Big Data Cosmology meets AI

Carol Cuesta-Lazaro

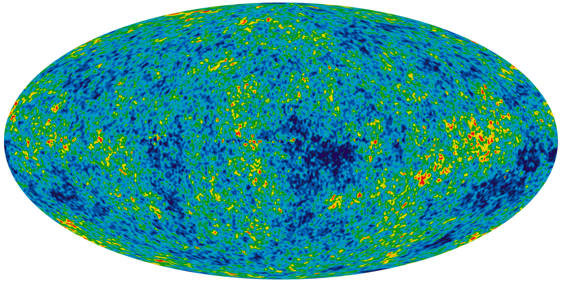
IAIFI Fellow

LBL - 3rd May 2024

Video Credit: N-body simulation Francisco Villaescusa-Navarro

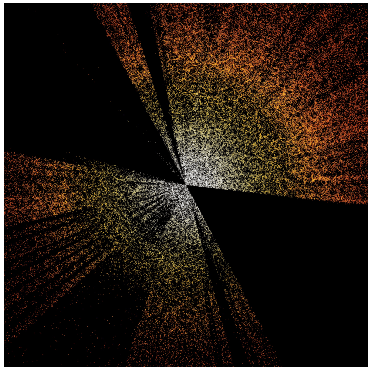
Simons Observatory

CMB-S4



Secondary anisotropies

DESI, DESI-II, Spec-S5



Galaxy formation

Euclid

LSST

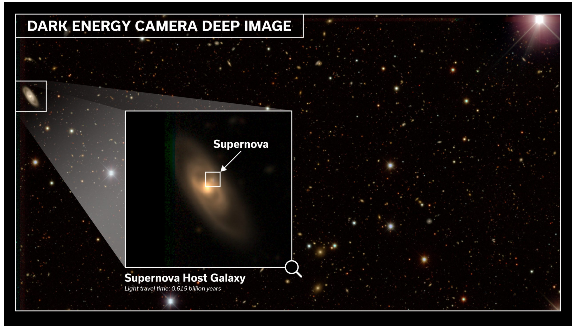
Intrinsic alignments

Machine Learning

The era of Big Data Cosmology

xAstrophysics

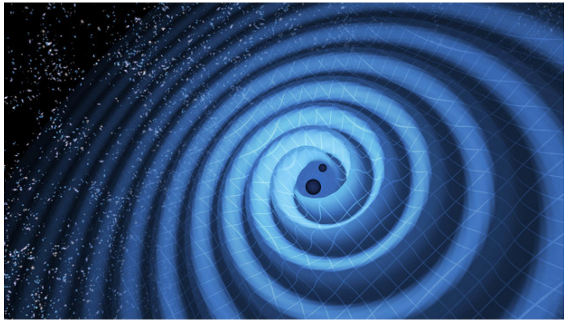
LSST



Dust

Ligo

Einstein

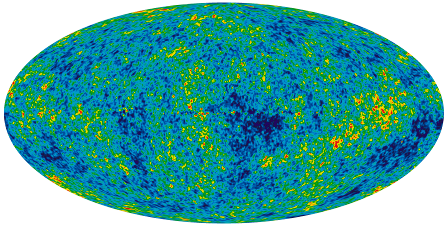


Energy and matter content

Early Universe

Inflation

δ_{Initial}



f_{NL}

Non-Gaussianity

Multifield Inflation

n_s

Tilt power spectrum

Evolution

Hubble tension

Ω_m

Dark matter

H_0

Hubble Constant

w_0, w_a

Dark energy

Ω_b

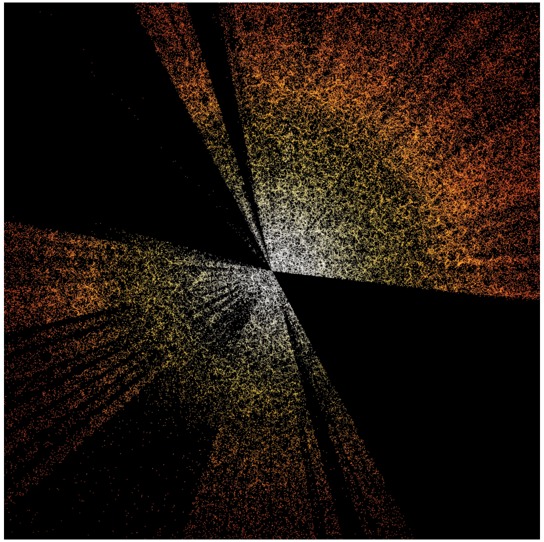
Baryons

$\sum m_\nu$

Neutrino masses

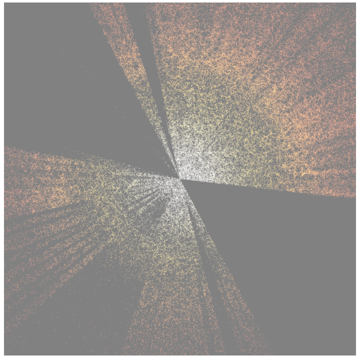
Beyond the Standard Model

Late Universe



δ_{Final}

1

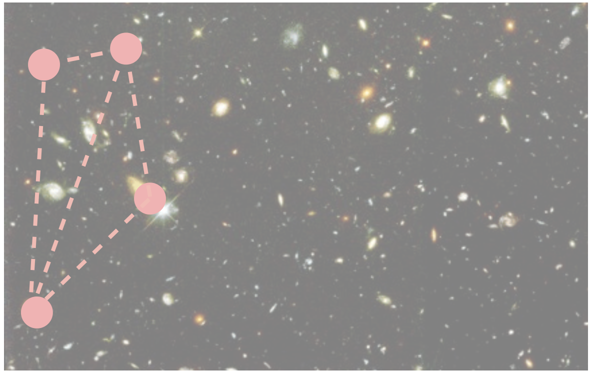


Cosmological *(field level)* Inference for Galaxy Surveys

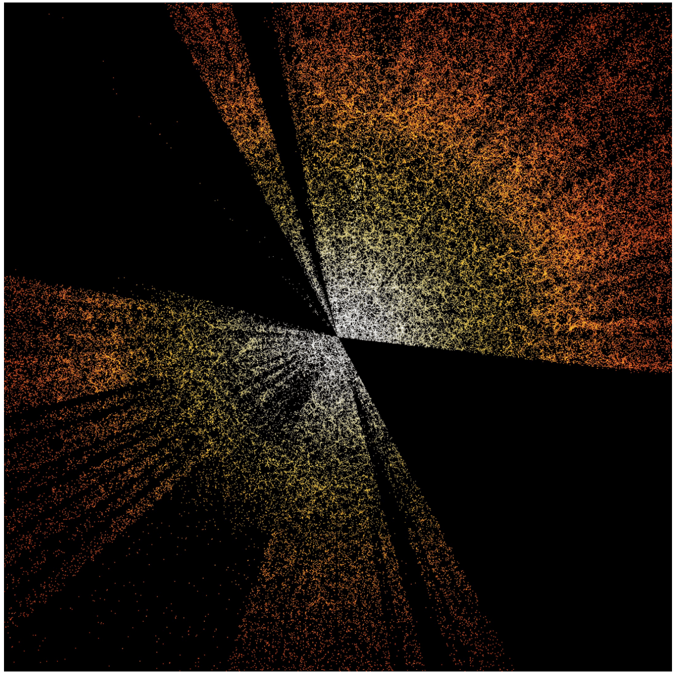
2

Hybrid ML - Physics Simulators

3

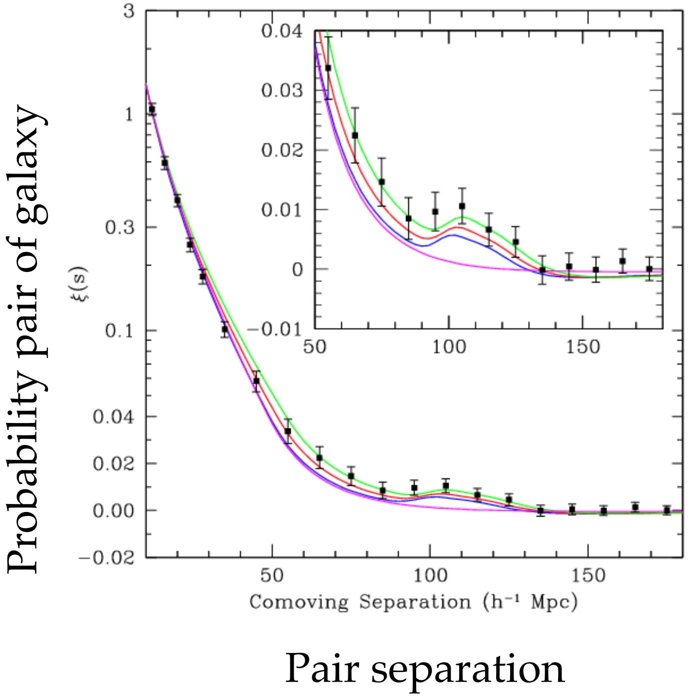


Unsupervised searches



High dimensional data x

Unknown $p(x|\mathcal{C})$



Simple summary statistic s

$p(s|\mathcal{C})$ estimated with Perturbation Theory

$$p(x|\theta)$$

Parameters

$$p(\theta|x)$$

Simulator + Density Estimation

Observable

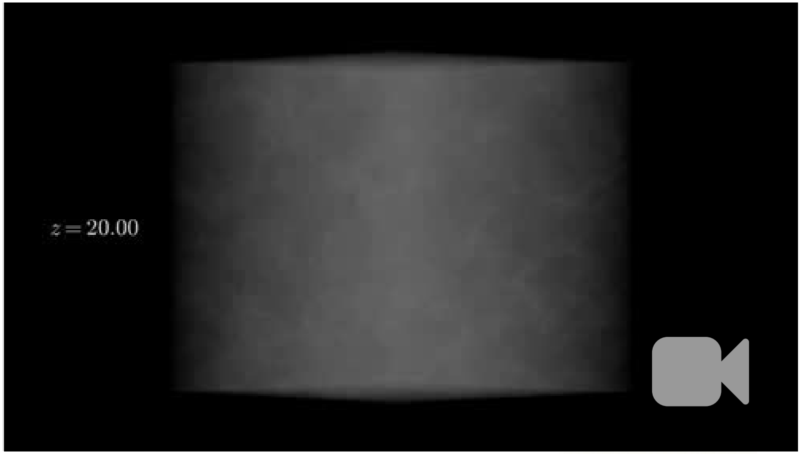
$$x$$

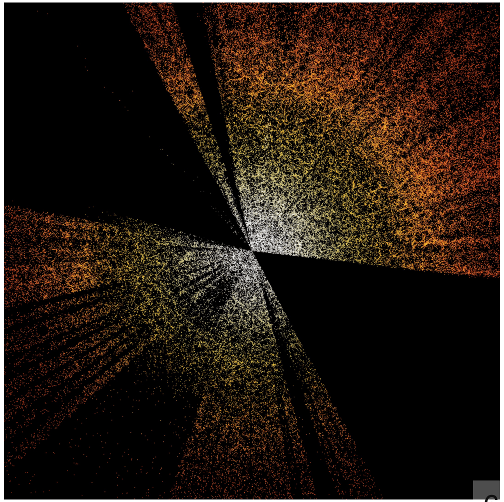
Ω_m, w_0, w_a, $f_{\rm NL}$ ...

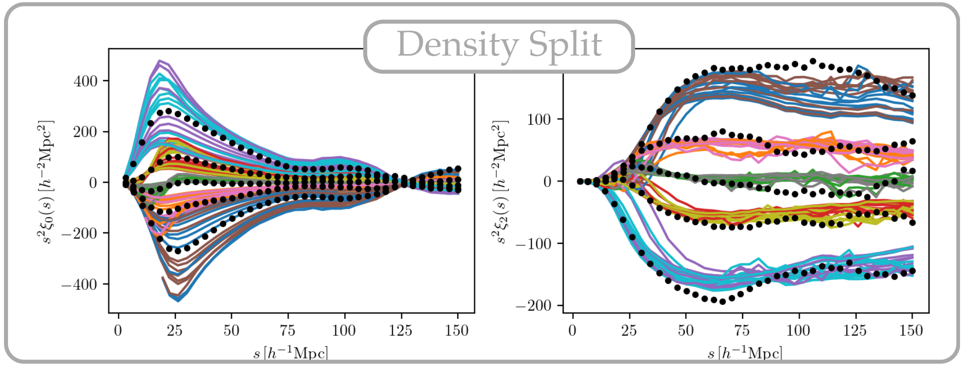
Dark matter

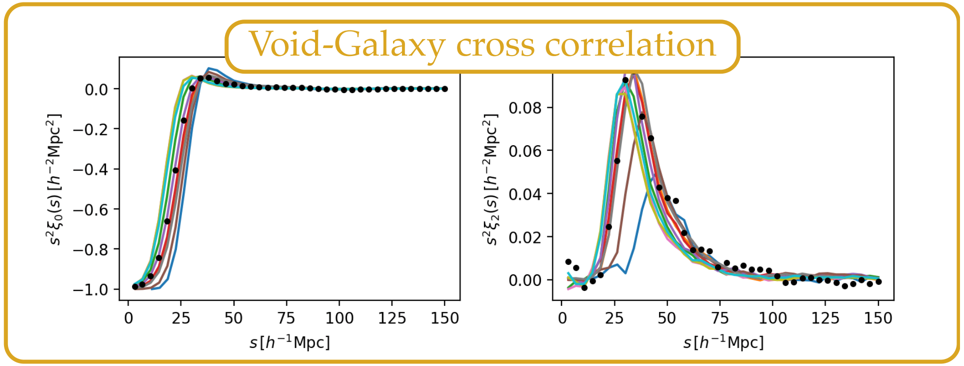
Dark energy

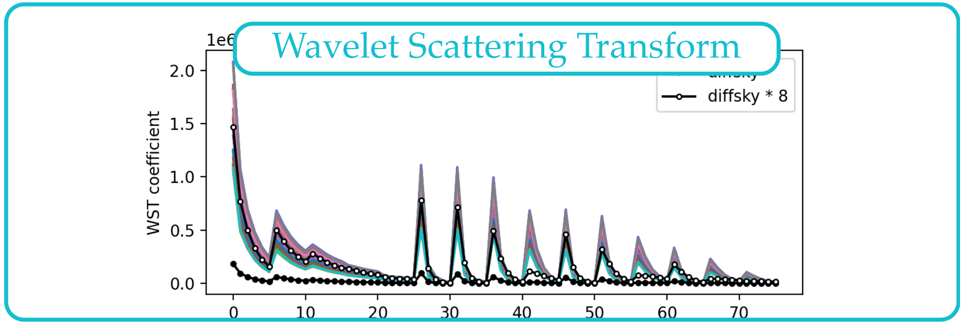
Inflation

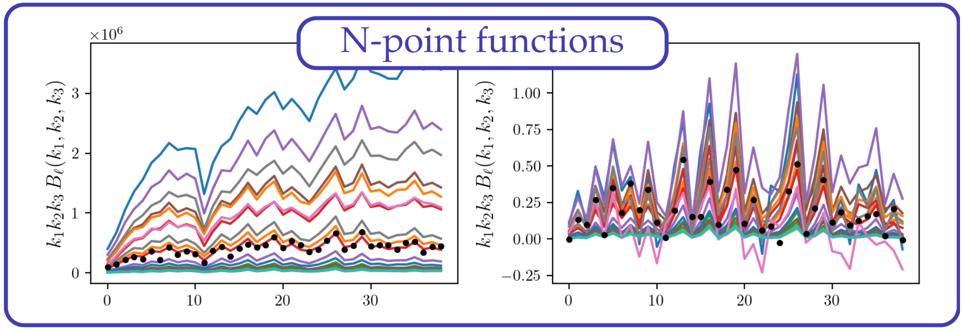


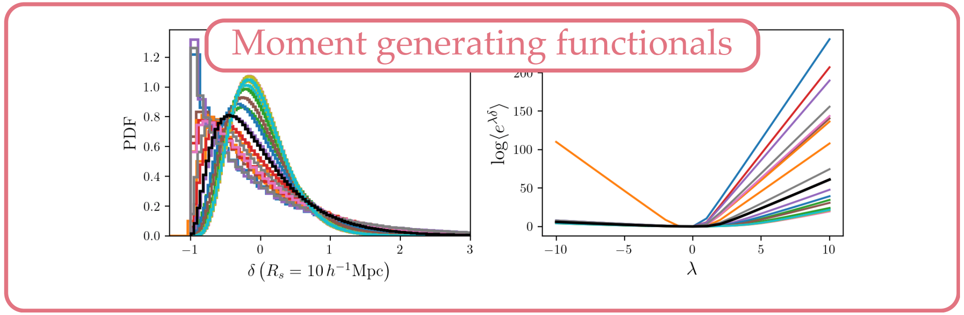


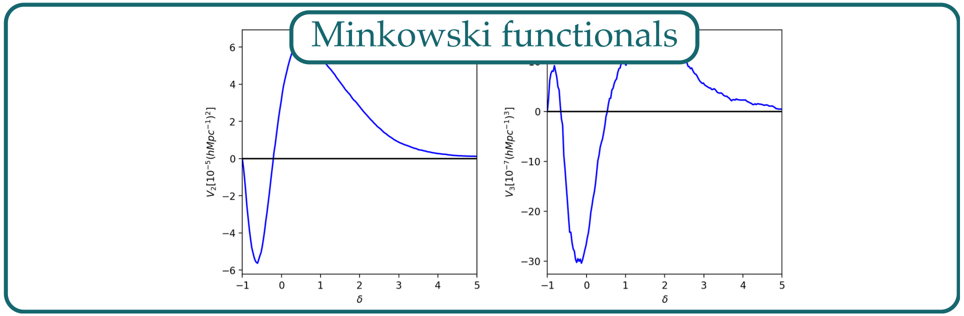












Credit: Alternative Clustering Methods. Enrique Paillas, Wei Liu, Mathilde Pinon, Gillian Beltz-Mohrmann, Georgios Valogiannis)

A forward model ***samples*** the likelihood

Parameters

θ $\longrightarrow$ **Forward Model** $\longrightarrow$ x

Observable

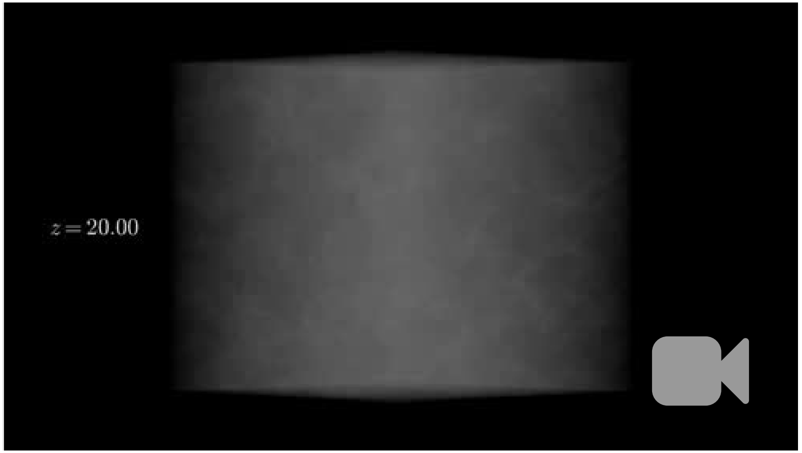
$$p(x|\theta) = \int dz p(x, z|\theta)$$

$\Omega_m, w_0, w_a, f_{\rm NL} \dots$

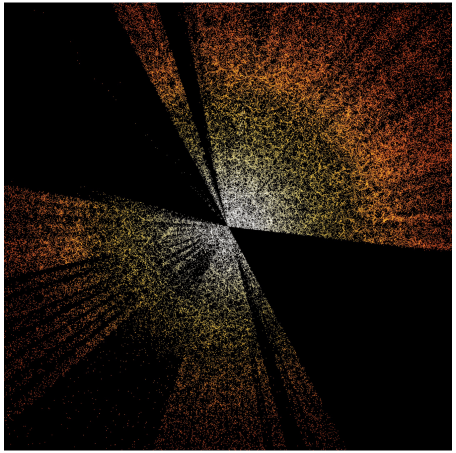
Dark matter

Dark energy

Inflation



Observed galaxy pointcloud

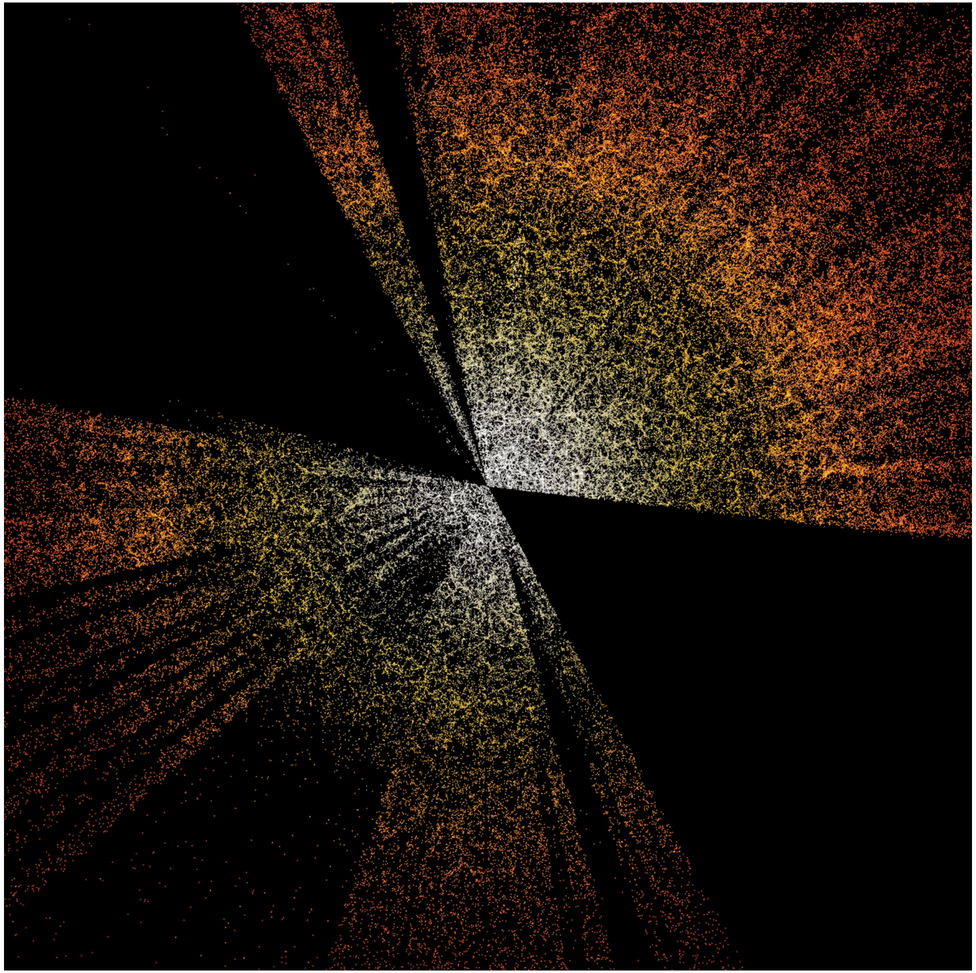


1024x1024

A 2D animation of a folk music band composed of anthropomorphic autumn leaves, each playing traditional bluegrass instruments, amidst a rustic forest setting dappled with the soft light of a harvest moon

Image credit: DALL·E 3

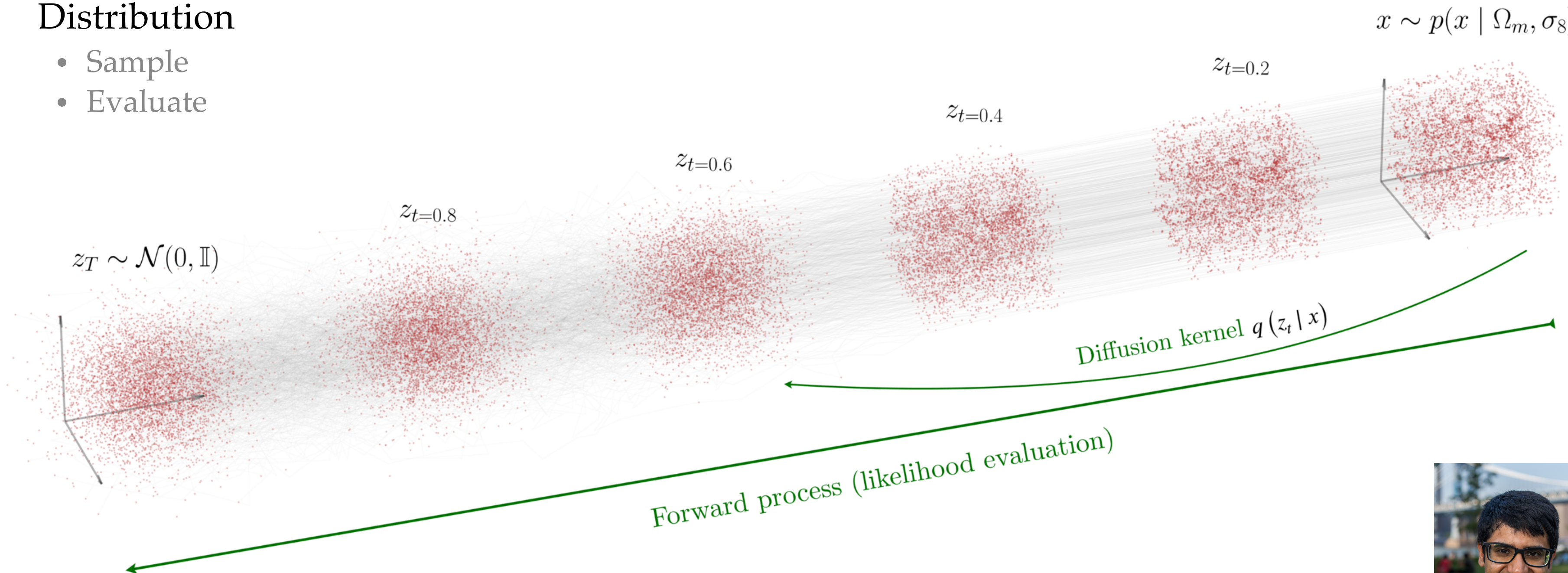
$$p(x|\mathcal{C})?$$



Target Distribution

Base Distribution

- Sample
- Evaluate

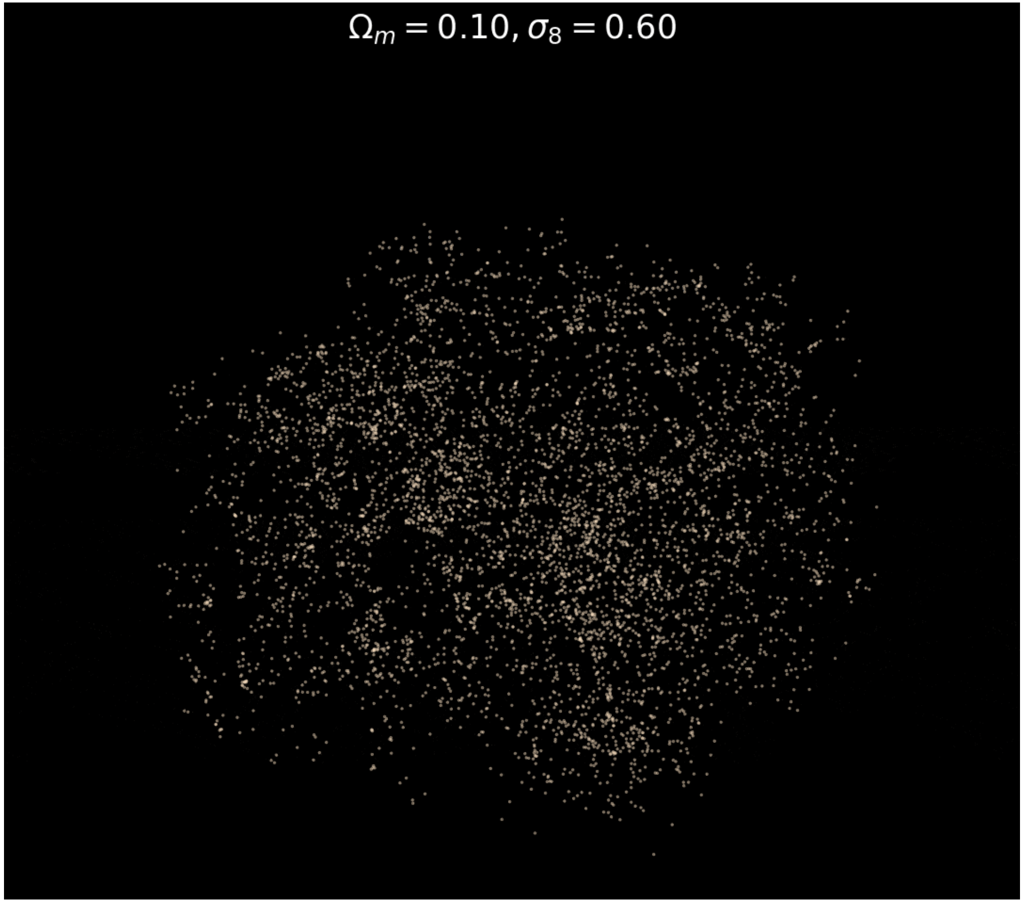


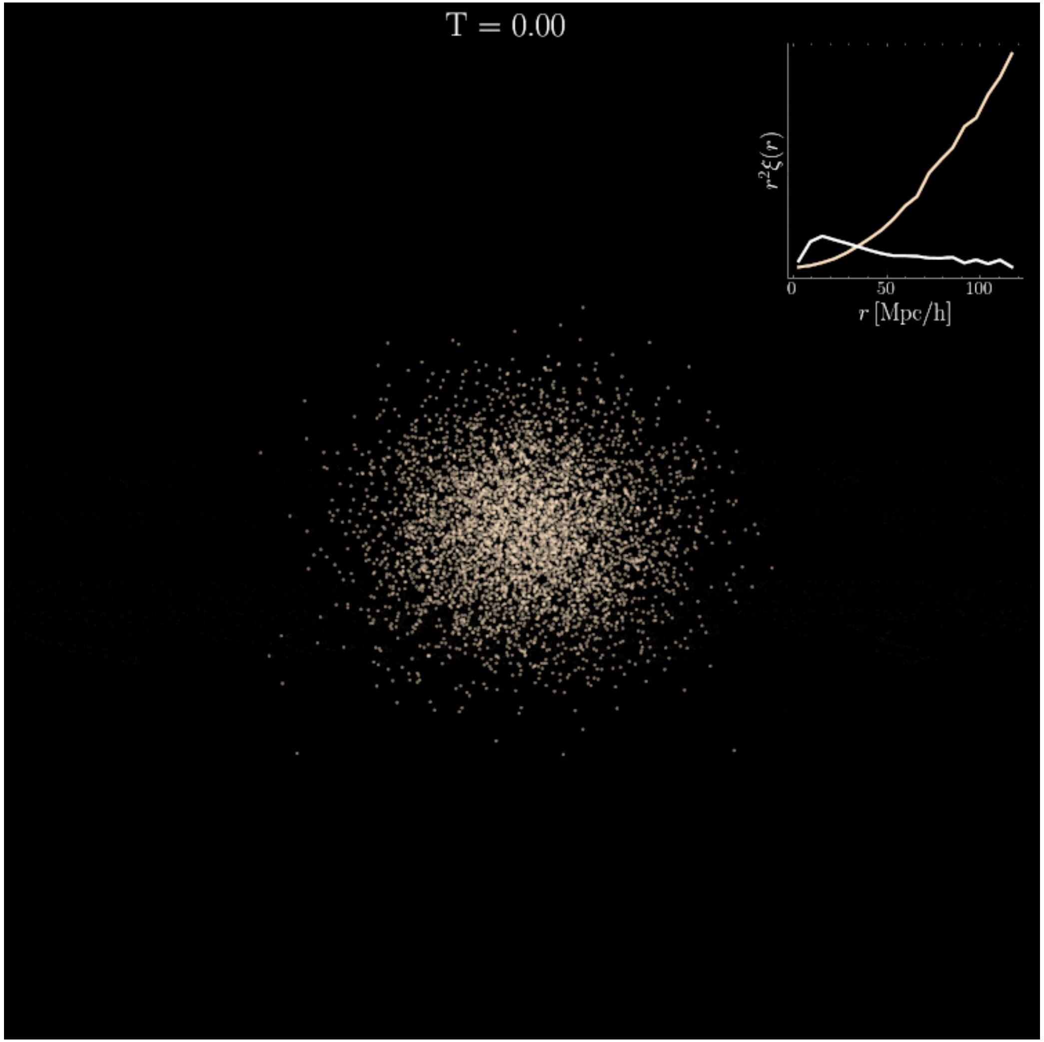
"A point cloud approach to generative modeling for galaxy surveys at the field level"
Cuesta-Lazaro and Mishra-Sharma
arXiv:2311.17141

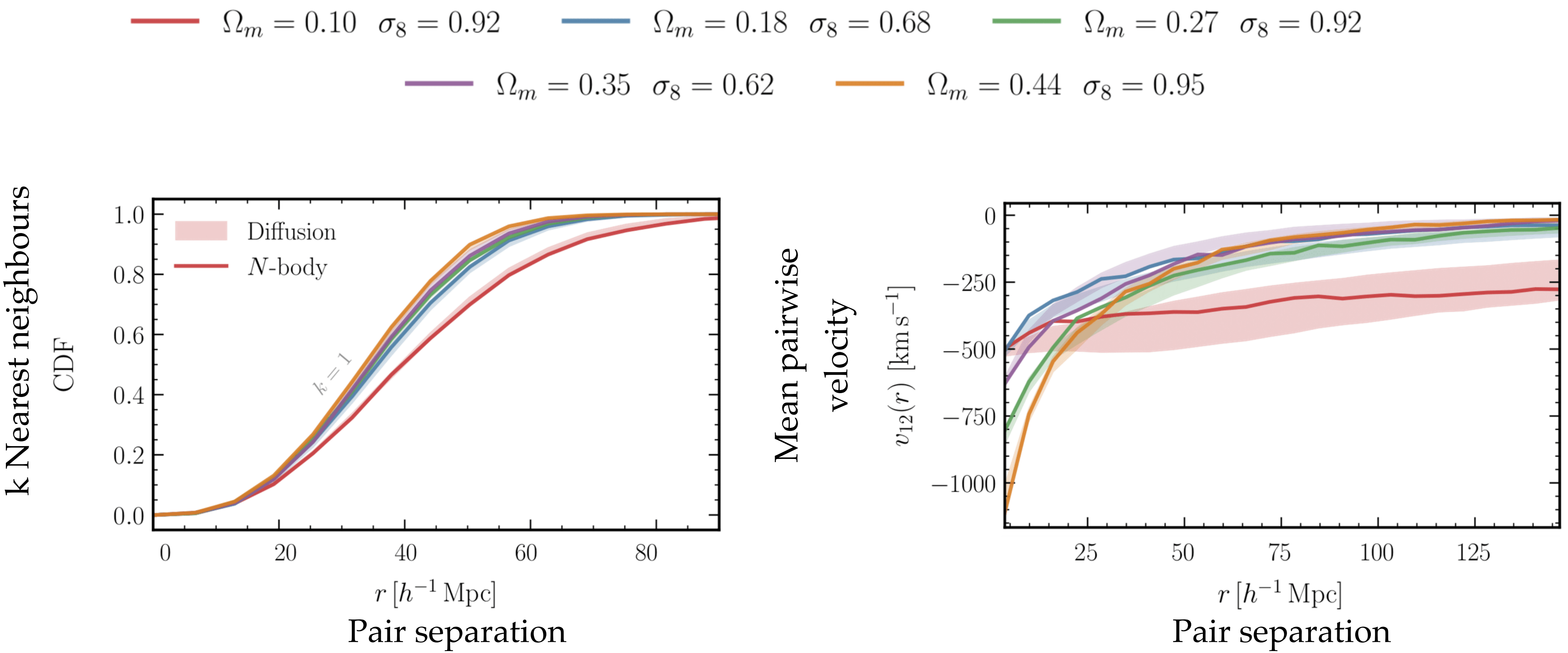
Siddharth Mishra-Sharma

Fixed Initial Conditions

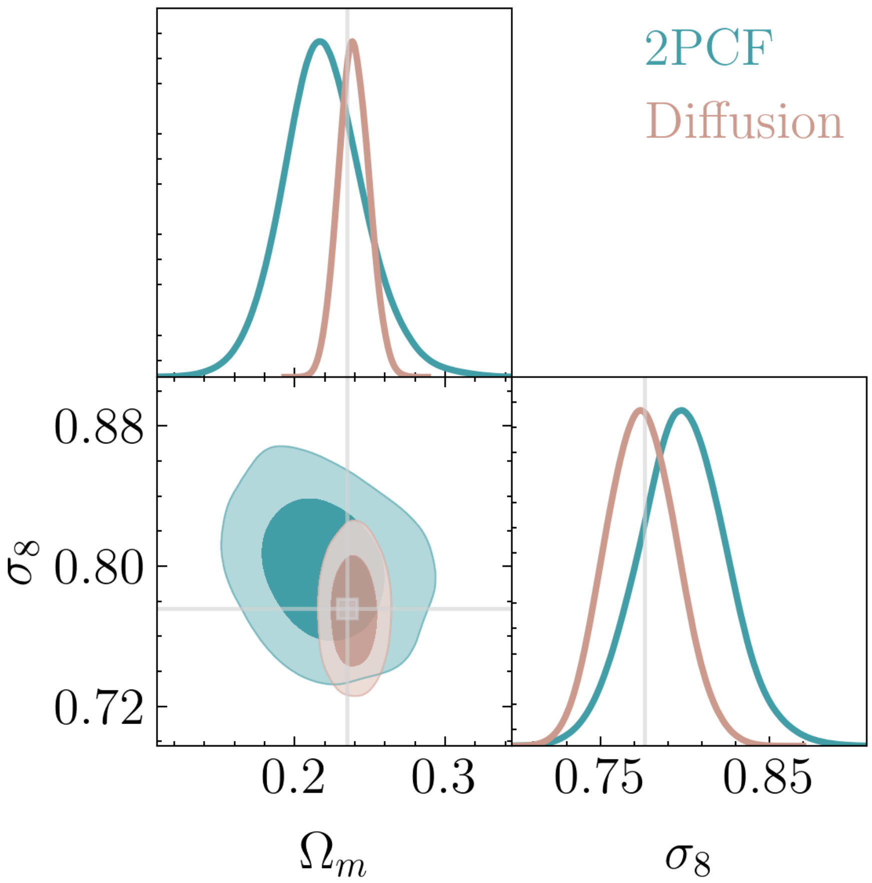
Varying Cosmology



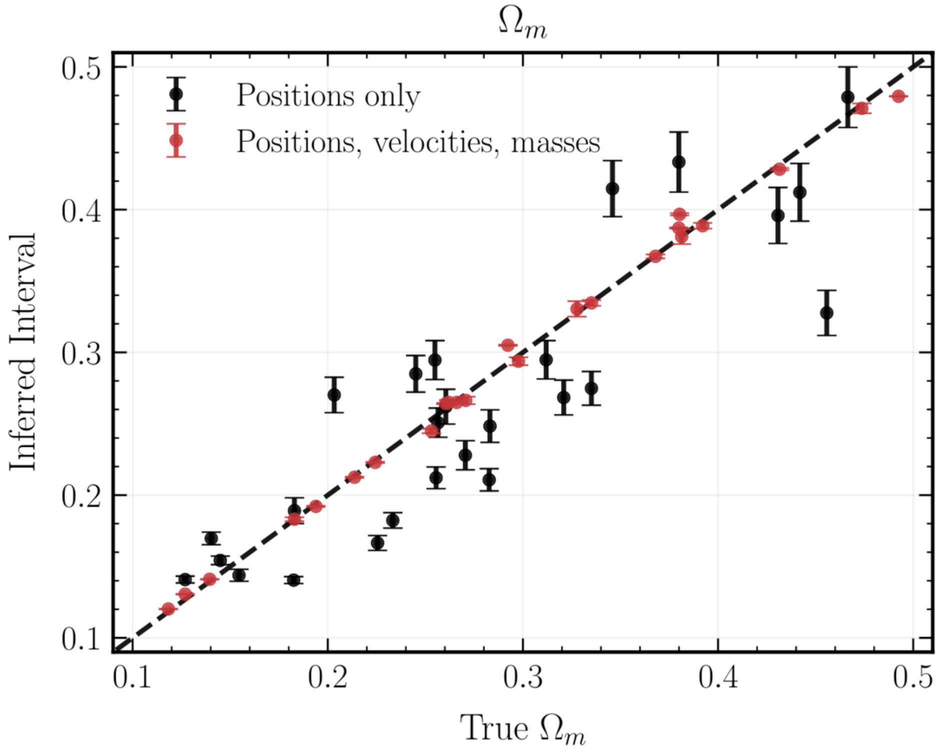

T = 0.00
$r^2\xi(r)$
0
50
100
r [Mpc/h]



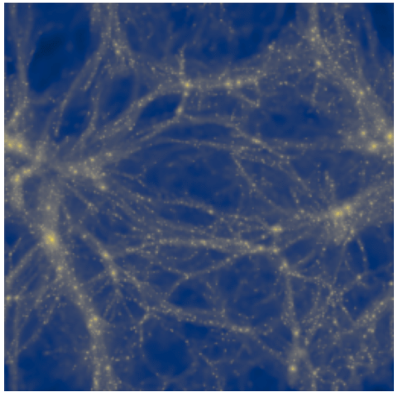
$\Omega_m = 0.10 \quad \sigma_8 = 0.92$
$\Omega_m = 0.18 \quad \sigma_8 = 0.68$
$\Omega_m = 0.27 \quad \sigma_8 = 0.92$
$\Omega_m = 0.35 \quad \sigma_8 = 0.62$
$\Omega_m = 0.44 \quad \sigma_8 = 0.95$
k Nearest neighbours
CDF
Diffusion
N-body
$k = 1$
$r\,[h^{-1}\,\mathrm{Mpc}]$
Pair separation
Mean pairwise velocity
$v_{12}(r)\ [\mathrm{km\,s^{-1}}]$
$r\,[h^{-1}\,\mathrm{Mpc}]$
Pair separation



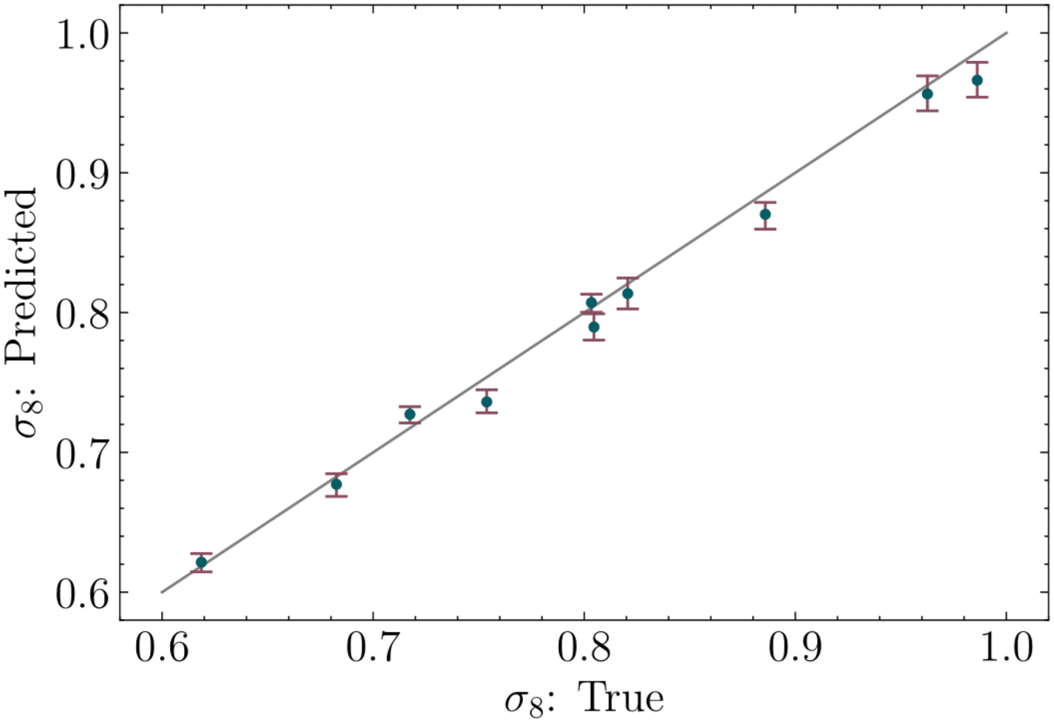
Trained on only 5000 positions!



Learning in 5000 dimensions with only 2000 simulations



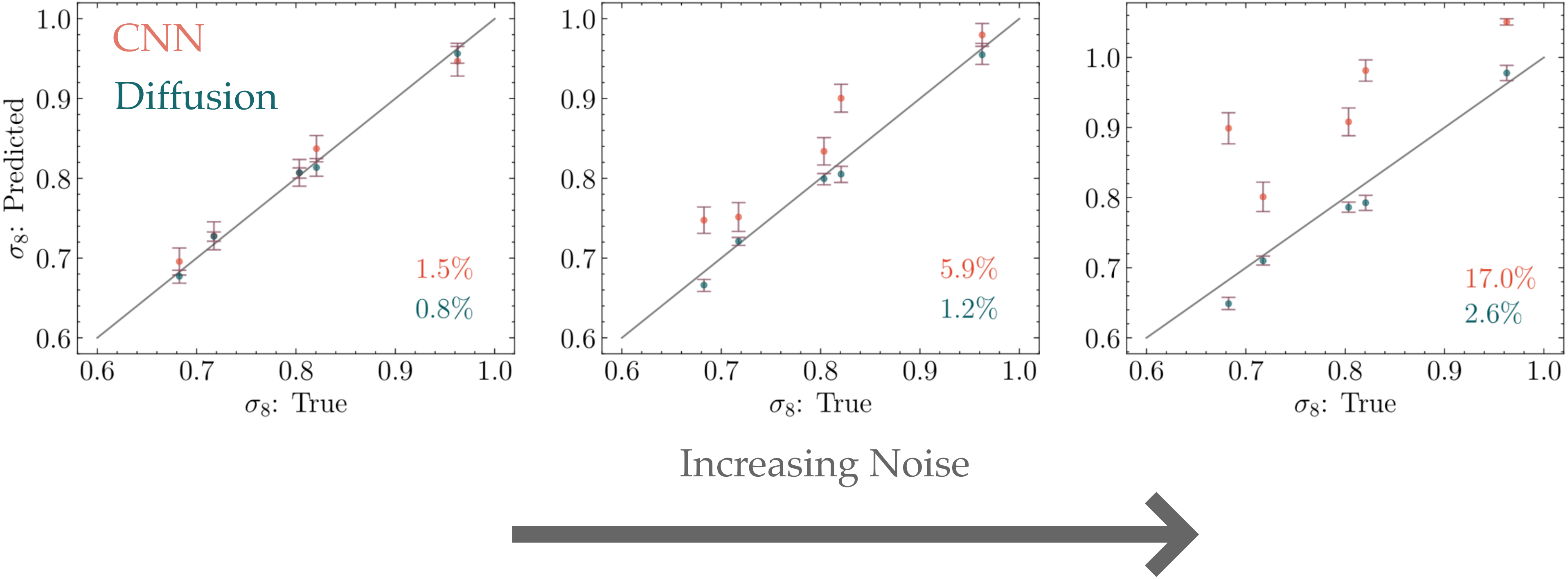
$$p(\delta_m | \Omega_m, \sigma_8)$$



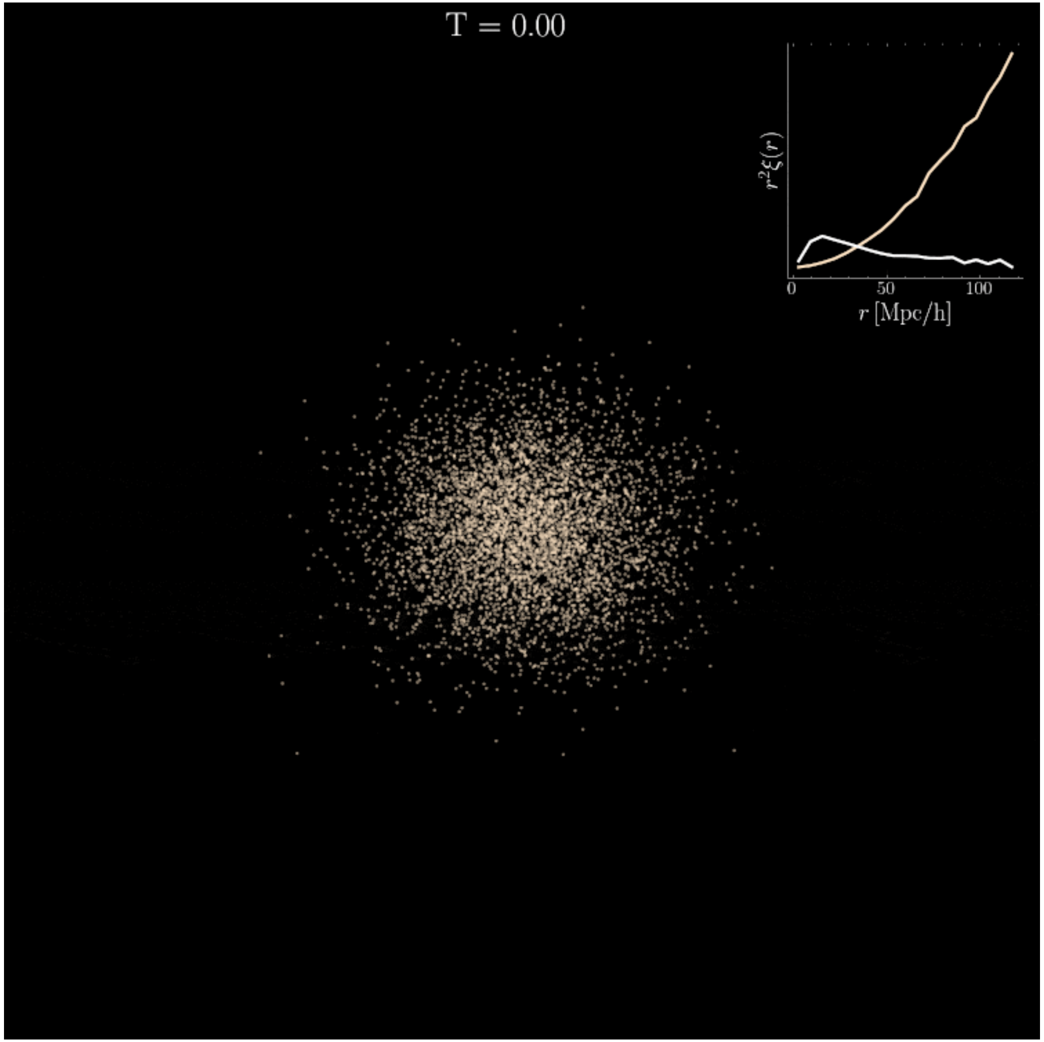
Nayantara
Mudur

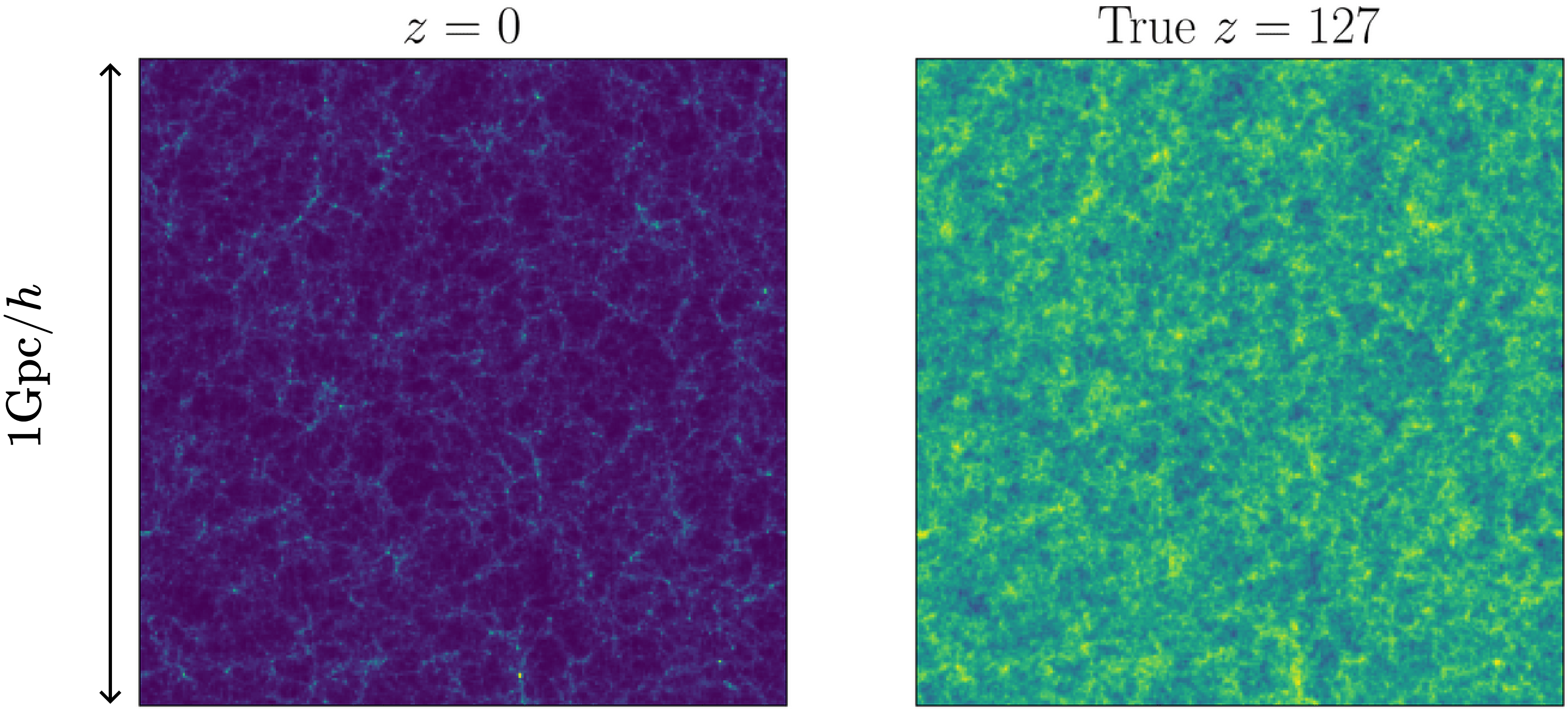
"Diffusion-HMC: Parameter Inference with Diffusion Model driven Hamiltonian Monte Carlo"
Mudur, **Cuesta-Lazaro** and Finkbeiner
in prep

$$p(\sigma_8|\delta_m) \qquad p(\sigma_8|\delta_m + 0.01\epsilon) \qquad p(\sigma_8|\delta_m + 0.02\epsilon)$$



"Your diffusion model is secretly a certifiably robust classifier"
Chen et al
arXiv:2402.02316

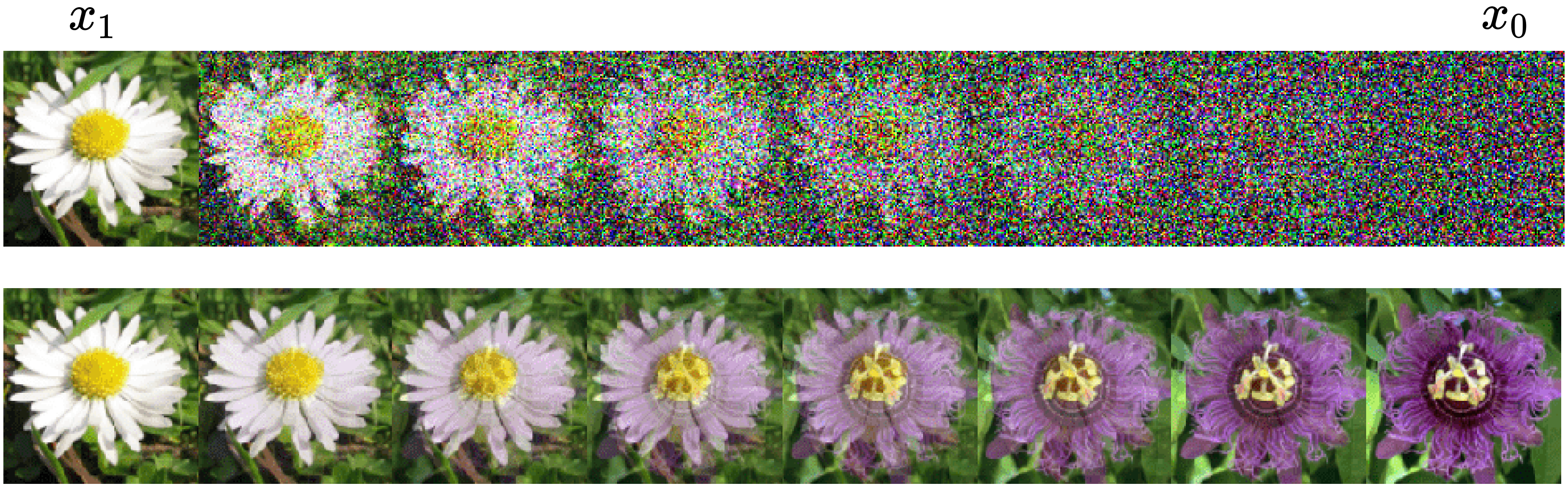

T = 0.00
$r^2\xi(r)$
0
50
100
r [Mpc/h]



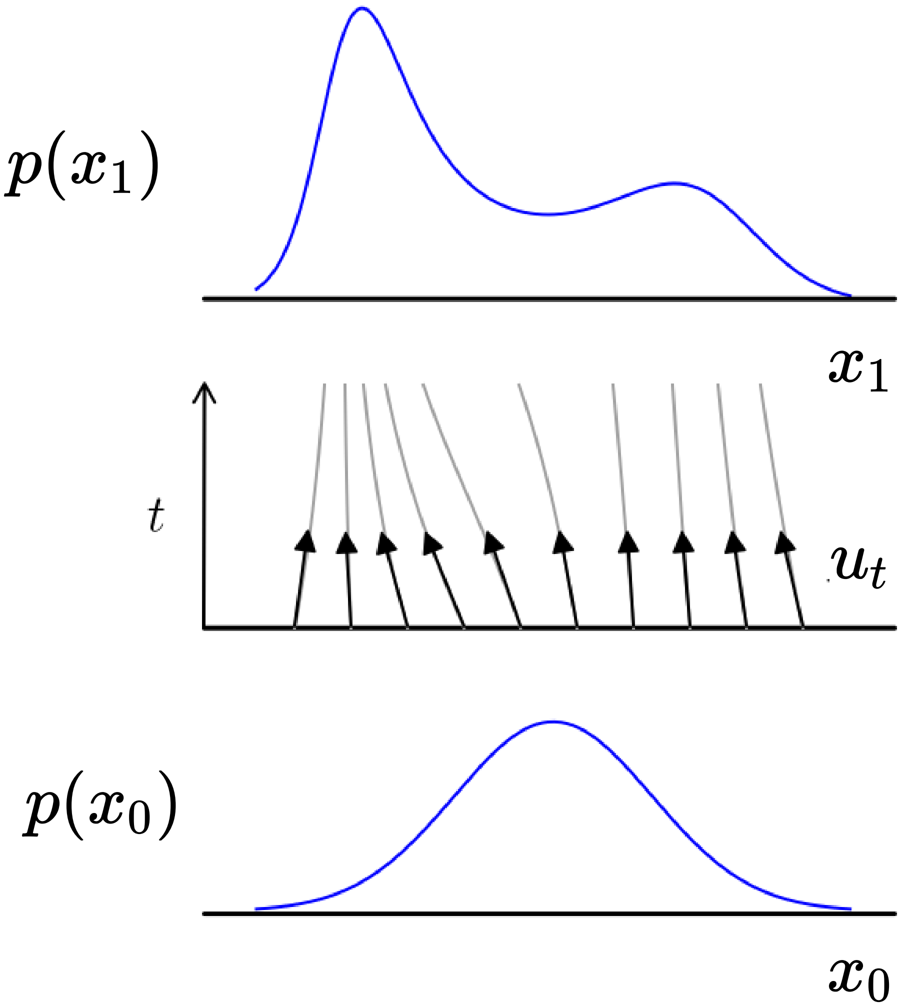
arXiv:2312.09271
arXiv:2307.09504

1 to Many: $p_\phi\left(\delta_{z=127} \middle| \delta_{z=0}\right)$

Stochastic Interpolants: Bridging arbitrary densities



"Stochastic Interpolants: A Unifying Framework for Flows and Diffusions"
Albergo, Boffi, Vanden-Eijnden
arXiv:2303.08797



Flow ODE

$$\frac{dx_t}{dt} = u_t(x_t)$$

Continuity Equation

$$\frac{dp_t}{dt} = -\left(\nabla u_t p_t\right)(x_t)$$

Regress the velocity field

Unknown!

$$L_{\text{FM}} = \min \mathbb{E}_{t,\boxed{p_t(x|x_0)}} \left|\left| v_t^\theta(x|x_0) - \boxed{u_t(x|x_0)} \right|\right|^2$$

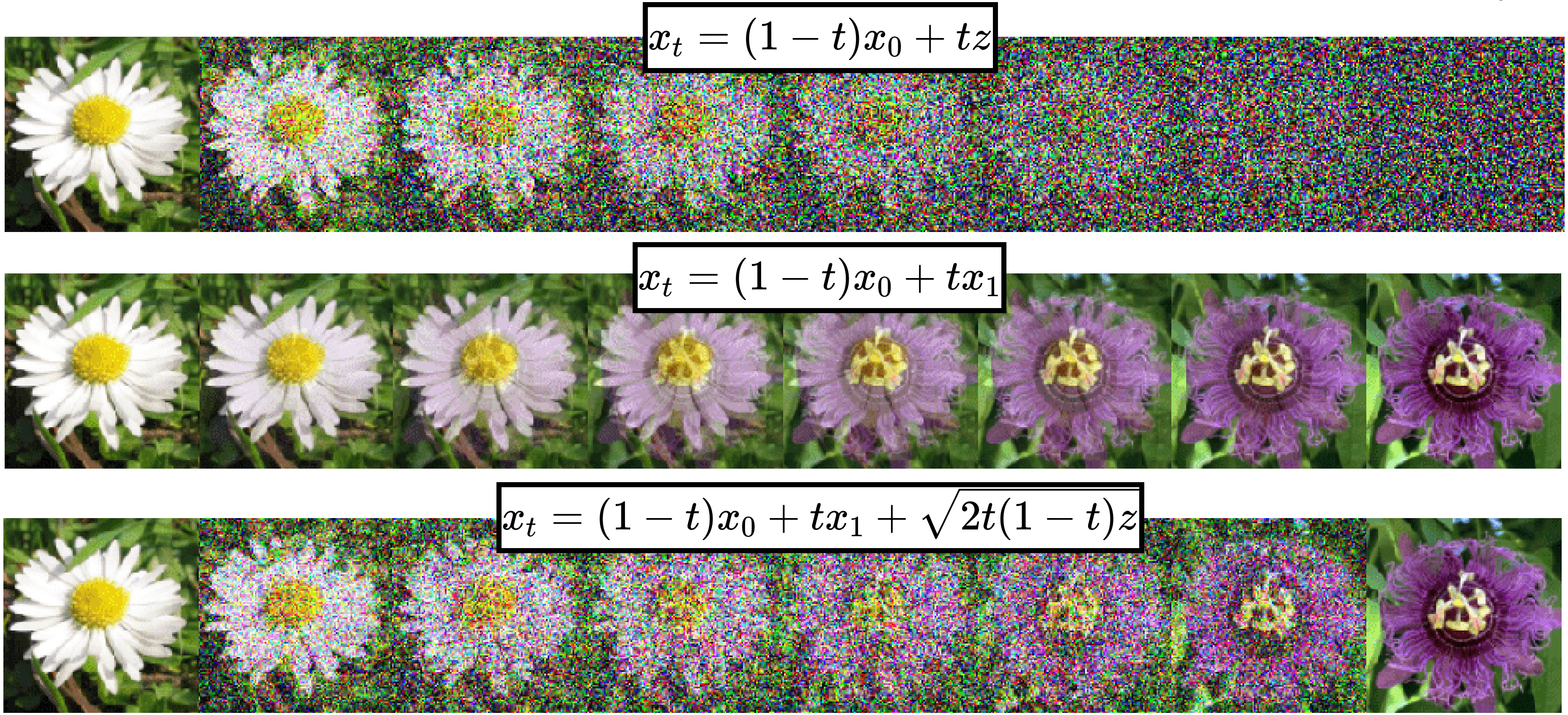
$$\boxed{x_t = (1-t)x_0 + tx_1}$$

$$p_t(x|x_0) \quad \begin{matrix} p_0(x|x_0) = \delta_{x_0} \\ p_1(x|x_0) = p \end{matrix}$$

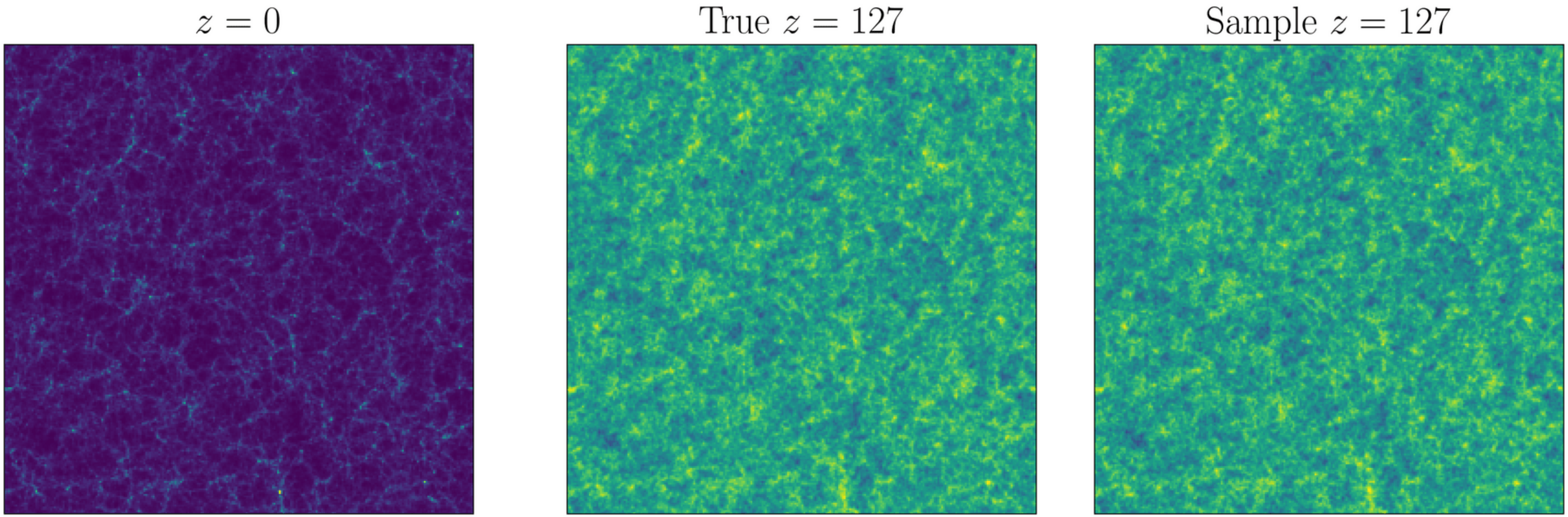
Boundary Conditions

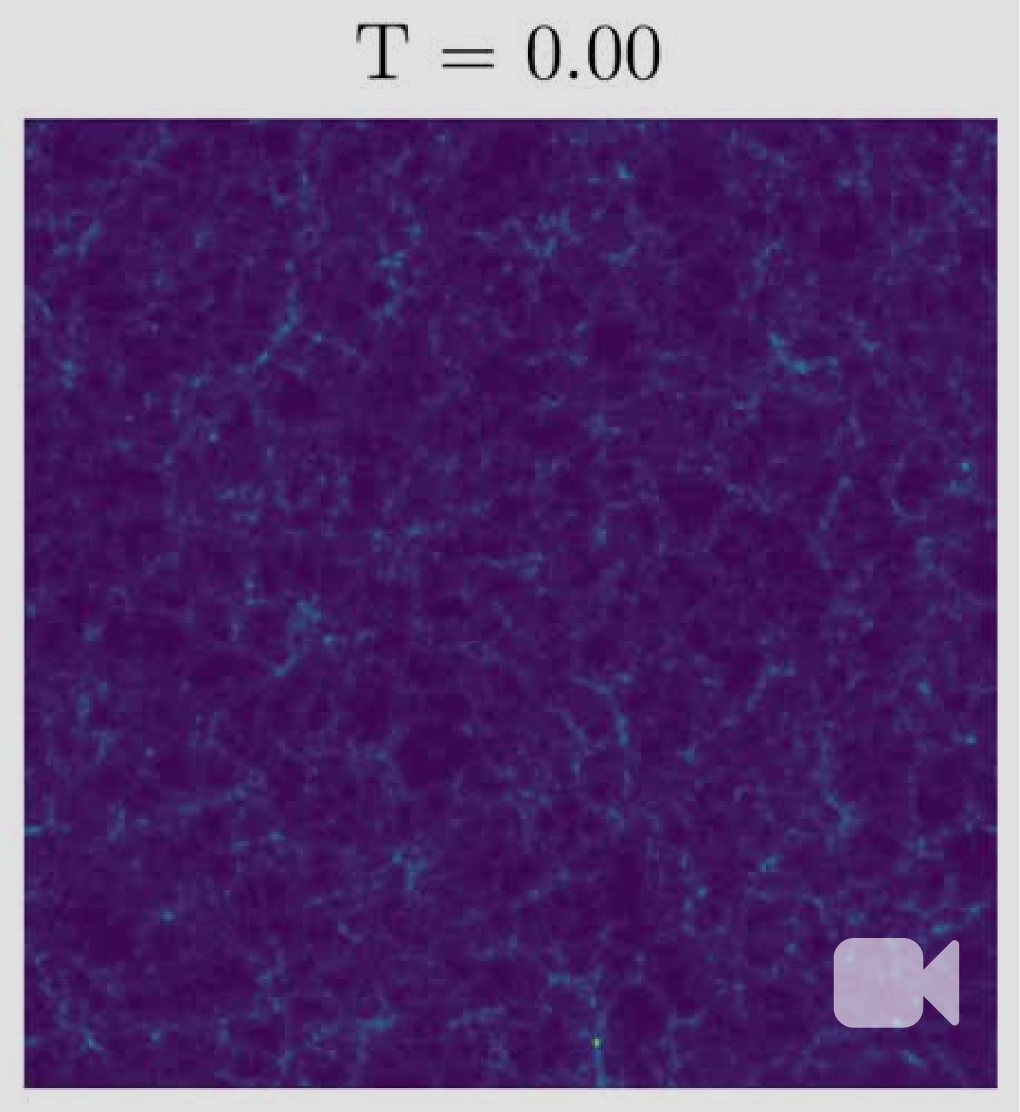
Stochastic Interpolants: Bridging arbitrary densities

x_1 x_0

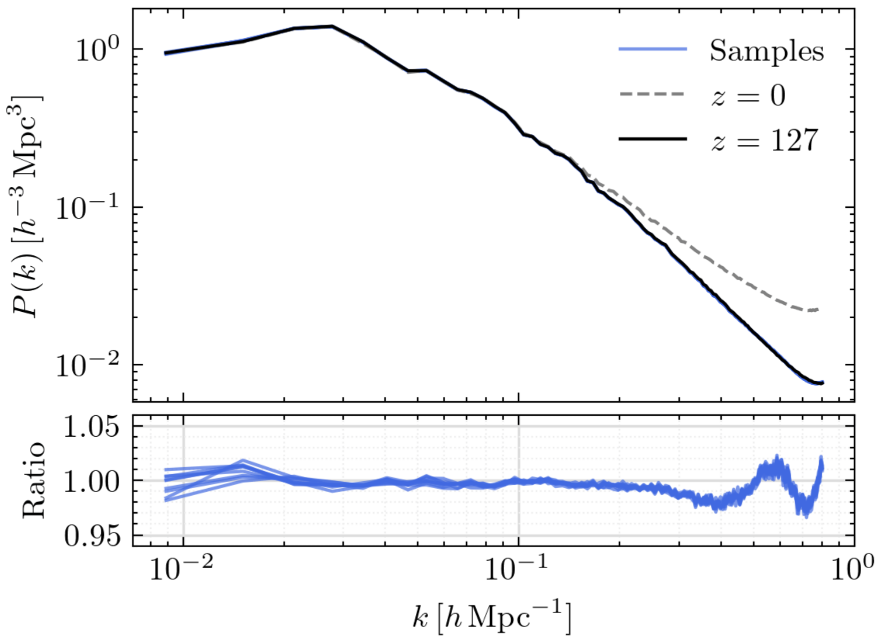


"Stochastic Interpolants: A Unifying Framework for Flows and Diffusions"
Albergo, Boffi, Vanden-Eijnden
arXiv:2303.08797


$z = 0$
True $z = 127$
Sample $z = 127$


$T = 0.00$

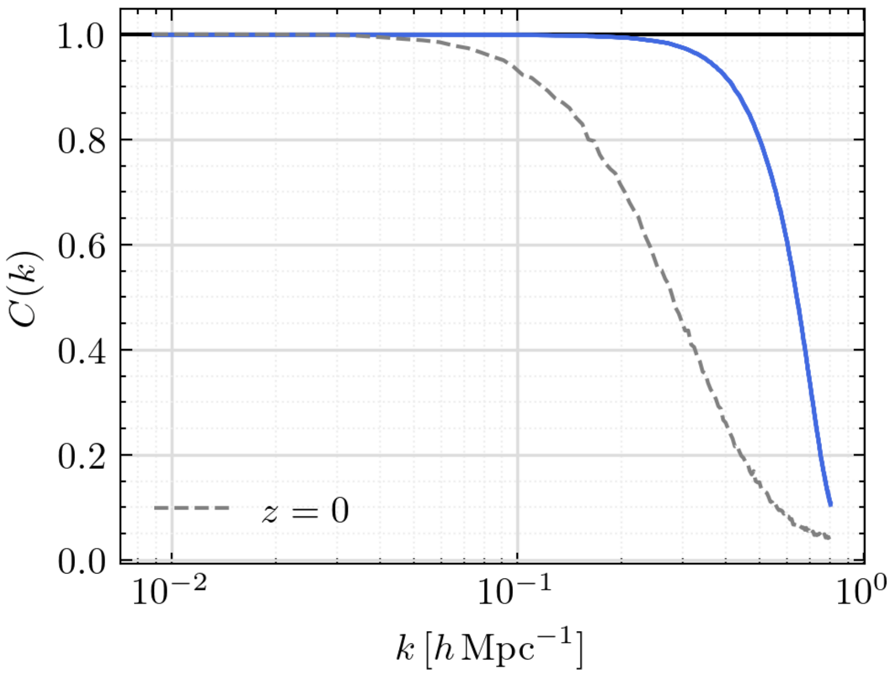
Power Spectrum



Scale (k)

← Large   Small →

Cross correlation



Scale (k)

← Large   Small →

All this works depends on simulations, but...

Can we run larger simulations? (DESI volumes)

Thousands of them?

At high resolution?

Faster?

Hybrid Physical / ML simulators

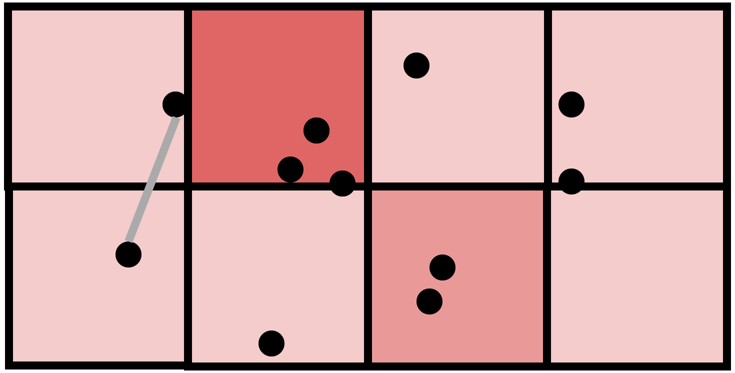
Gravitational evolution ODE

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}a} = \frac{1}{a^3 E(a)}\mathbf{v}$$

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}a} = \frac{1}{a^2 E(a)}\mathbf{F}(\mathbf{x}, a)$$

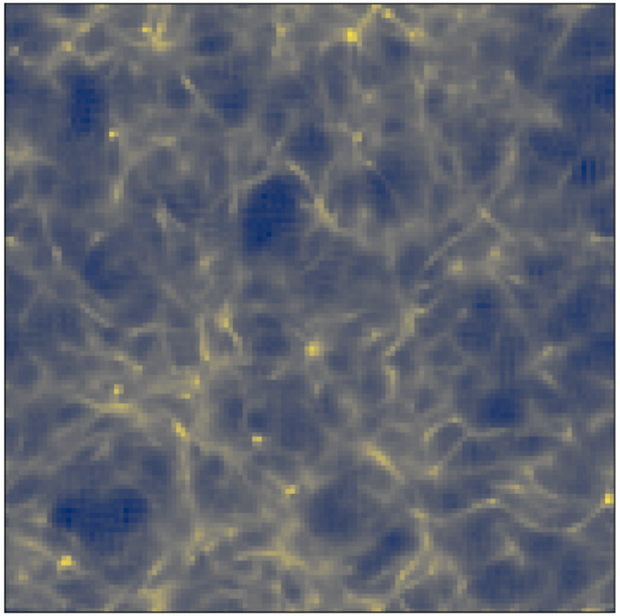
Particle-mesh

$$\mathbf{F}(\mathbf{x}, a) = \frac{3\Omega_m}{2}\nabla\phi^{\mathrm{PM}}(\mathbf{x})$$

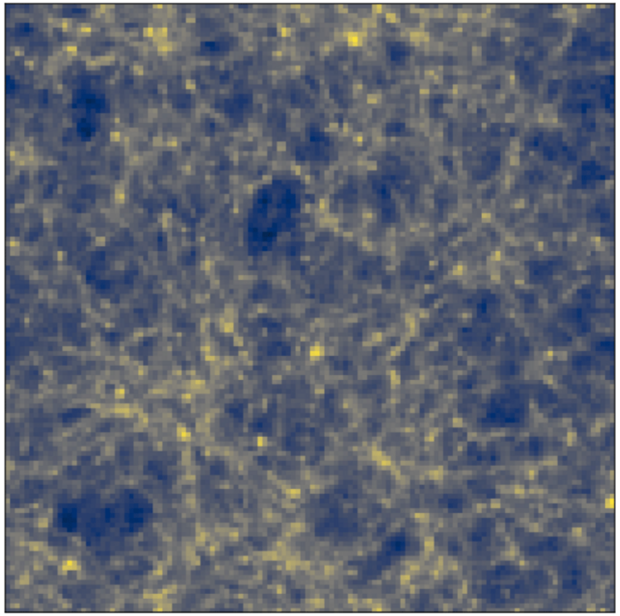


"Nbodyify: Adaptive mesh corrections for PM simulations" **Cuesta-Lazaro**, Modi *in prep*

Particle-mesh



Full Nbody



Gravitational evolution ODE

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}a} = \frac{1}{a^3 E(a)}\mathbf{v}$$

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}a} = \frac{1}{a^2 E(a)}\mathbf{F}(\mathbf{x}, a)$$

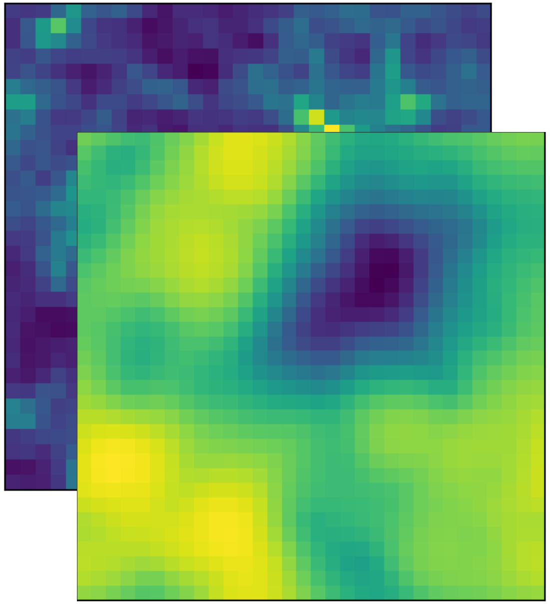
Hybrid Simulator - on the fly

$$\mathbf{F}_\theta(\mathbf{x}, a) = \frac{3\Omega_m}{2}\nabla\left[\phi^{\mathrm{PM}}(\mathbf{x}) + \phi_\theta^{\mathrm{corr}}(\mathbf{x}, a, \phi^{\mathrm{PM}}, \delta^{\mathrm{PM}})\right]$$

Trained to match particle velocities and positions:
DIFFERENTIABLE

Density

δ_{LR}



ϕ_{LR}

Gravitational Potential

2. Read features at position using attention

1. CNN

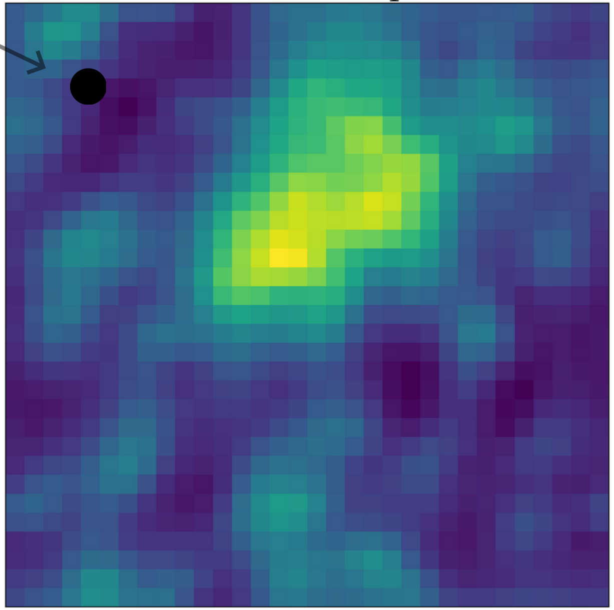
Learn features h

3. Compute force correction

$$\mathbf{F}_\theta(\mathbf{x}, a) = f_\theta(h_\theta(\mathbf{x}), a)$$

$h_\theta(\mathbf{x})$

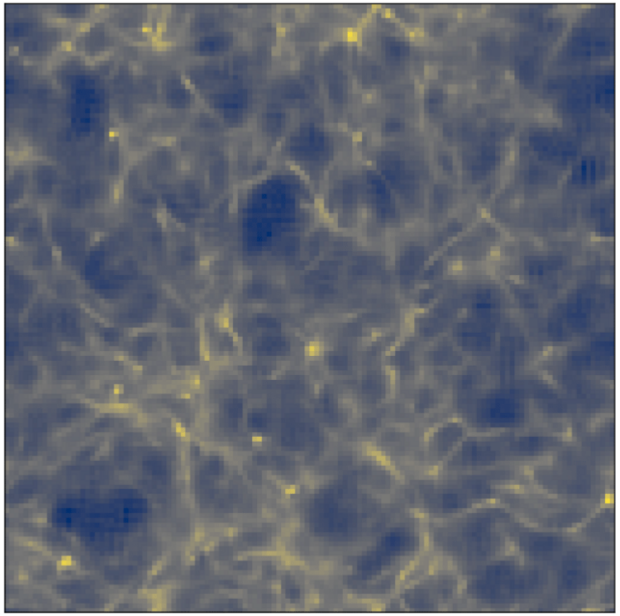
Feature maps



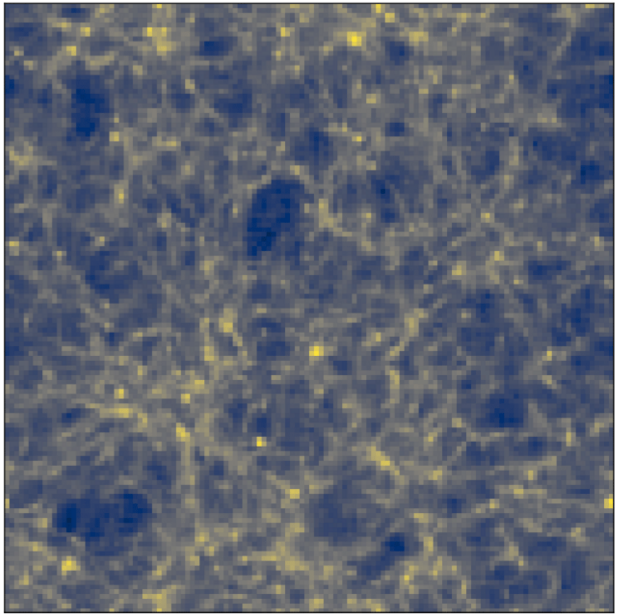
$$\mathcal{L} = \sum_t \left(x_t^{\mathrm{pred}} - x_t^{\mathrm{HR}}\right)^2$$

4. Run corrected simulation

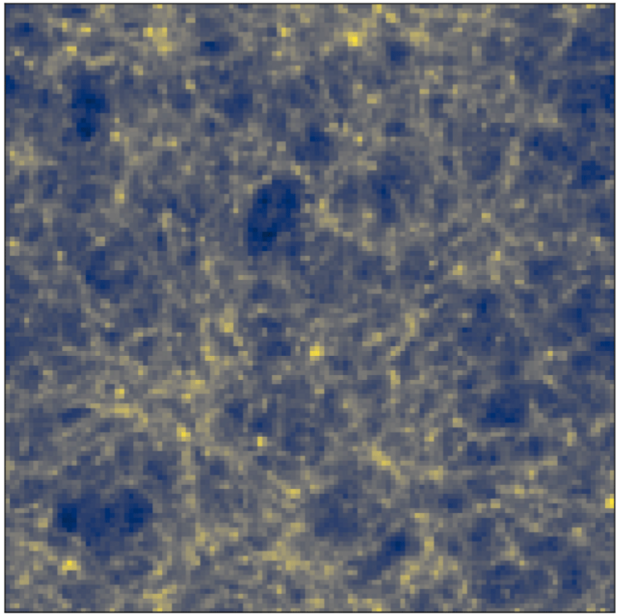
Particle-mesh



Hybrid ML-Simulator



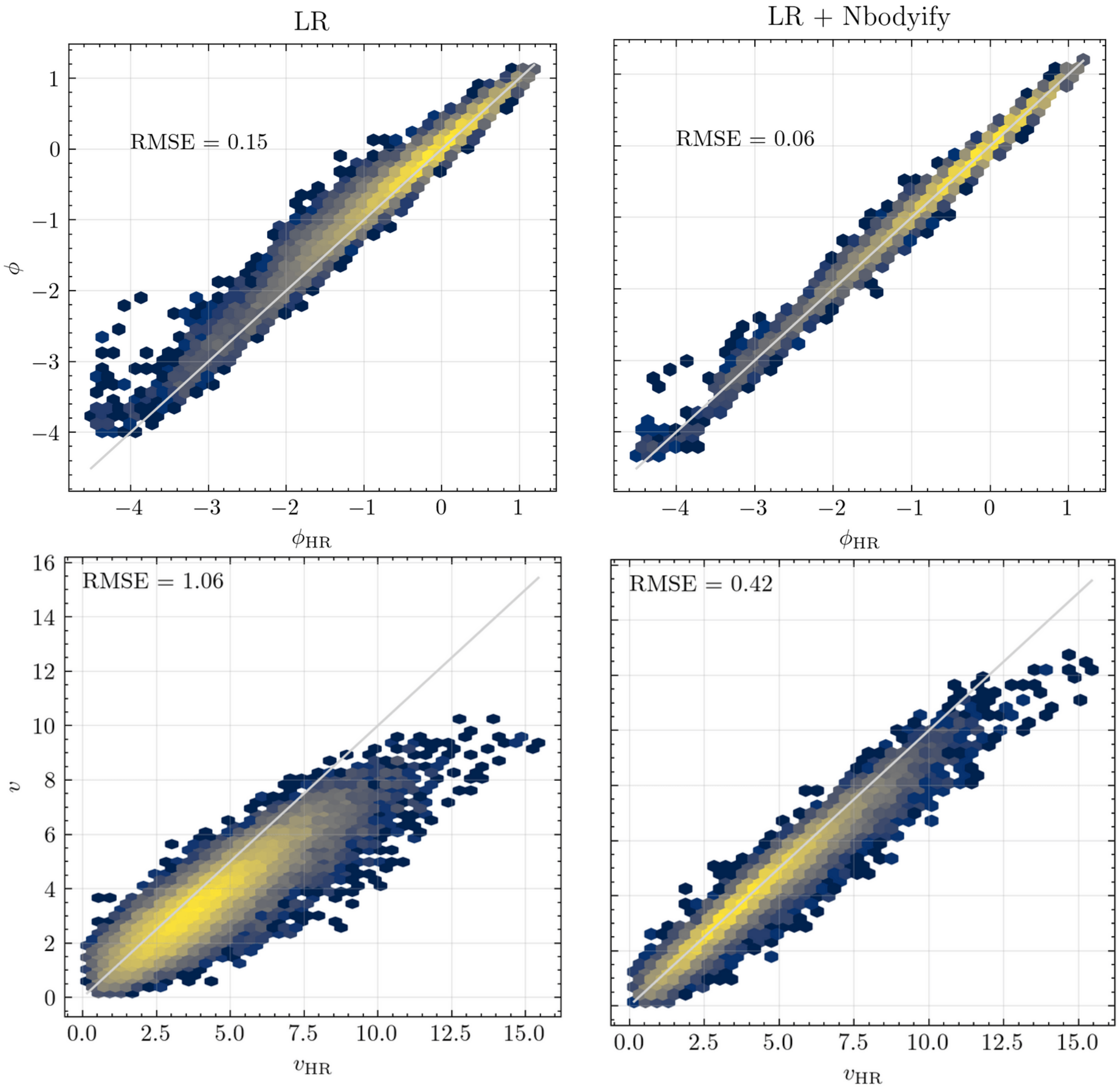
Full Nbody

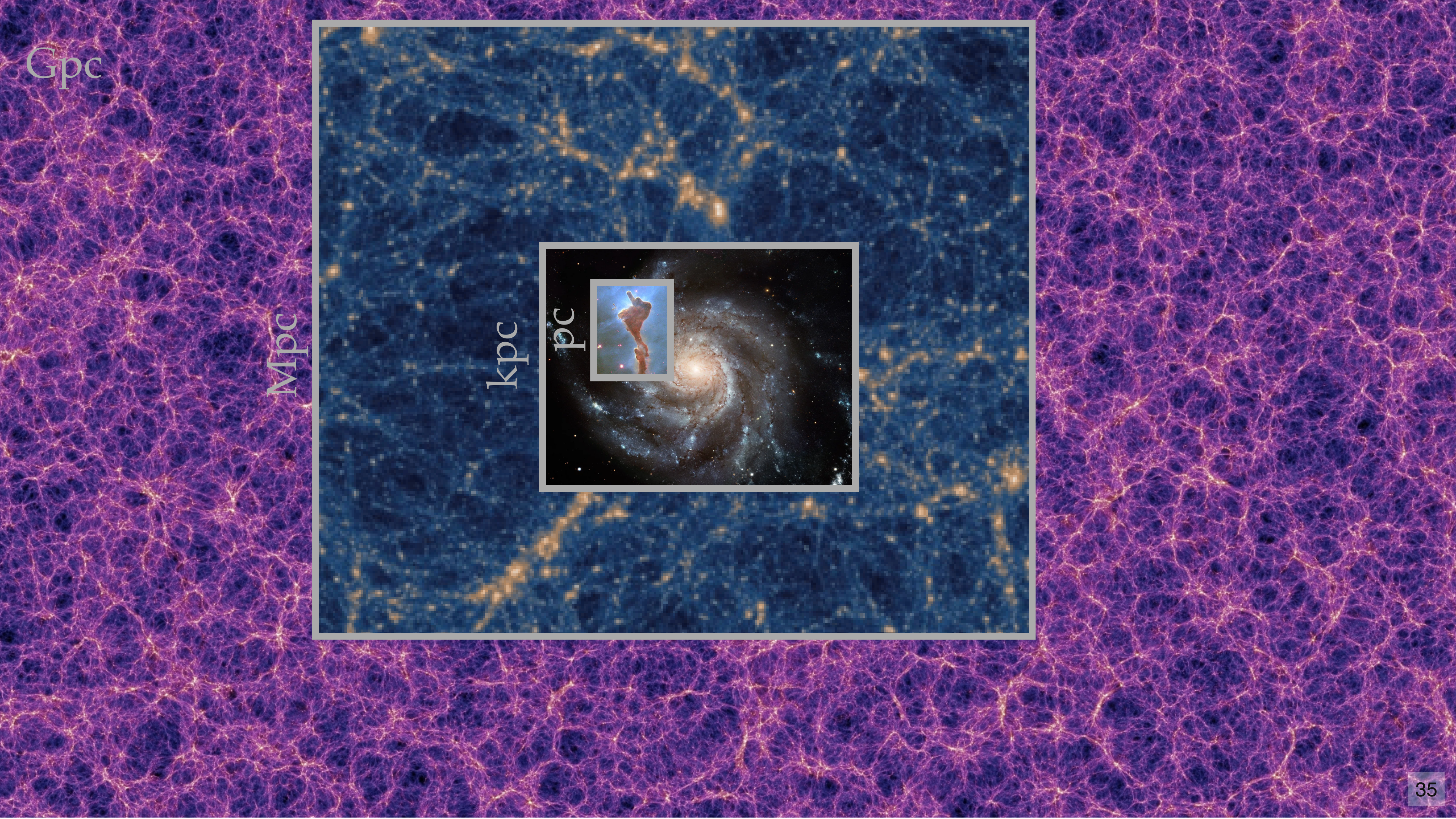


"Nbodyify: Adaptive mesh corrections for PM simulations" **Cuesta-Lazaro**, Modi *in prep*

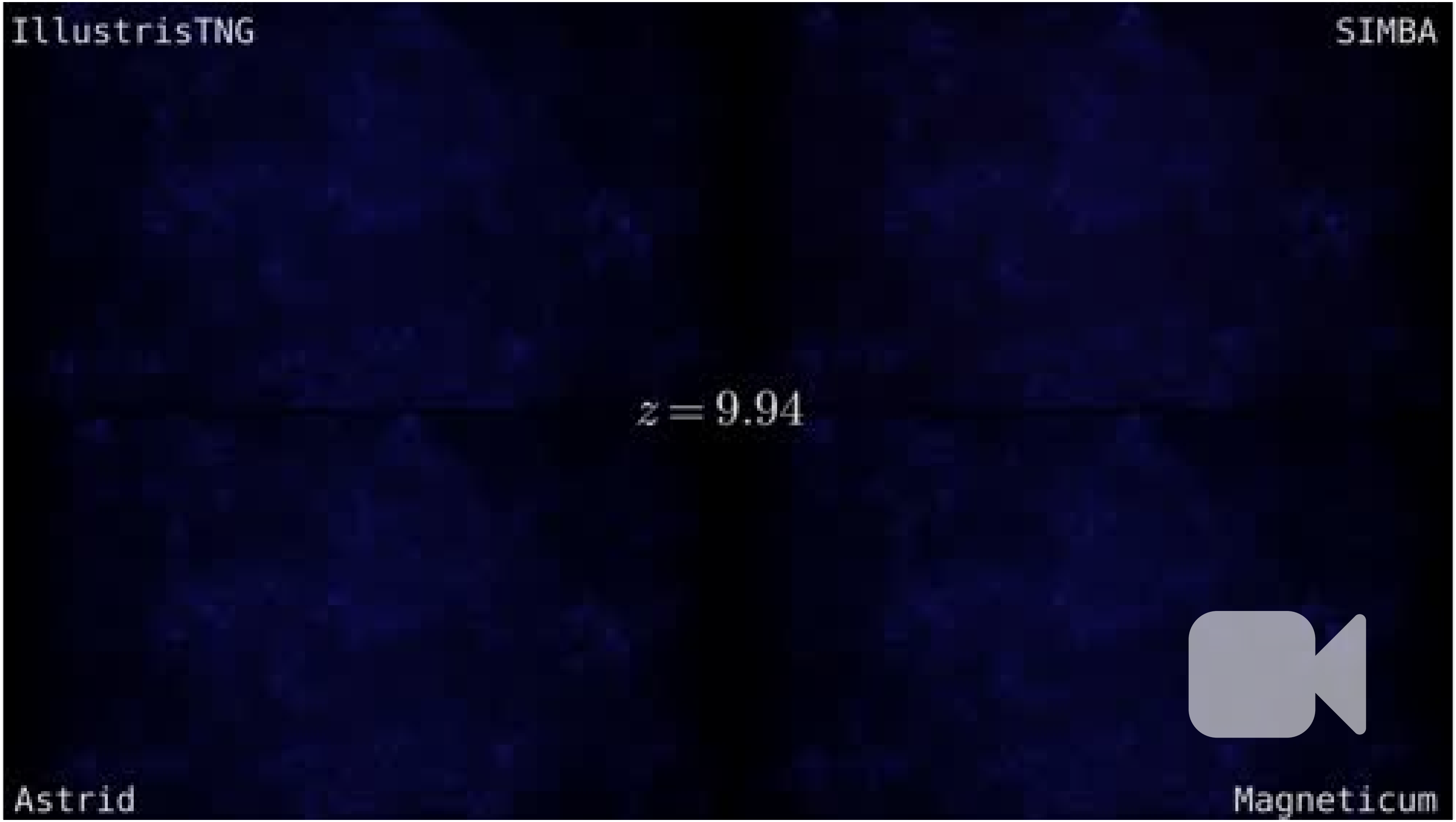
Gravitational potential

Particle velocities




Gpc
Mpc
kpc
pc

Gas density
Gas temperature

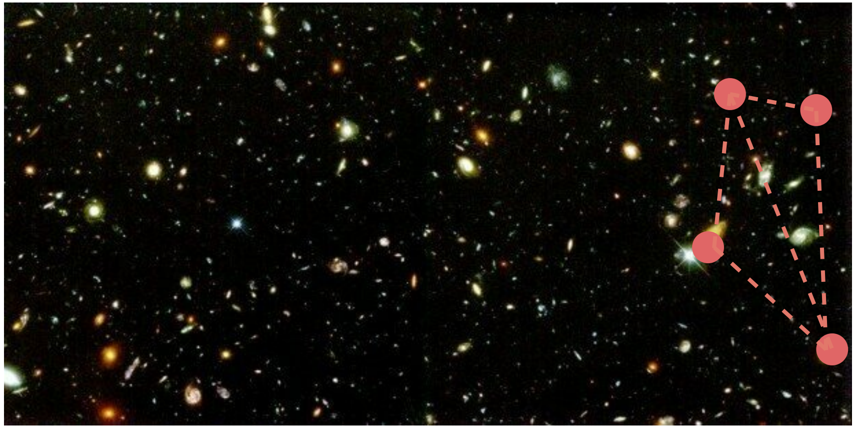


Video credit: Francisco Villaescusa-Navarro

Are there problems in cosmology that bypass a forward model?

x

$\mathrm{Mirror}(x)$



"Could sample variance be responsible for the parity-violating signal seen in the BOSS galaxy survey?"
Philcox, Ereza arXiv:2401.09523

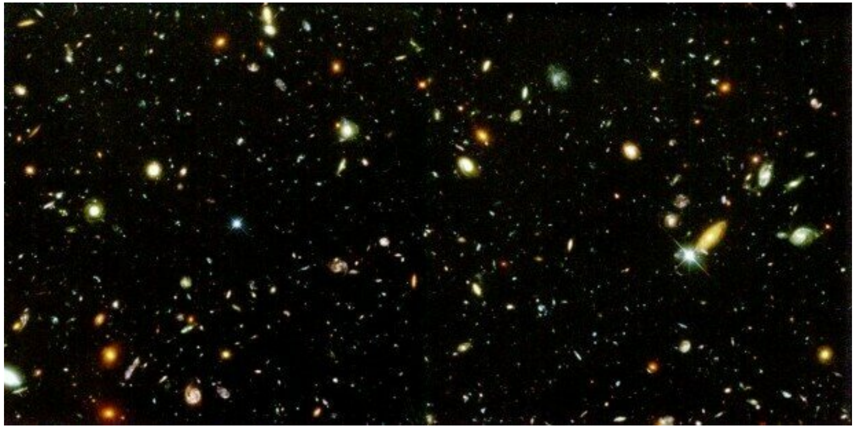
7σ 1σ?

Parity violation cannot be originated by gravity

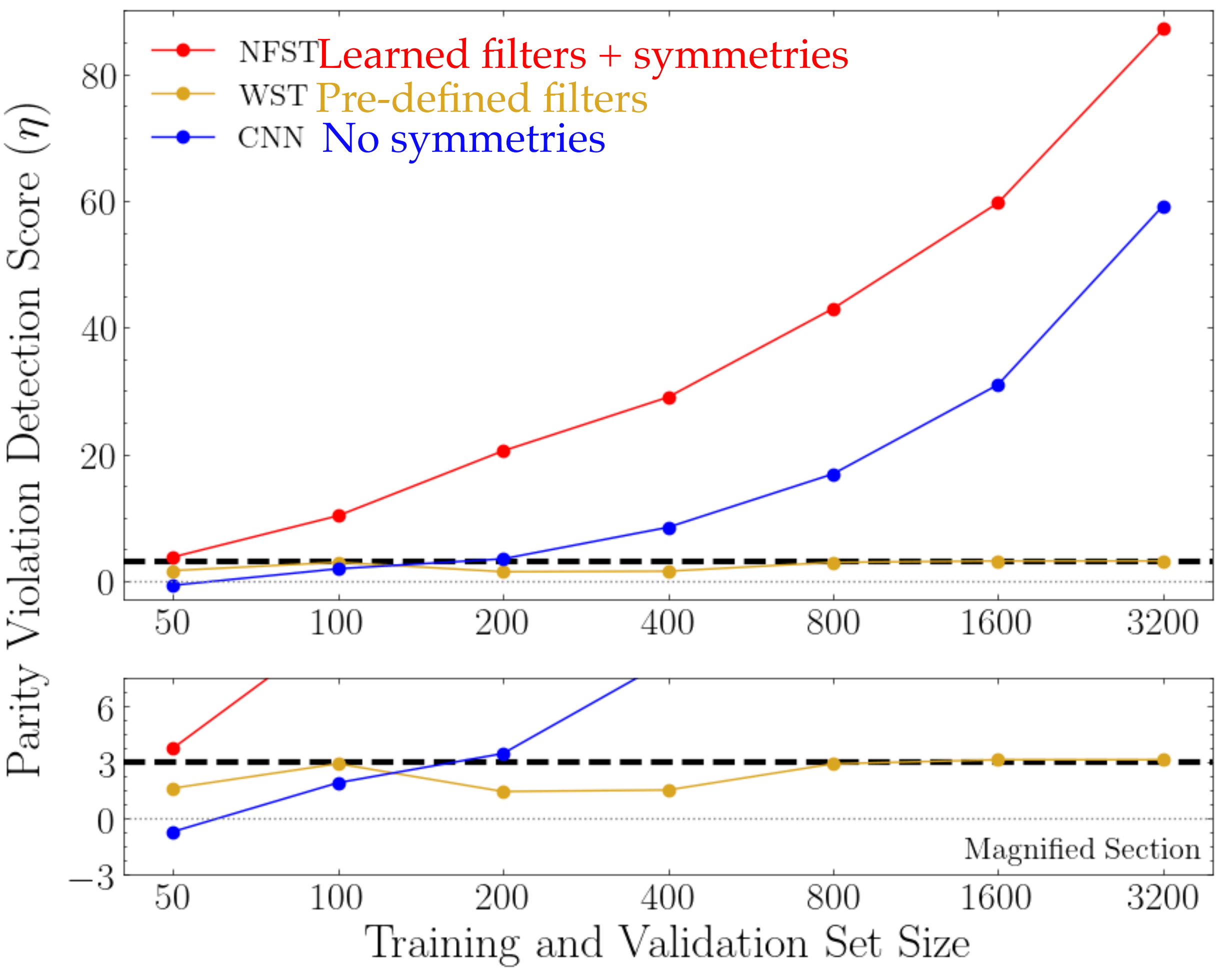
"Measurements of parity-odd modes in the large-scale 4-point function of SDSS..."
Hou, Slepian, Chan arXiv:2206.03625

x

$\mathrm{Mirror}(x)$



$$\max\ (f_\theta(x) - f_\theta(\mathrm{Mirror}(x)))$$

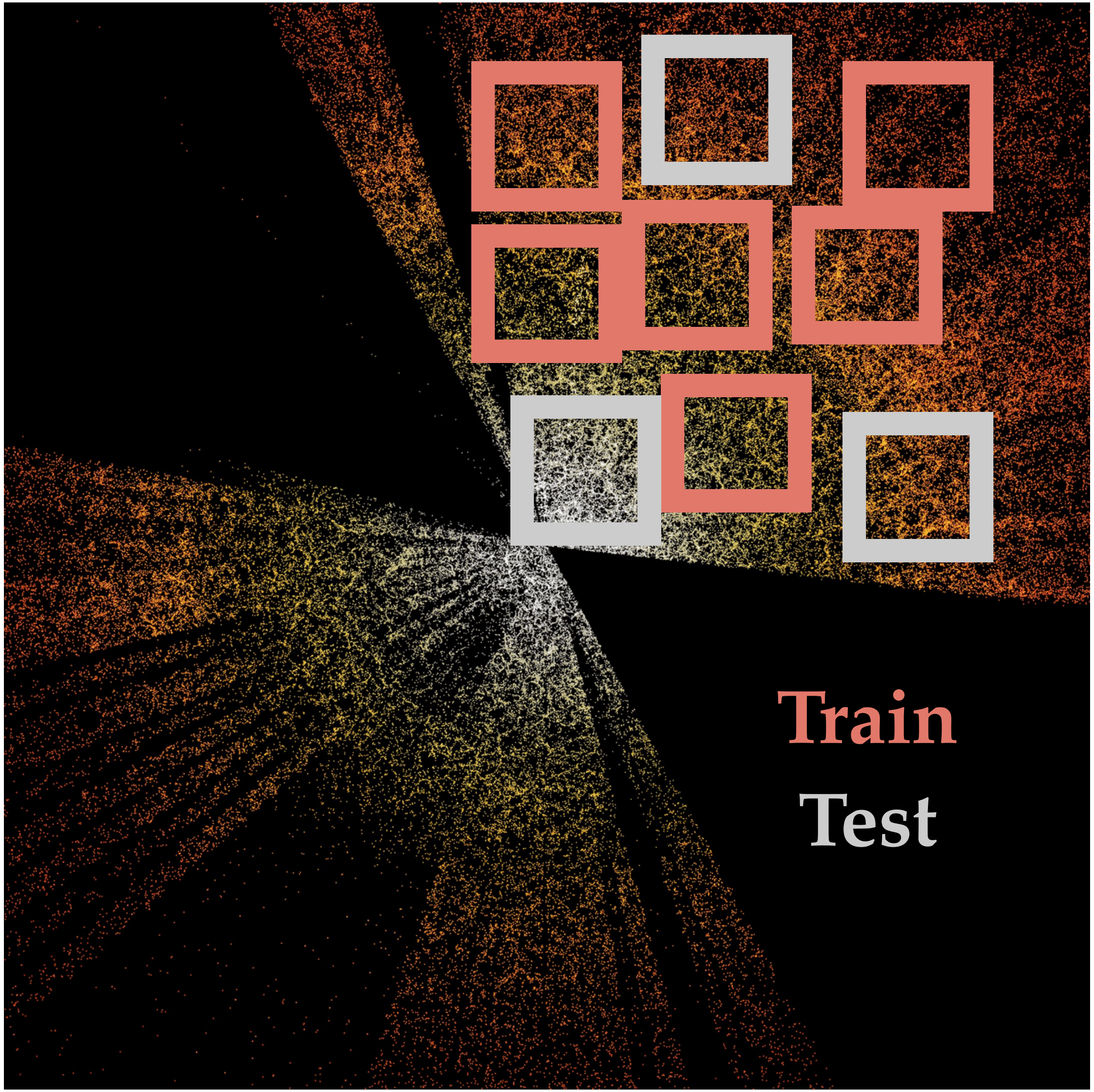


Matthew Craigie

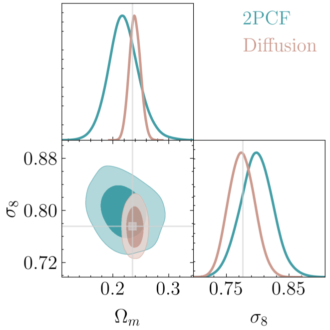
Peter Taylor

Yuan-Sen Ting

Me: I can't wait to work with observations
Me working with observations:

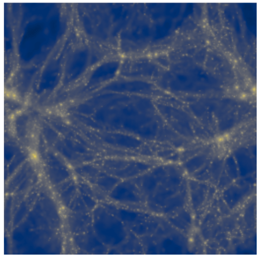


Conclusions



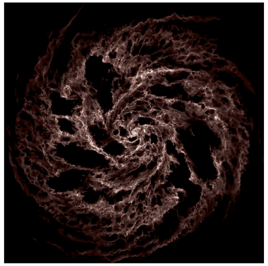
1. There is a lot of information in galaxy surveys that ML methods can access

Field level inference



2. We can tackle high dimensional inference problems so far unatainable

Dark matter density reconstruction, Initial Conditions, let's get creative!



3. Our ability to simulate will limit the amount of information we can extract

Hybrid simulators, forward models, robustness

Unsupervised problems: parity violation