all)

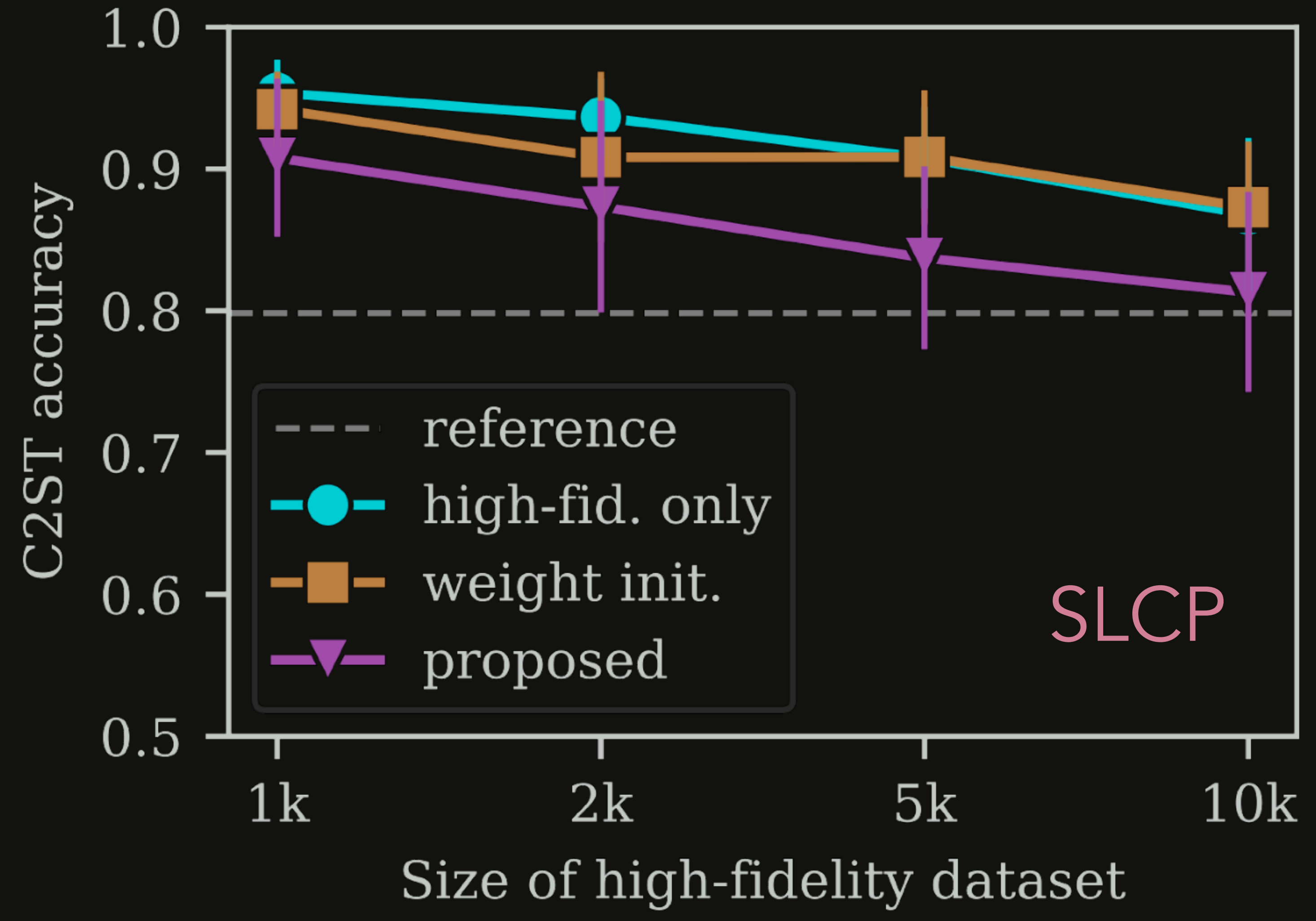




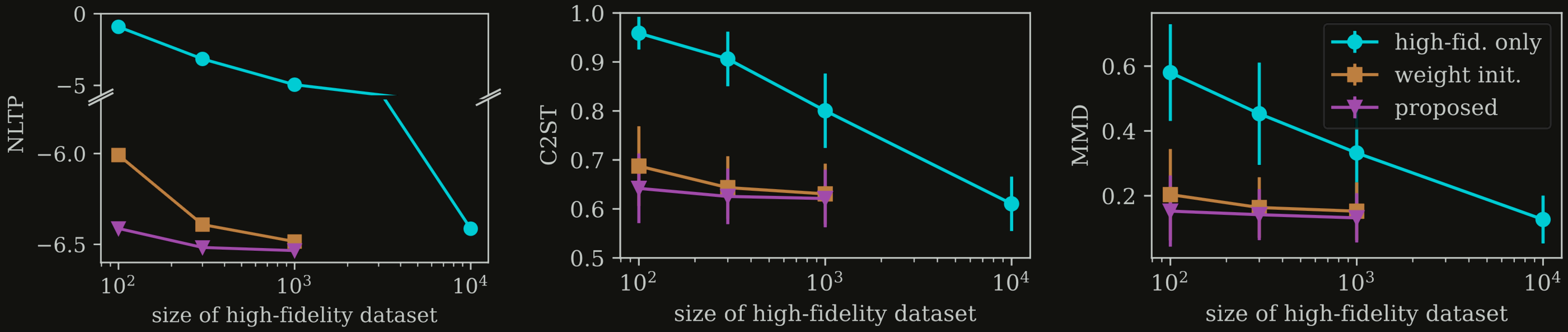
Multi-fidelity: results

synthetic examples
(lower is better for all)





cosmology example (lower is better for all)



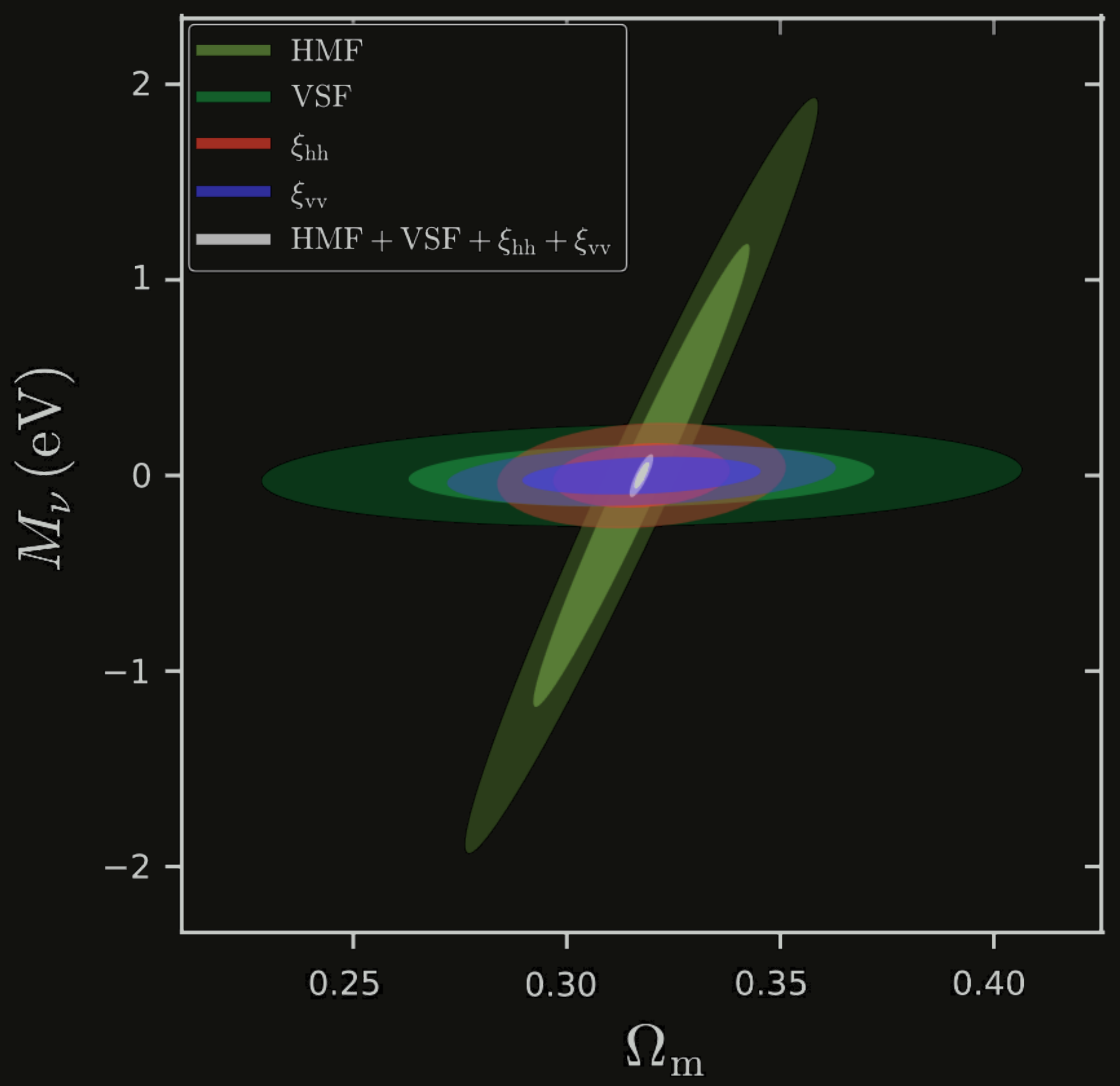
Summary

- Simulation-based inference = machine learning method to solve inverse problems defined implicitly through a simulator
- Simulation-based inference has proven viability in simple examples:
  - weak lensing (e.g., HSC Y1)
  - galaxy clustering (e.g., SDSS CMASS)
- In order to make it a standard tool, need to increase simulation quality while reducing training cost
  - sequential methods
  - learning corrections
  - combine with traditional approaches for large scales
  - multi-fidelity training
- Have demonstrated a regularization method towards multi-fidelity training

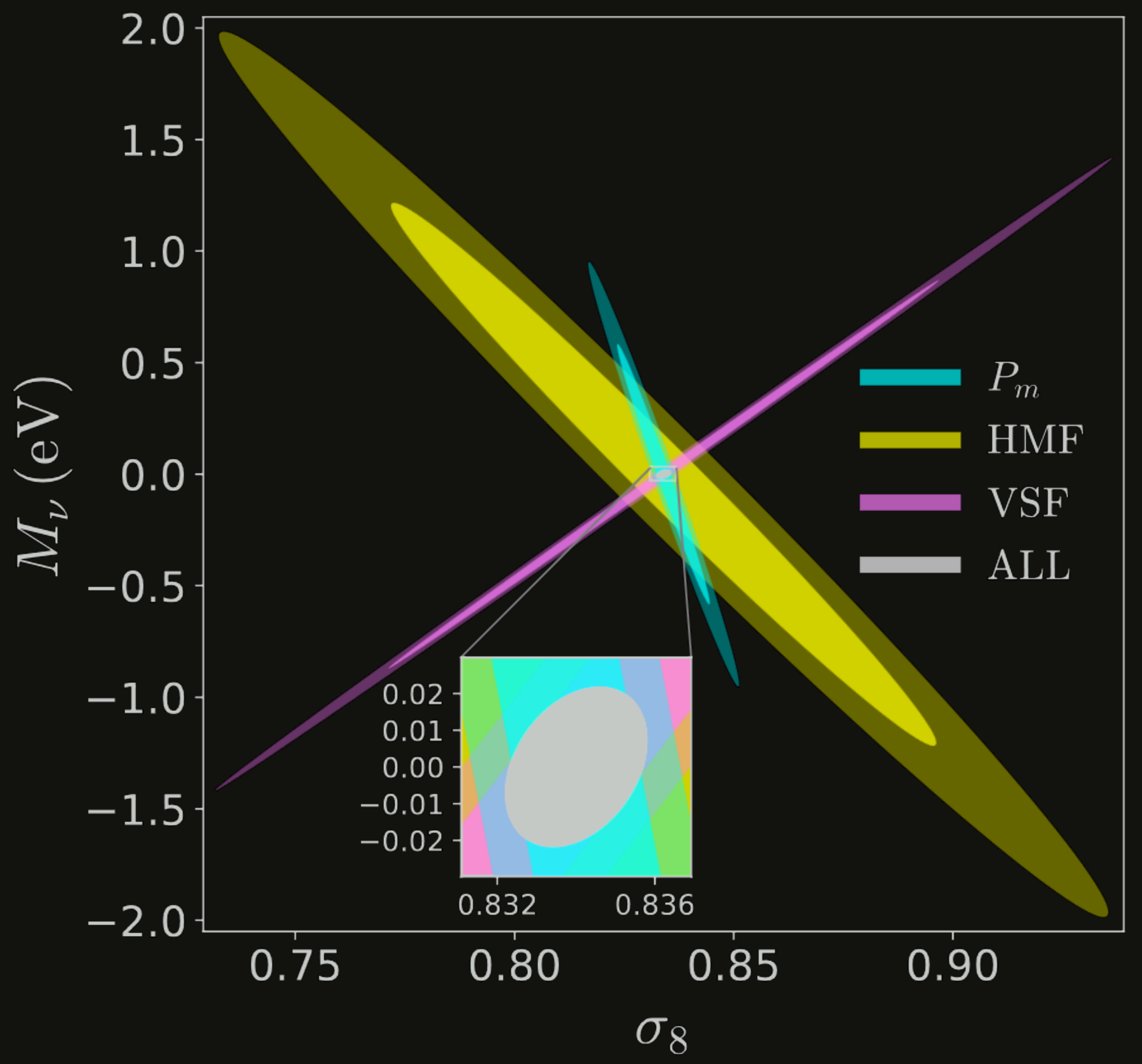
Voids & Neutrino Mass: Forecasts

QUIJOTE halos (1 Gpc/h)

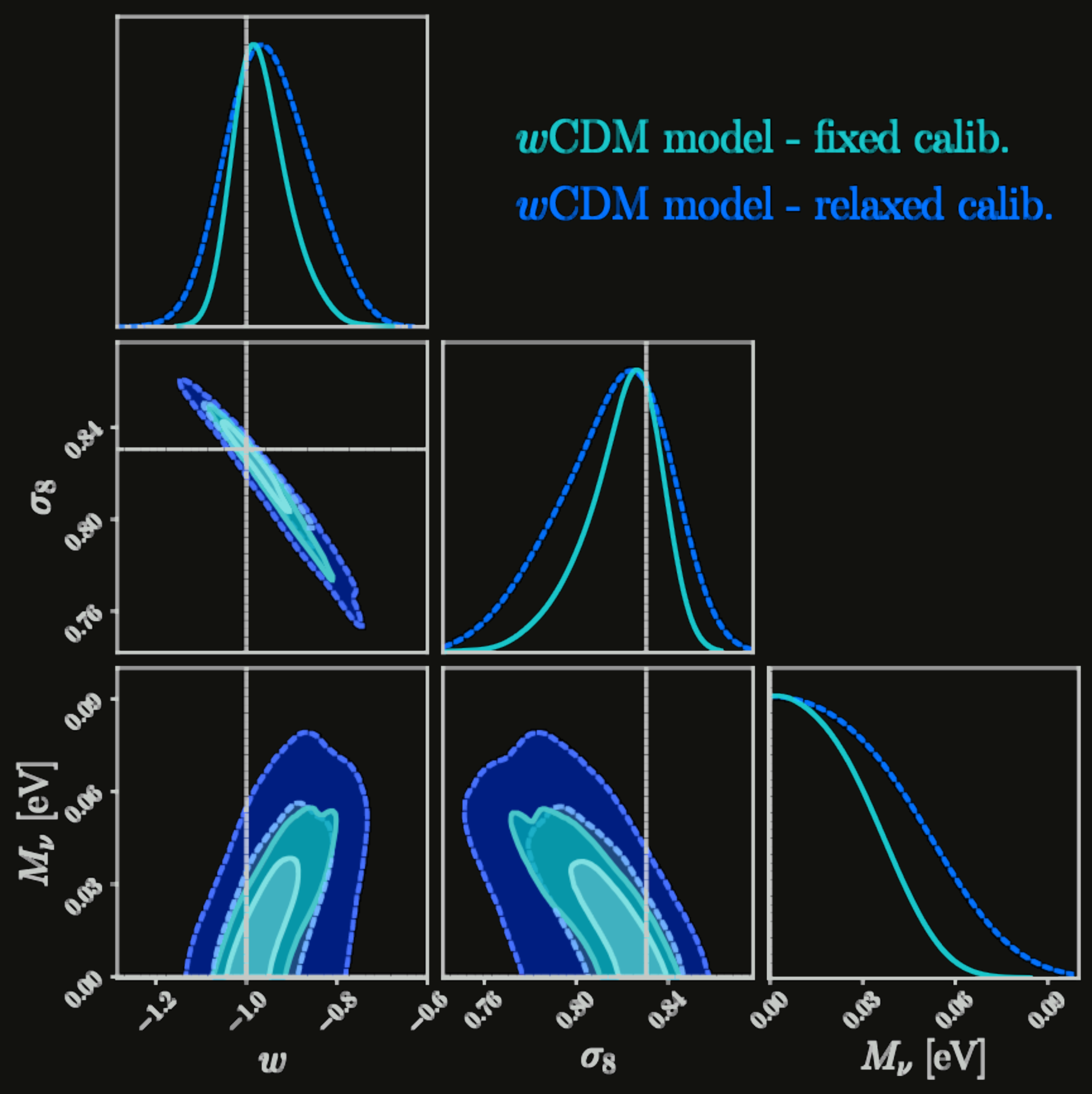
"Euclid" spec-z galaxies



[Kreisch+2021]



[Bayer+2021]



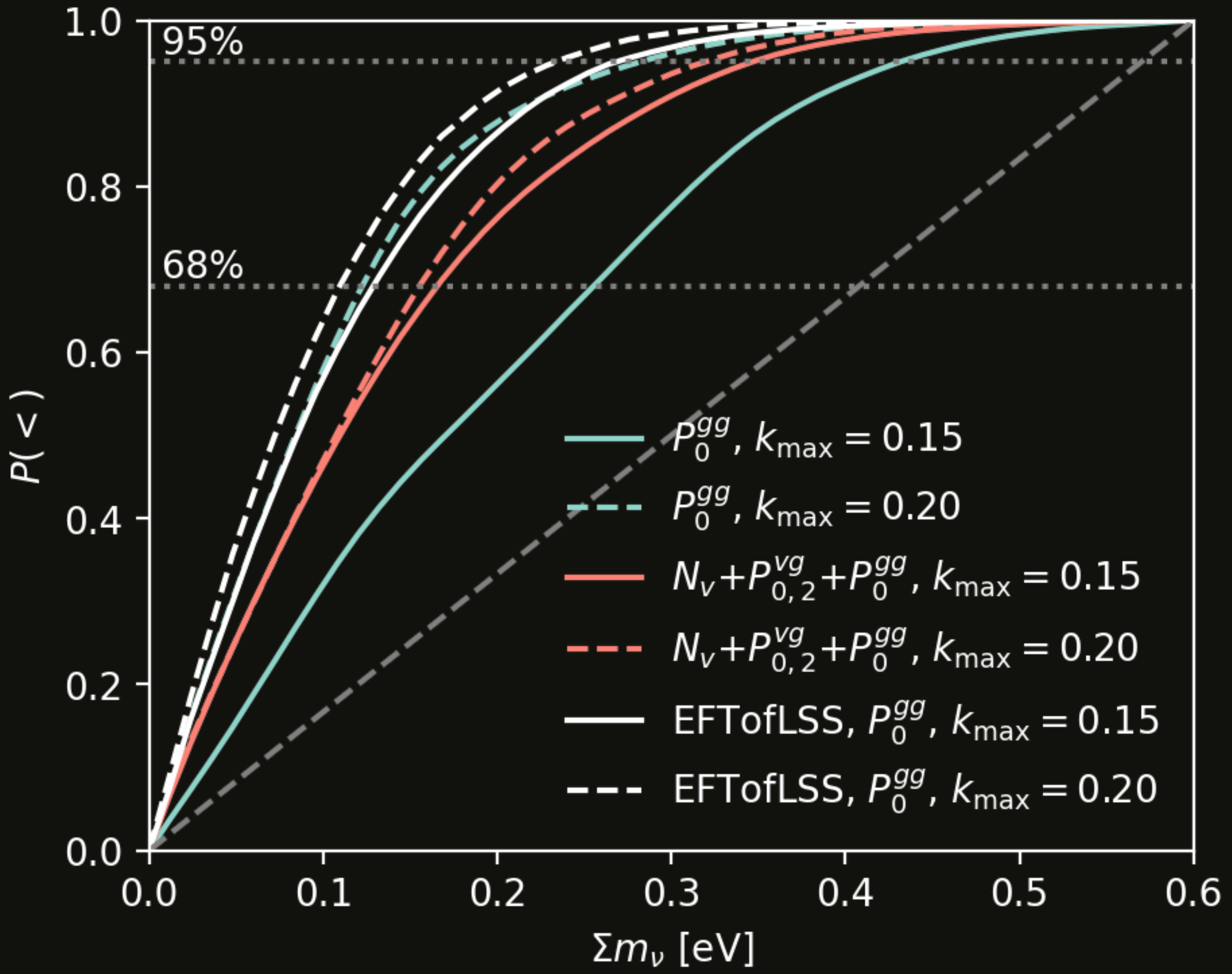
[Contarini+2021]

→ take this to real data from BOSS,
including complications from galaxy formation & survey systematics

Comparison to EFTofLSS

- EFT possible issues:
  - prior dependence, volume effects
  - small scales
  - redshift evolution?

- our possible issues:
  - prior dependence, volume effects
  - approximate simulations
  - "augmentations"



Neural ratio estimation

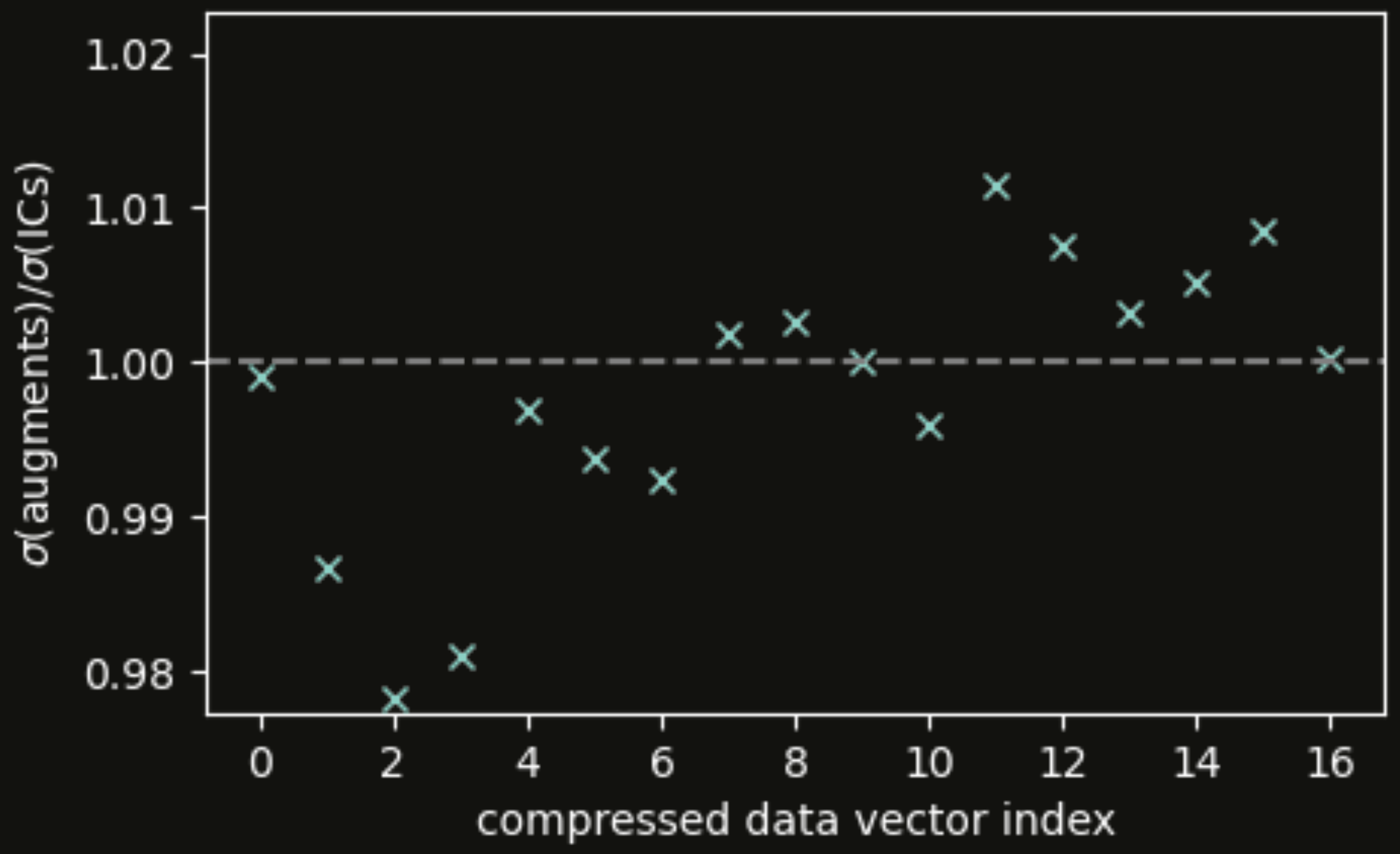
$$P(\text{parameters}\,|\,\text{data}) = \frac{\boxed{\dfrac{P(\text{data}\,|\,\text{parameters})}{P(\text{data})}}\; P(\text{parameters})}{}$$

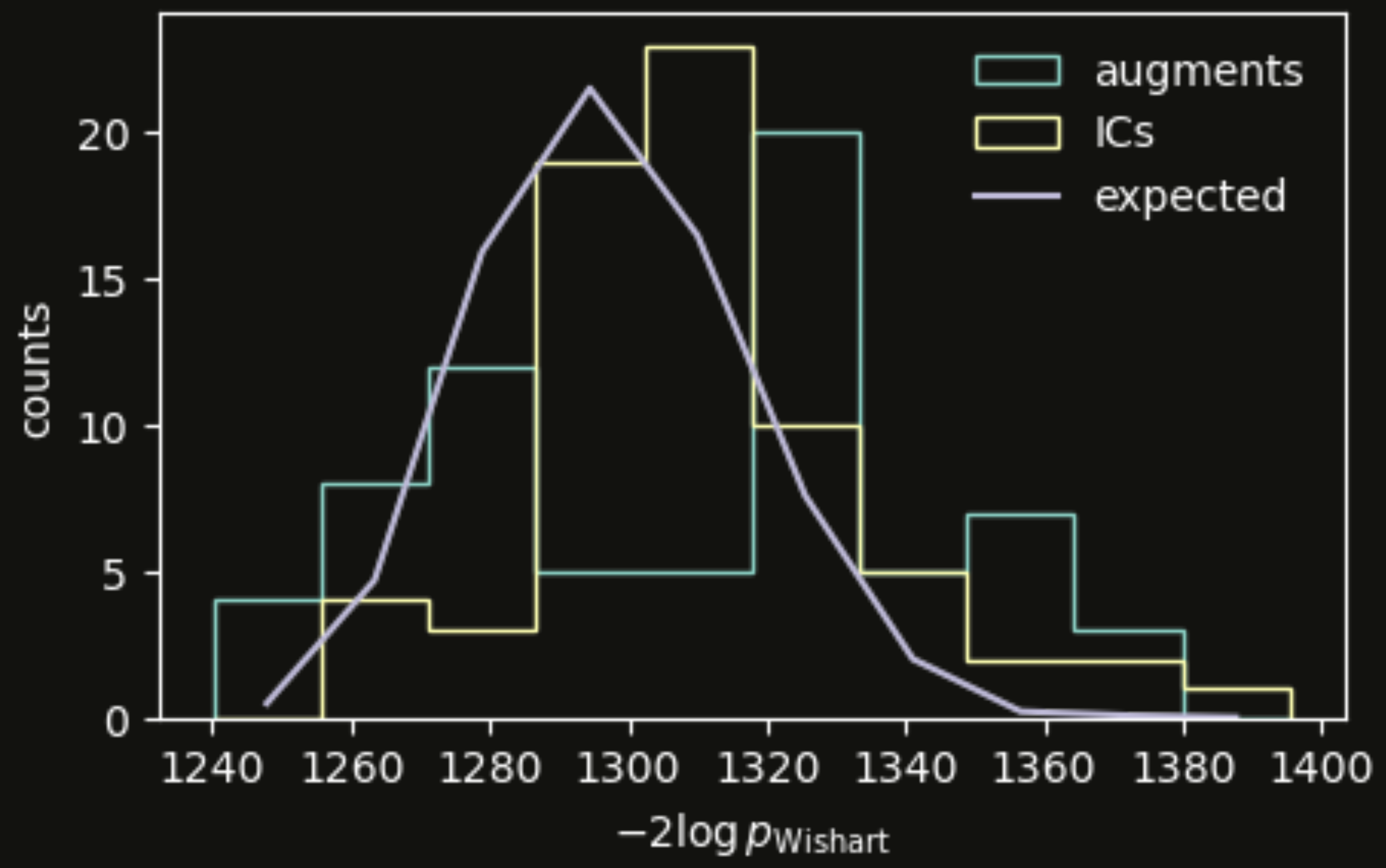
- draw θ, θ' ~ P(parameters)
- simulate x_{sim} ~ P(data | θ) [drawing η, ζ]
- evaluate neural net: $y=f(x_{sim}, \theta)$, $y'=f(x_{sim}, \theta')$
- classification loss, e.g. $\mathcal{L} = -\log(y) - \log(1-y')$
- train $f^* = \text{argmin} \int D\theta\, D\theta'\, Dx_{sim}\, \mathcal{L}$
- → $P(x \mid \theta) / P(x) = f^*(x, \theta) / [\, 1 - f^*(x, \theta)\,]$

- any simulatable effect can be incorporated
- rephrase as classification problem → sophisticated machinery exists
- no formal difference between nuisance parameters and initial conditions

Augmentations

- cosmo-varied simulations have same random seed
- can generate quasi-independent realizations from single simulation
- how independent? Check with fiducial simulations





Putting galaxies in the simulations

Using halo occupation distribution, tuned on preliminary tests with QUIJOTE sims and data.

Standard 5-dim HOD plus

- velocity biases
- linear redshift evolution
- secondary/assembly bias based on T/U

= 11-dim in total

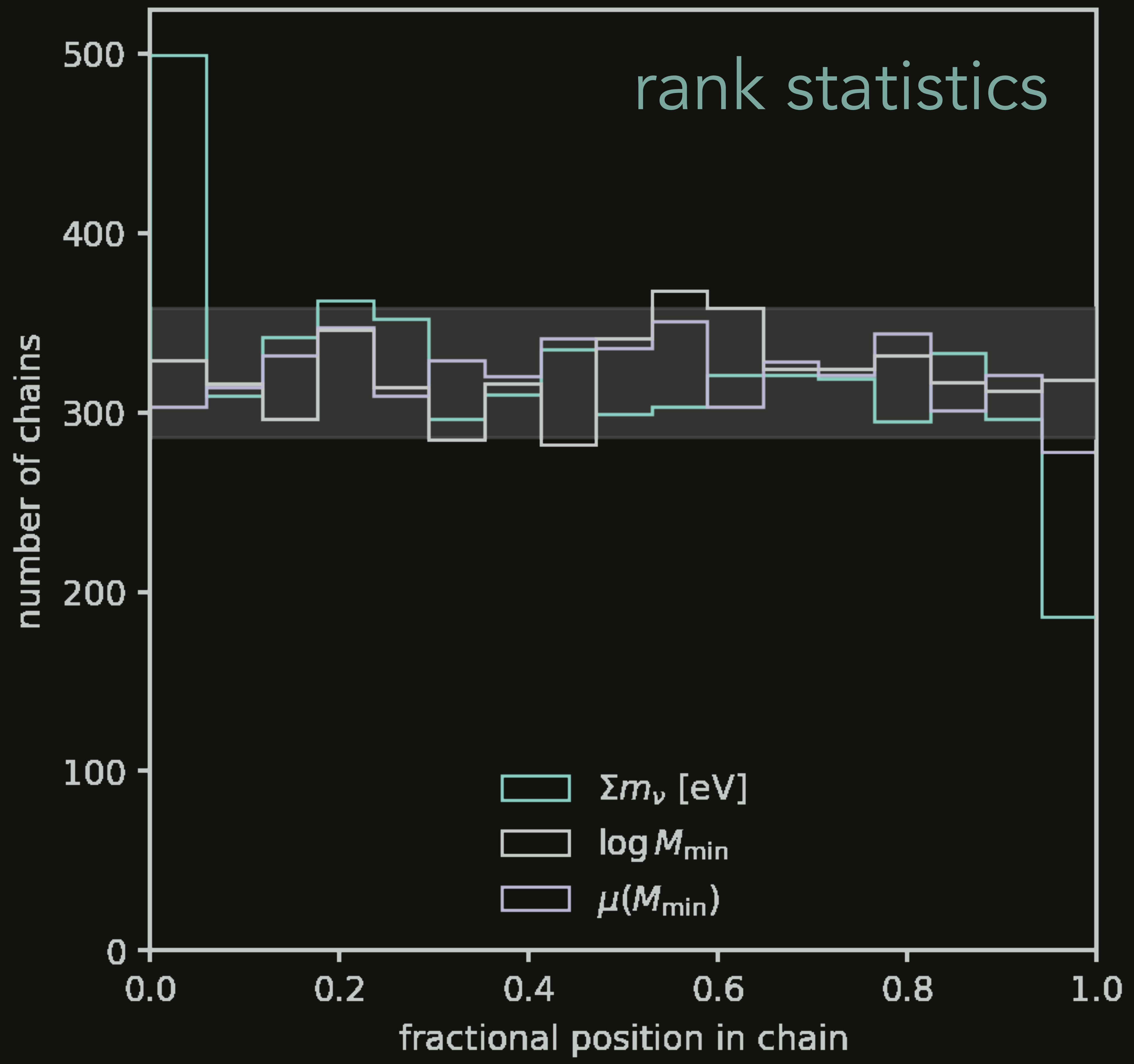
Likely not ideal!

Project on lightcone and implement survey realism

- mask
- fiber collisions
- n(z)

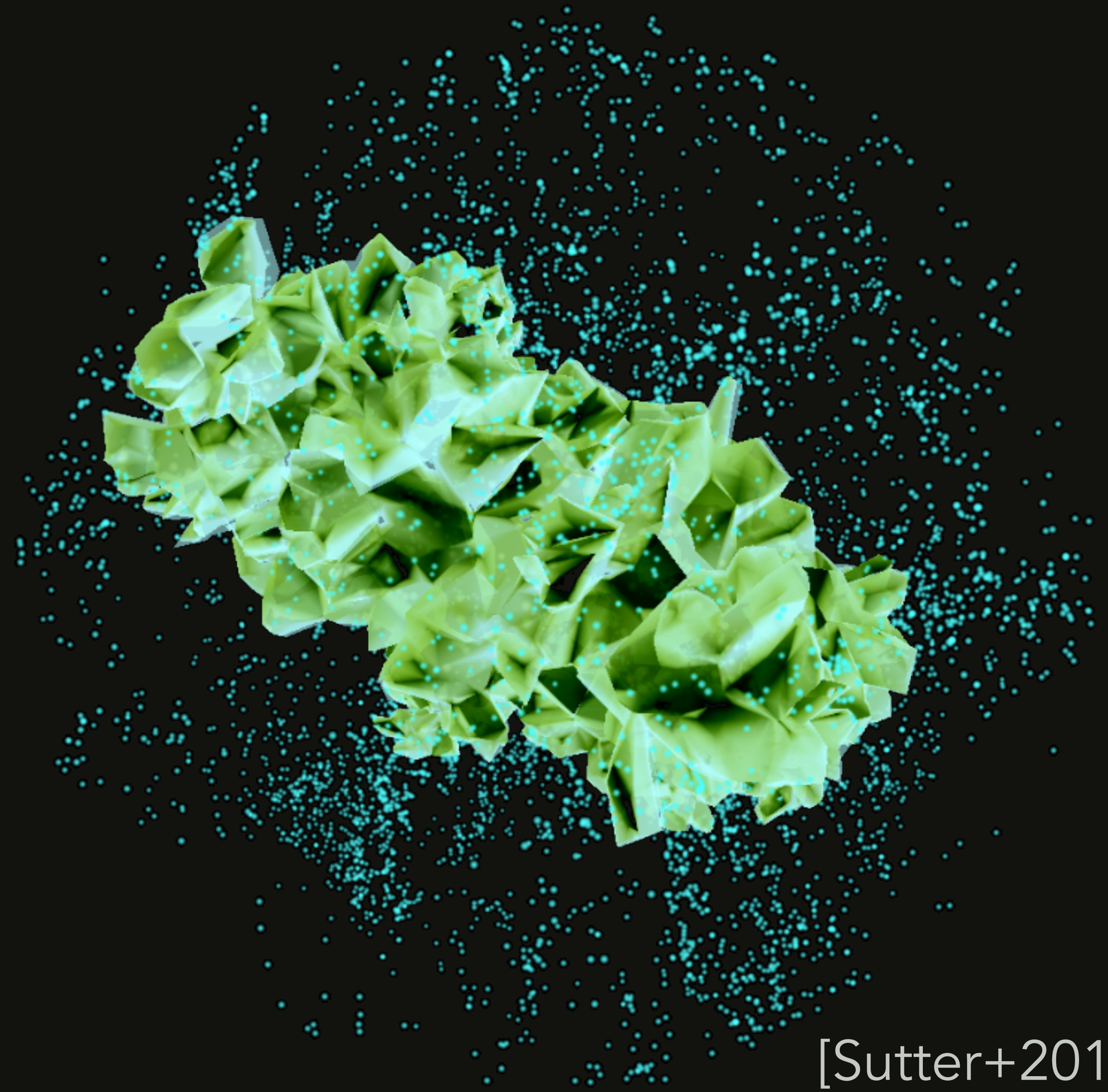
Check internal consistency

test posterior calibration by running inference on mocks from prior

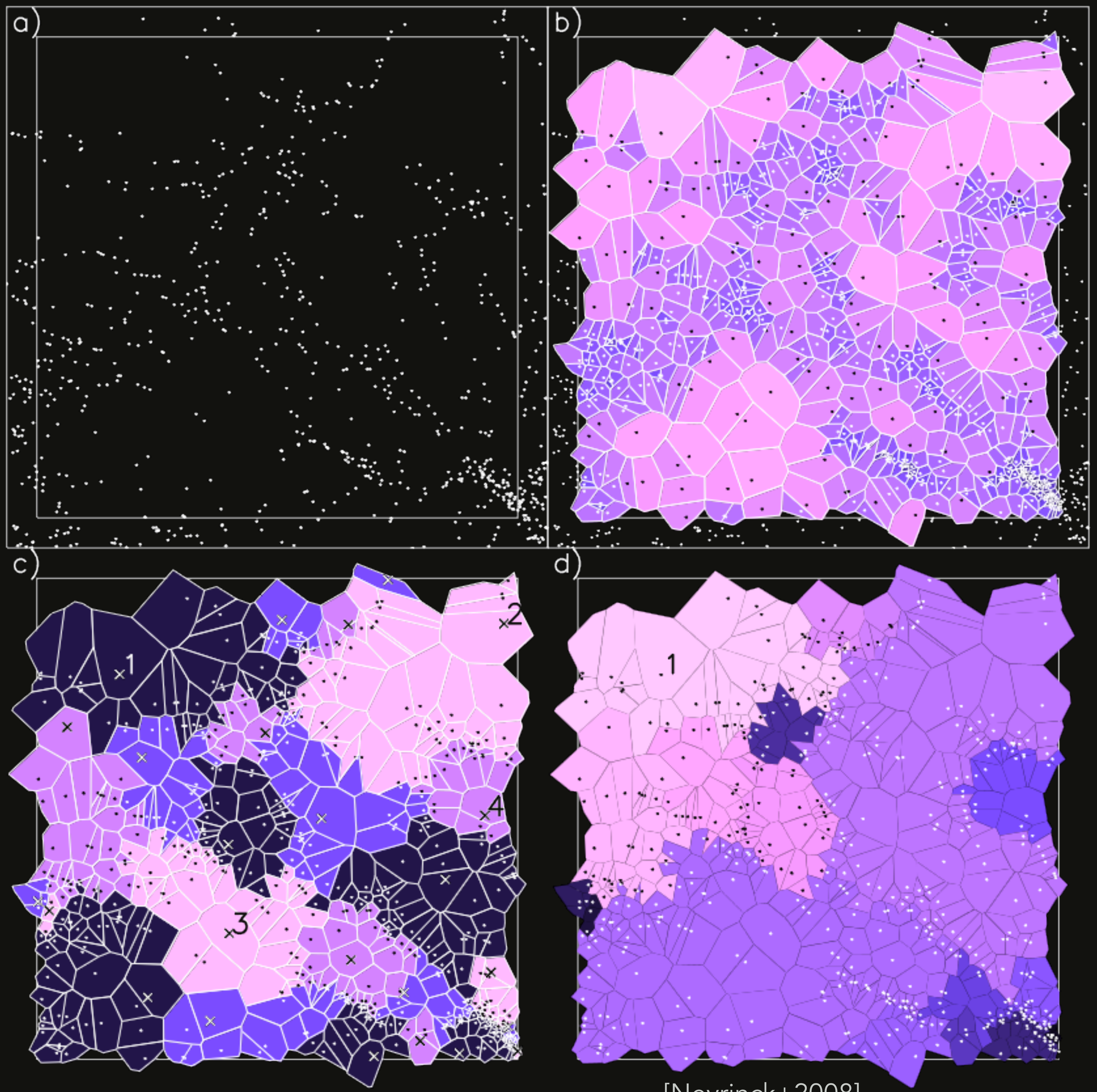


Void finding

- currently no universally accepted void definition
- simulation based analysis → can choose without biasing
- ZOBOV/VIDE → Voronoi tesselation

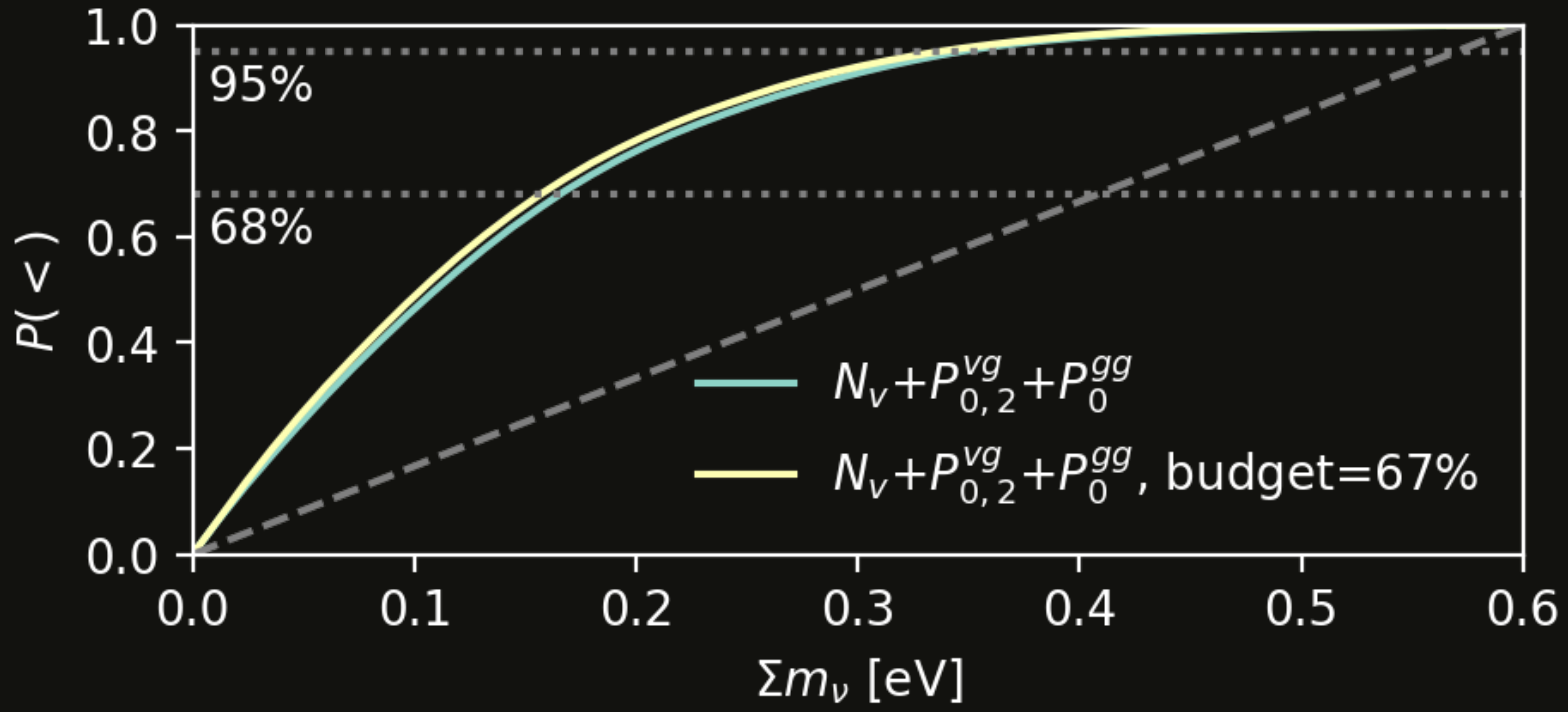


[Sutter+2012]

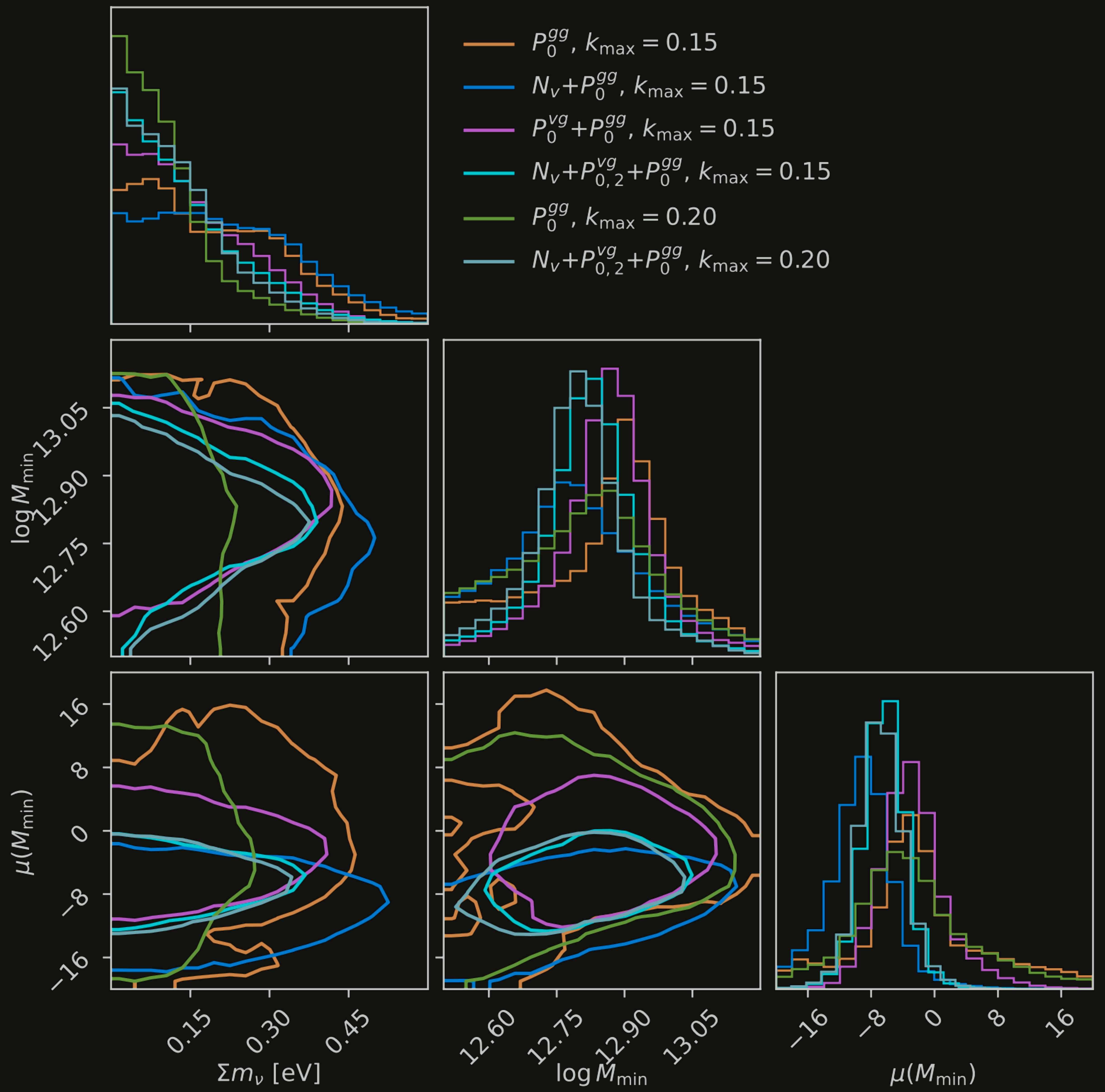


[Neyrinck+2008]

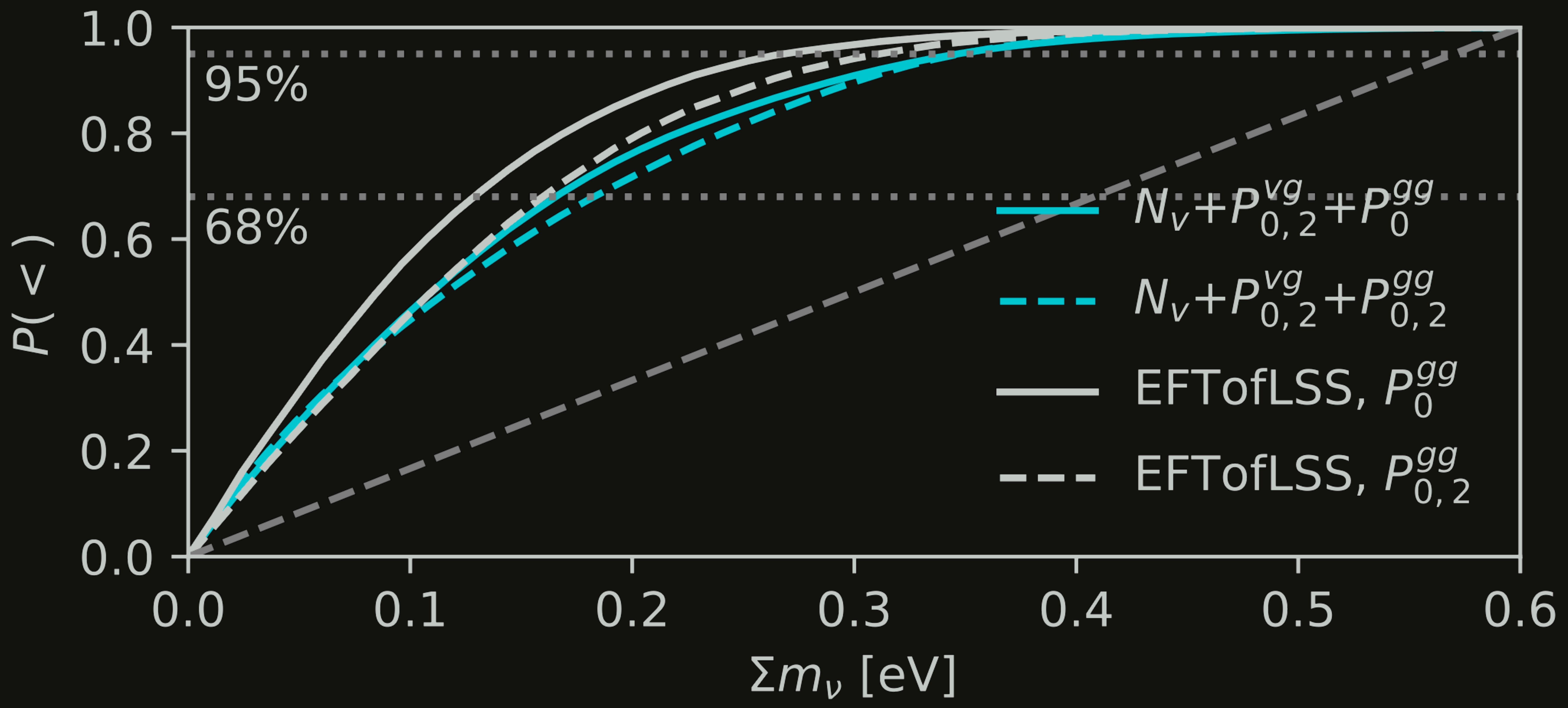
Simulation budget



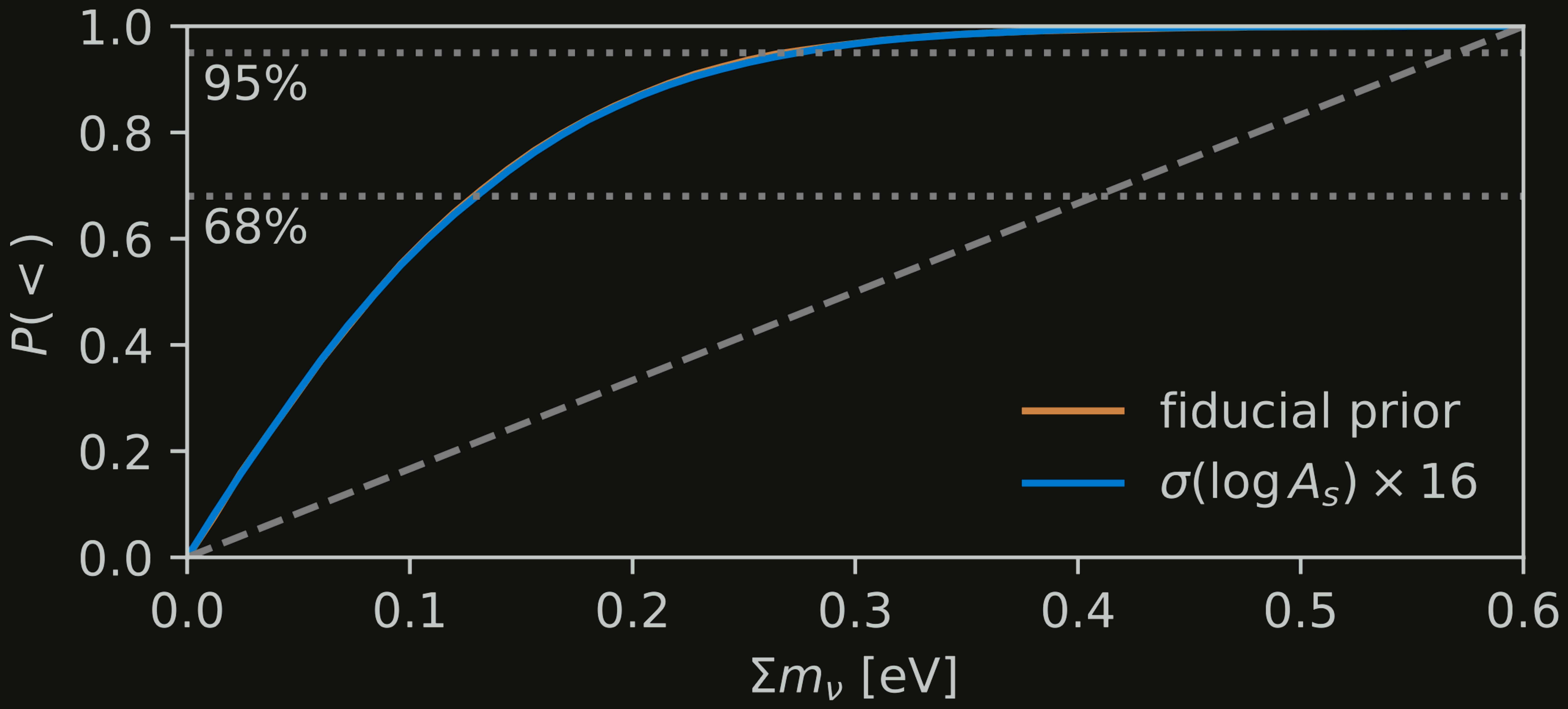
Full posterior



Adding quadrupole

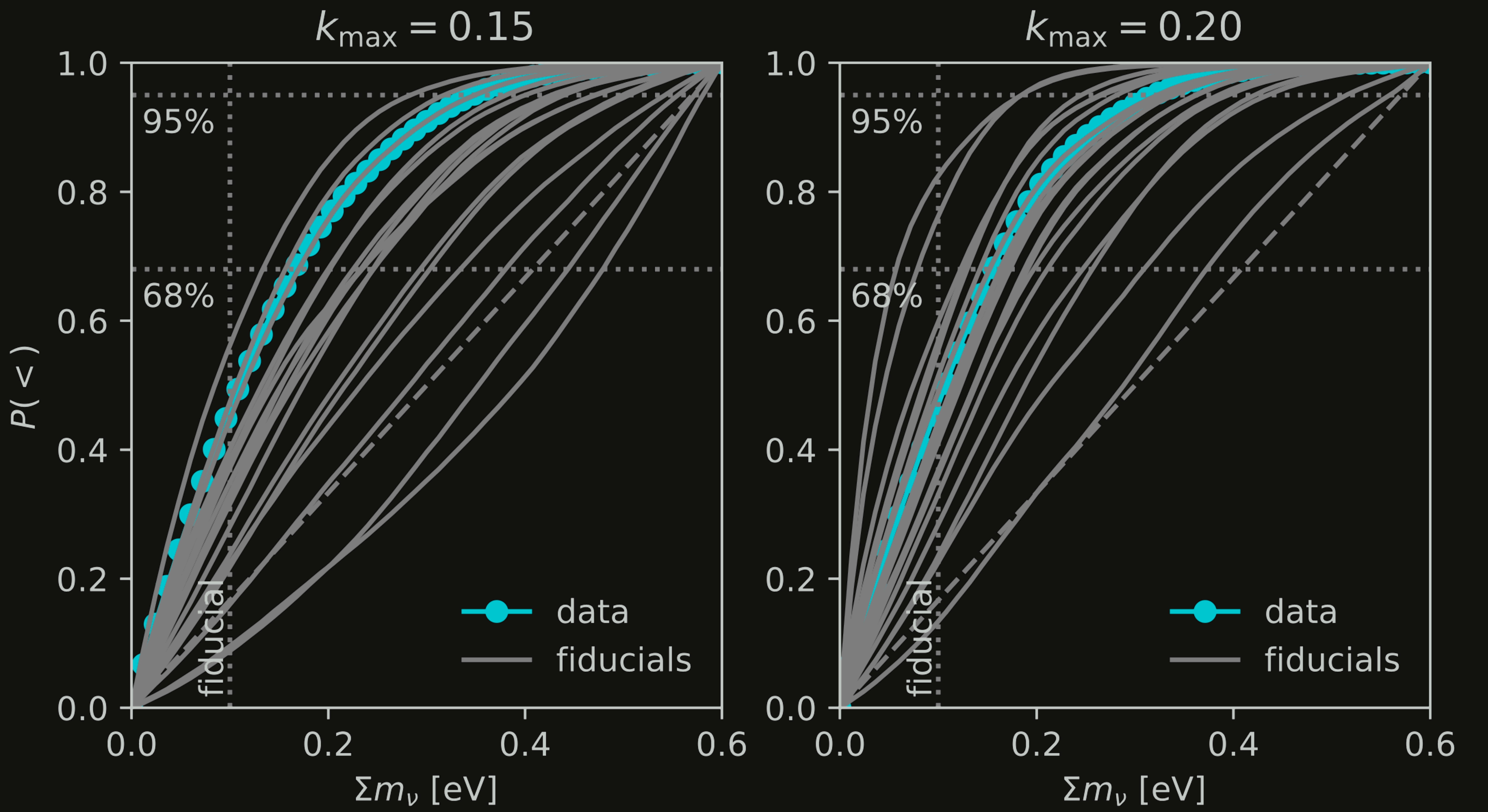


A_s prior

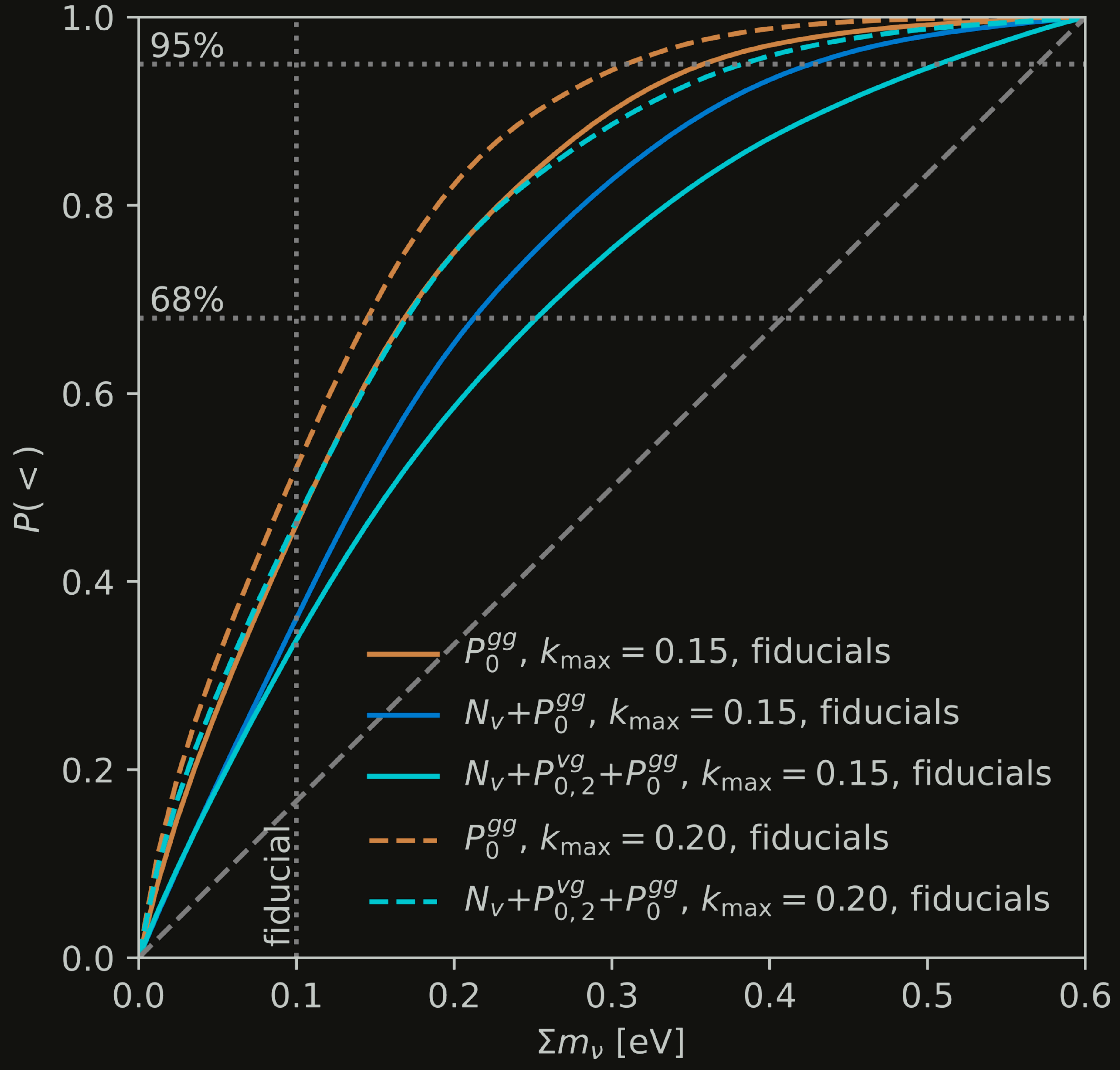


1.0
0.8
0.6
0.4
0.2
0.0
95%
68%
$P(<)$
fiducial prior
$\sigma(\log A_s) \times 16$
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Σm_ν [eV]

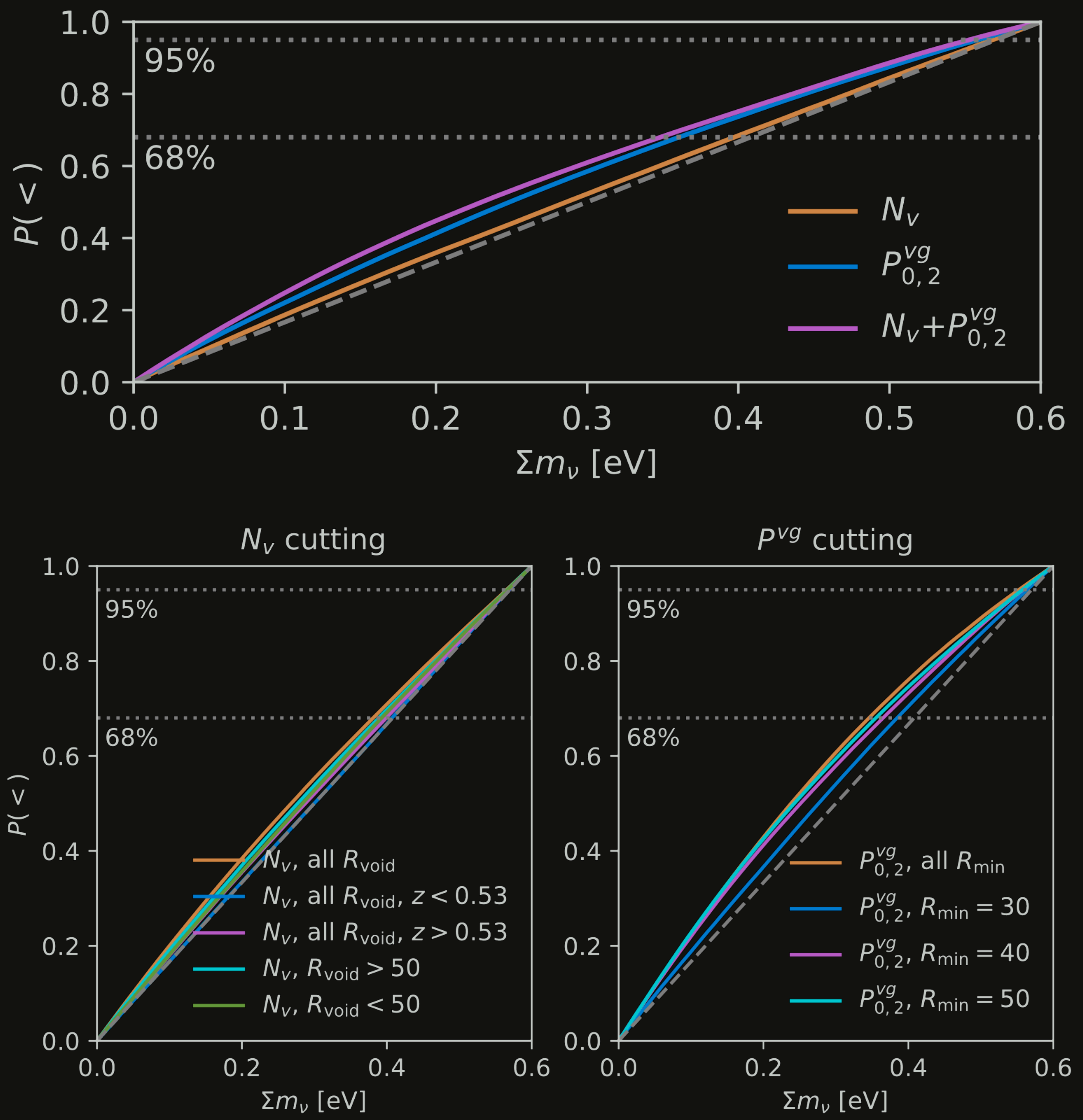
Compare to posteriors on fiducial mocks



Posterior broadening

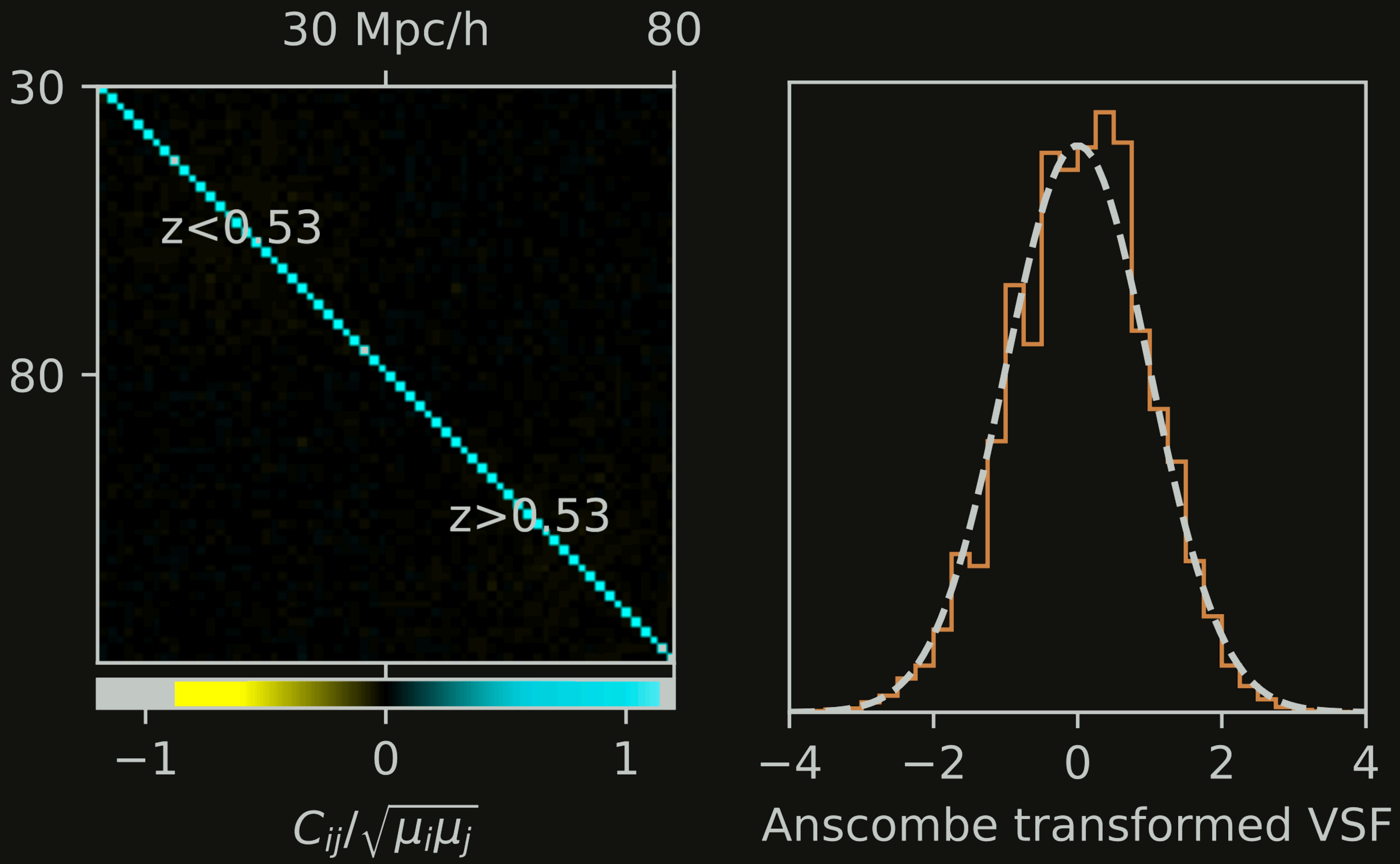


Voids only

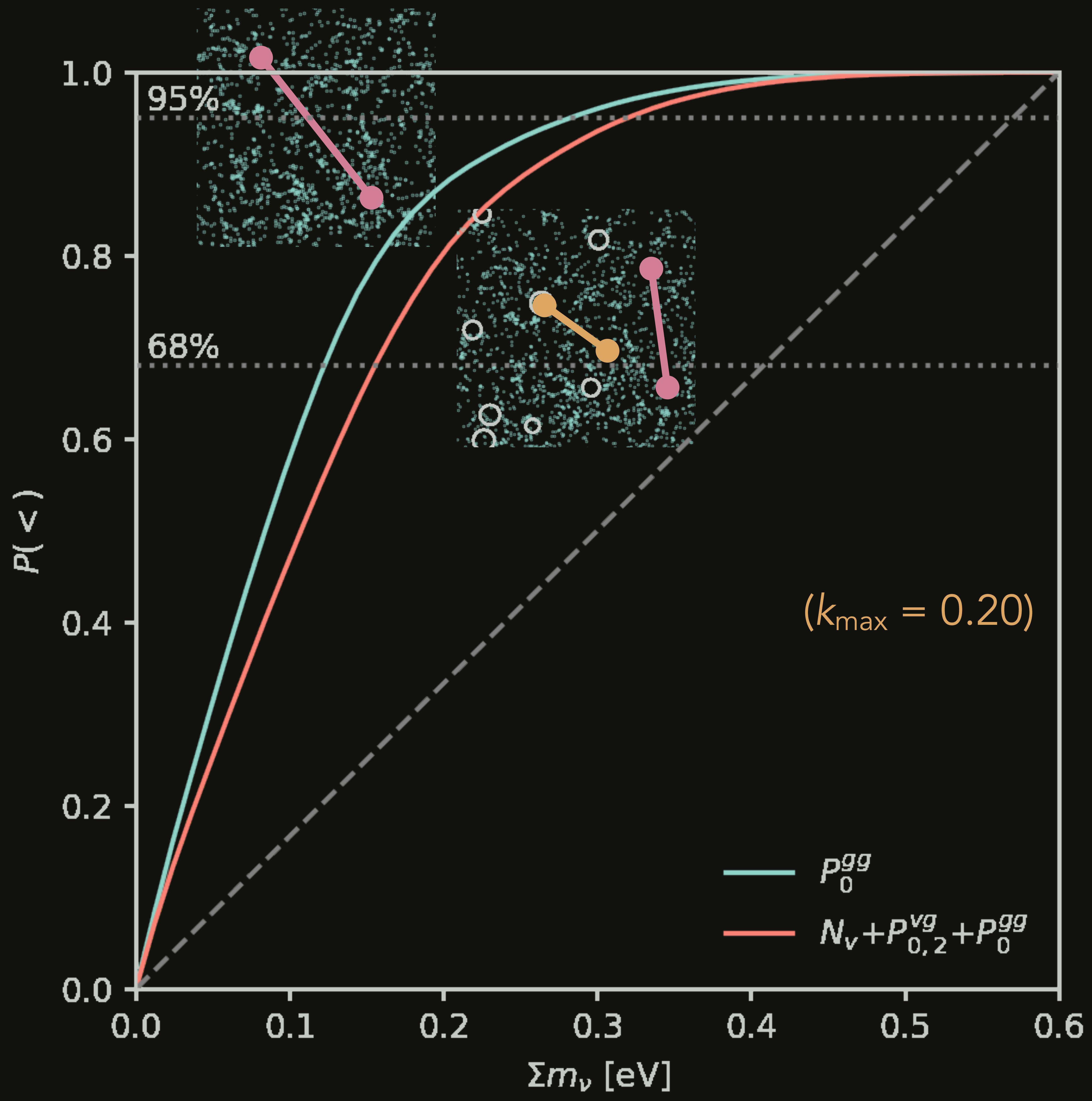


1.0
0.8
0.6
0.4
0.2
0.0
95%
68%
$P(<)$
N_v
$P^{vg}_{0,2}$
N_v+$P^{vg}_{0,2}$
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Σm_ν [eV]
N_v cutting
P^{vg} cutting
N_v, all $R_{\rm void}$
N_v, all $R_{\rm void}$, $z<0.53$
N_v, all $R_{\rm void}$, $z>0.53$
N_v, $R_{\rm void}>50$
N_v, $R_{\rm void}<50$
$P^{vg}_{0,2}$, all $R_{\rm min}$
$P^{vg}_{0,2}$, $R_{\rm min}=30$
$P^{vg}_{0,2}$, $R_{\rm min}=40$
$P^{vg}_{0,2}$, $R_{\rm min}=50$
0.0 0.2 0.4 0.6

Poisson void size function

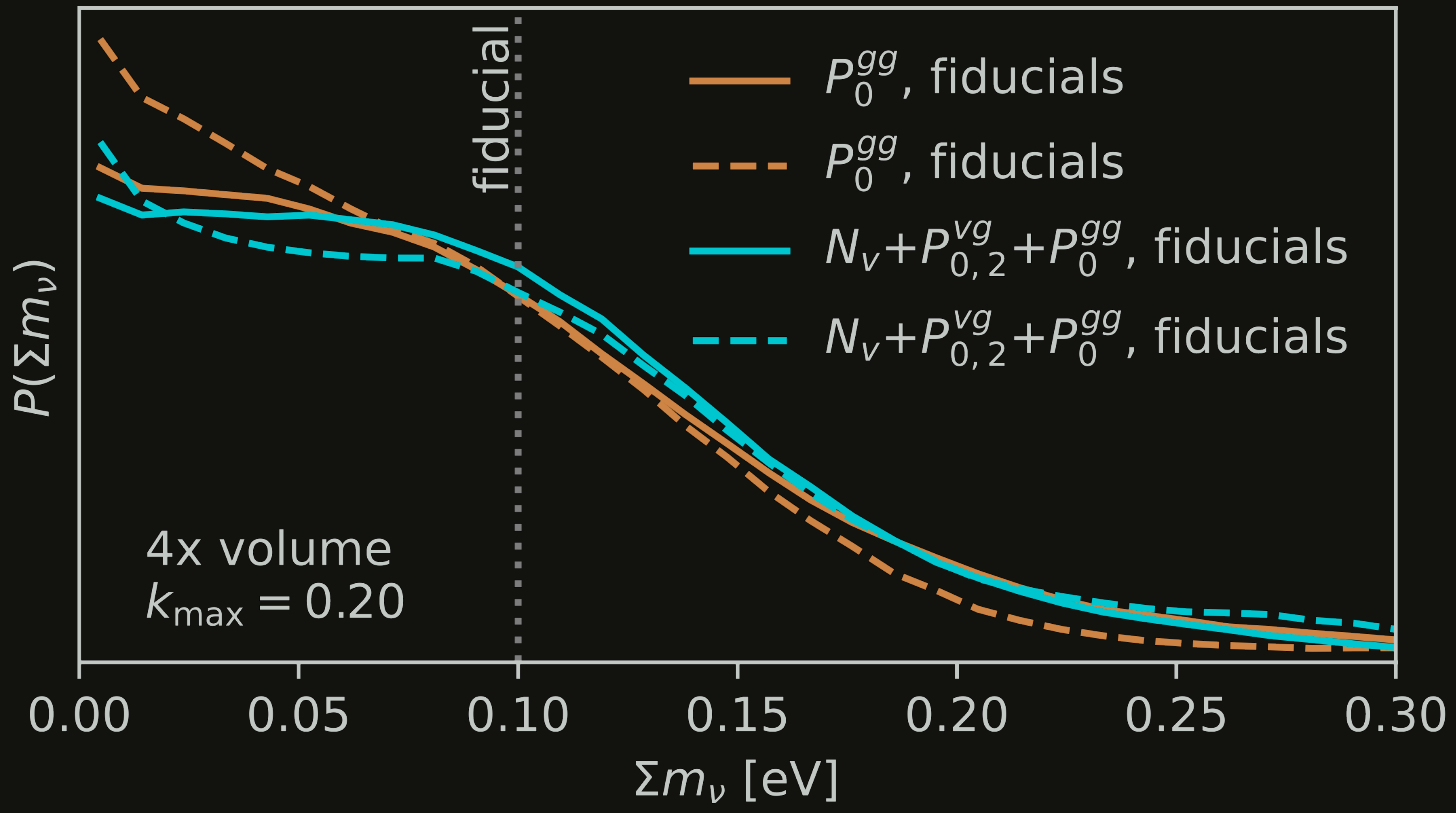


Including smaller scales



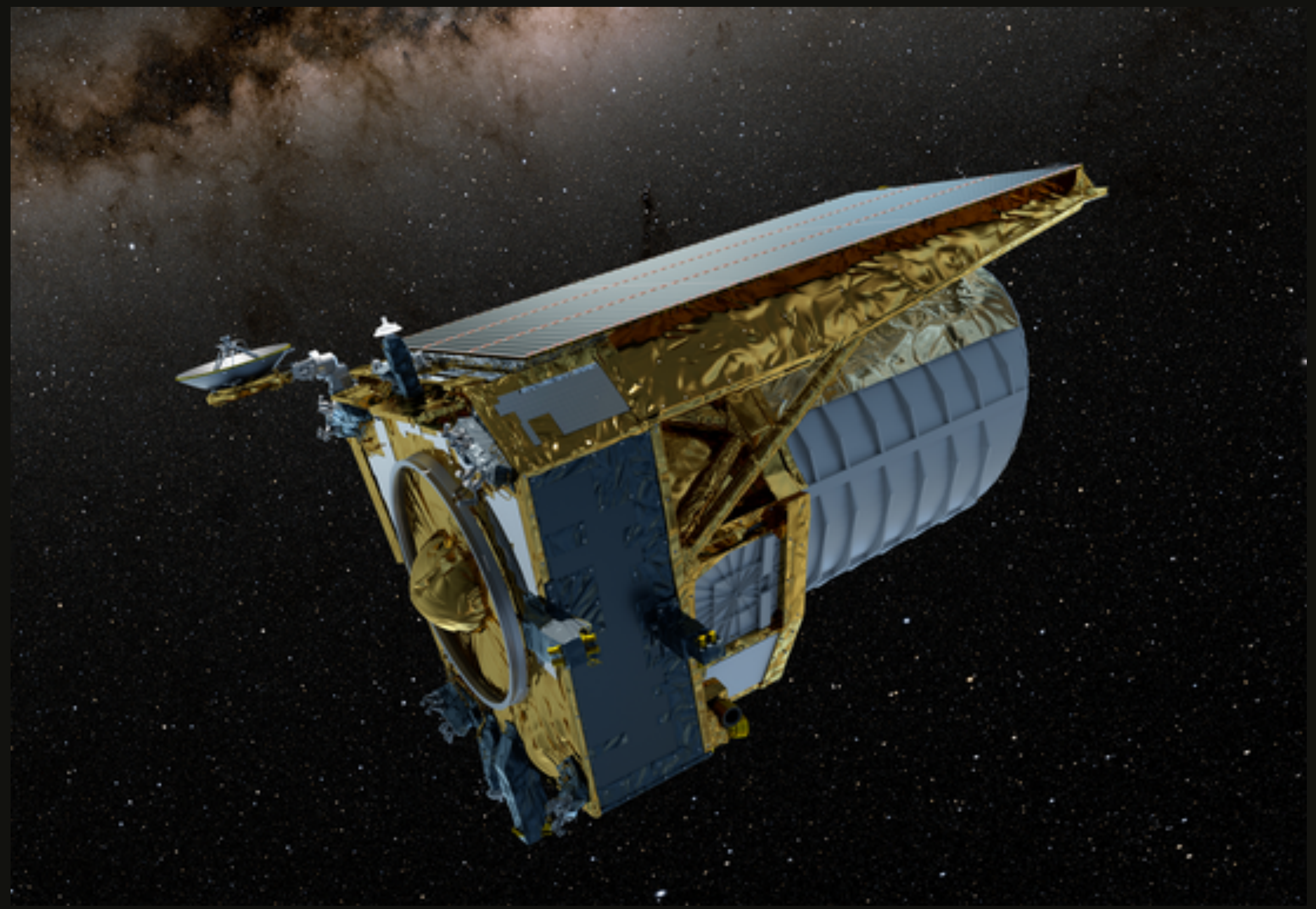
Situation reversed at k_{max}=0.2 hMpc^{-1}.
→ One-sided posteriors are different from Gaussian ones.

Larger volume, on mocks

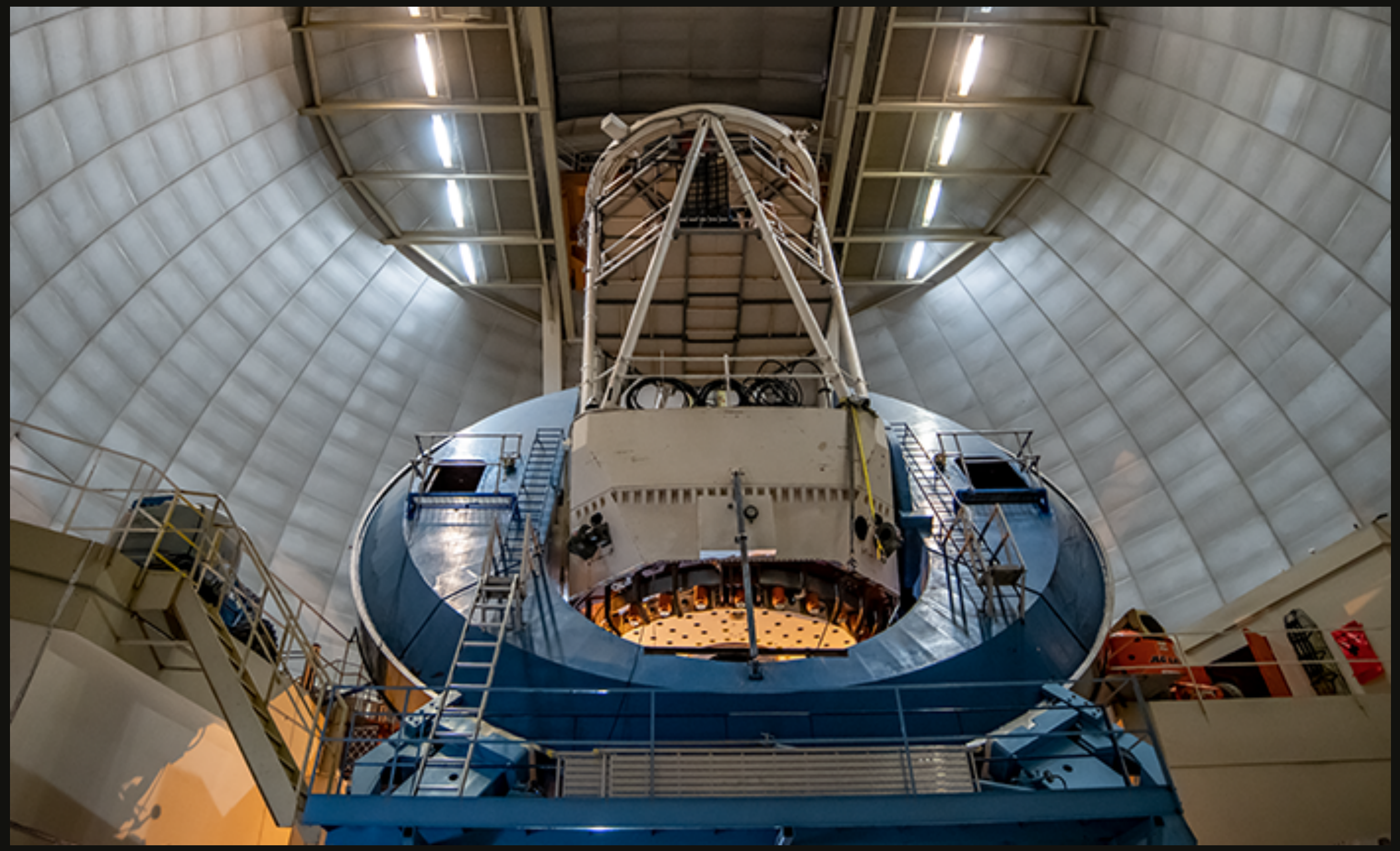


The future for voids

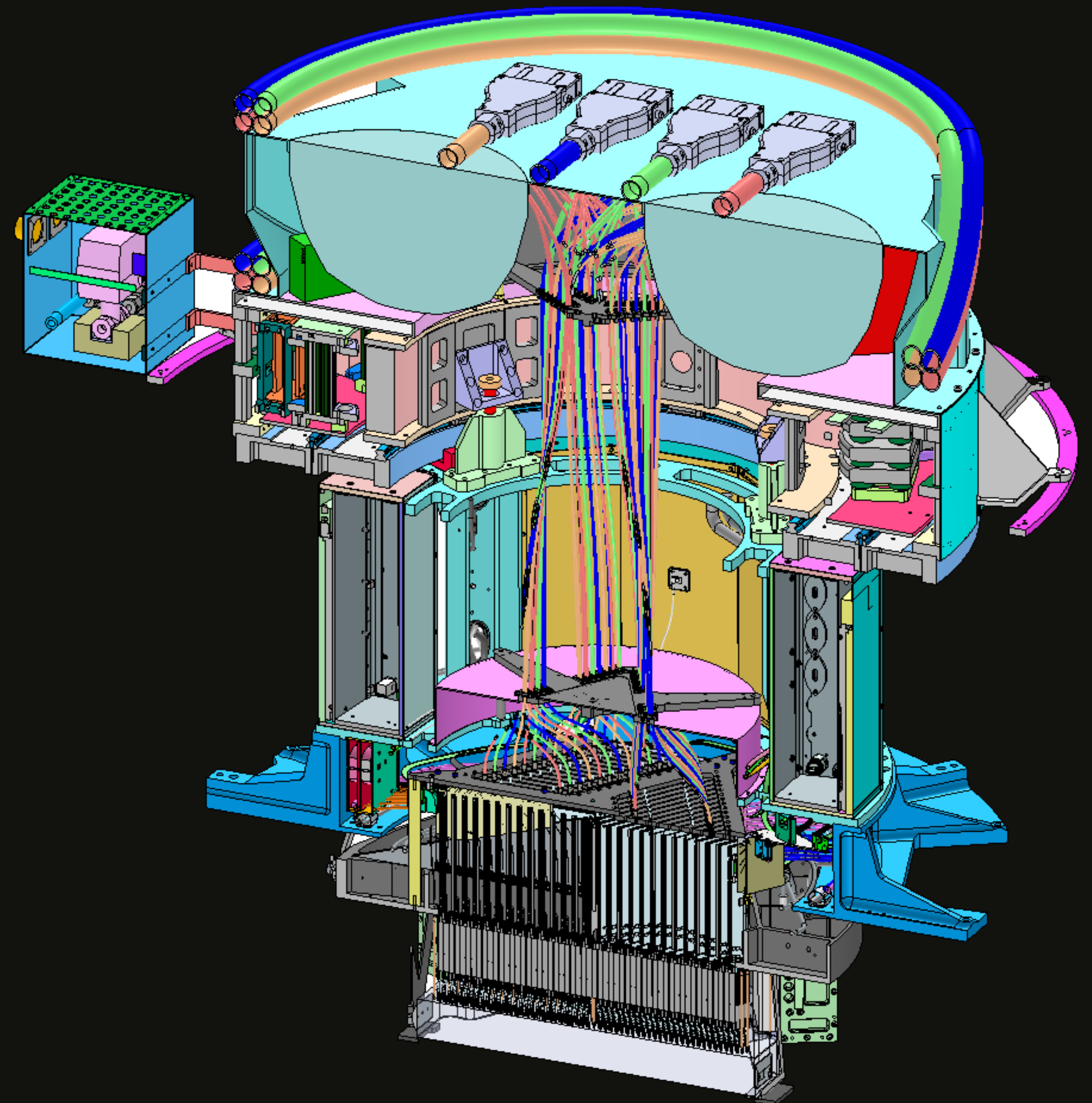
interplay between density contrast and number



Euclid: lots of volume

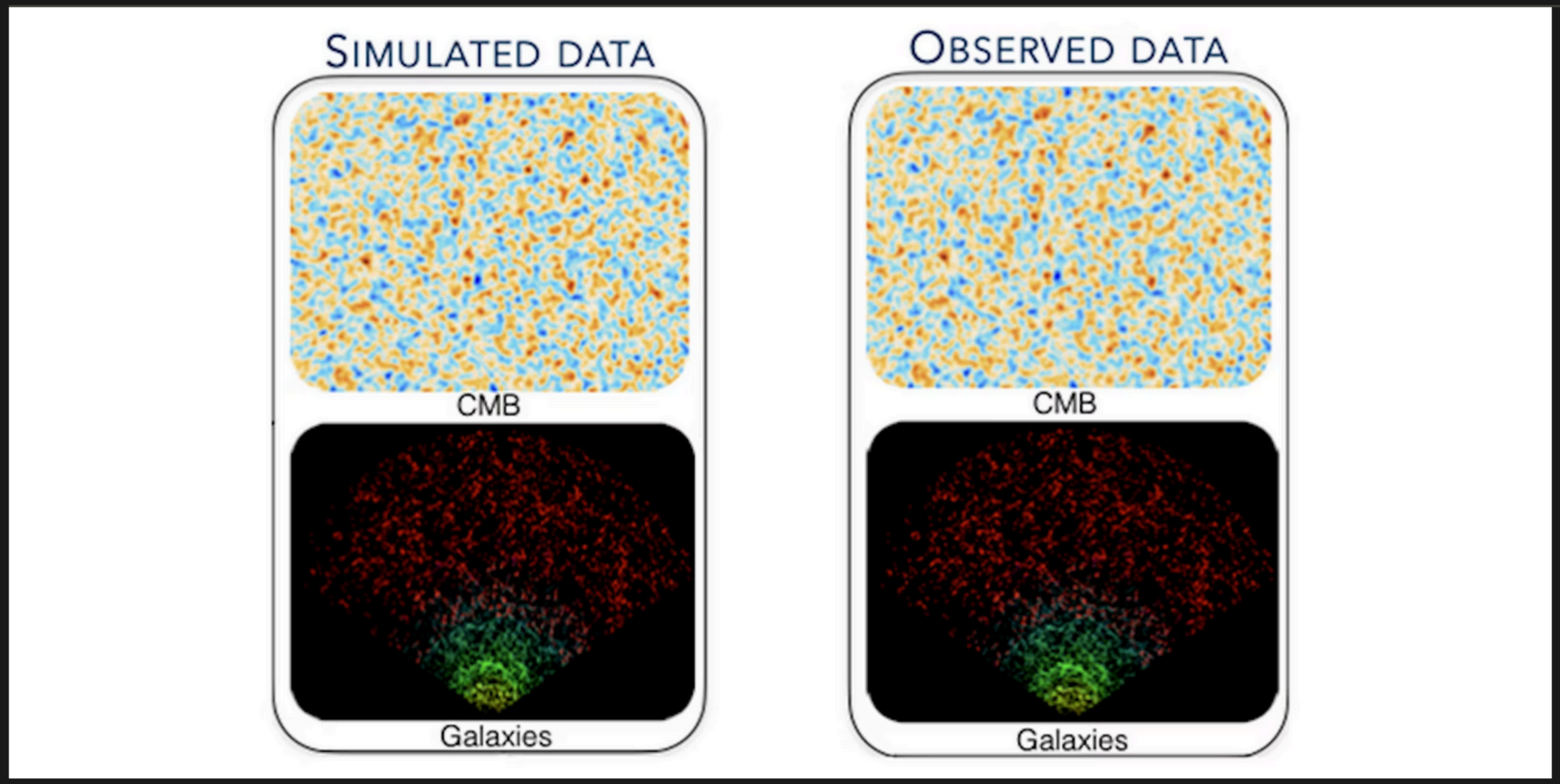


DESI: supercharged BOSS

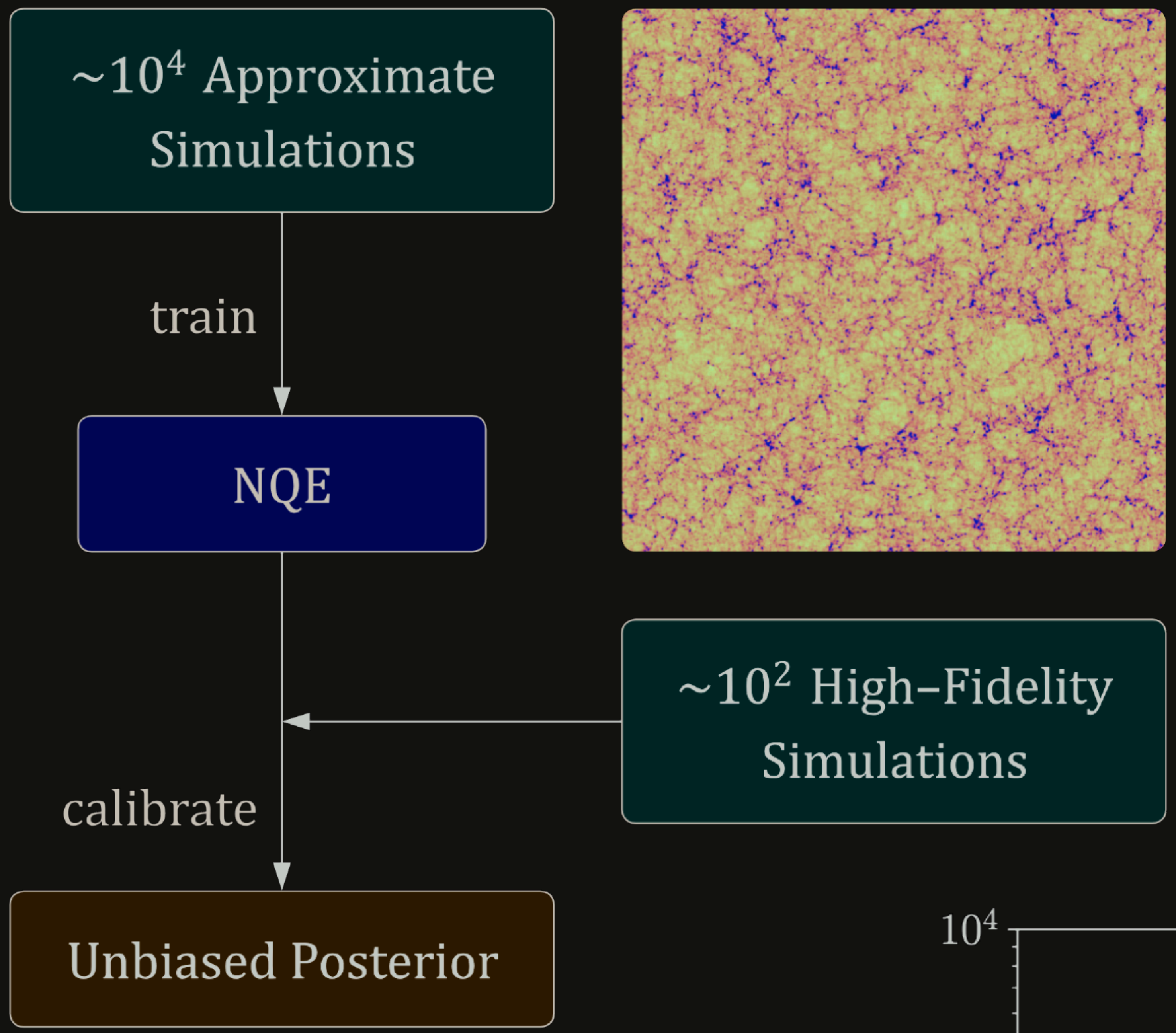


PFS Lyα: clean small high-z voids

Learning the Universe



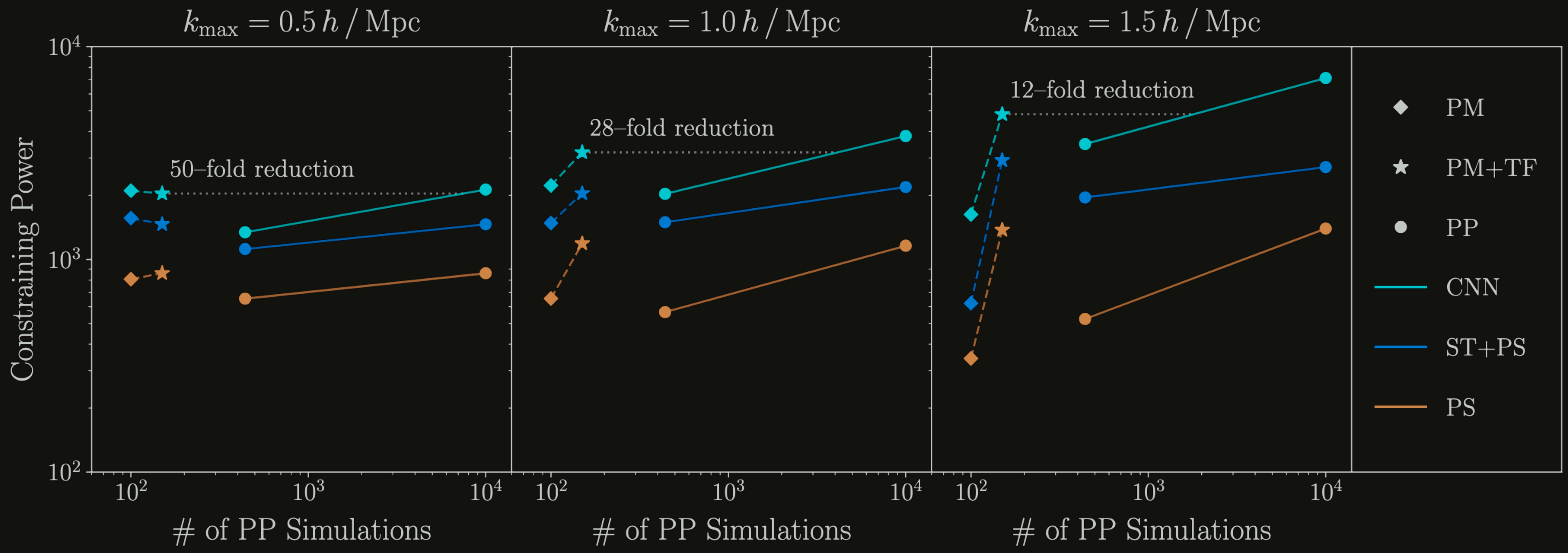
Multi-fidelity



We know that it can work from re-calibration using quantile estimation.

Can we make it optimal?

[He Jia 2024]



Neural ratio estimation

$$P(\text{parameters}\,|\,\text{data}) = \boxed{\frac{P(\text{data}\,|\,\text{parameters})}{P(\text{data})}}\,P(\text{parameters})$$

- draw θ, θ' ~ P(parameters)
- simulate x_{sim} ~ P(data | θ) [drawing η, ζ]
- evaluate neural net: $y=f(x_{sim}, \theta)$, $y'=f(x_{sim}, \theta')$
- classification loss, e.g. $\mathcal{L} = -\log(y) - \log(1-y')$
- train $f^* = \text{argmin} \int D\theta\, D\theta'\, Dx_{sim}\, \mathcal{L}$
- → $P(x \,|\, \theta) / P(x) = f^*(x, \theta) / [\, 1 - f^*(x, \theta)\,]$

- any simulatable effect can be incorporated
- rephrase as classification problem → sophisticated machinery exists
- no formal difference between nuisance parameters and initial conditions