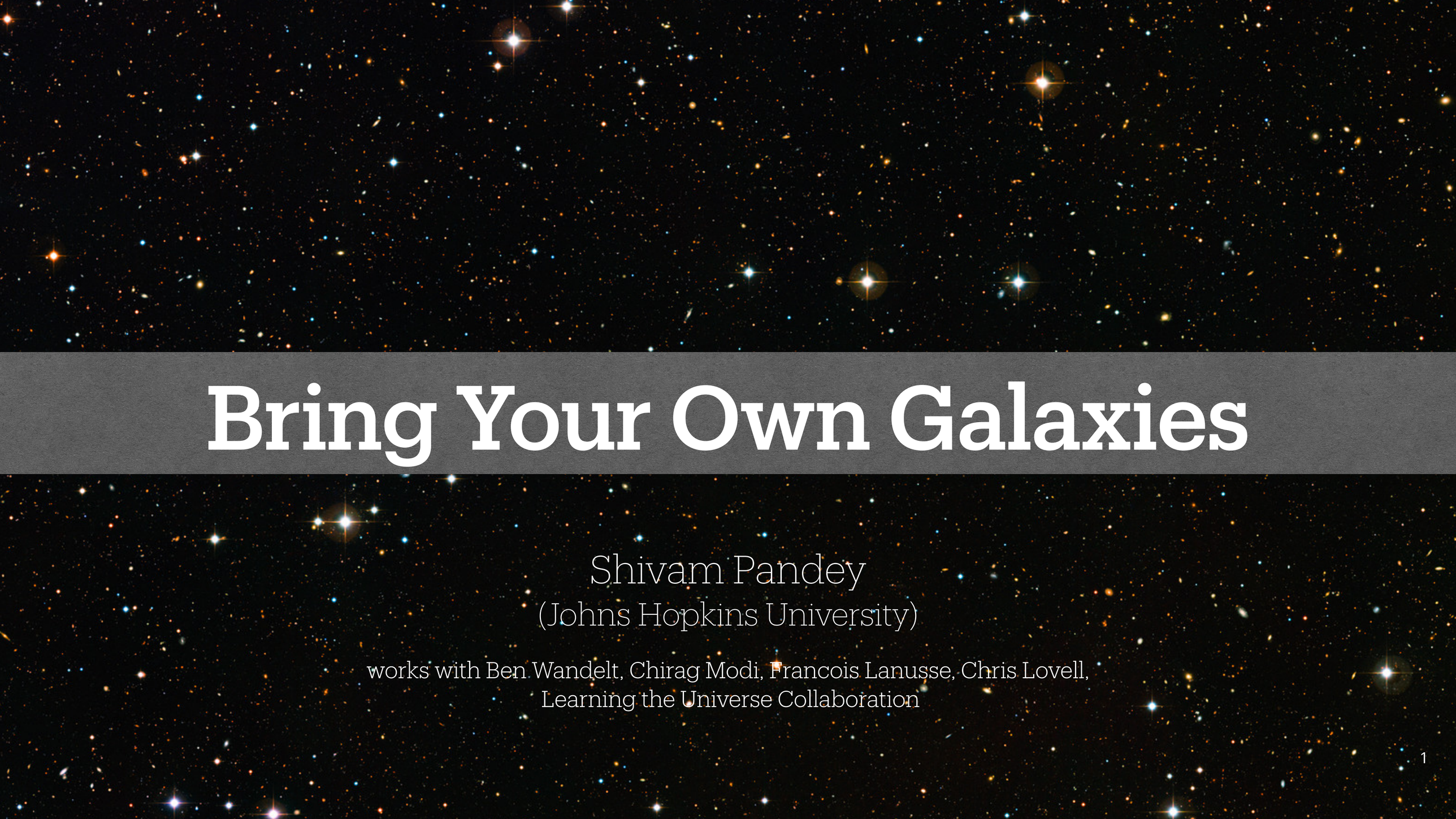


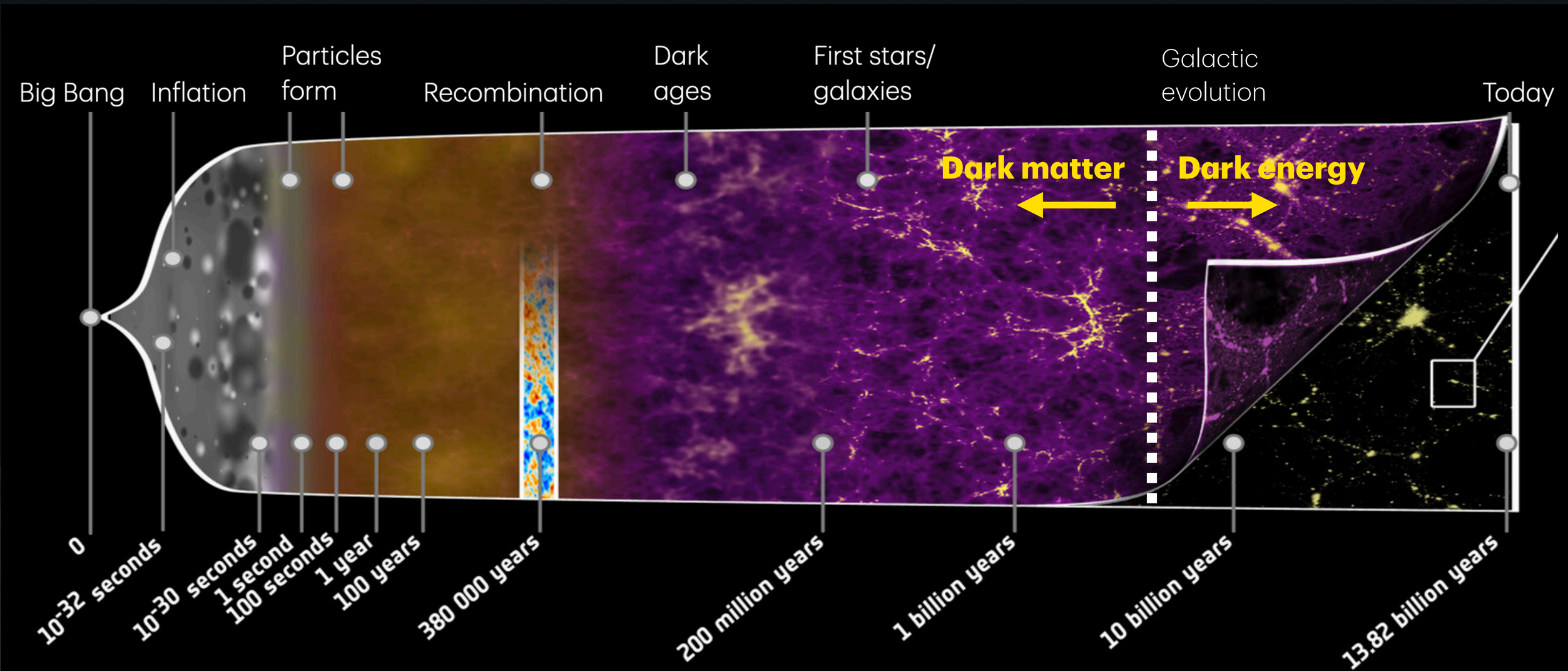
Bring Your Own Galaxies

Shivam Pandey
(Johns Hopkins University)

works with Ben Wandelt, Chirag Modi, Francois Lanusse, Chris Lovell,
Learning the Universe Collaboration

What we want to understand…

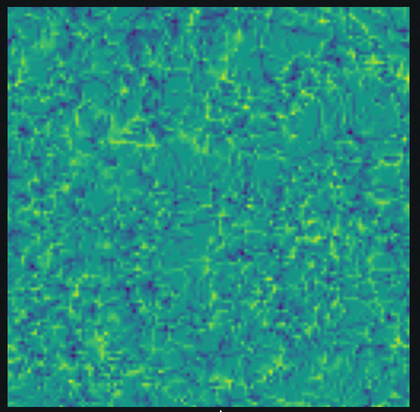
Evolution of the Large Scale Structure



Credits: ESA

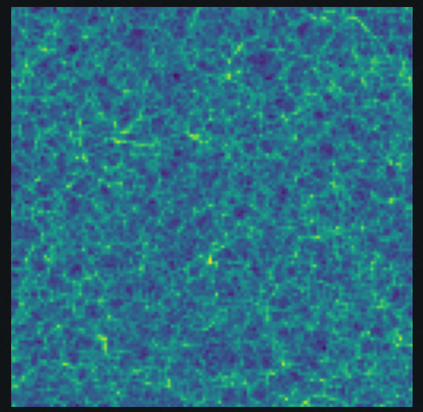
Traditional Inference pipeline

Initial conditions



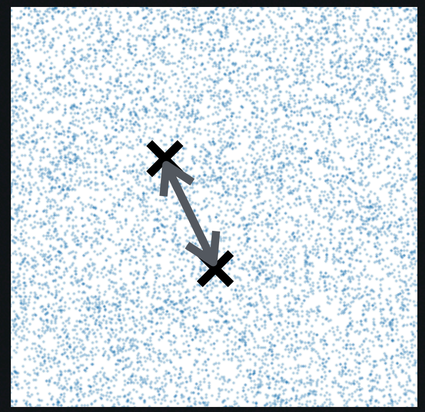
Non-linear Evolution

Matter field

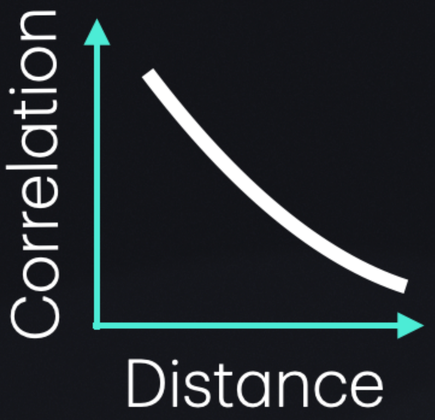


Matter-observable connection

Observables



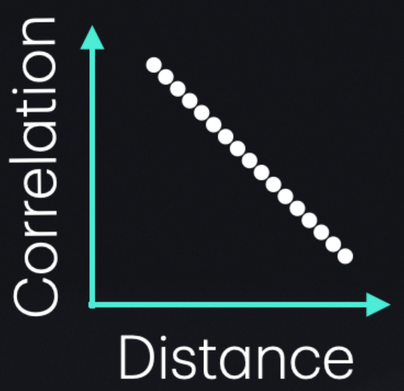
2pt function



$\vec{\theta} = \{\text{Cosmology, ...}\}$

Data

Measurements



Analysis

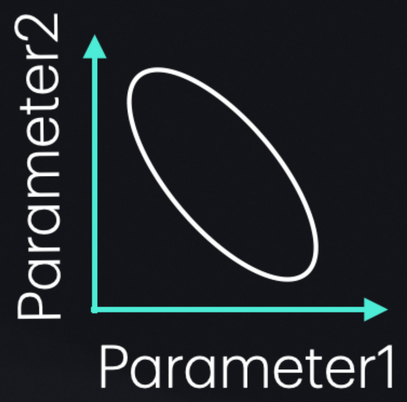
Covariance

+

Likelihood

+ MCMC

Constraints



Traditional Inference pipeline

Initial conditions

Non-linear Evolution

Matter field

Matter-observable connection

Observables

2pt function

Correlation

Distance

Analytic Modeling

$\vec{\theta} = \{\text{Cosmology, ...}\}$

Data

Measurements

Correlation

Distance

Analysis

Covariance

+

Likelihood

+ MCMC

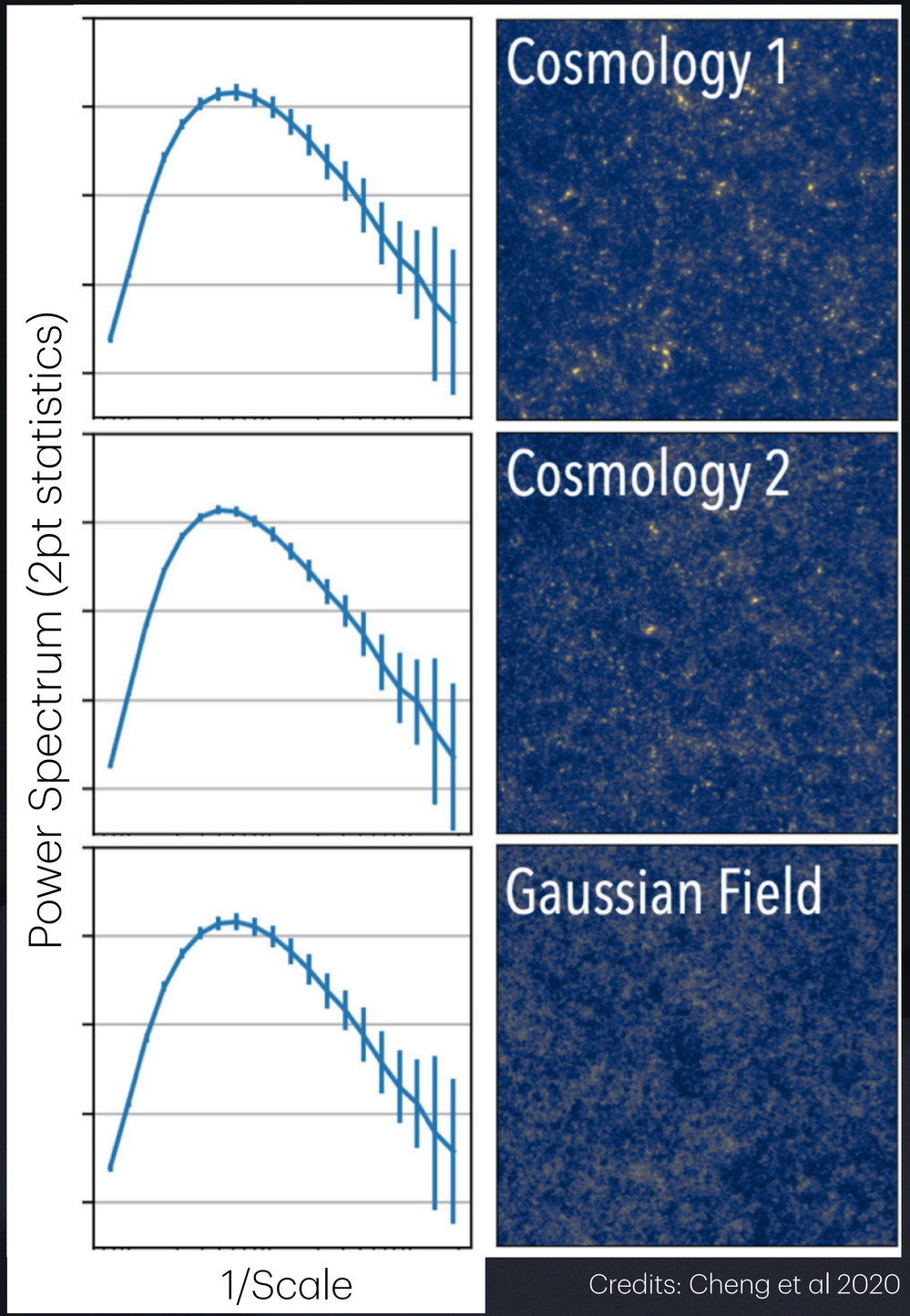
Constraints

Parameter2

Parameter1

There is much more information in the observation!

2pt correlation alone does not capture all the information in a non-gaussian field.



Credits: Cheng et al 2020

How do we do better?

- Go beyond 2pt statistics!
- Multiple options:
  - Marked correlation function
  - Wavelets transforms
  - Graph neural networks
  - Convolutional neural networks
  - ..... (many more)

How do we do better?

- Go beyond 2pt statistics!
- Multiple options:
  - Marked correlation function
  - Wavelets transforms
  - Graph neural networks
  - Convolutional neural networks
  - ..... (many more)

(Too non-linear)

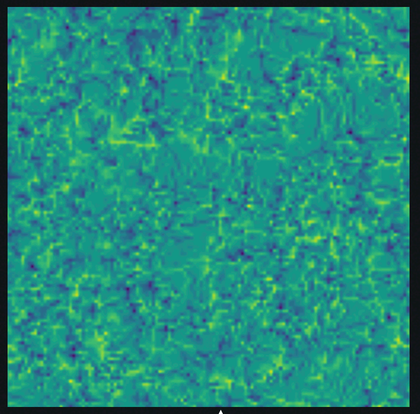
(Too non-linear)

(Need too many simulations)

(Away from central limit theorem)

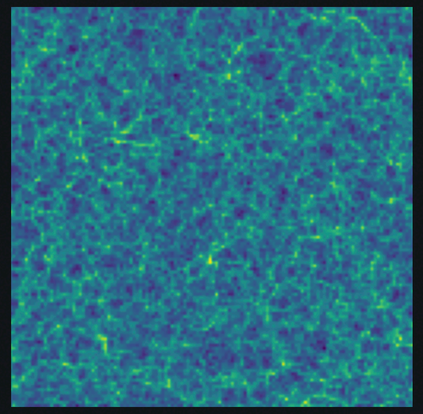
Simulation Based Inference (SBI) pipeline

Initial conditions



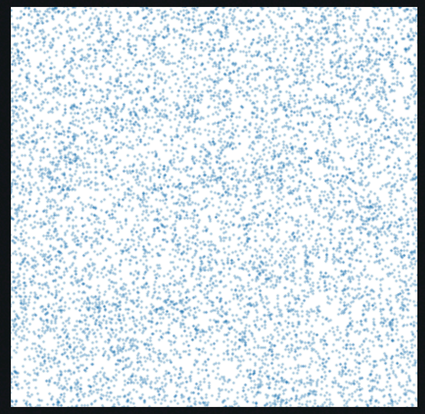
Non-linear Evolution

Matter field



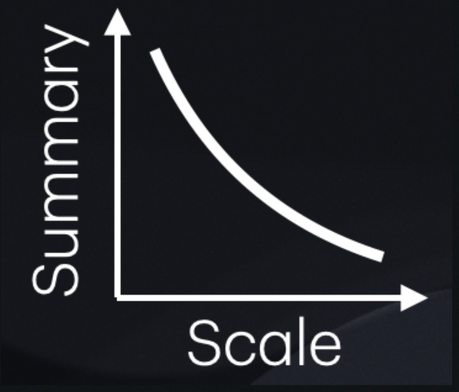
Matter-observable connection

Observables



$\mathcal{F}$(observable)

Summary statistics



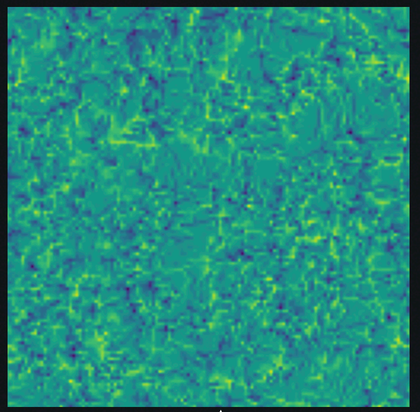
$\vec{\theta} = \{\text{Cosmology, ...}\}$

Neural networks

$p(\theta, \text{summary})$

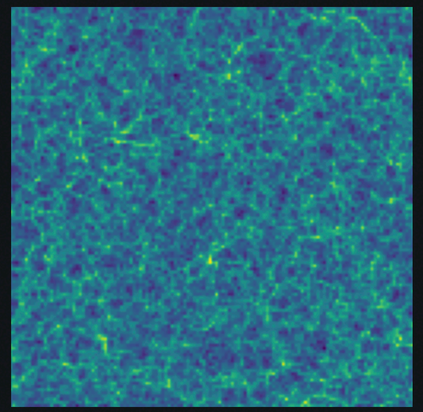
Simulation Based Inference (SBI) pipeline

Initial conditions



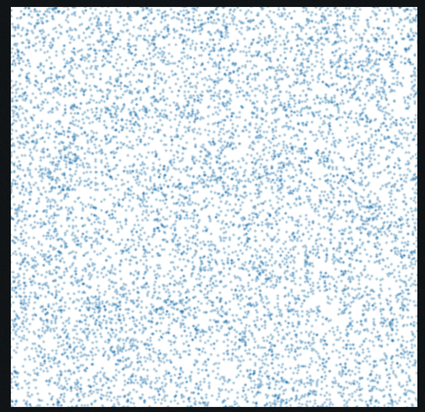
Non-linear Evolution

Matter field



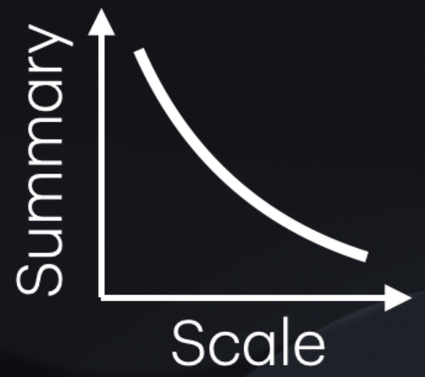
Matter-observable connection

Observables



$\mathcal{F}$(observable)

Summary statistics

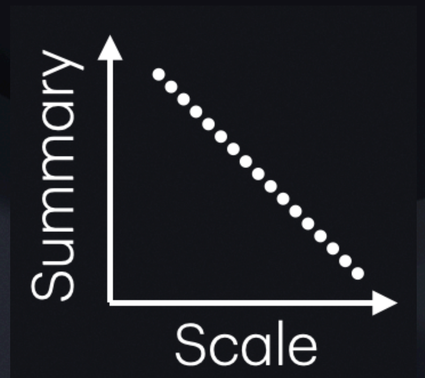


$\vec{\theta} = \{\text{Cosmology, ...}\}$

Data

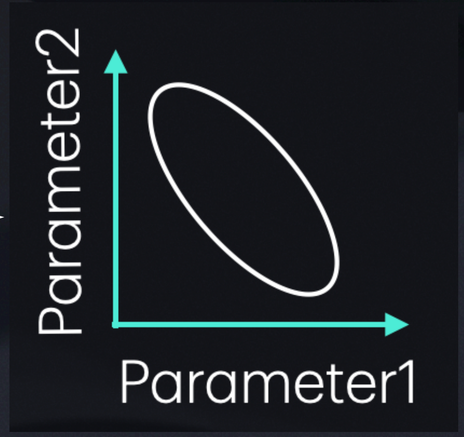
$\mathcal{F}$(observation)

Measurements



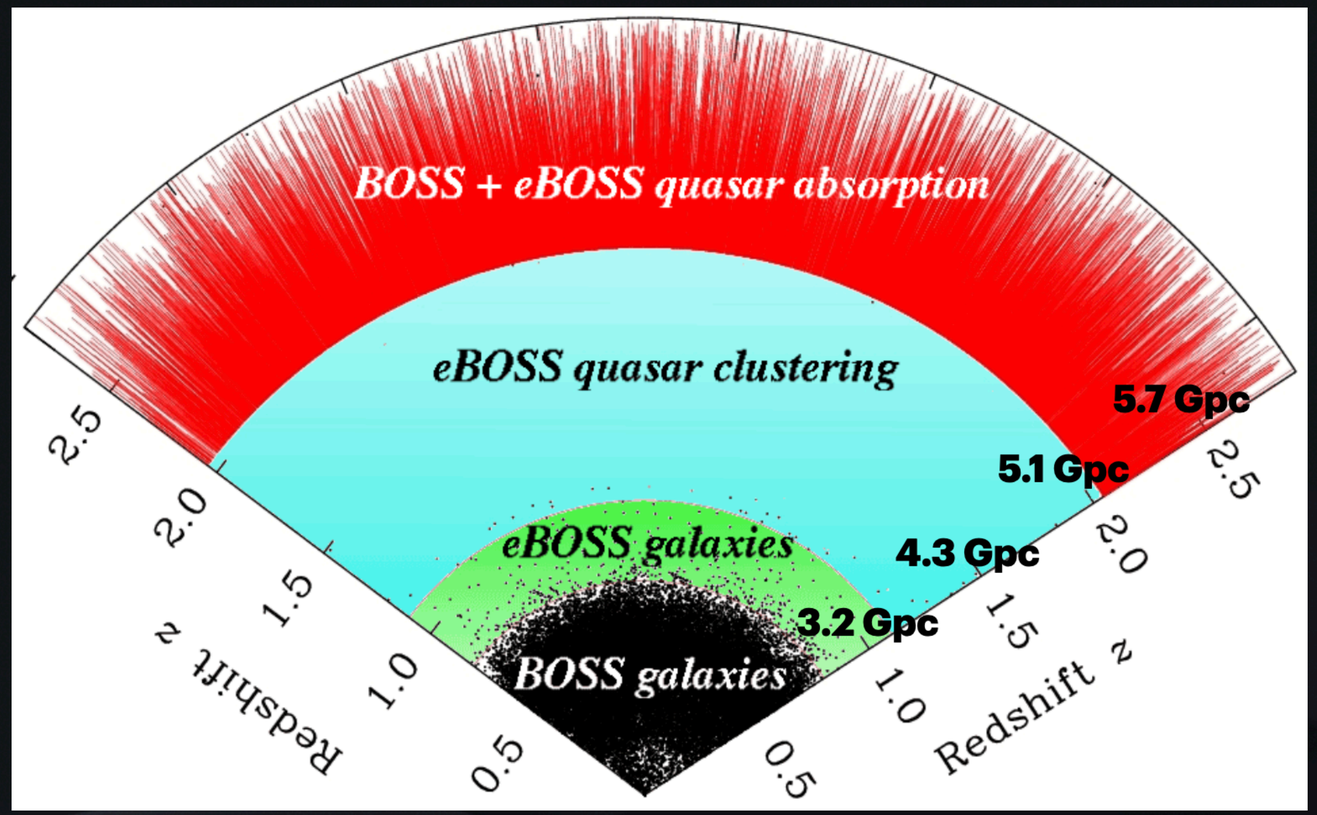
Neural networks

Neural posterior
$p(\theta, \text{summary})$
$p(\theta \mid \text{data})$



What are the main bottlenecks in this pipeline?

Simulation volume



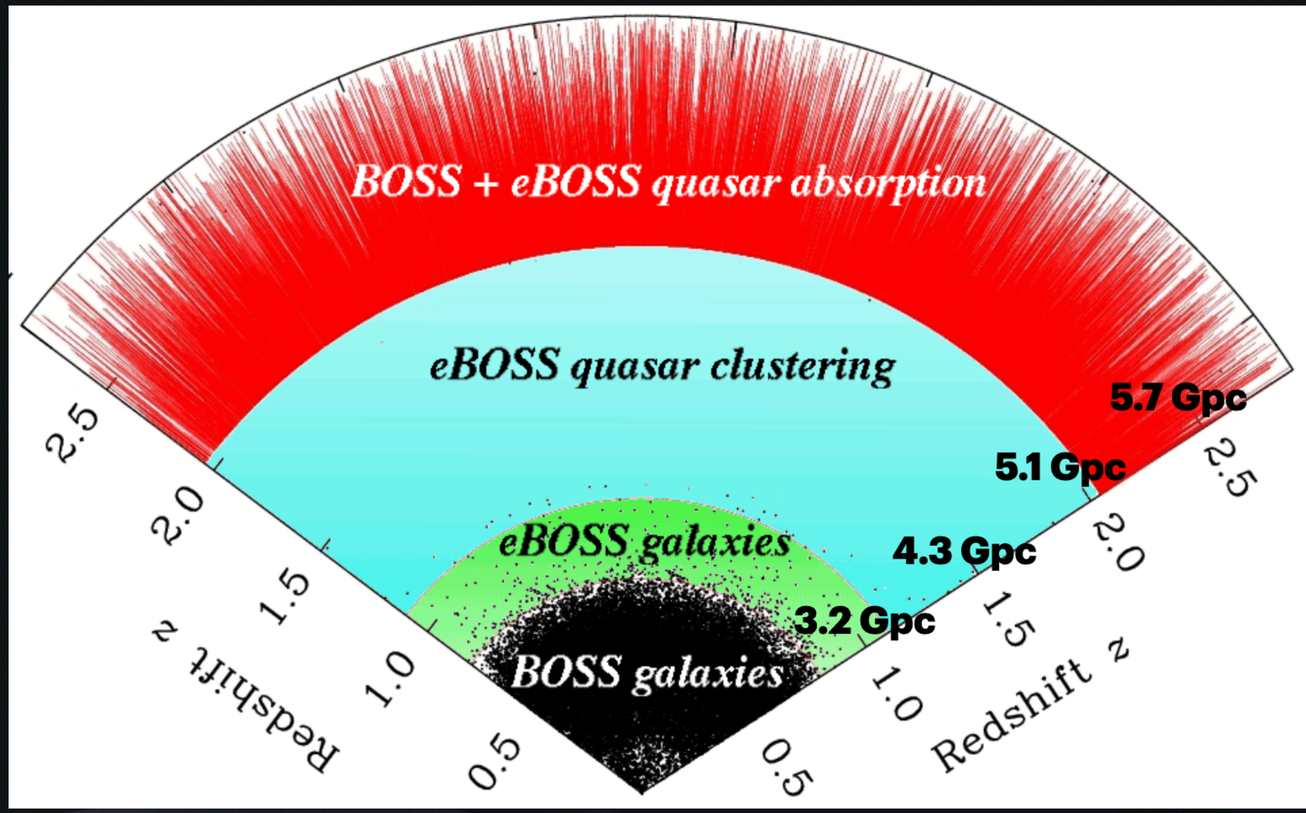
5.7 Gpc

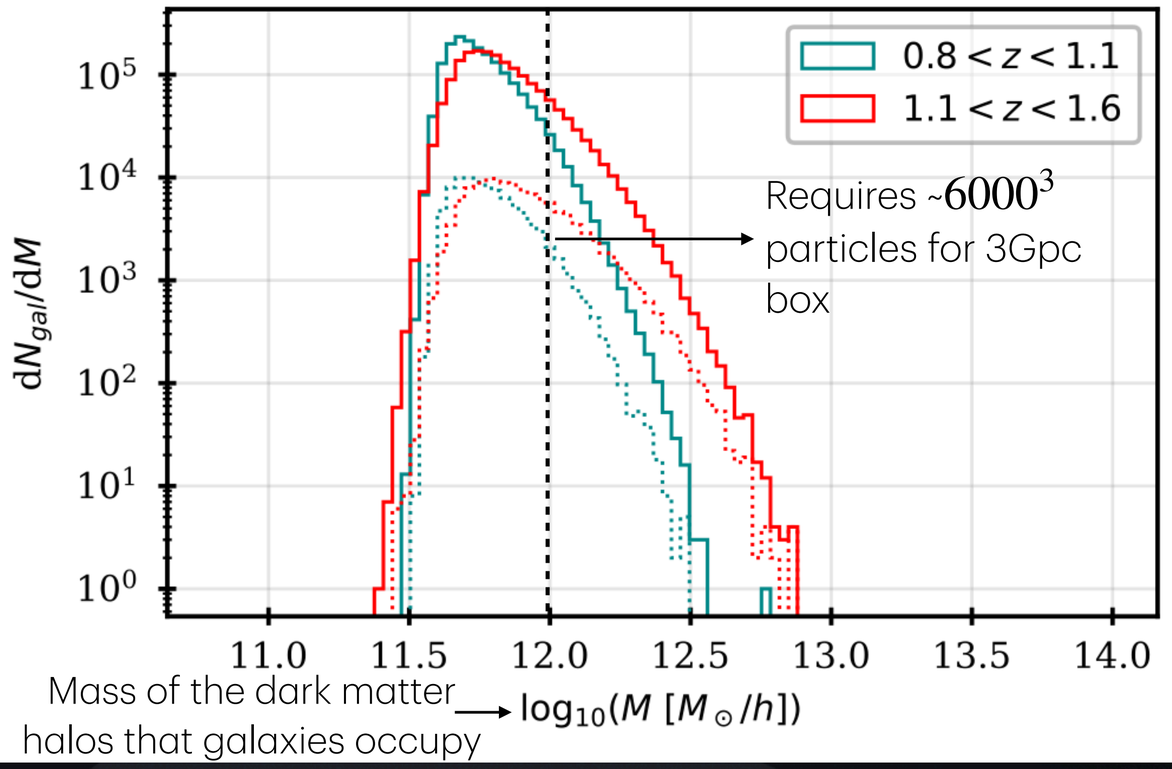
5.1 Gpc

4.3 Gpc

3.2 Gpc

Simulation volume *and* resolution



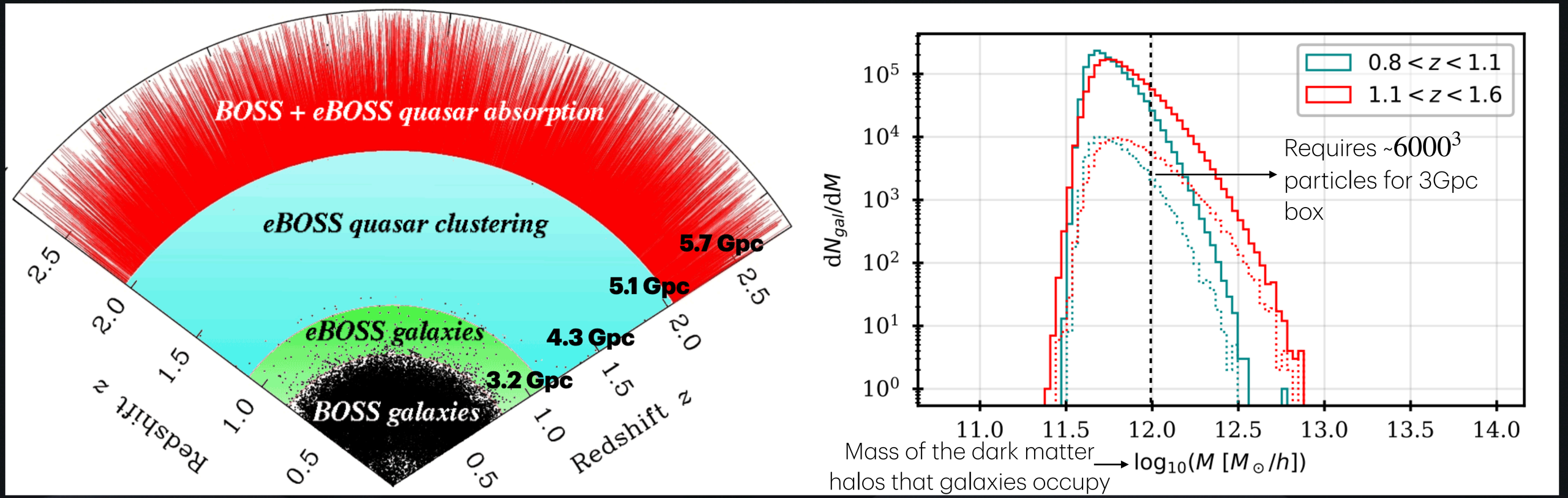


Requires ~6000^3 particles for 3Gpc box

Mass of the dark matter halos that galaxies occupy

Rocher et al 24

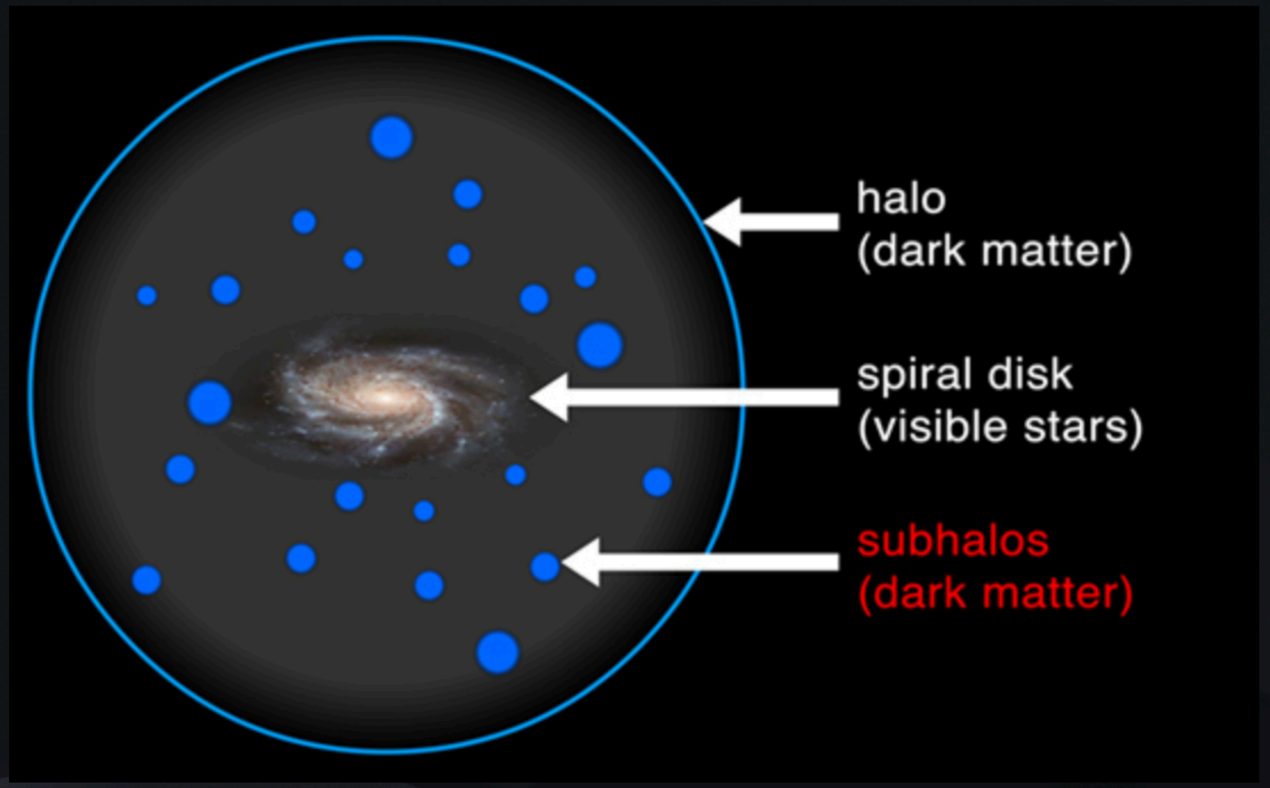
Simulation volume *and* resolution



Rocher et al 24

Typically need to run a lot of them to compare against the observations using techniques like SBI as mentioned before

Let's start with accelerating halo finding in the most extensive simulation suite we have

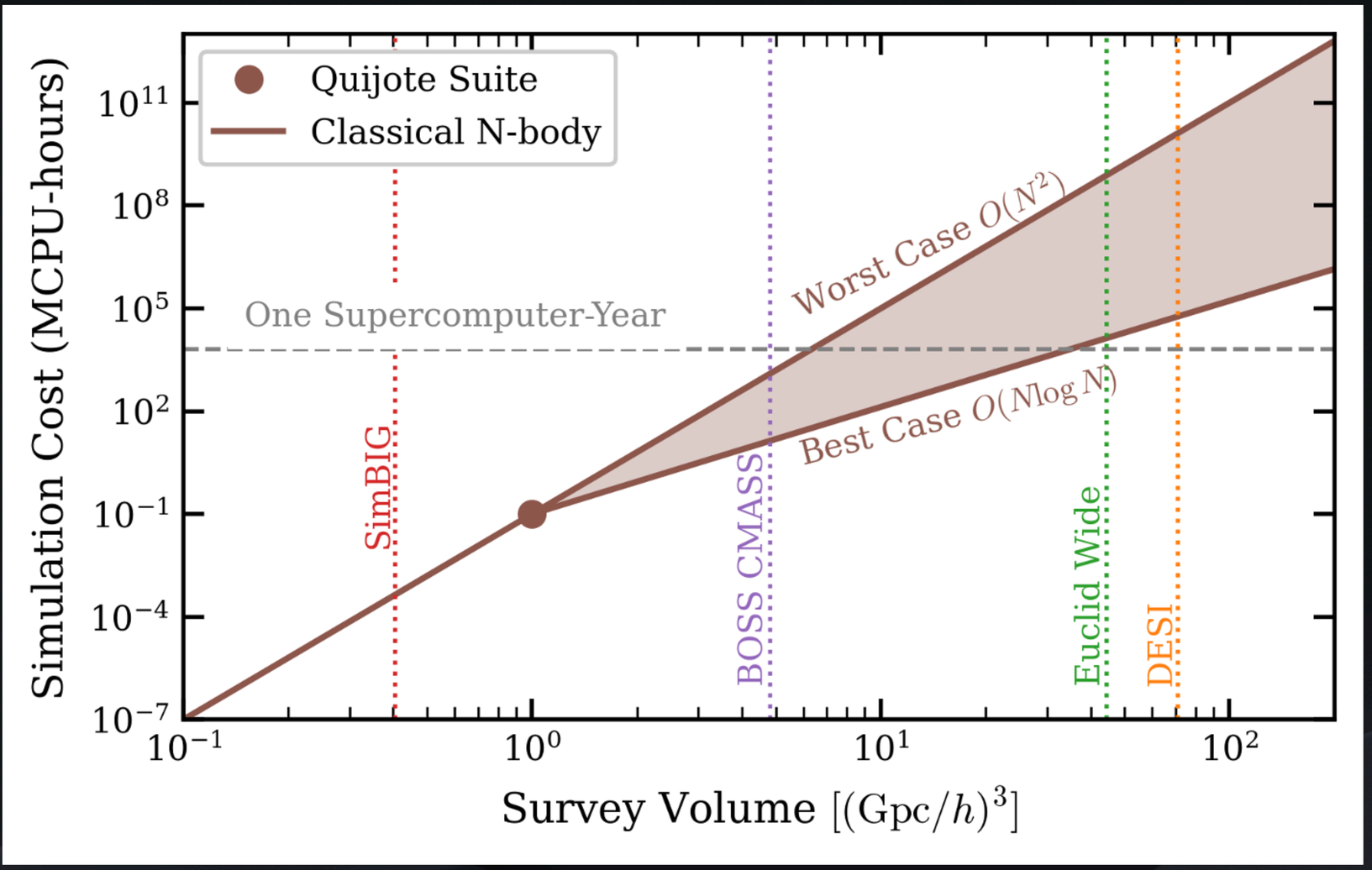


Credit: Luna Zagorac + Astrobites

Quijote N-body

- $1\,(\mathrm{Gpc}/h)^3$ volume, 1024^3 particles
- ~2000 sims varying cosmologies
- ~5000 CPU hours/sim
- $10^{13}\,M_\odot$ halo mass resolution

Quijote N-body

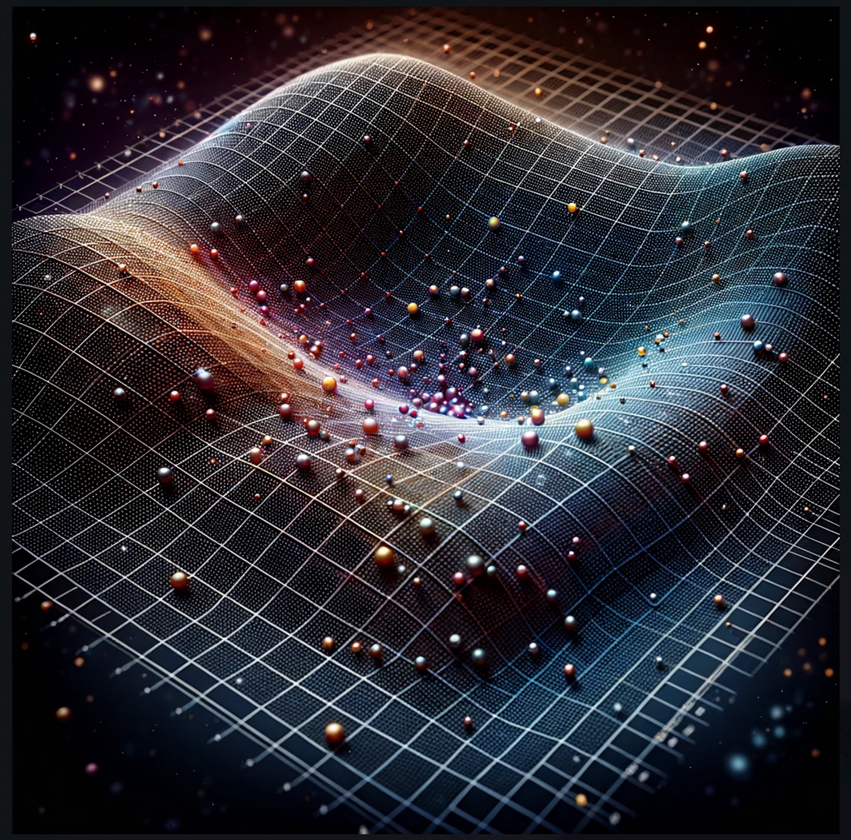


Credit: Matt Ho + LtU-ILI team

Naively scaling Quijote simulations to the needed volume will not work!

- $1\,(\mathrm{Gpc}/h)^3$ volume, 1024^3 particles
- ~2000 sims varying cosmologies
- ~5000 CPU hours/sim
- $10^{13}\,M_\odot$ halo mass resolution
- *>270 million CPU hours needed for full SDSS-like volume LH set*

Particle mesh

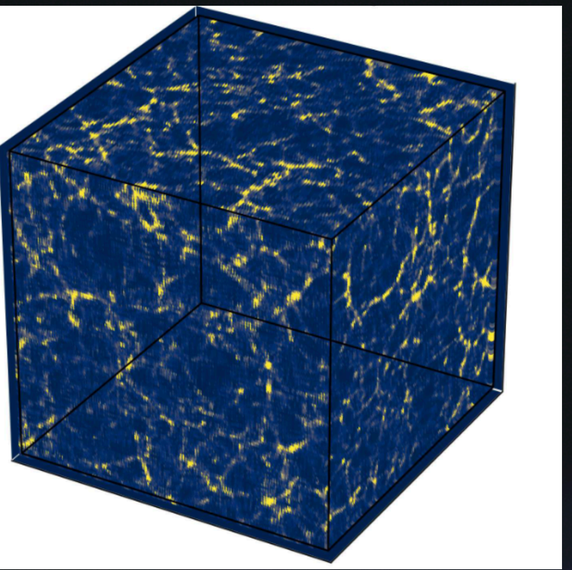


- $1\,(\mathrm{Gpc}/h)^3$ volume
- 384^3 particles
- ~5 CPU hours/sim or 10 GPU sec/sim
- IC —> Matter can be differentiable and much faster with GPUs

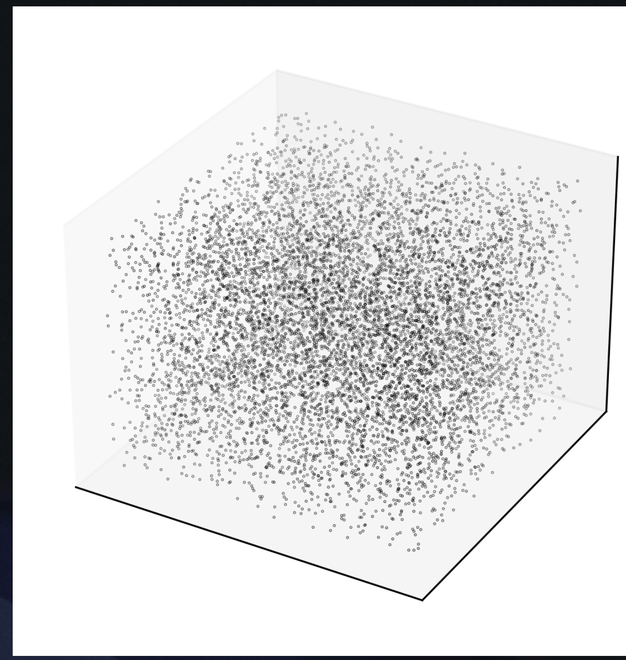
Quijote N-body

- $1\,(\mathrm{Gpc}/h)^3$ volume
- 1024^3 particles
- ~5000 CPU hours/sim
- Not-differentiable
- *>270 million CPU hours needed for full SDSS-like volume LH set*

Connect PM to N-body halos



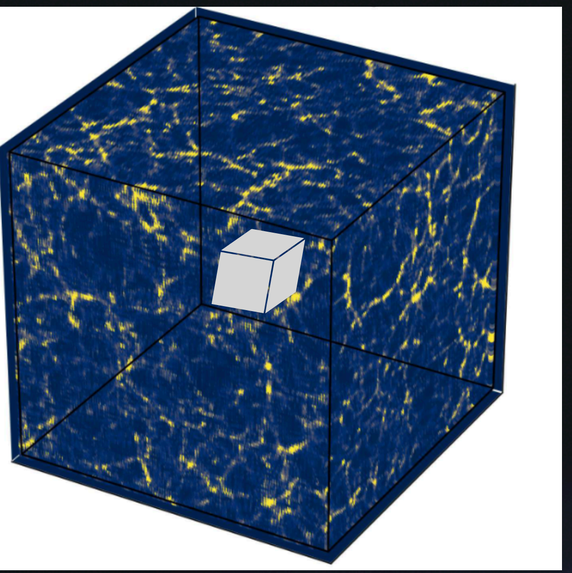
?



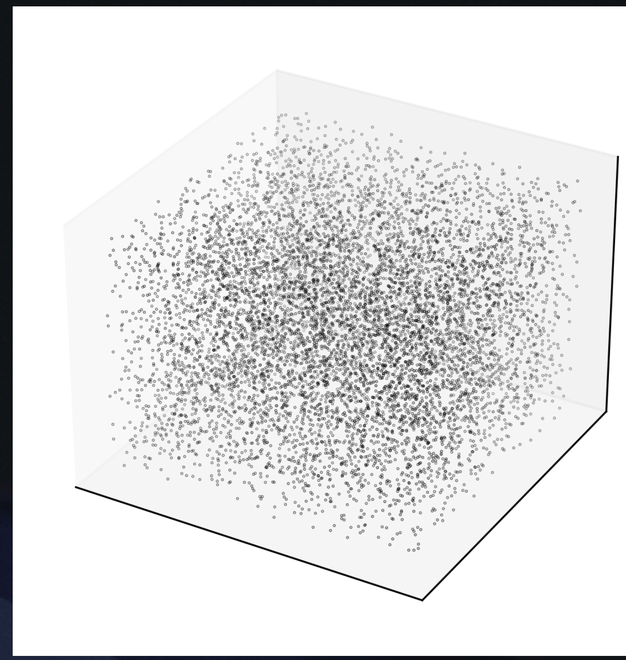
3D
Matter density field
(PM Simulation)
Input

3D
Halo distribution
(N-body Simulation)
Target

Voxelize the 3D volume (~8Mpc voxels)



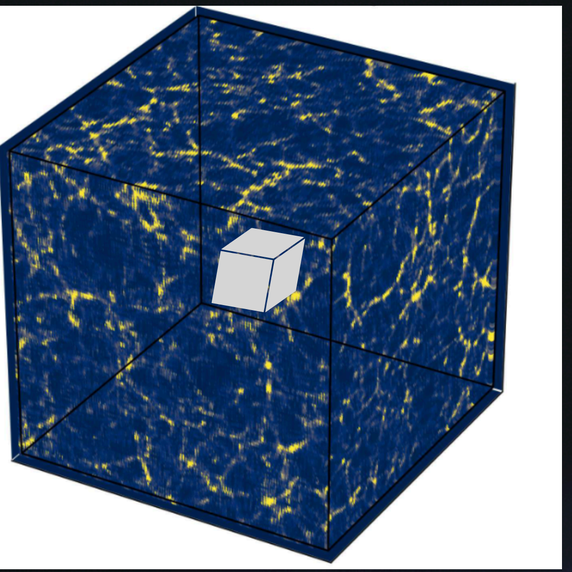
$$p(\{M_1, M_2, \ldots, M_{N_j}\} \mid \vec{\delta}_j)$$



3D
Matter density field
(PM Simulation)
Input

3D
Halo distribution
(N-body Simulation)
Target

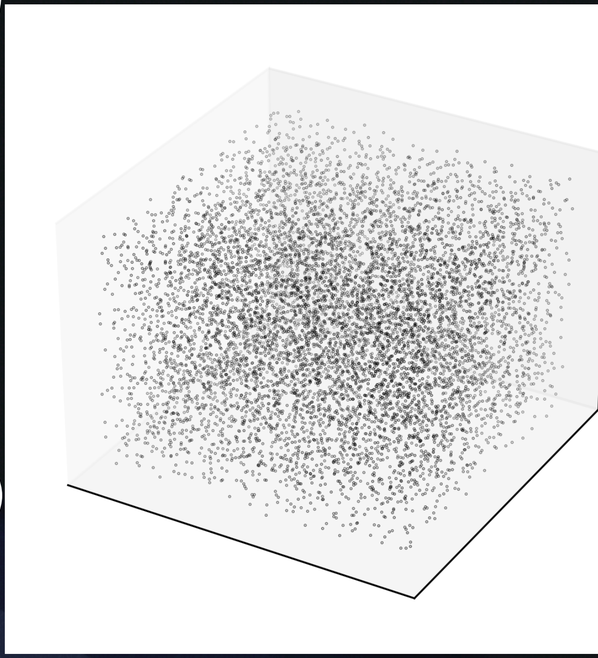
Creating Halos with Auto-Regressive Multi-stage networks (CHARM, arxiv:2409.09124)



$$p(\{M_1, M_2, \ldots, M_{N_j}\} \mid \vec{\delta}_j) =$$

$$p(N_j \mid \vec{\delta}_j) \times p(M_1 \mid \vec{\delta}_j, N_j) \times p(M_2 \mid M_1, \vec{\delta}_j, N_j)$$

$$\times p(M_{N_j} \mid \{M_1, \ldots N_{j-1}\}, \vec{\delta}_j, N_j)$$

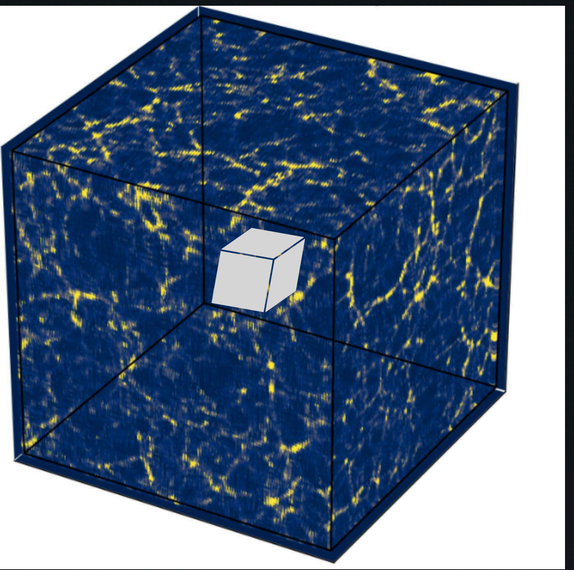


3D
Matter density field
(PM Simulation)
Input

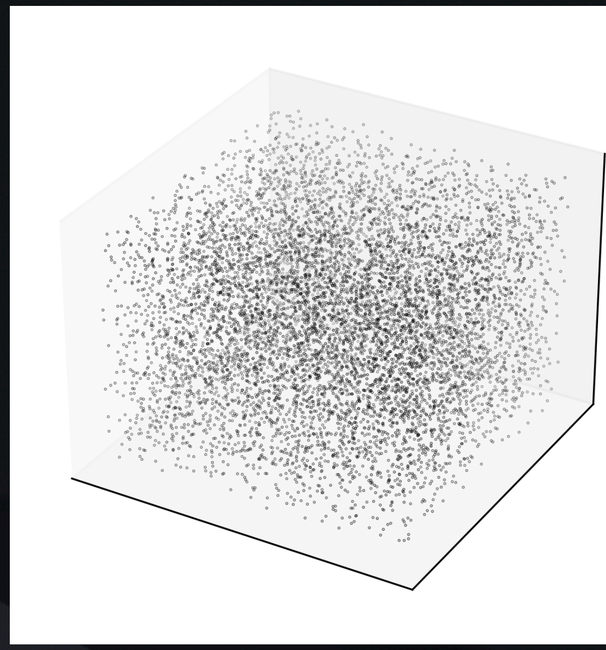
3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

arxiv:2409.09124



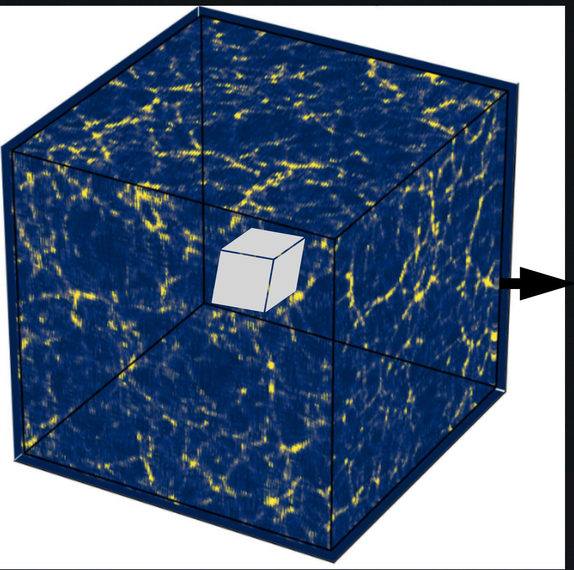
3D
Matter density field
(PM Simulation)
Input



3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

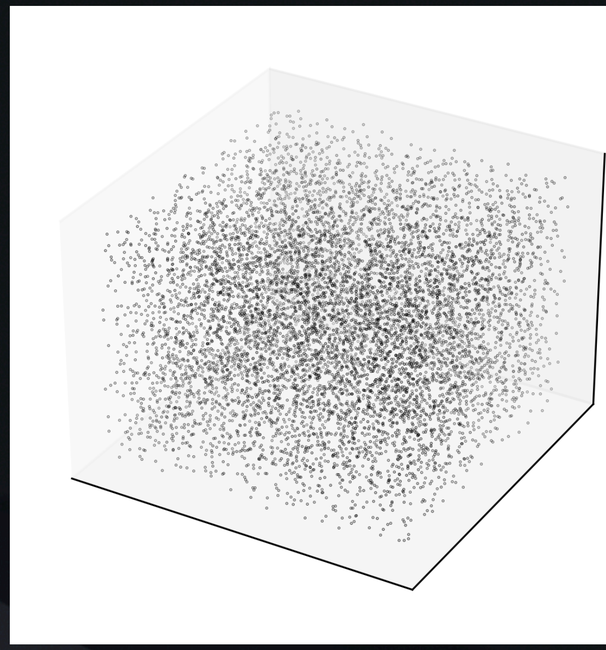
arxiv:2409.09124



Feature extraction

Learned Features

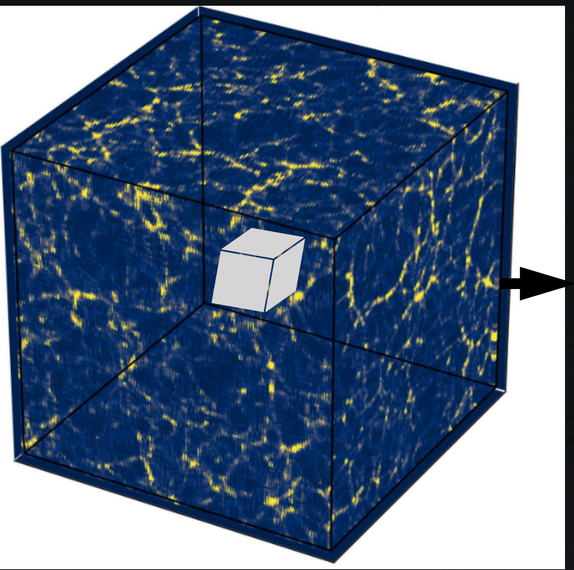
3D
Matter density field
(PM Simulation)
Input



3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

arxiv:2409.09124



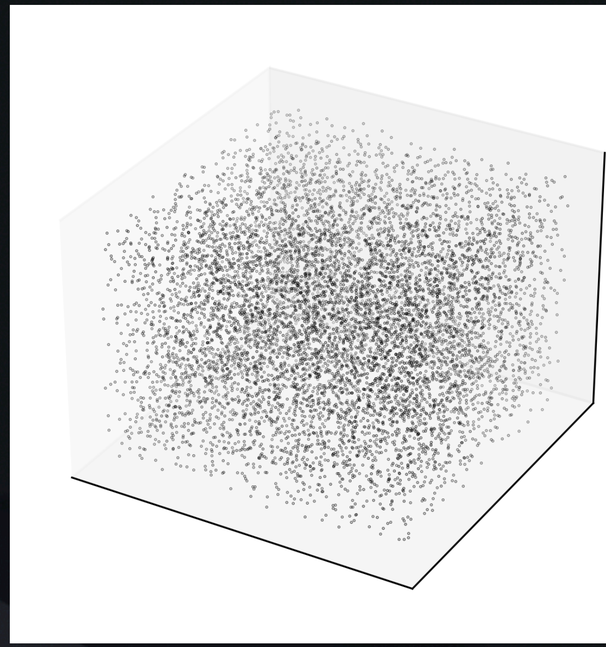
3D
Matter density field
(PM Simulation)
Input

ResNet x2

Feature extraction

Learned Features

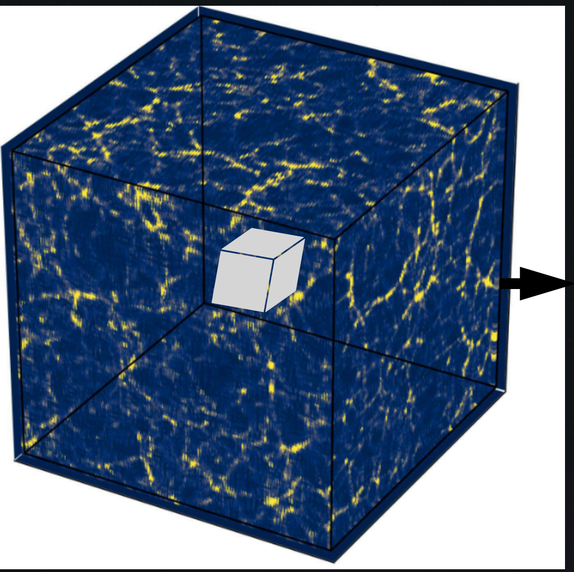
Cosmological Parameters



3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

arxiv:2409.09124



3D
Matter density field
(PM Simulation)
Input

ResNet x2

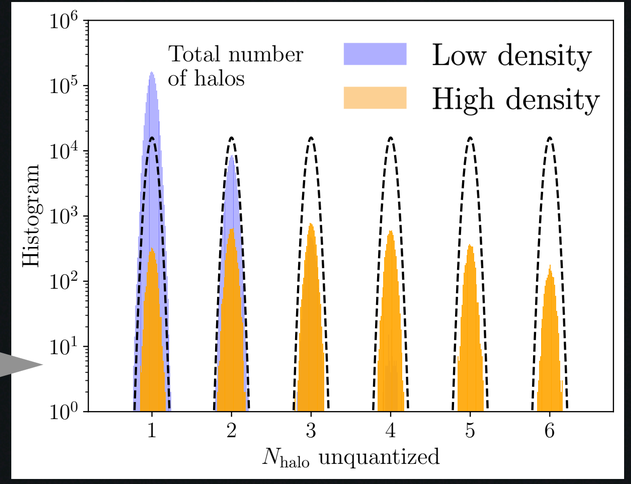
Feature extraction

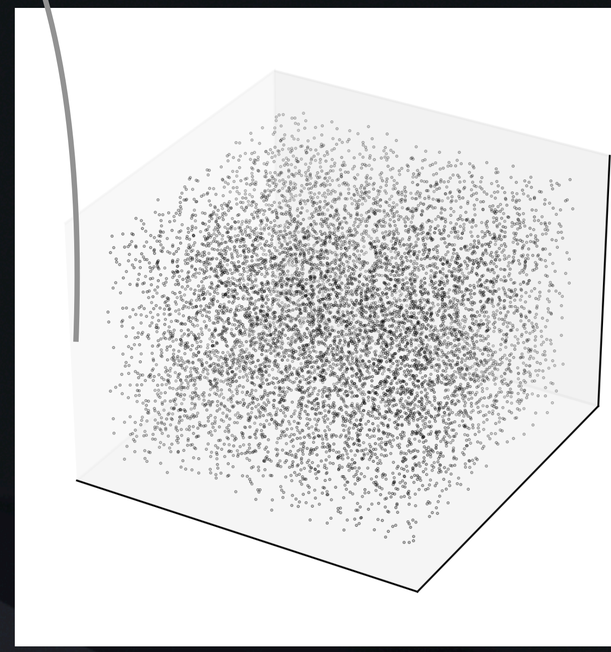
Learned Features

Cosmological Parameters

1. Mixture Density Network ($N_{\rm tot}$)

Generative Model

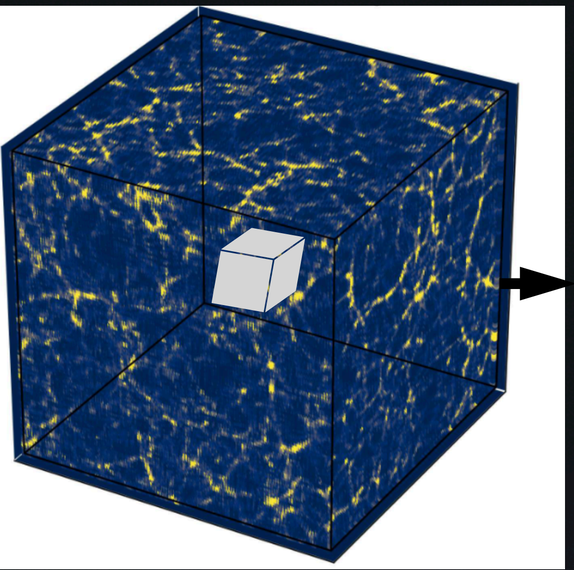




3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

arxiv:2409.09124



3D
Matter density field
(PM Simulation)
Input

ResNet
x2

Feature
extraction

Learned
Features

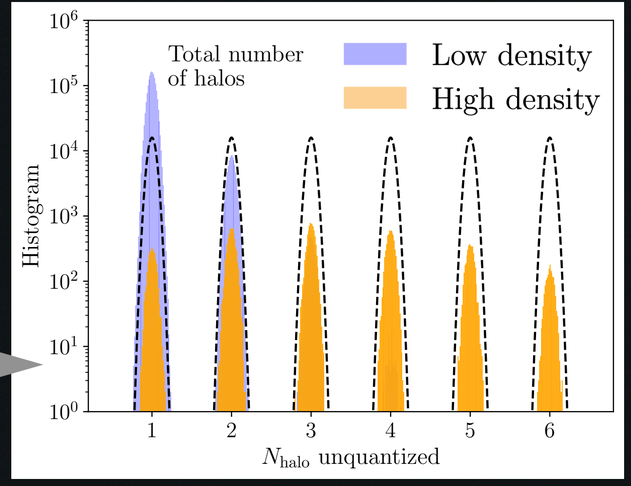
Cosmological
Parameters

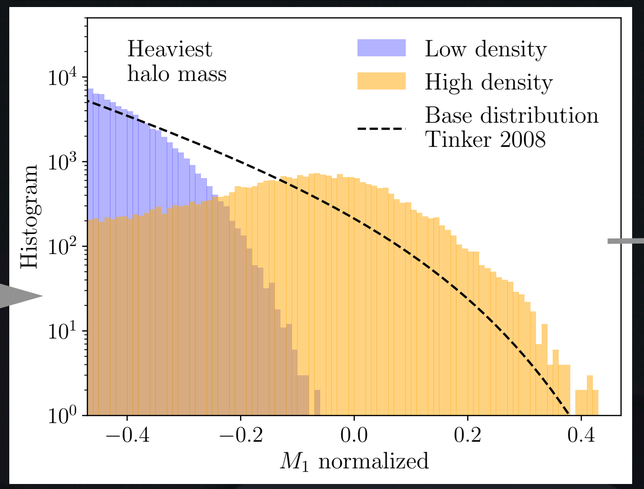
1. Mixture
Density
Network
($N_{\rm tot}$)

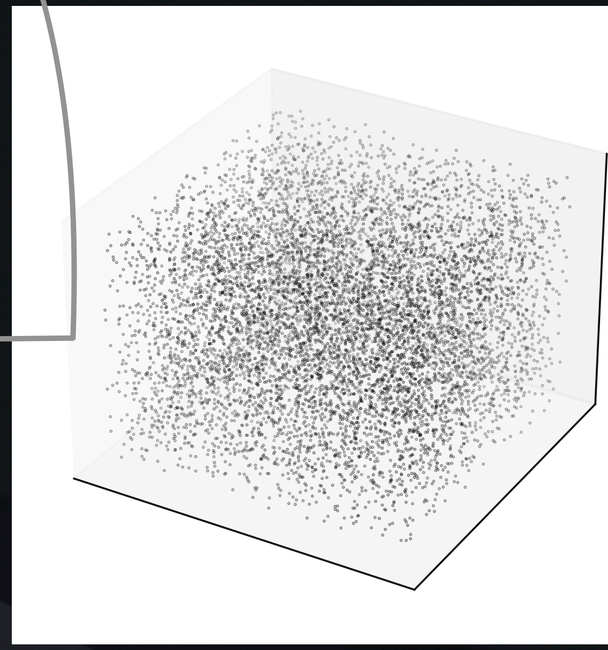
+

2. Spline
Transformation
x5
(M_1)

Generative
Model



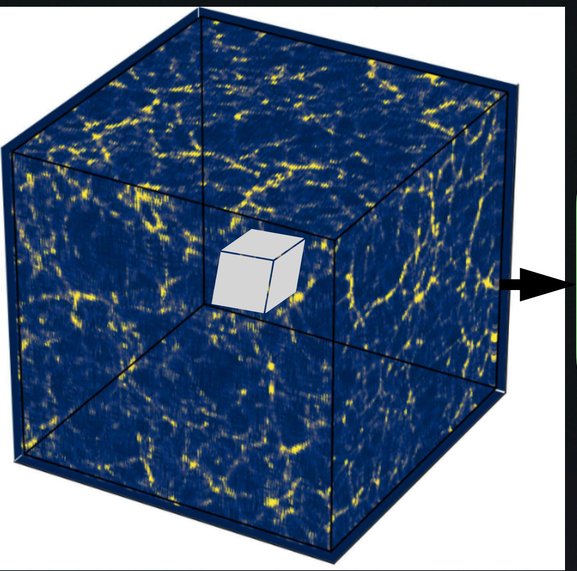




3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

arxiv:2409.09124



3D
Matter density field
(PM Simulation)
Input

ResNet x2

Feature extraction

Learned Features

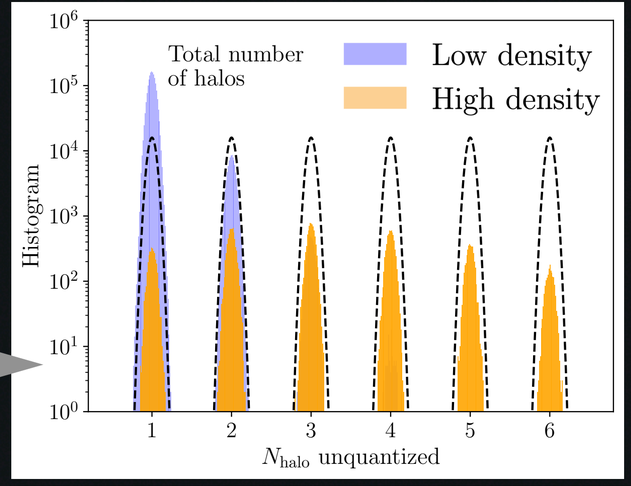
Cosmological Parameters

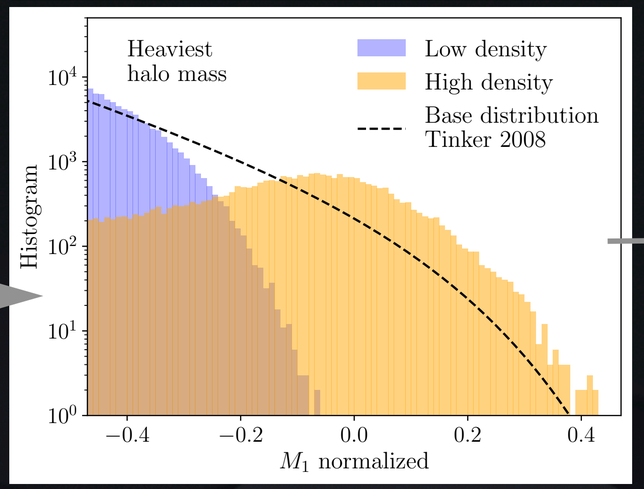
1. Mixture Density Network (N_{tot})

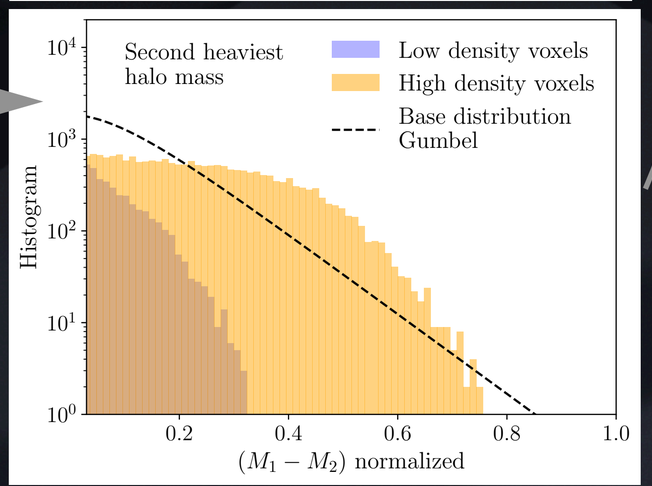
2. Spline Transformation x5 (M_1)

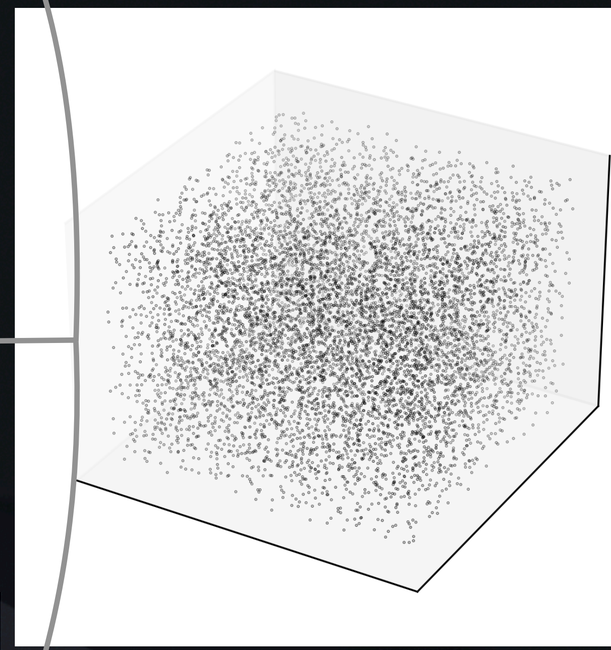
3. Neural Spline Flows x2 $(M_2, \ldots, M_{N_{\text{tot}}})$

Generative Model





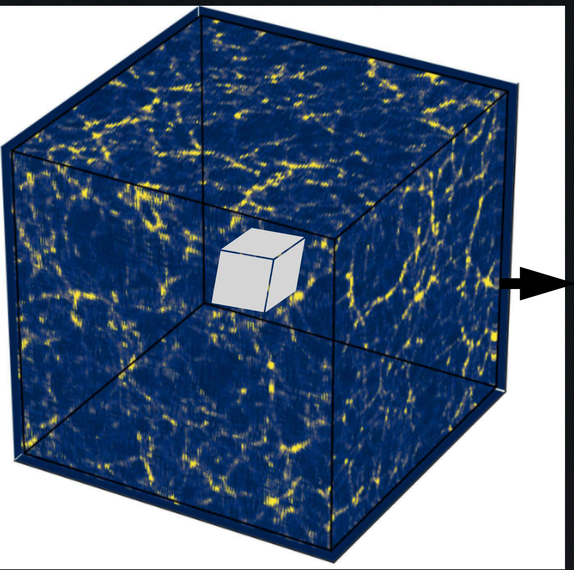




3D
Halo distribution
(N-body Simulation)
Target

CHARM Overview

arxiv:2409.09124



3D
Matter density field
(PM Simulation)
Input

ResNet x2

Feature extraction

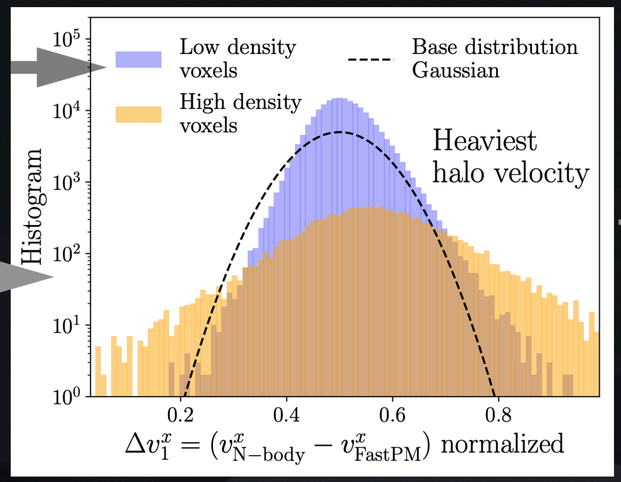
Learned Features

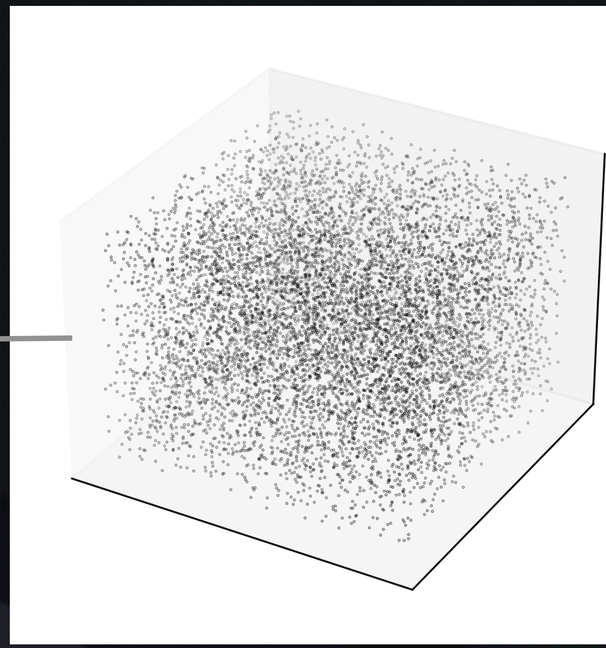
+

Cosmological Parameters

2. Spline Flow x2 ($\vec{v}$)

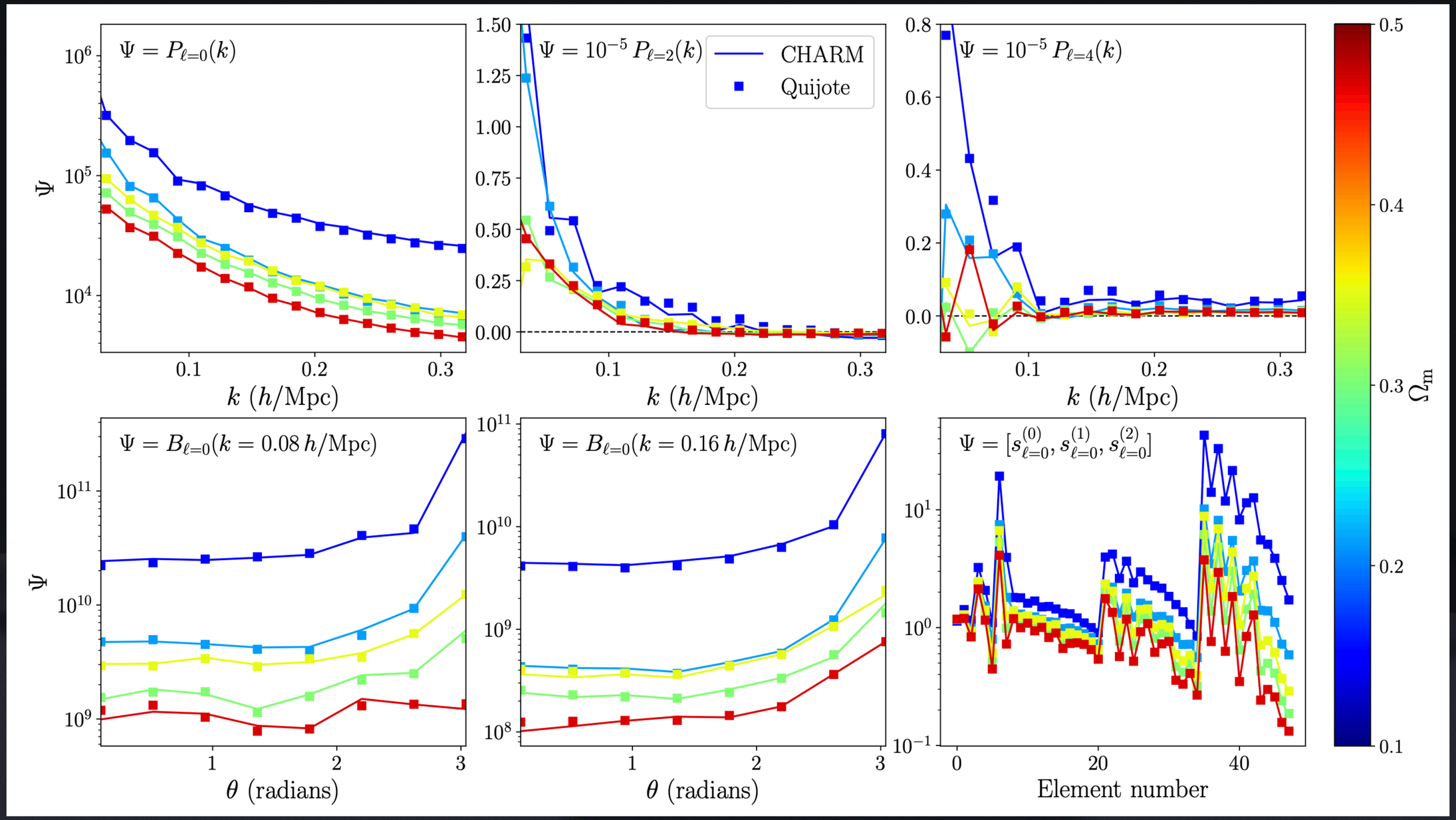
Generative Model



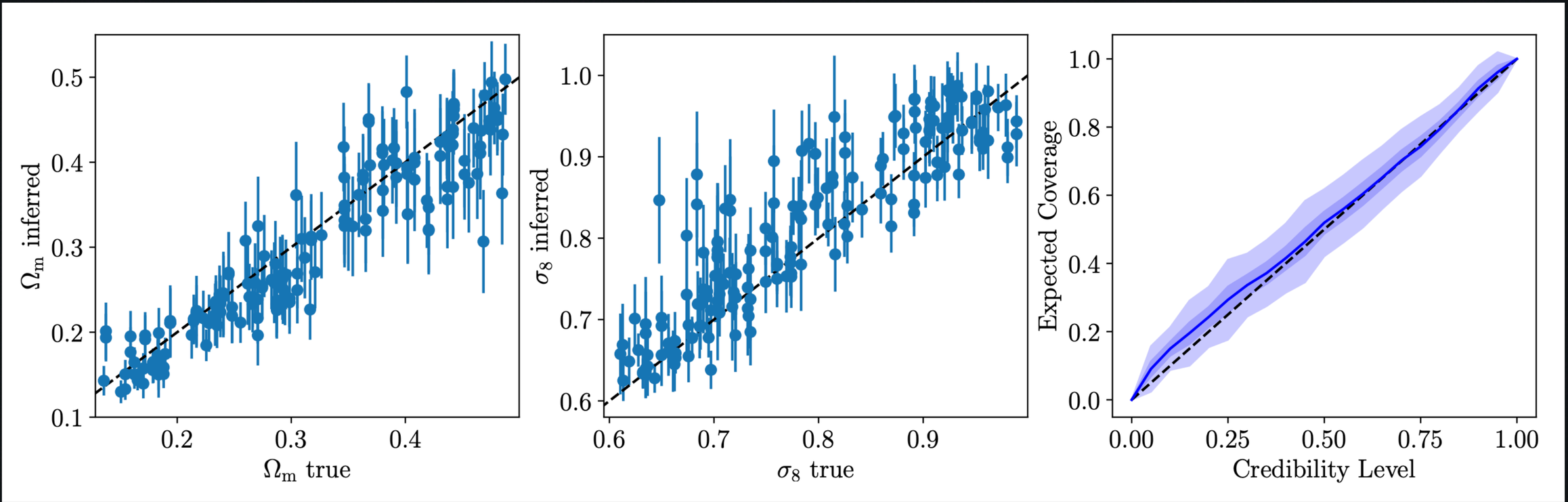


3D
Halo distribution
(N-body Simulation)
Target

N-pt performance — Redshift space



Cosmology inference performance with SBI

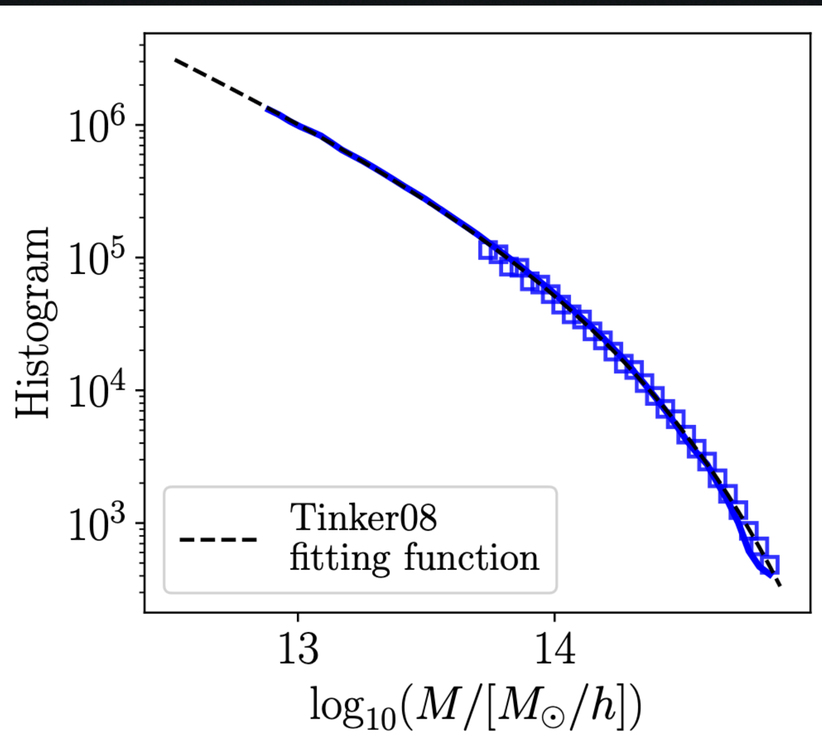


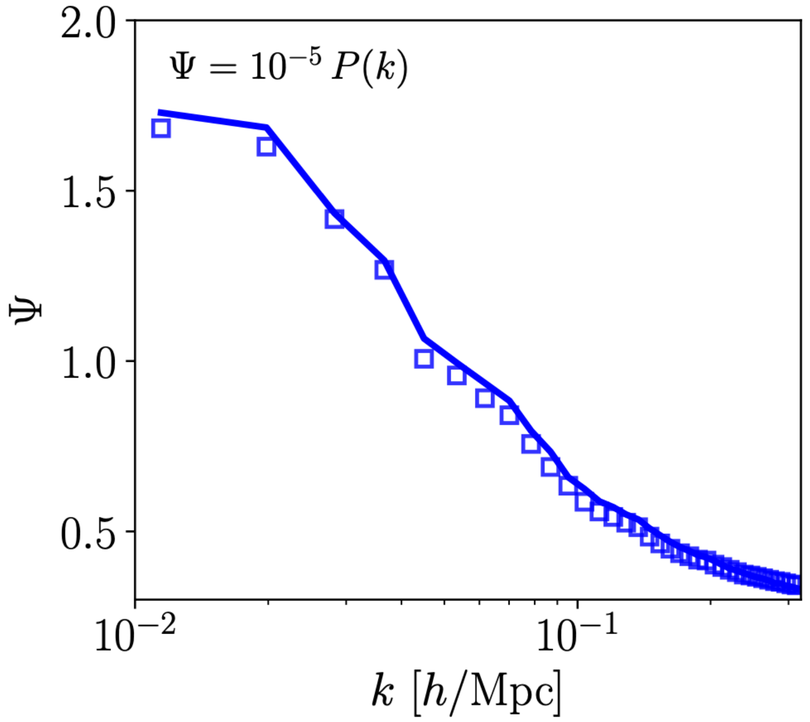
Accounts for halo-galaxy connection with analytical frameworks (eHOD)

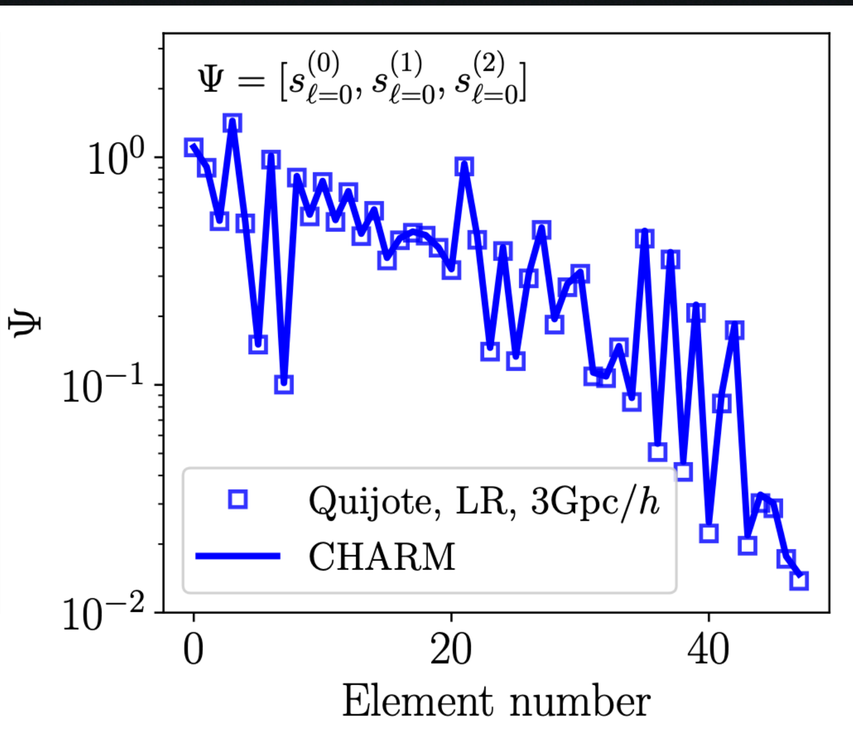
Currently being developed further to analyze the SDSS-NGC data within the LtU.

Being further developed to generalize to redshifts to get halo lightcone

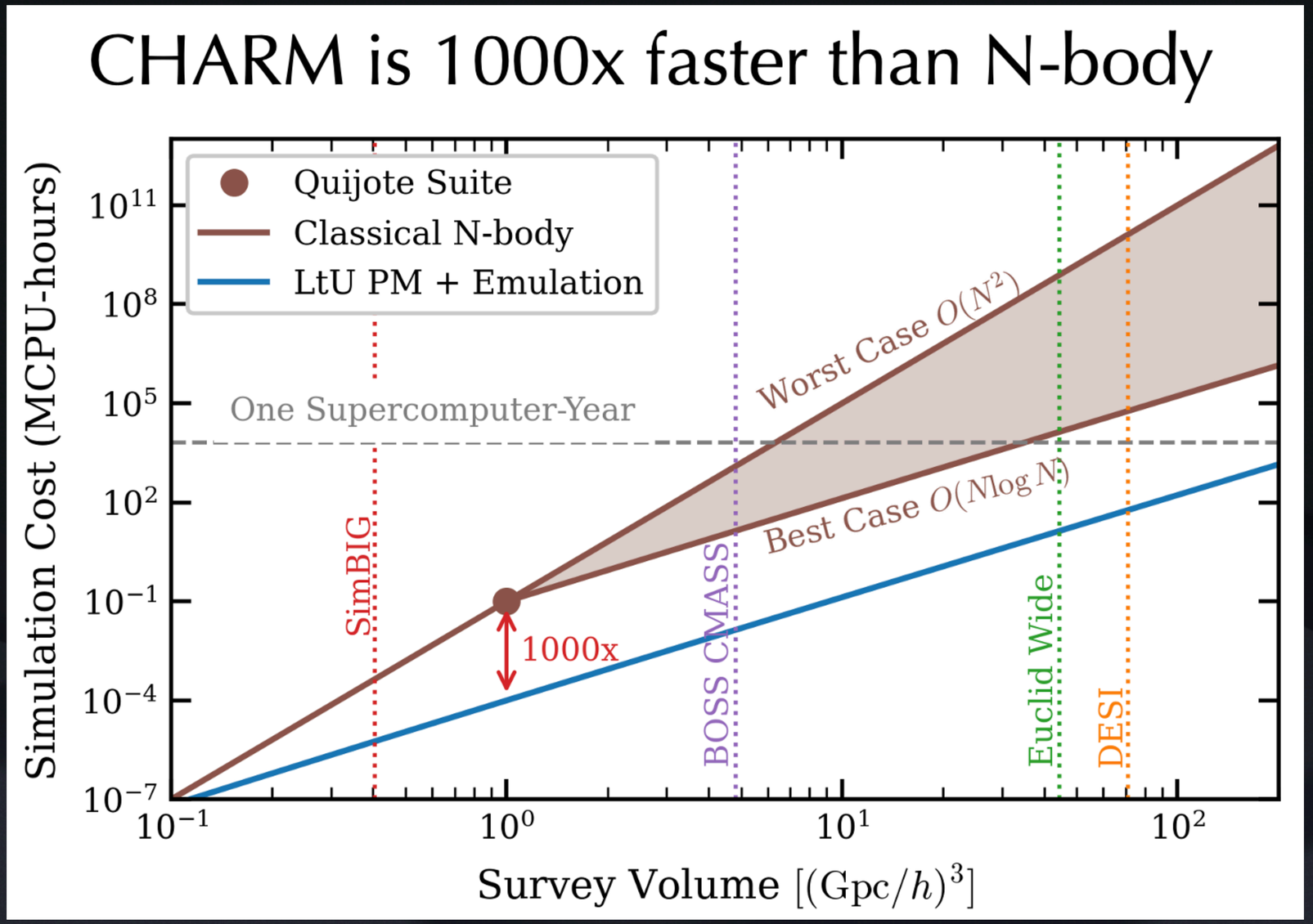
Generalizes to large volumes







Can now run suite of large simulations

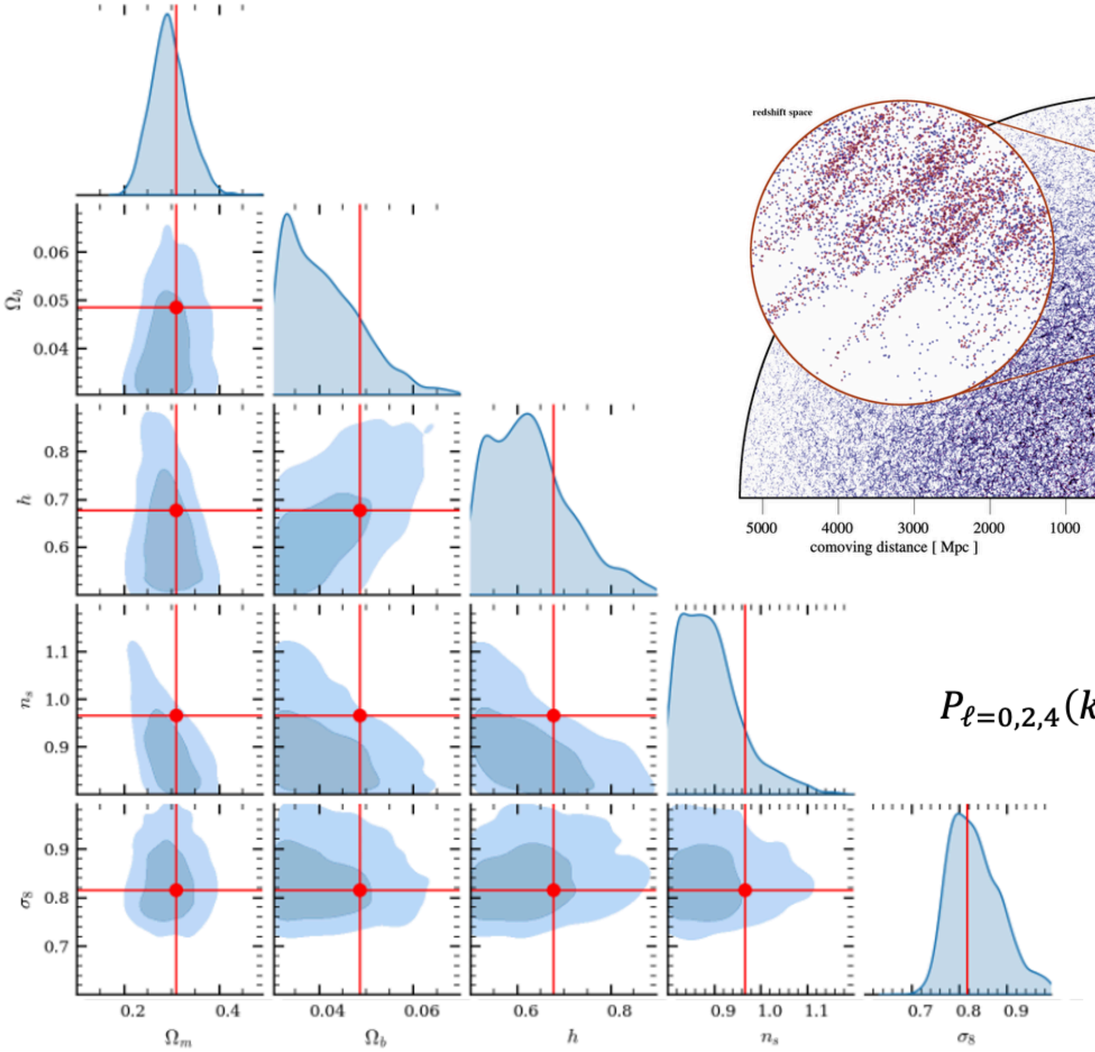


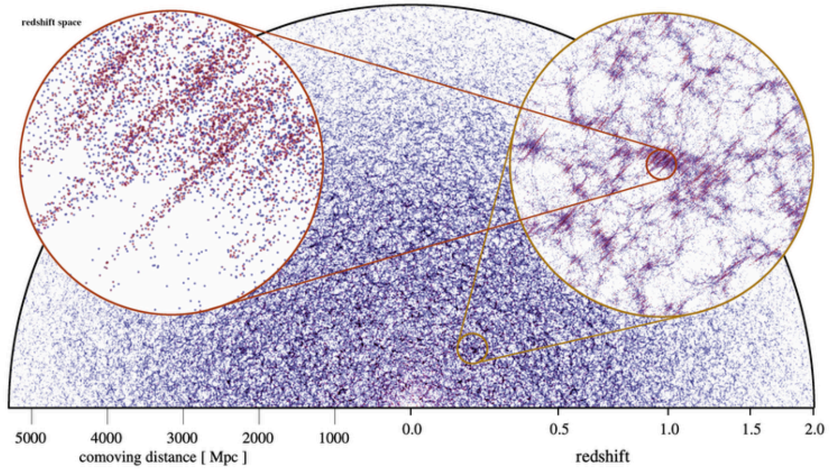
Credit: Matt Ho + LtU-ILI team

Robustness

- Validating on completely independent simulation:
  - Different halo finder
  - Using a completely different way of populating the galaxies
- Also similarly validated on the Abacus simualtions:
  - Different gravity solver, different mass definitions, multiple cosmologies

CHARM can generalize to MTNG + L-Galaxies



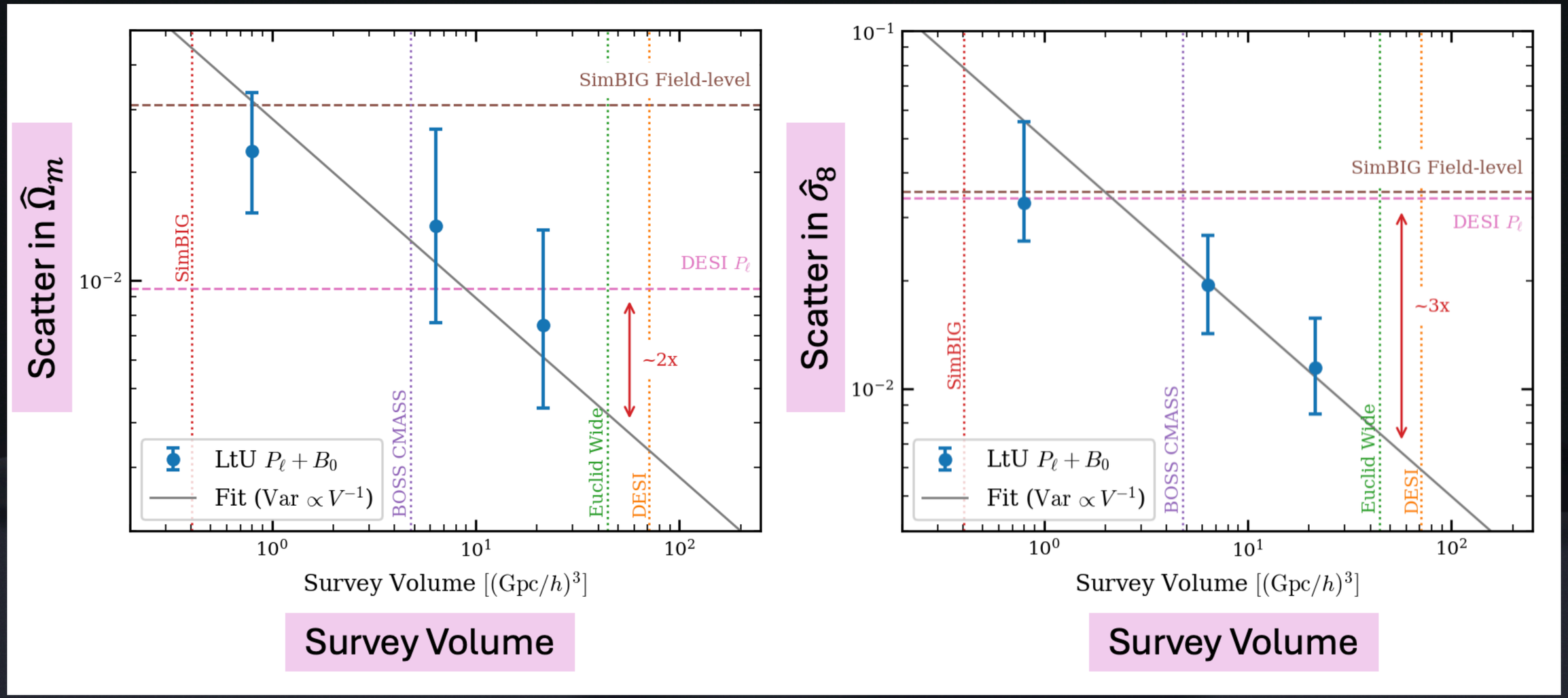


$$P_{\ell=0,2,4}(k) + B_0(k), \quad k \leq 0.4\ h/\mathrm{Mpc}$$

Credit: Matt Ho + LtU-ILI team

Forecasted constraints

SDSS data analysis ongoing within LtU



Credit: Matt Ho + LtU-ILI team

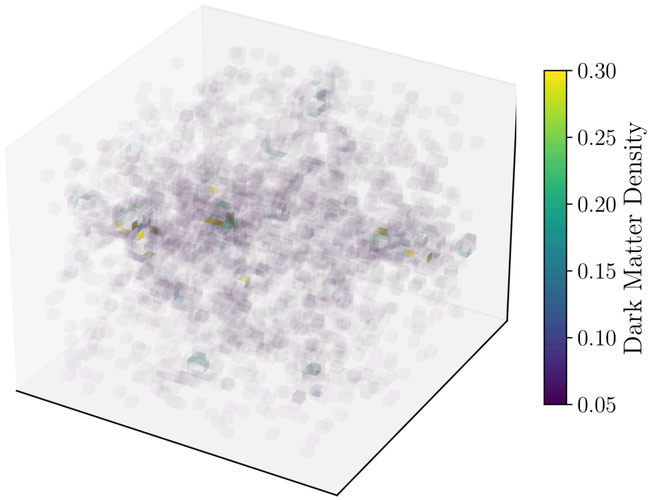
Here halo locations are localized to ~8Mpc.
But we want to go to smaller scales!
Can treat positions as properties as well.

Here halo locations are localized to ~8Mpc.
But we want to go to smaller scales!
Can treat positions as properties as well.

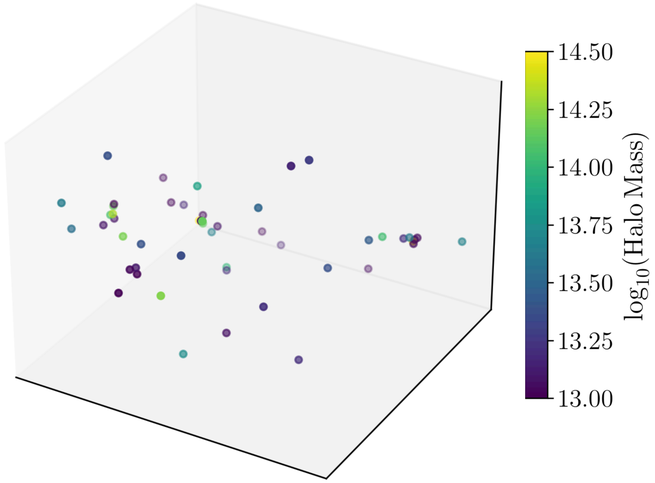
We want an architecture that is very *efficient* and *scalable* for *long-length* auto-regressive dependencies.

How else to go from left to right distribution?

(PM Simulation)
Input



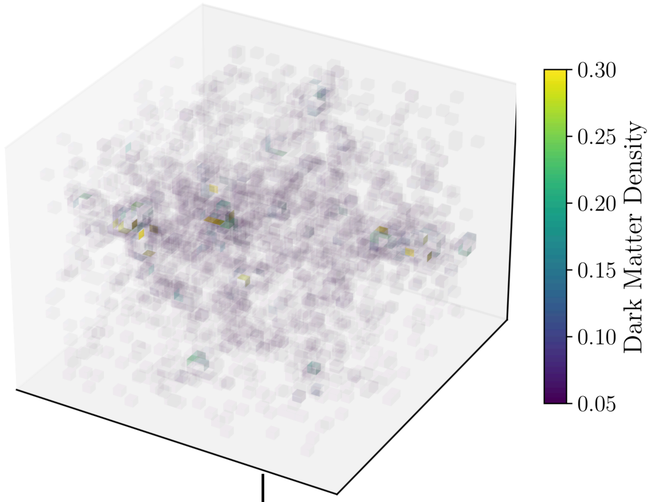
Halo distribution
(N-body Simulation)
Target

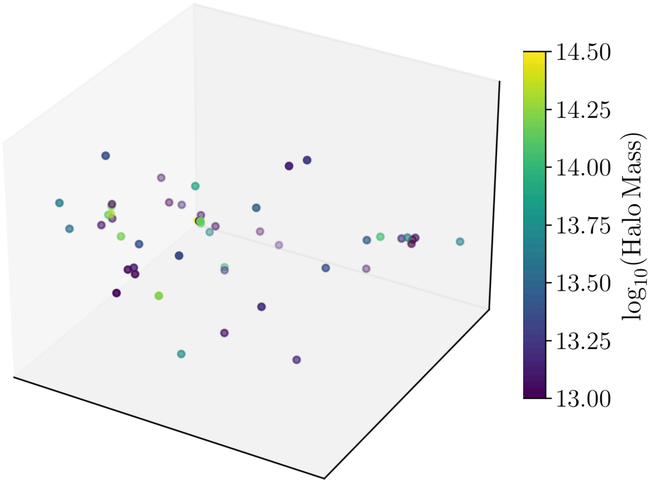


Matter density + velocity fields
z=0.5, 1.0
(PM Simulation)
Input

} ~15 conditional fields

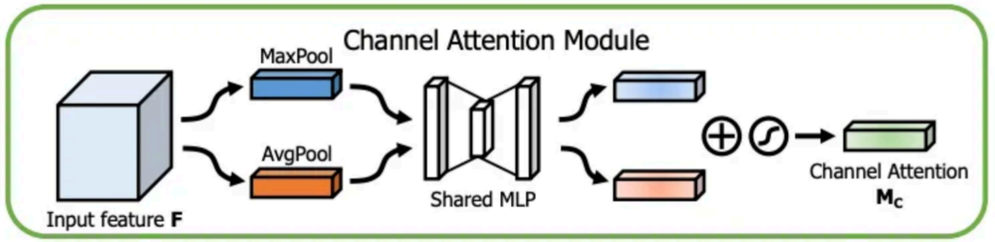
Halo distribution
(N-body Simulation)
Target

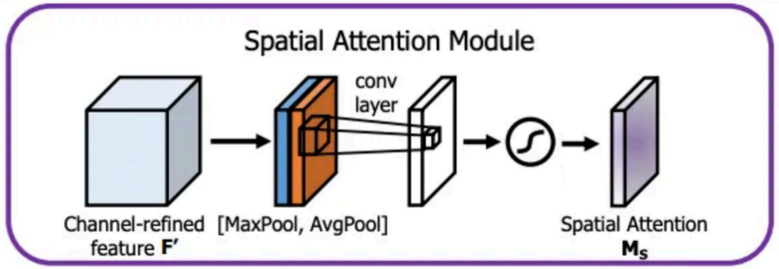


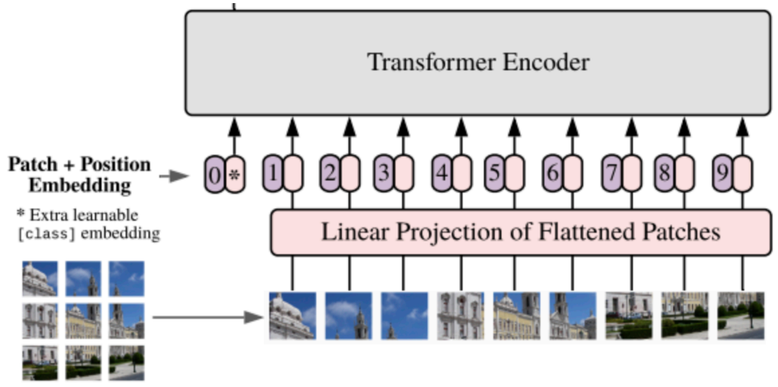


CBAM

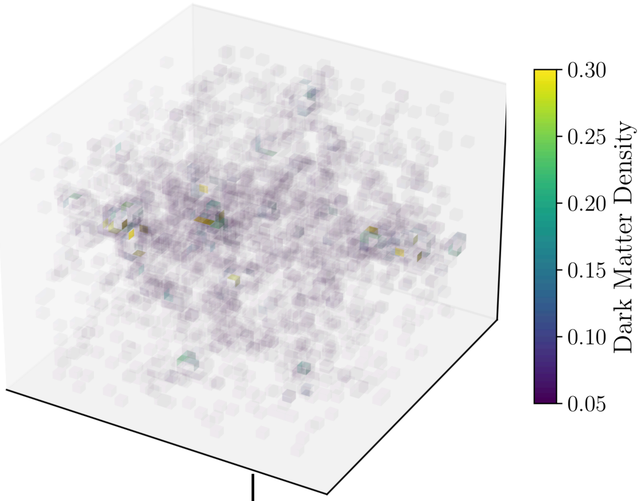
Vision Transformer





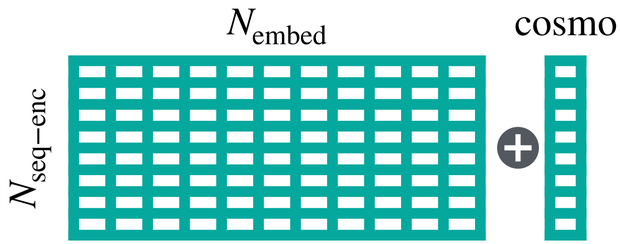


Matter density + velocity fields
z=0.5, 1.0
(PM Simulation)
Input

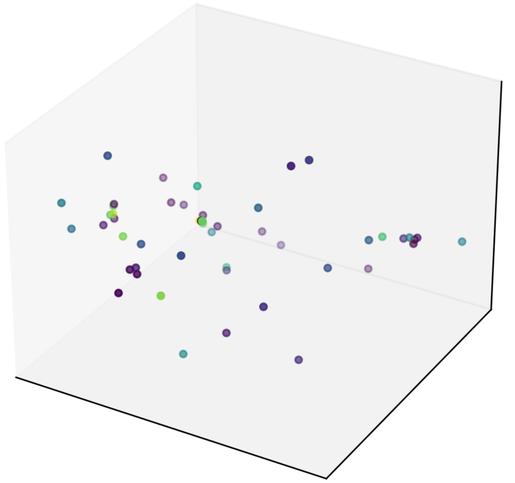


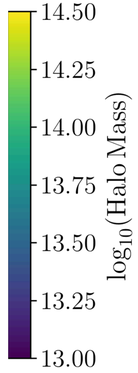
CBAM

Vision Transformer

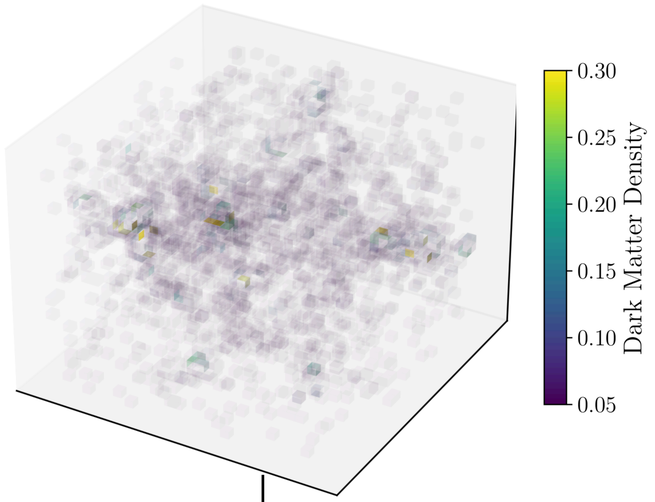


Halo distribution
(N-body Simulation)
Target





Matter density + velocity fields
z=0.5, 1.0
(PM Simulation)
Input



CBAM

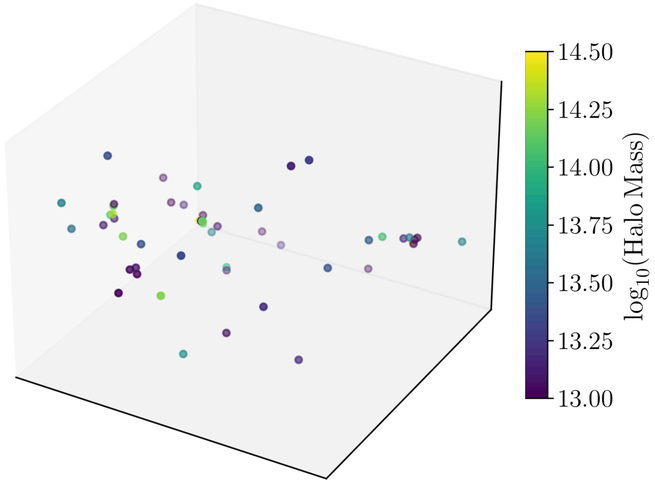
Vision Transformer

N_{embed}

$N_{\text{seq-enc}}$

cosmo

Halo distribution
(N-body Simulation)
Target



Tokenization

Embedding

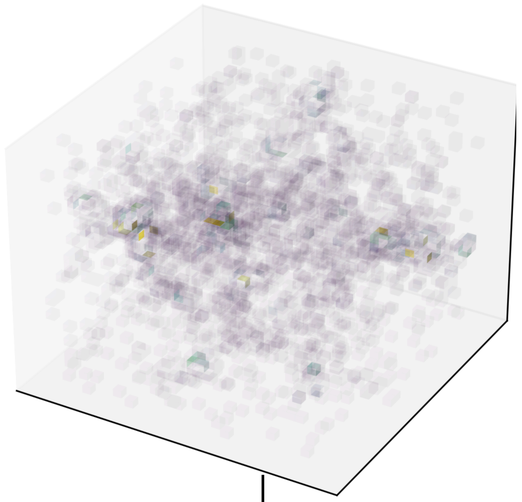
Heaviest halo property tokens

3D position tokens

Second heaviest halo property tokens

[<START>, 42, 21, 12, 32, 45, 32, <SPACE>, 37, 31, 12, 27, 19, 39, <SPACE>,..., <END>, <PAD>,..., <PAD>]

Matter density + velocity fields
z=0.5, 1.0
(PM Simulation)
Input



CBAM

Vision Transformer

N_{embed}

$N_{\text{seq-enc}}$

cosmo

Softmax

Output Probabilities

Linear

Layer Norm

MLP Feed-forward

Layer Norm

Multi-head Cross-Attention

Layer Norm

Masked Multi-head Self-Attention

Layer Norm

$\times N_{\text{block}}$

Positional Encoding

Embedding

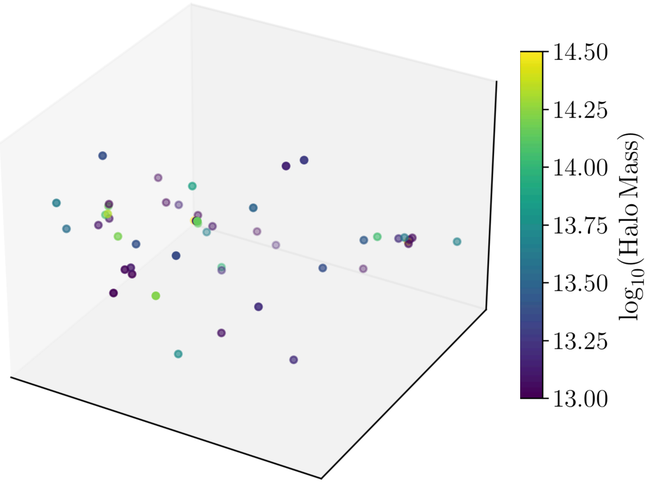
Heaviest halo property tokens

Second heaviest halo property tokens

3D position tokens

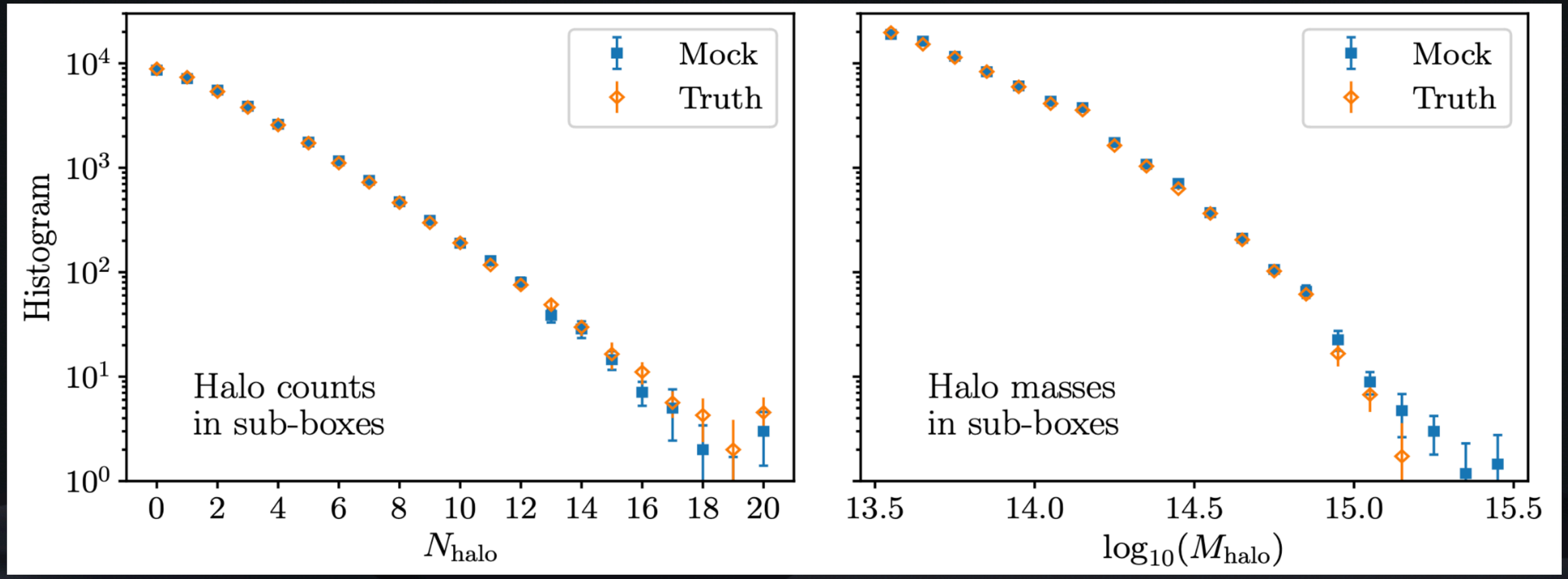
[<START>, 42, 21, 12, 32, 45, 32, <SPACE>, 37, 31, 12, 27, 19, 39, <SPACE>,..., <END>, <PAD>,..., <PAD>]

Halo distribution
(N-body Simulation)
Target



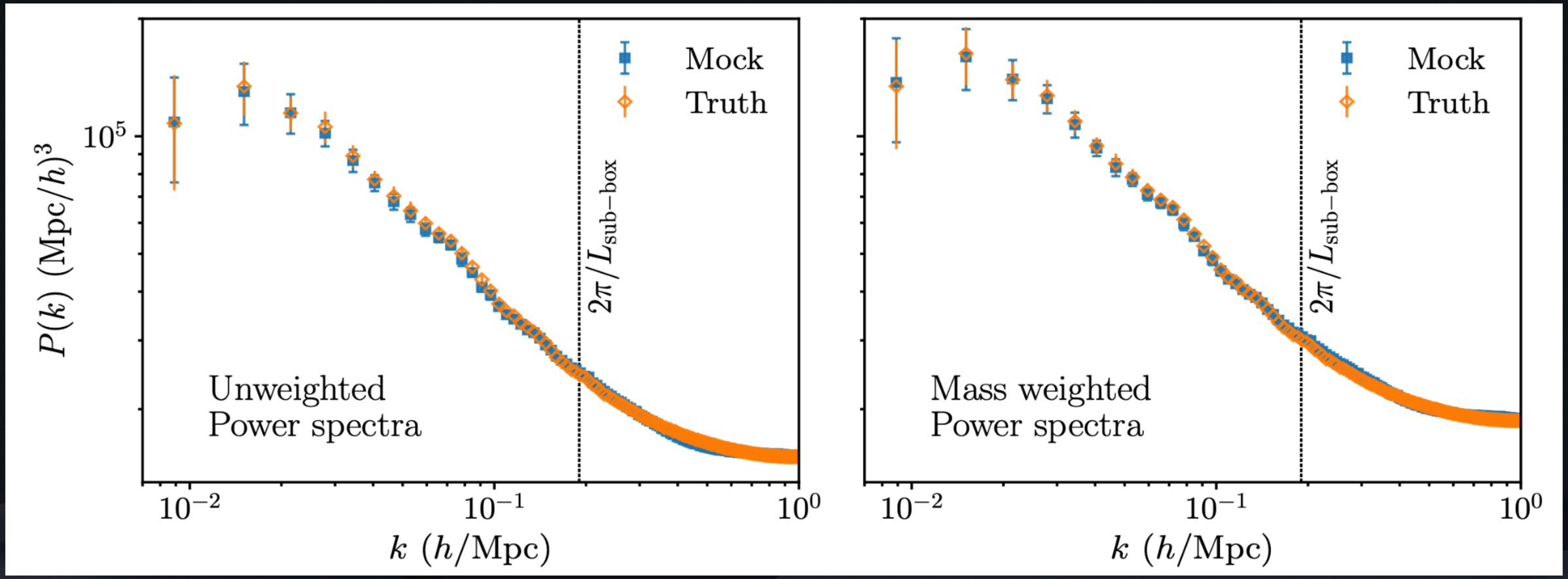
Tokenization

Fixed cosmology: 1-pt performance



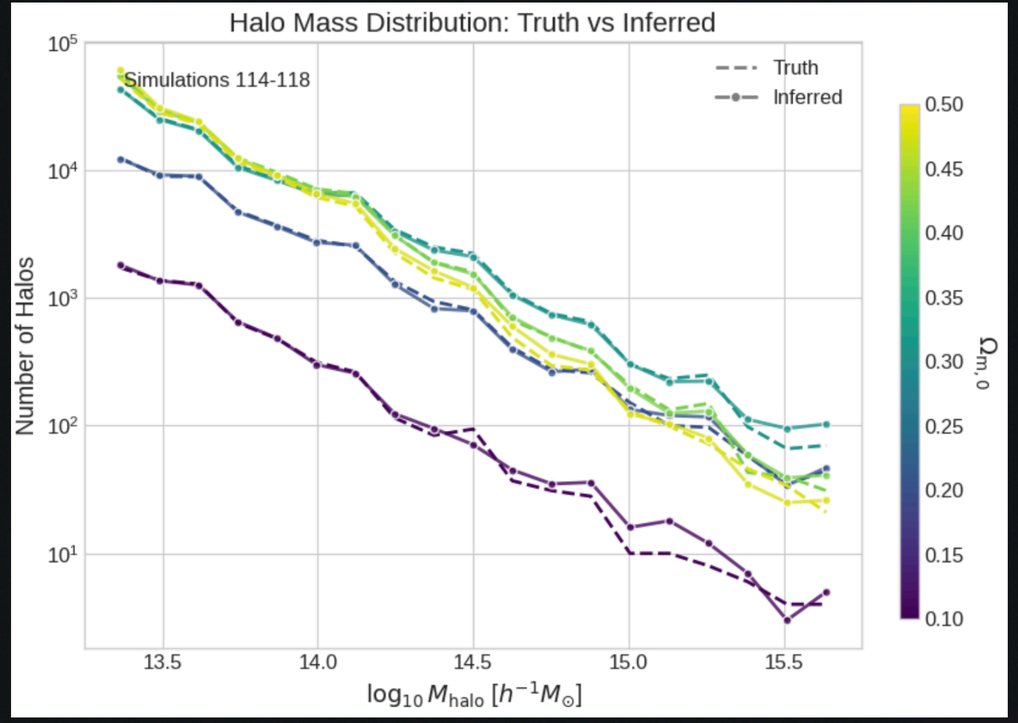
arXiv:2409.11401

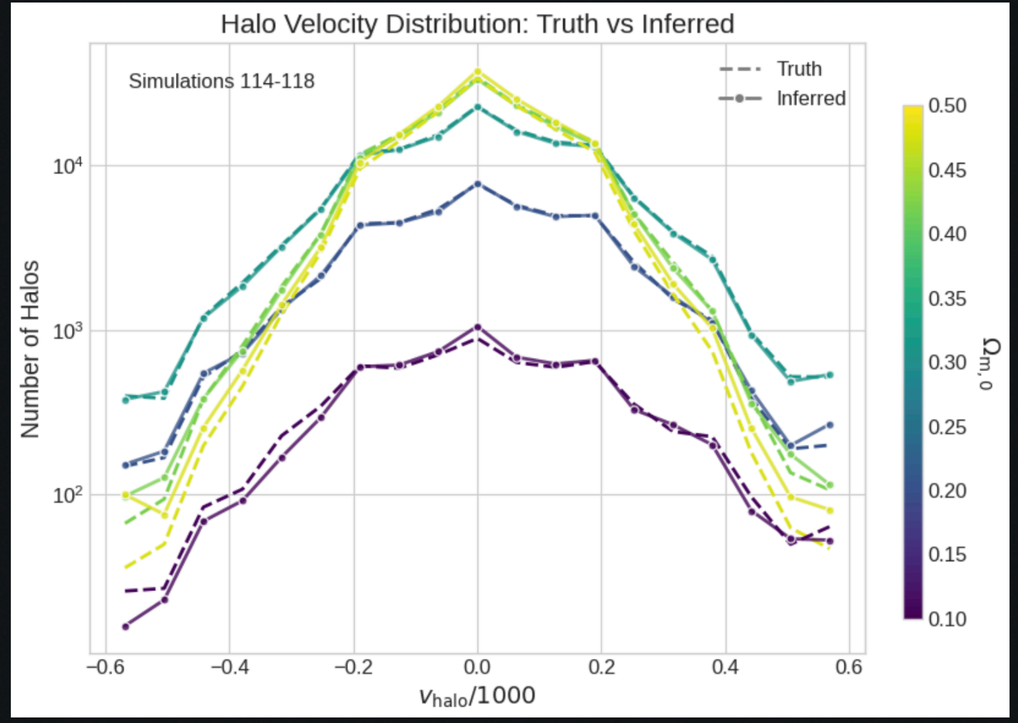
Fixed cosmology: 2-pt performance



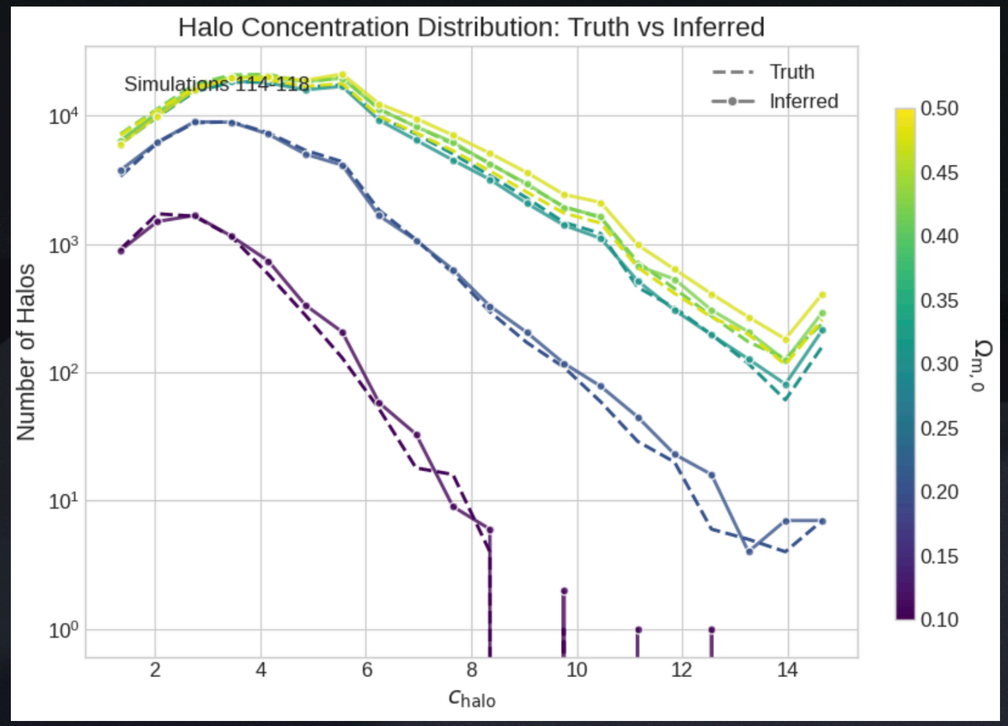
Can go to much smaller scales now!

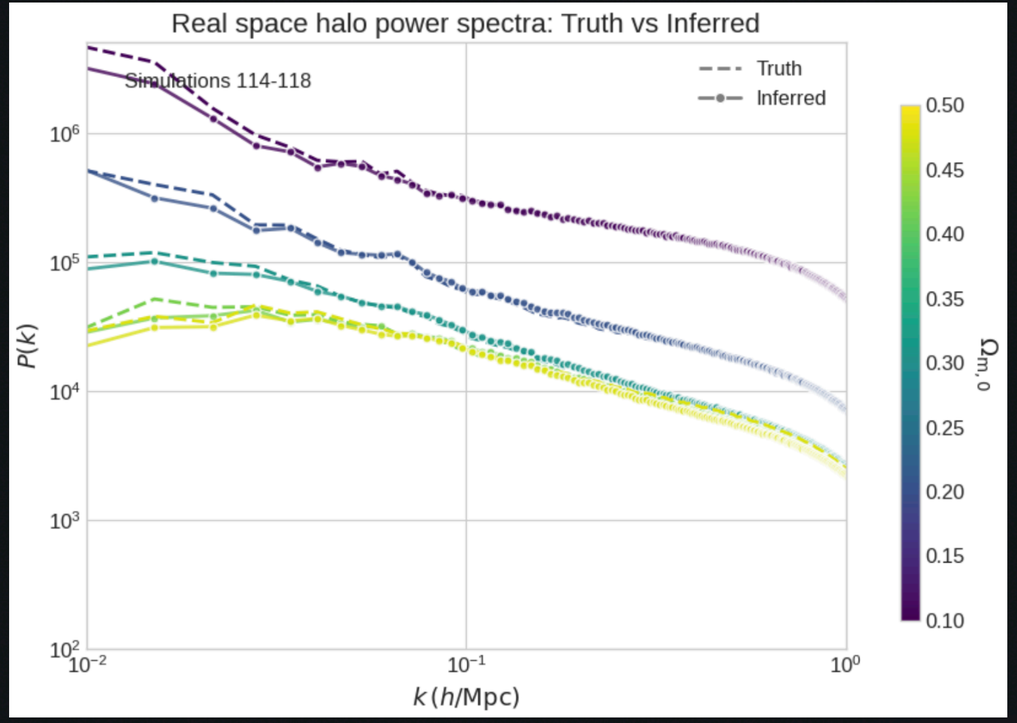
arXiv:2409.11401

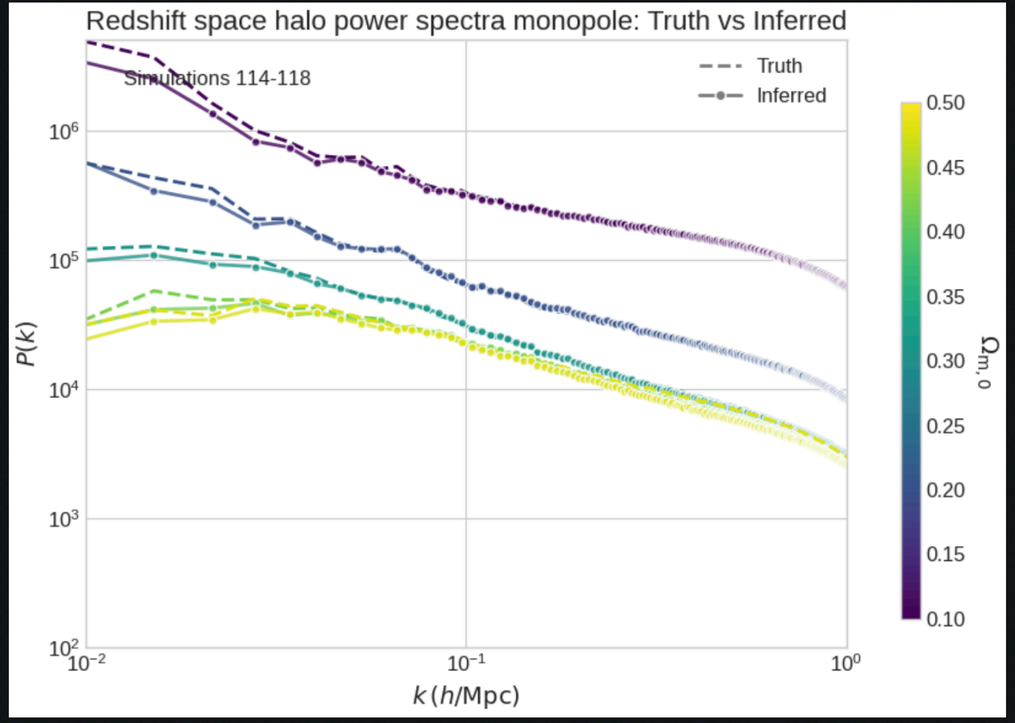


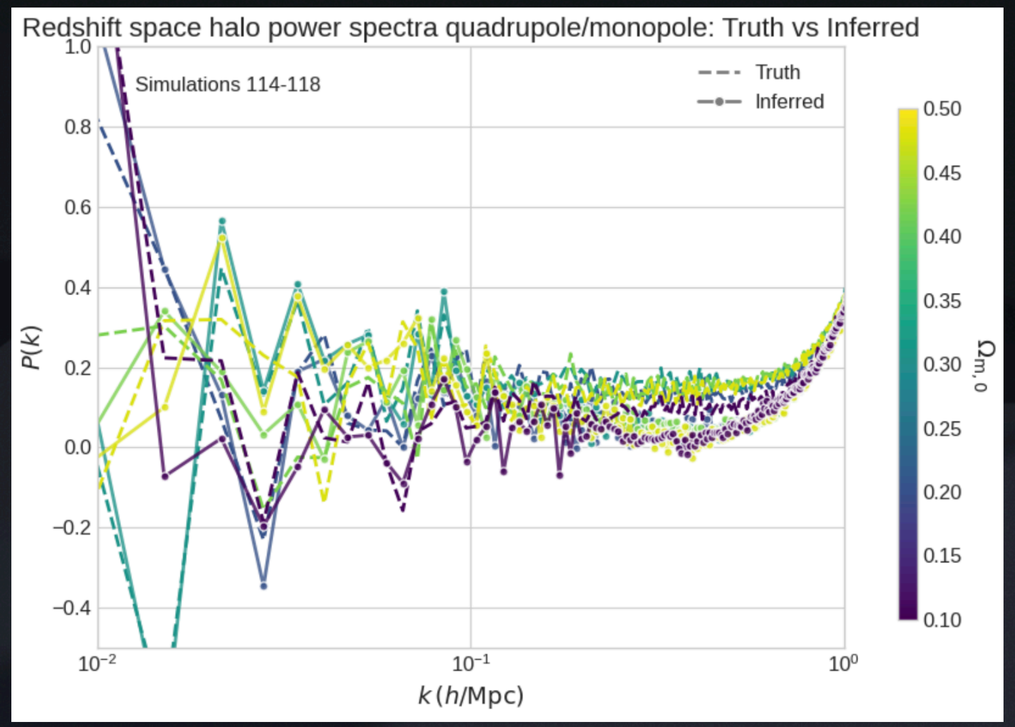


- Preliminary results
- 1-pt inference on test simulations
- Varying cosmologies have orders-of-magnitude different number of halos







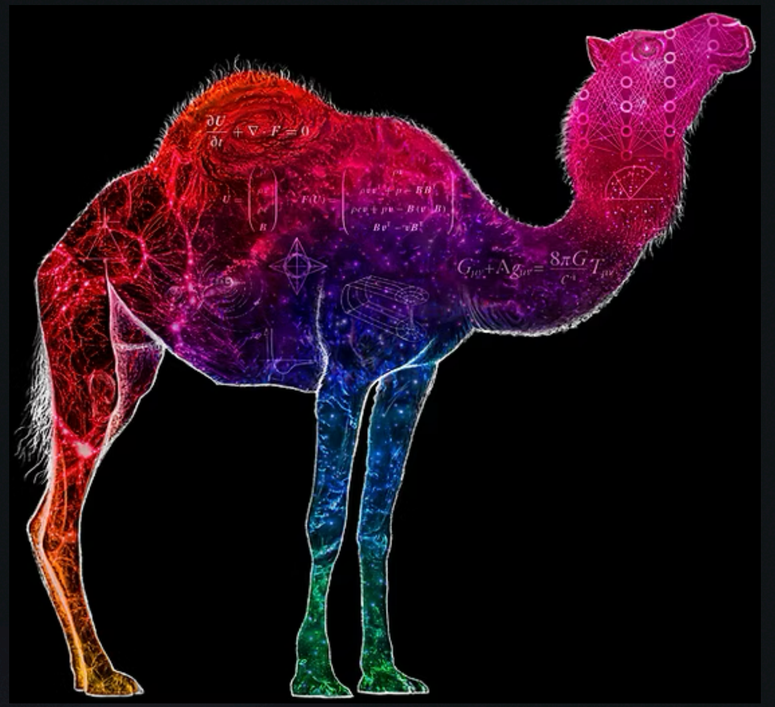


- Preliminary results
- 2-pt inference on test simulations
- Showing both real-space and redshift-space power spectra multipoles

This was for halos and their properties only,
requires halo-galaxy connection!

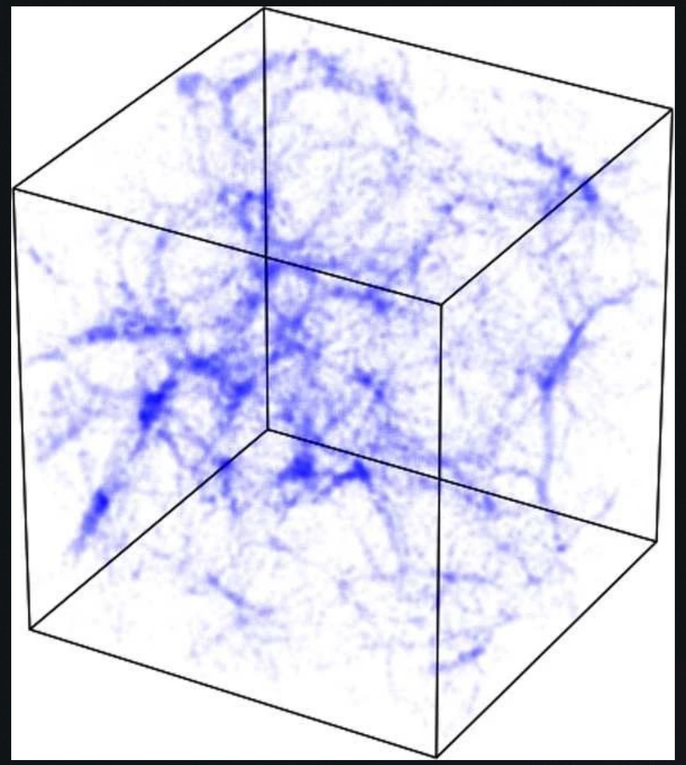
Can we go to galaxies and their observed properties directly while *marginalizing over cosmology*?

CAMELS Hydro-sims



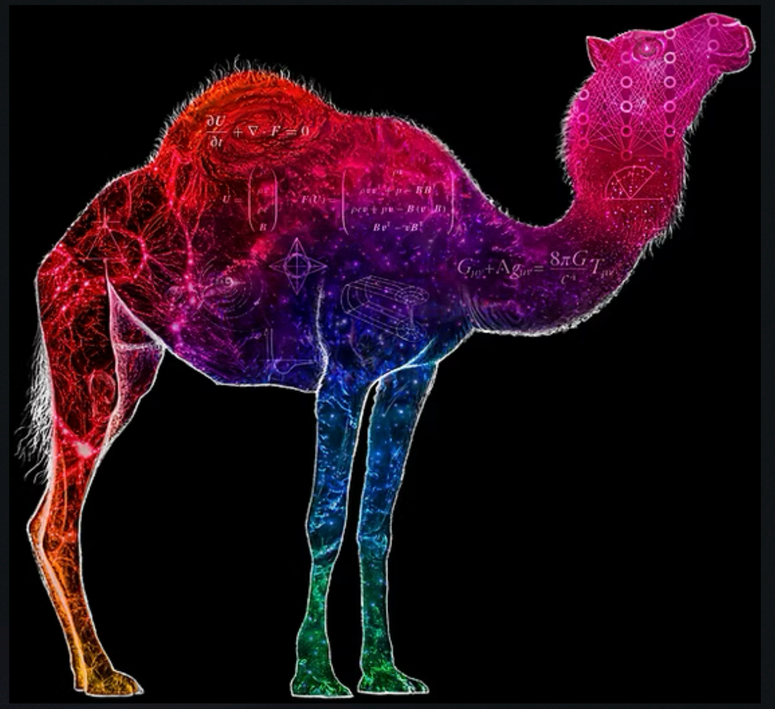
- $(25\ \mathrm{Mpc}/h)^3$ volume
- ~6000 CPU hours/sim
- Four astrophysical and two cosmological parameters varied
- 1000+ simulations in a latin-hypercube space of parameters

N-body sims



- Dark matter only
- Matched IC and cosmology
- ~100x cheaper/sim

CAMELS Hydro-sims

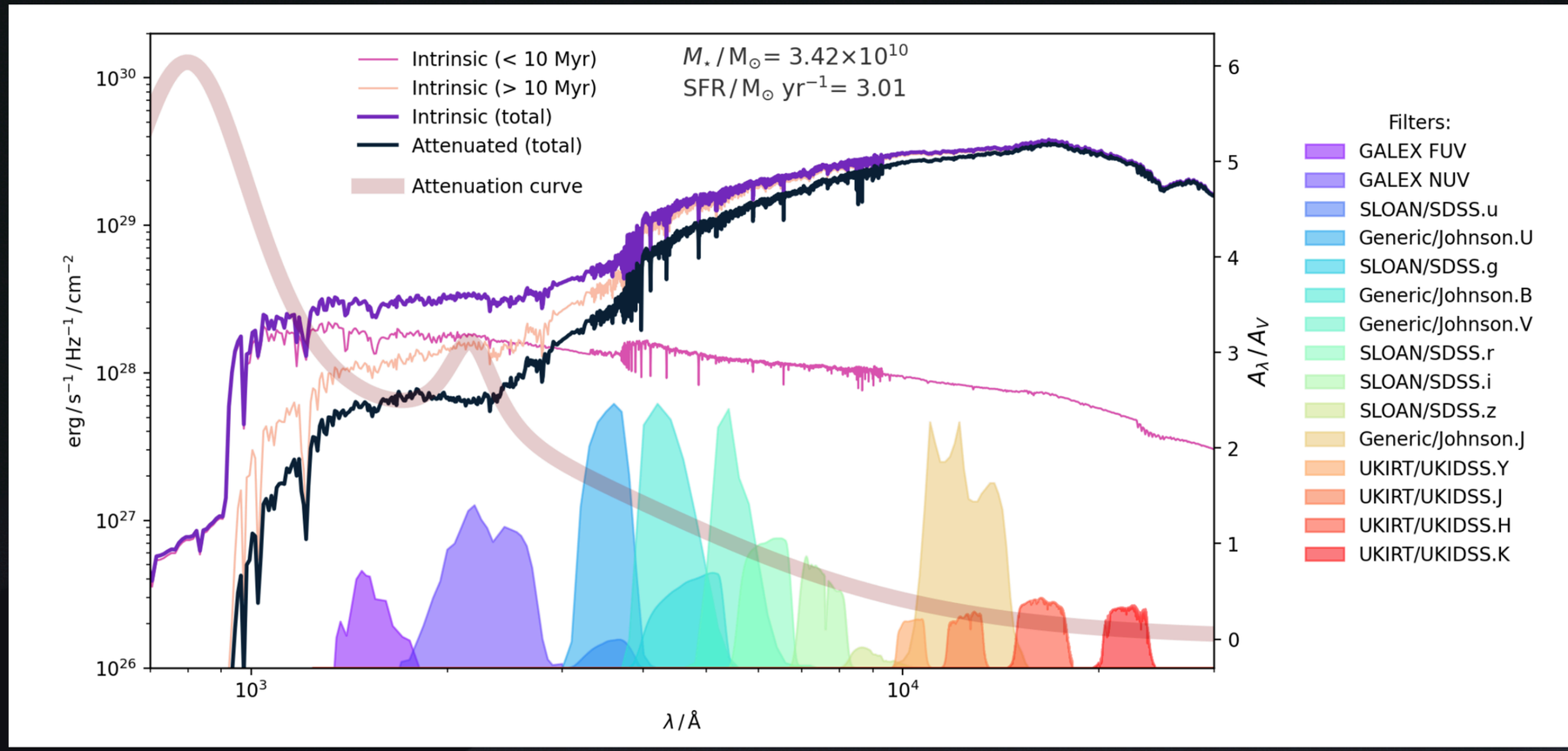


- $(25\ \mathrm{Mpc}/h)^3$ volume
- ~6000 CPU hours/sim
- Four astrophysical and two cosmological parameters varied
- 1000+ simulations in a latin-hypercube space of parameters

Galaxy properties

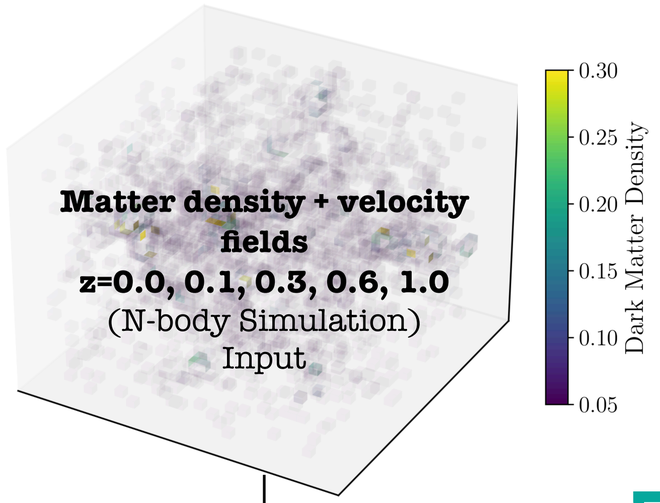
IllustrisTNG-only

- Stellar mass
- Galaxy line-of-sight velocity
- SDSS photometry (dust attenuated)



Lovell et al 24

Now we condition on lots of DM fields



CBAM

Vision Transformer

N_{embed}

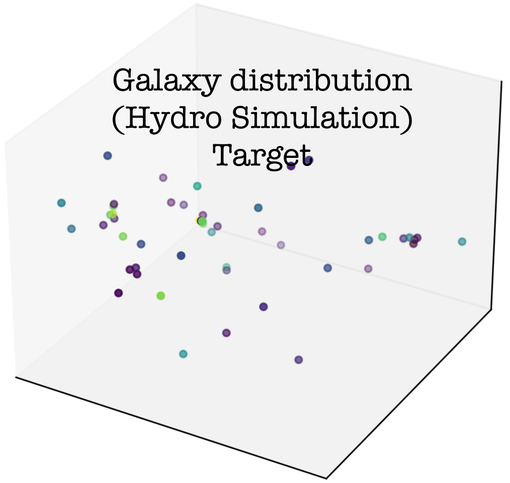
$N_{\text{seq-enc}}$

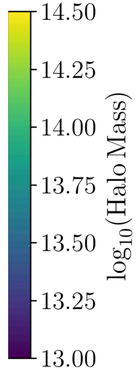
cosmo

Transformer

Softmax

Output Probabilities





Tokenization

Embedding

Heaviest galaxy property tokens

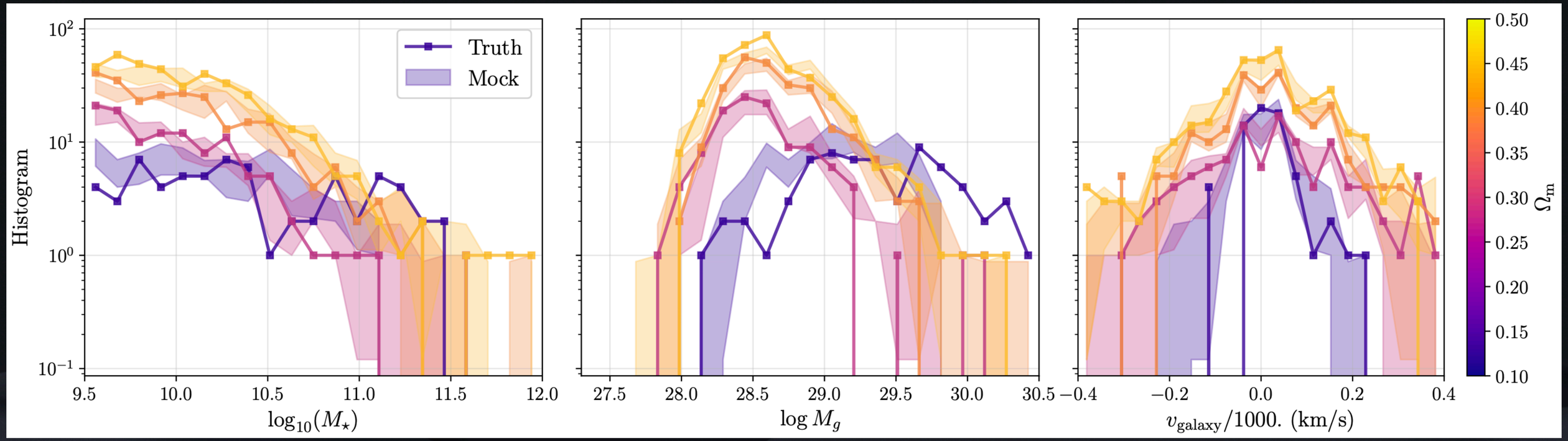
3D position tokens

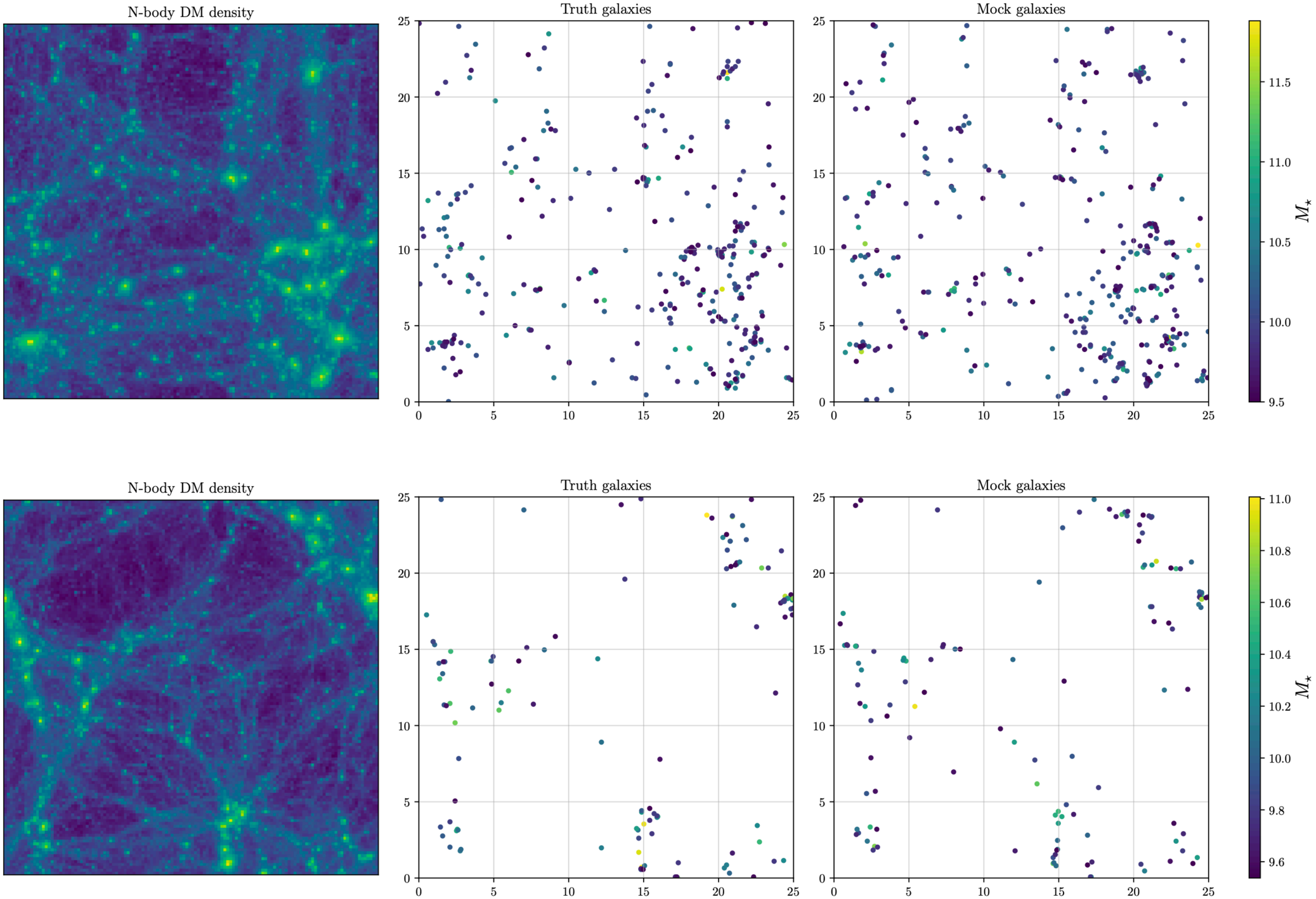
Second heaviest galaxy property tokens

[<START>, 42, 21, 12, 32, 45, 32, <SPACE>, 37, 31, 12, 27, 19, 39, <SPACE>,..., <END>, <PAD>,..., <PAD>]

1pt PDFs

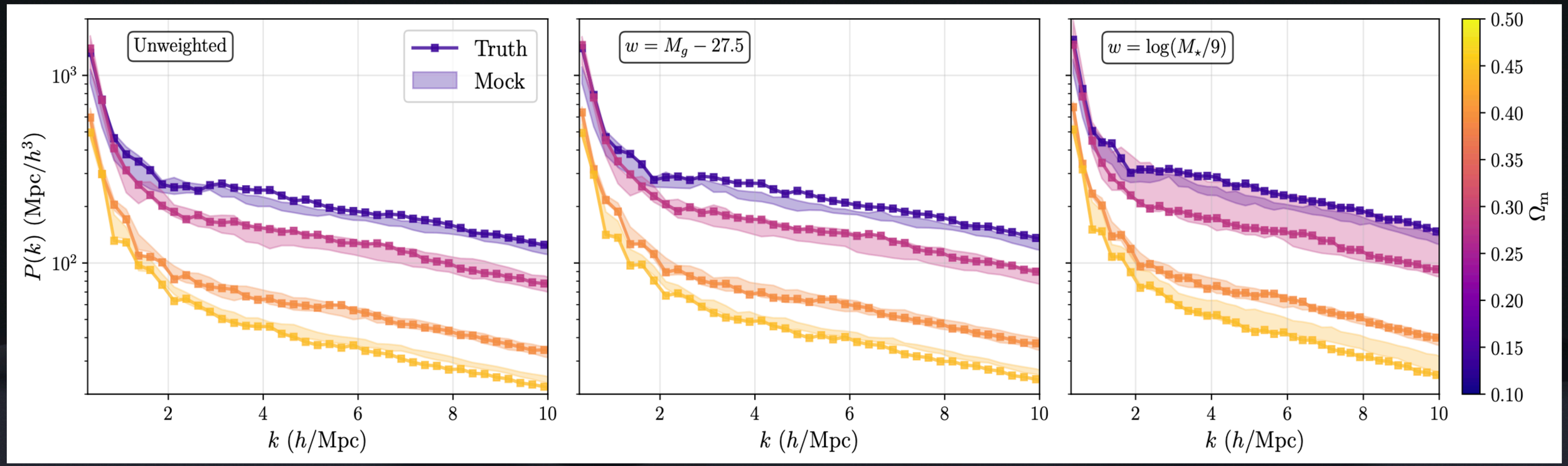
- Inference of galaxy properties on test LH simulations.




N-body DM density
Truth galaxies
Mock galaxies
$M_\star$
N-body DM density
Truth galaxies
Mock galaxies
$M_\star$

2pt (and beyond) correlations

- Inference of galaxy properties on test LH simulations.

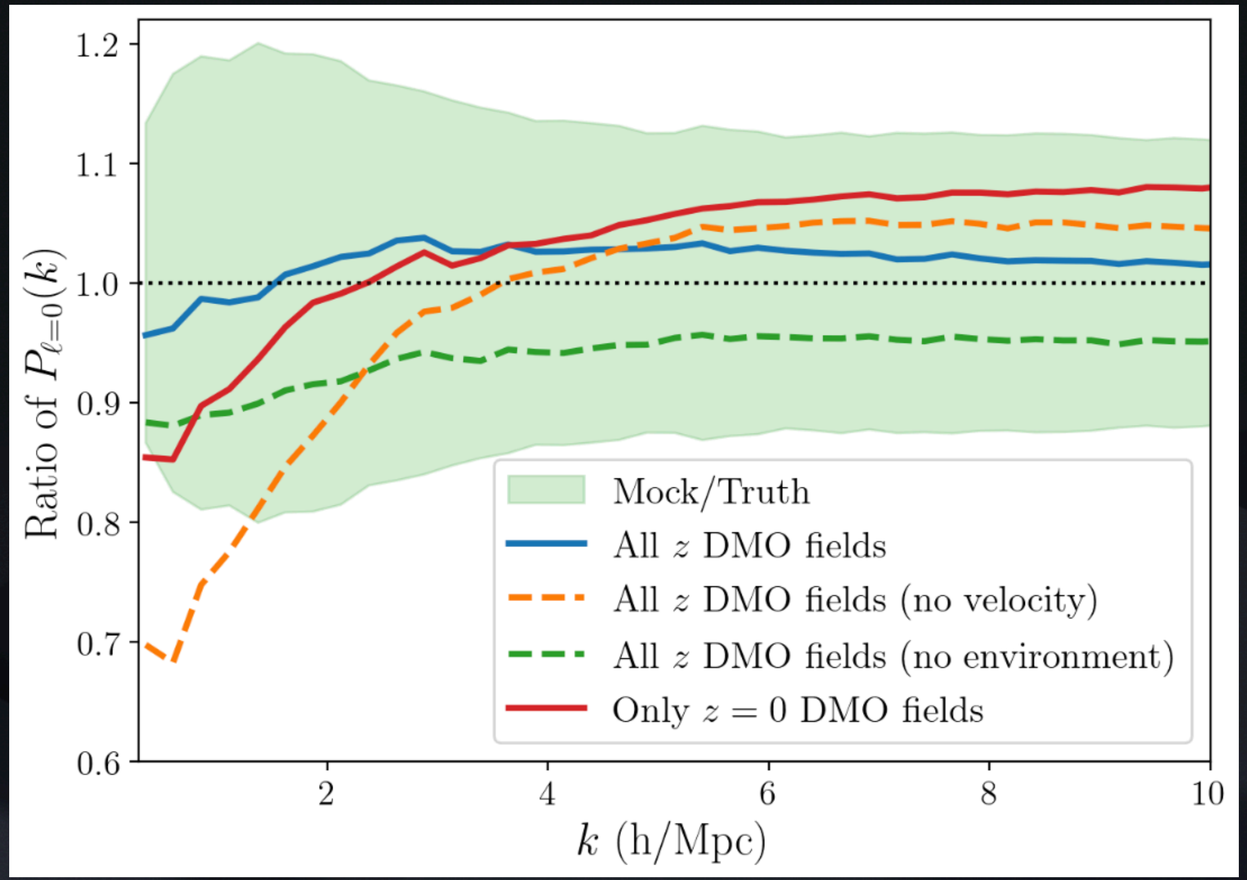


Interpreting the learning

It is nice to know that the computer understands the problem. But I would like to understand it too.

- Eugene Wigner

Interpreting the learning



Preliminary

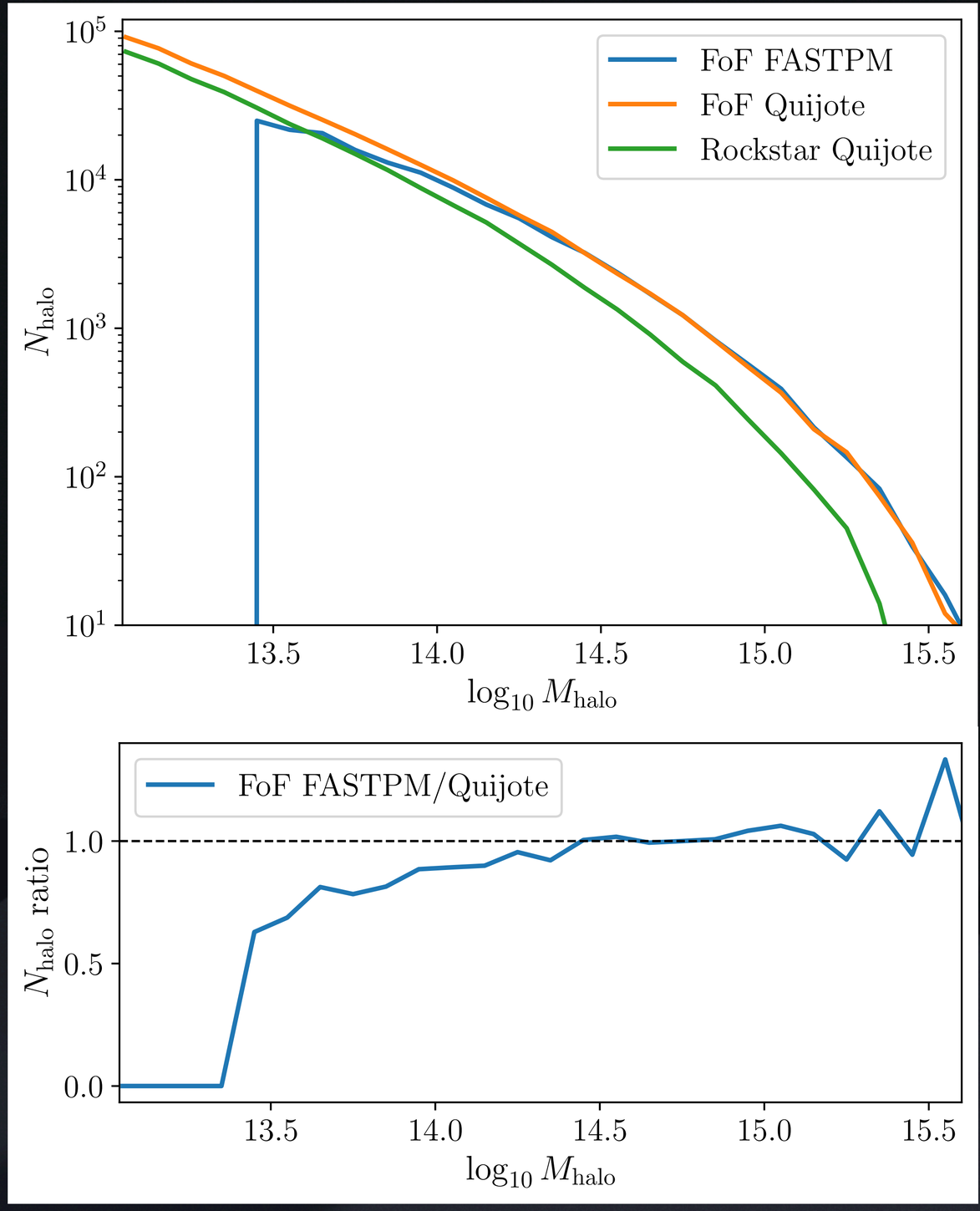
Conclusions

- *Accelerated forward models are needed* if we wish to extract full cosmological information from current/future surveys.
- The generalizability of *transformers to multi-modal inputs* can be extended to cosmological simulations as well!
- More work is needed to integrate it into final data pipeline and to interpret the performance of the model.

Extra slides

Halo finding in PM

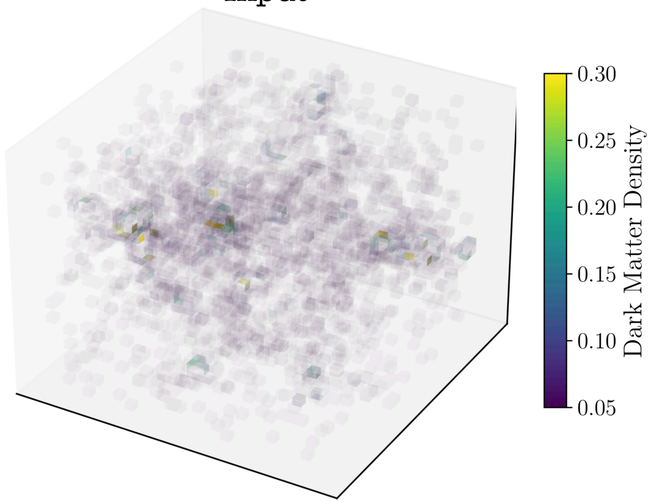
- PM can only reliably find friends-of-friends halos of high masses
  - Rockstar halos much more reliable.
  - SDSS/DESI LRG galaxies occupy halos with $M \sim 10^{13}\, M_\odot/h$



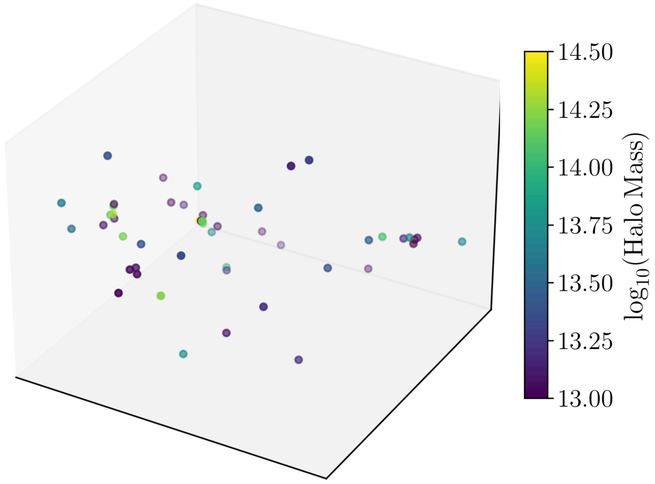
Can we go to small scales?

How else to go from left to right distribution?

Matter density field
(PM Simulation)
Input

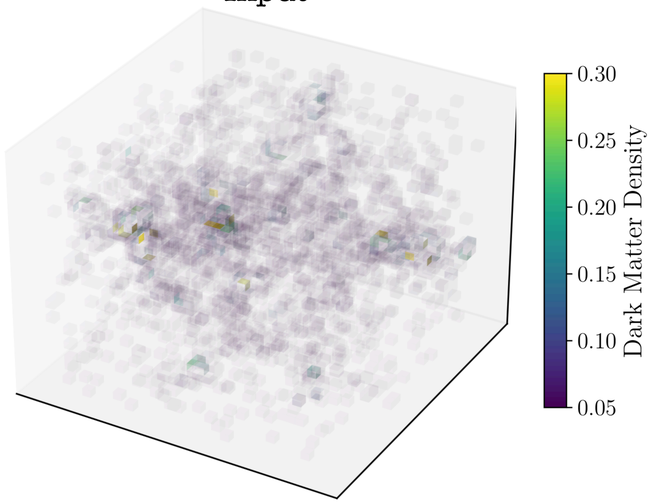


Halo distribution
(N-body Simulation)
Target

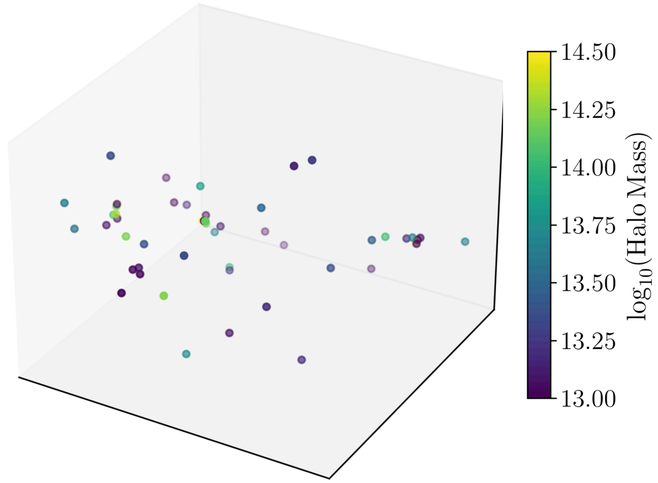


Can treat it as a language translation problem

Matter density field
(PM Simulation)
Input

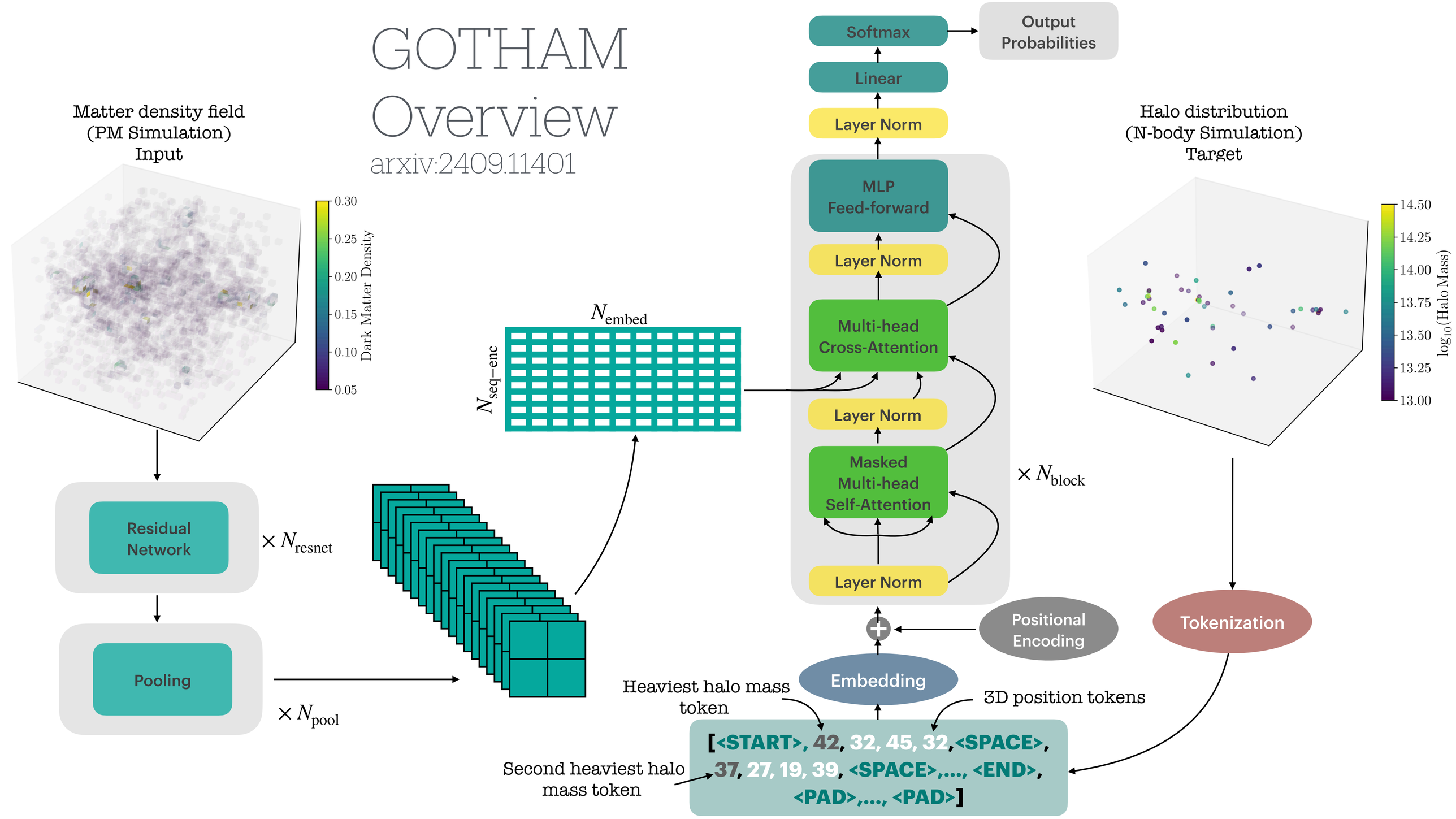


Halo distribution
(N-body Simulation)
Target

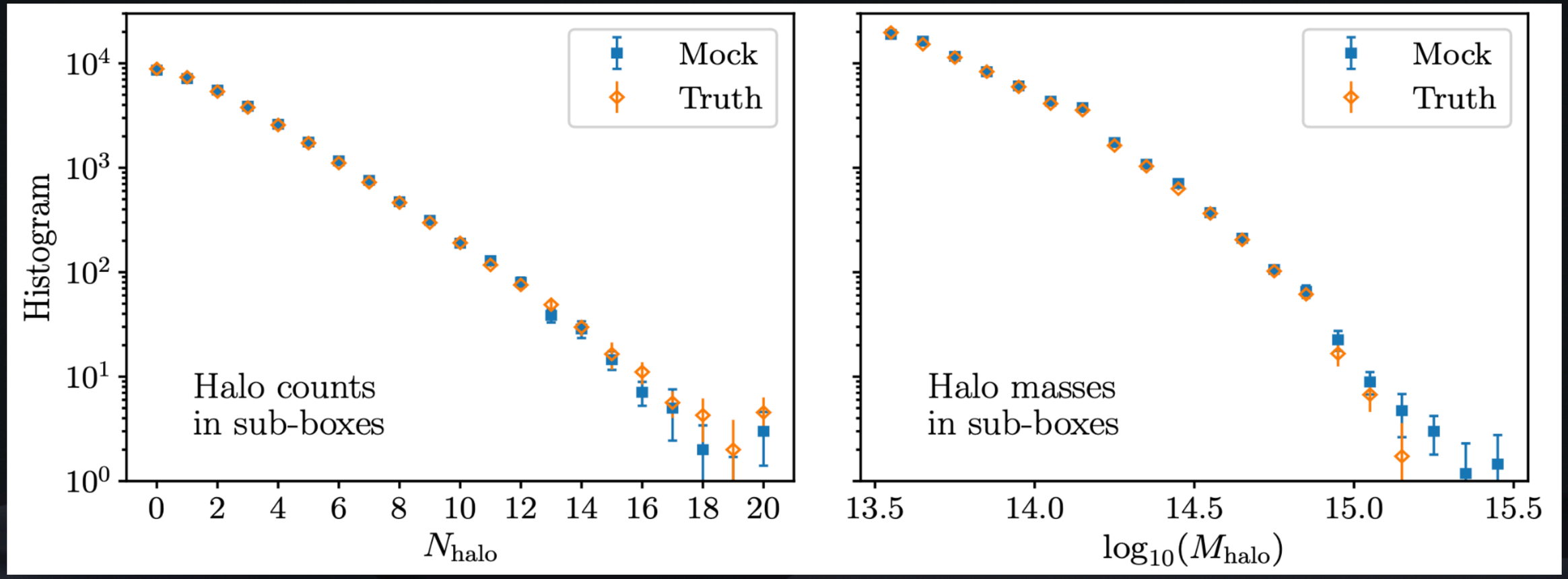


GOTHAM Overview

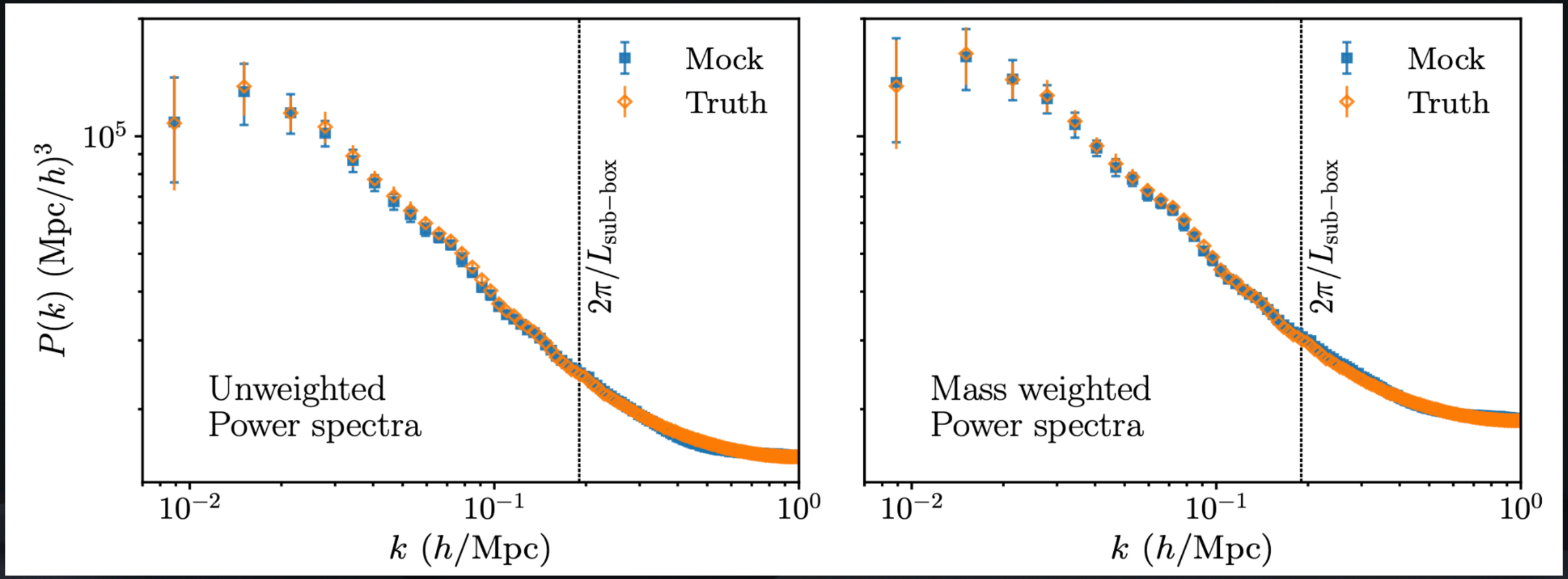
arxiv:2409.11401



1-pt performance



2-pt performance



Can go to much smaller scales now!