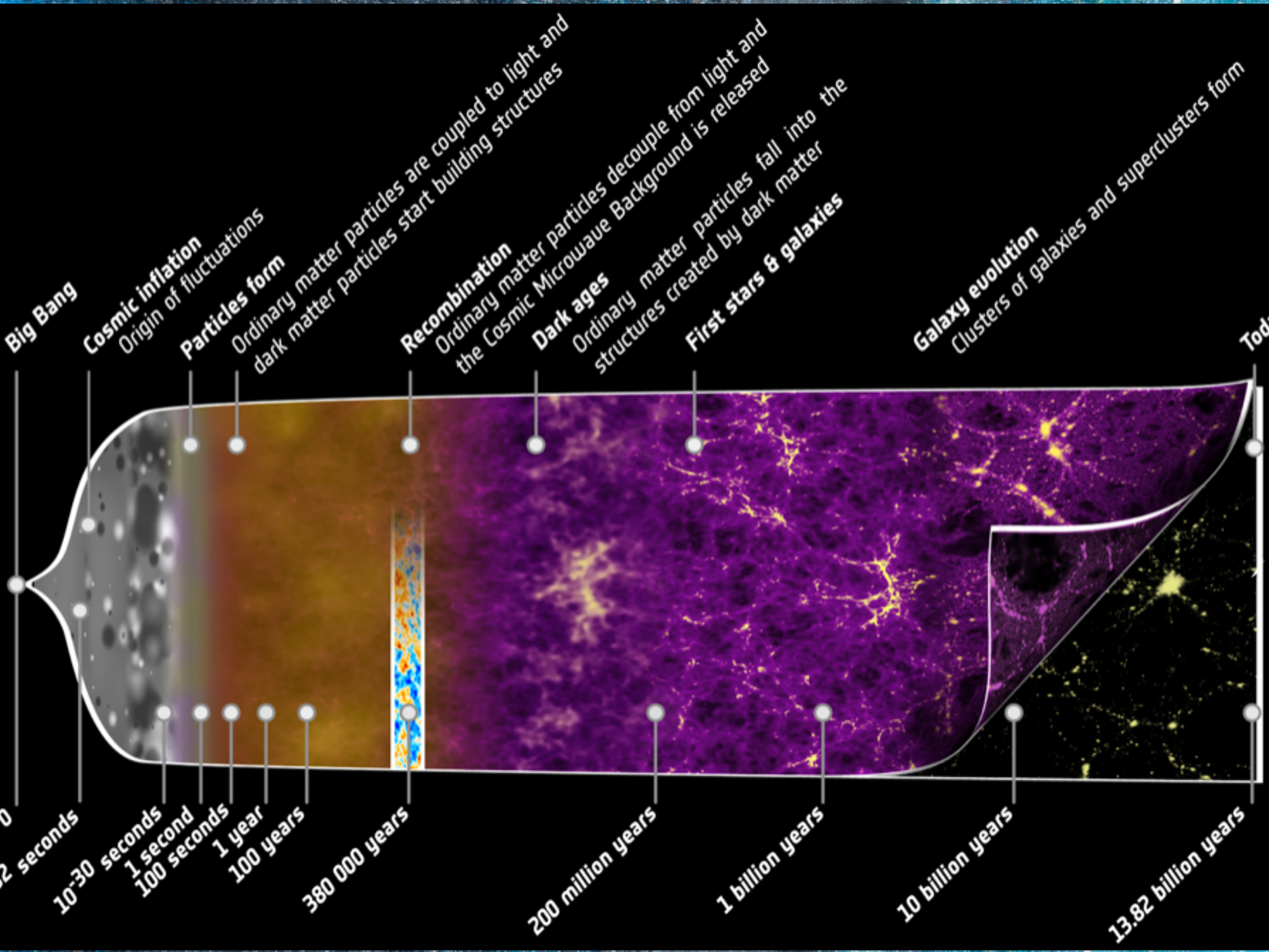
Persistent Homology of Cosmic Structures

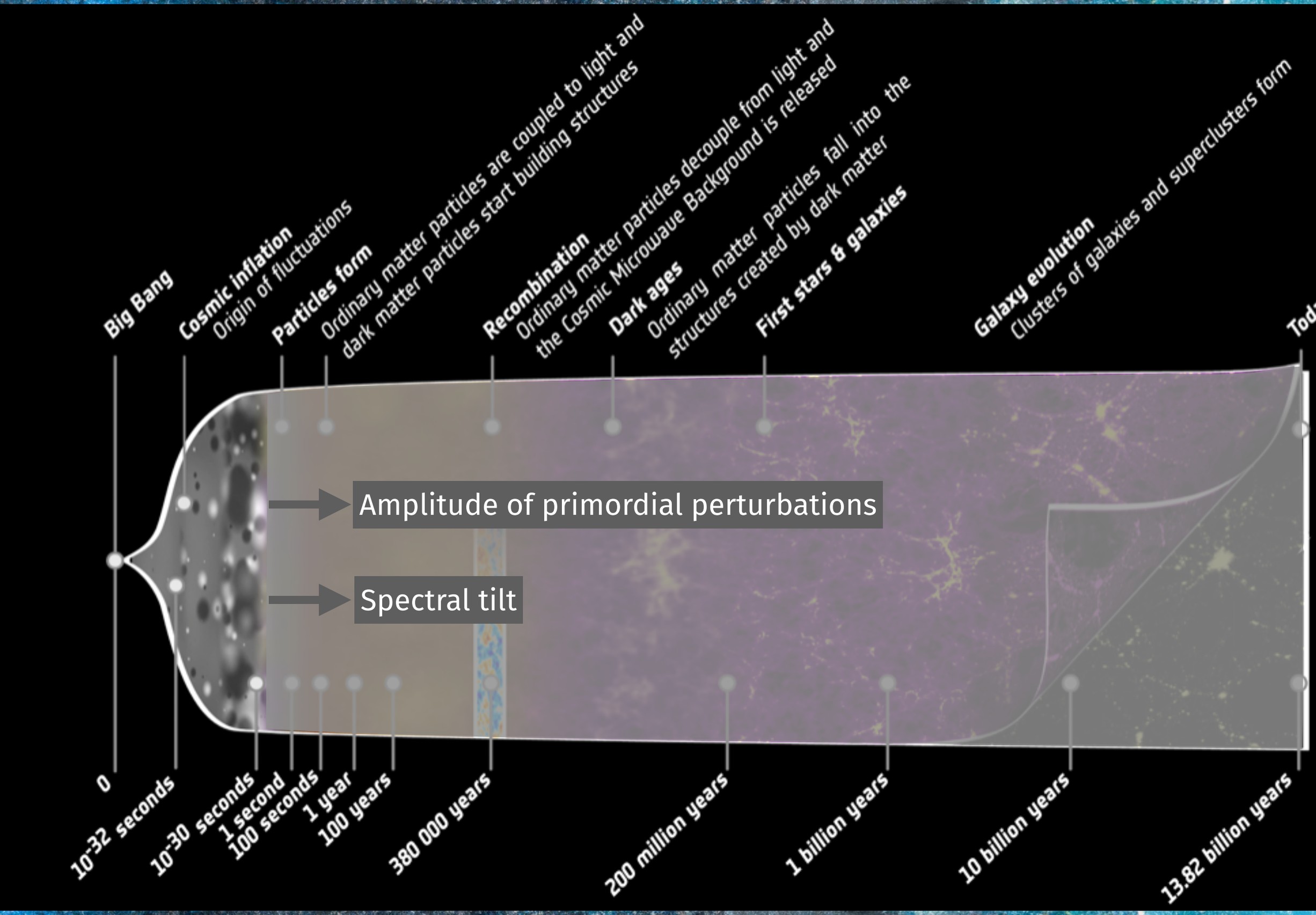
Matteo Biagetti

DataShape Seminar

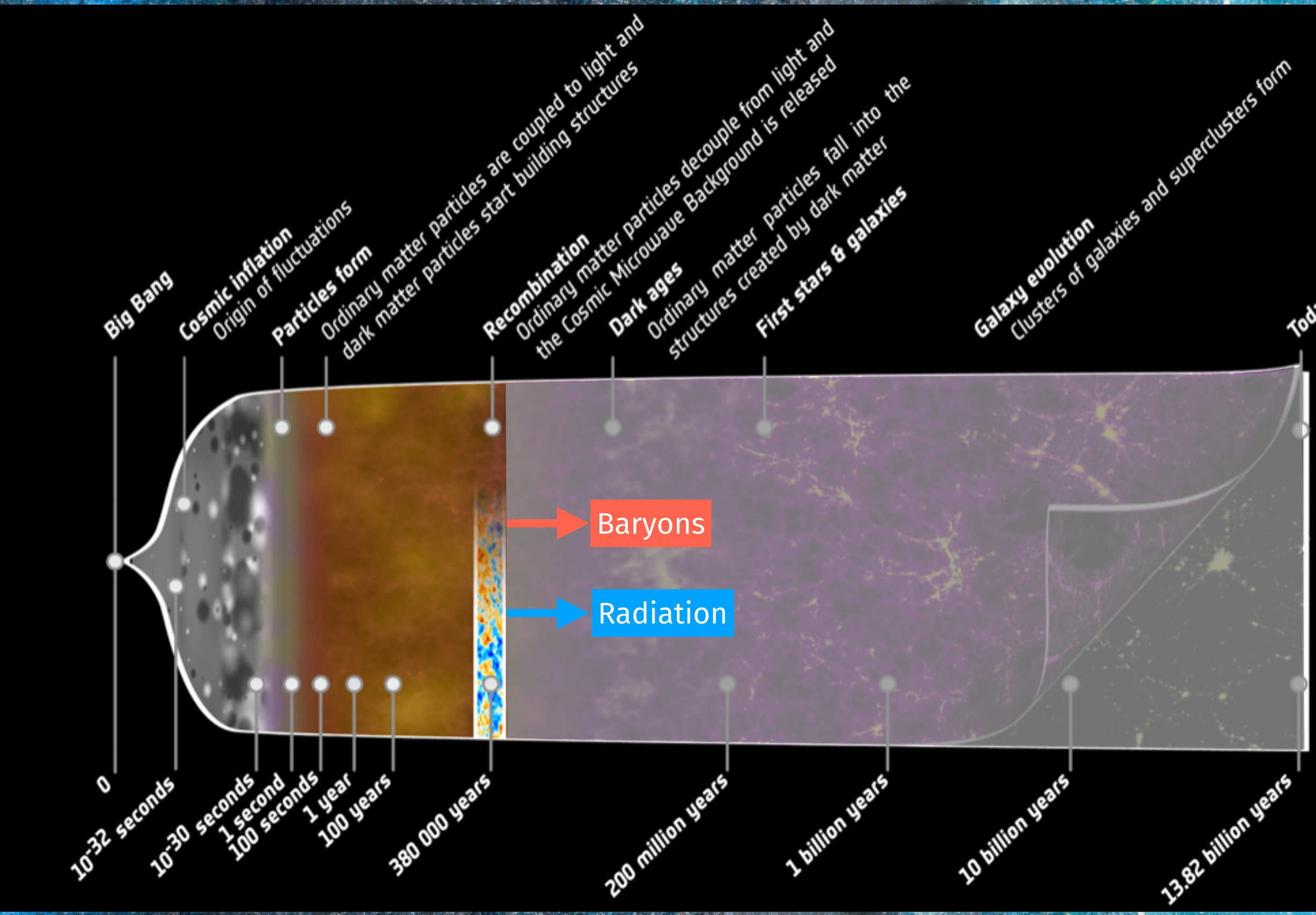
Persistent Homology of Cosmic Structures



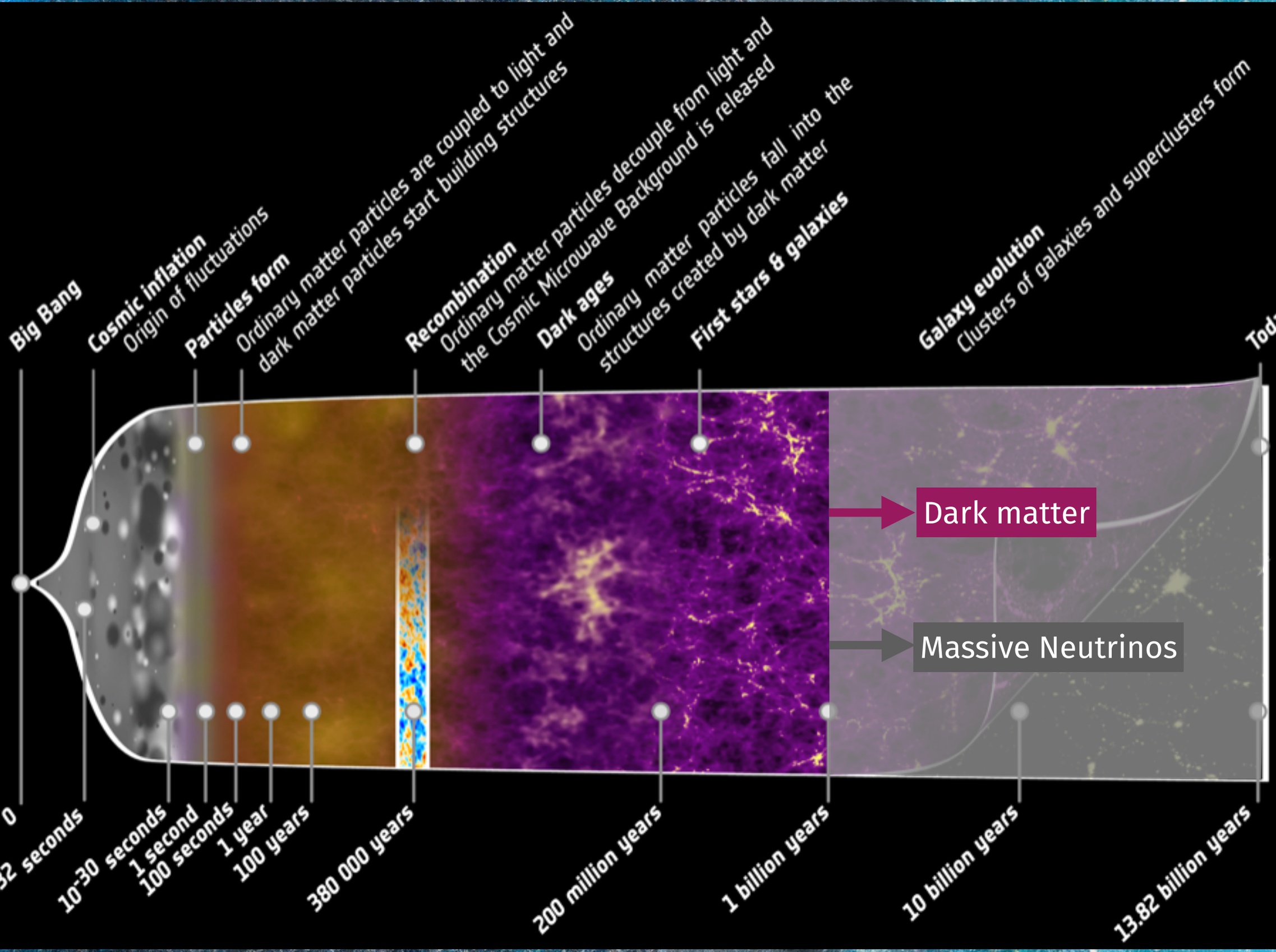
Persistent Homology of Cosmic Structures



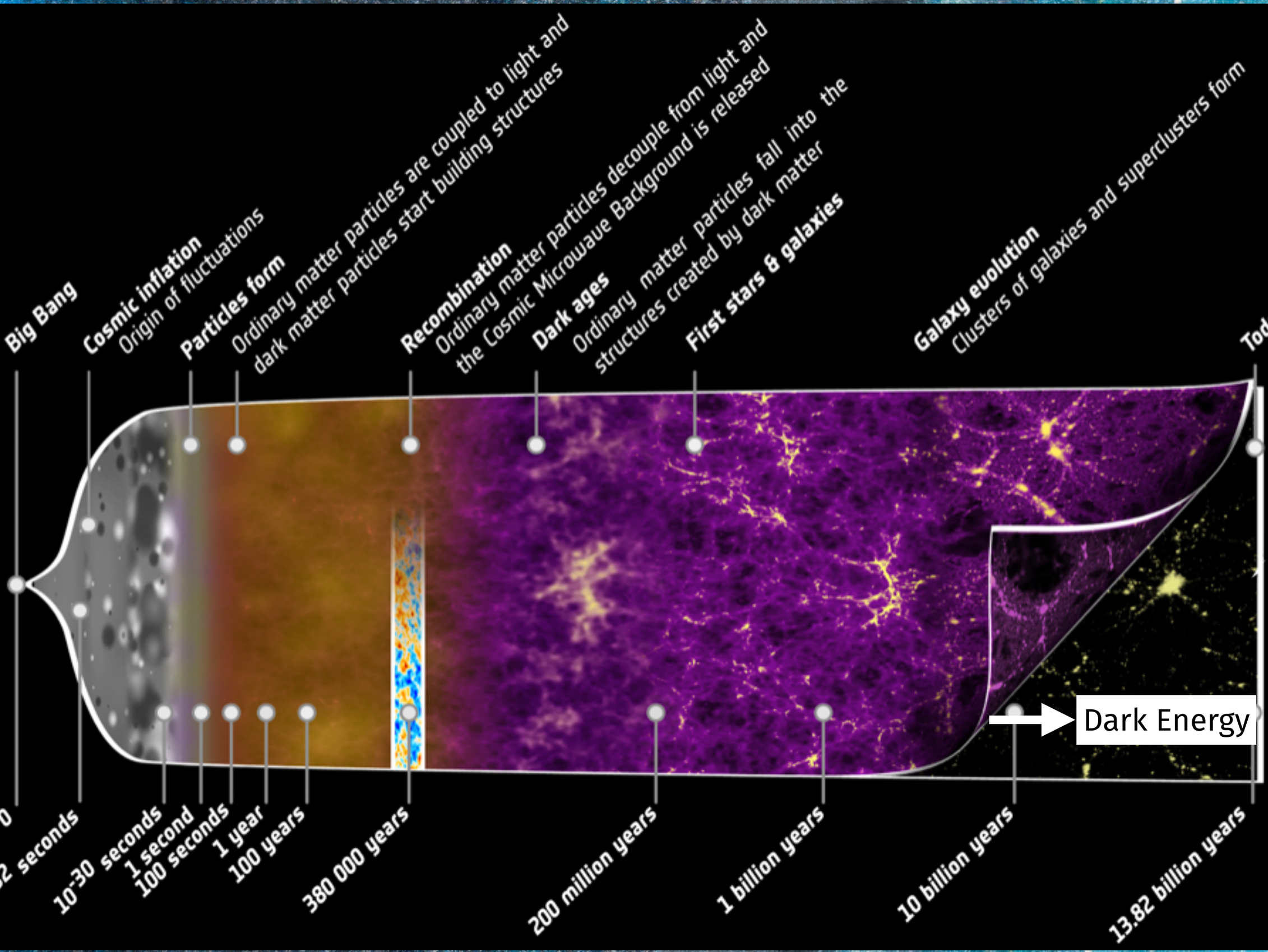
Persistent Homology of Cosmic Structures



Persistent Homology of Cosmic Structures



Persistent Homology of Cosmic Structures



Cosmological Challenges

- **Cosmic Inflation**
  how structures are formed at the origins of the universe
  Field content? Interactions at high energies?
  Primordial gravitational waves?

- **Dark Matter**
  Particle nature? Primordial black holes?
  Cold? Warm?
  Non gravitational interactions?

- **Dark Energy**
  Vacuum energy?
  Modified gravity?

- **Massive Neutrinos**
  Mass detection? Hierarchy?

- **Parameter Tensions**
  Expansion rate? Amplitude of matter fluctuations?

Cosmological Challenges

- Cosmic Inflation
  how structures are formed at the origins of the universe
  Field content? Interactions at high energies?
  Primordial gravitational waves?

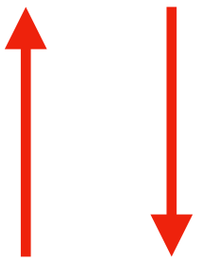
- Dark Matter
  Particle nature? Primordial black holes?
  Cold? Warm?
  Non gravitational interactions?

- Dark Energy
  Vacuum energy?
  Modified gravity?

- Massive Neutrinos
  Mass detection? Hierarchy?

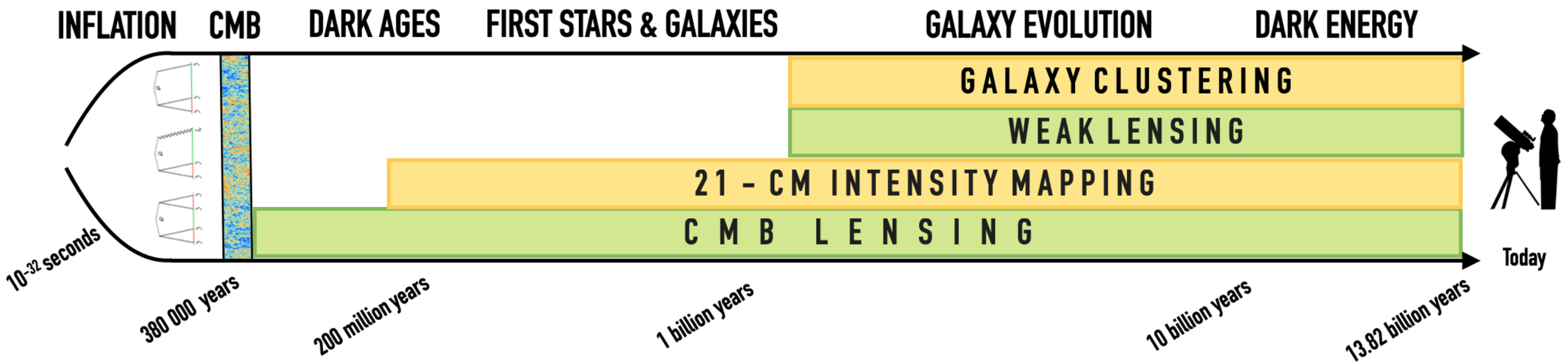
- Parameter Tensions
  Expansion rate? Amplitude of matter fluctuations?

Need more data!



Improve data analysis

Cosmological Observables



CMB	Statistical distribution of temperature anisotropies
Galaxy clustering	Statistical distribution of observed galaxies
Weak and CMB lensing	Light deflection from matter distribution
21-cm intensity mapping	Mapping the hydrogen distribution

Outline of the problem to solve

Given:

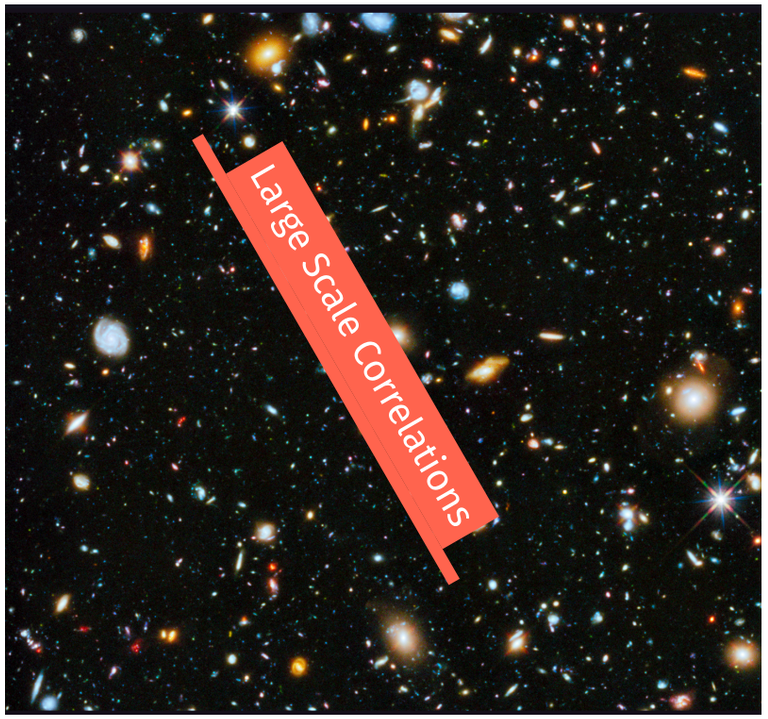
- a set of points (galaxies) on a three-dimensional volume at some final time t_f
- a set of assumptions on statistical properties at initial time t_0
- a model describing the evolution from t_0 to t_f as a function of a few (6) interesting parameters and several nuisance parameters

Given an observation at t_f, we would like to ***infer the most likely values*** of interesting parameters with a confidence interval (uncertainty)

Cosmology from galaxy surveys

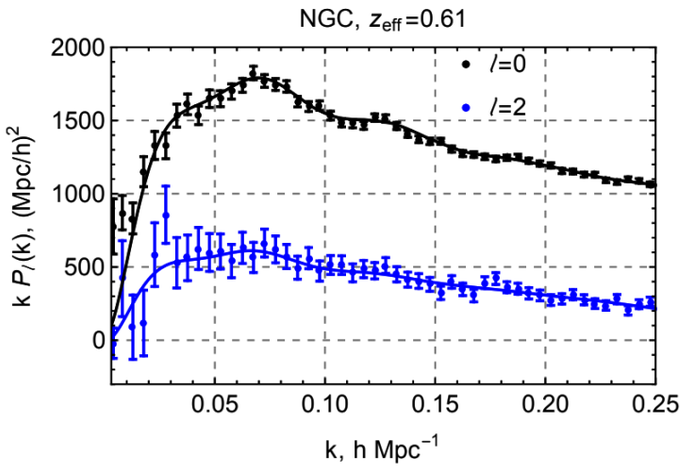
“Conventional” methodology

Observation: 10^7 galaxies



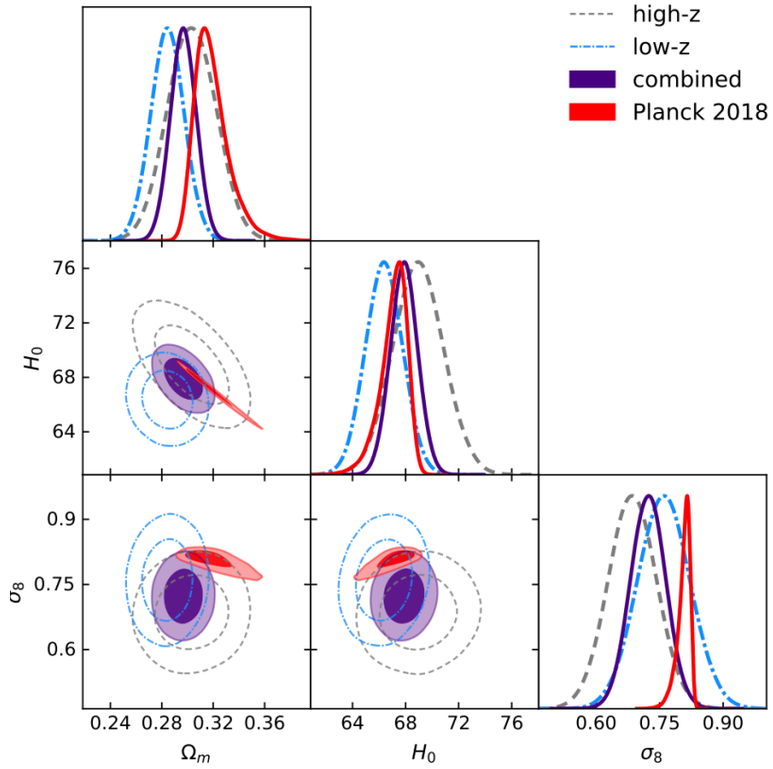
(Discrete points in a 3D volume)

Method: perturbative models



Measure correlation of galaxy pairs and ***compare to theoretical model*** based on matter dynamics, galaxy distribution, etc.

Goal: parameter inference



Cosmological parameters:
amount of matter in the Universe,
expansion rate,
amplitude of matter fluctuations,
…

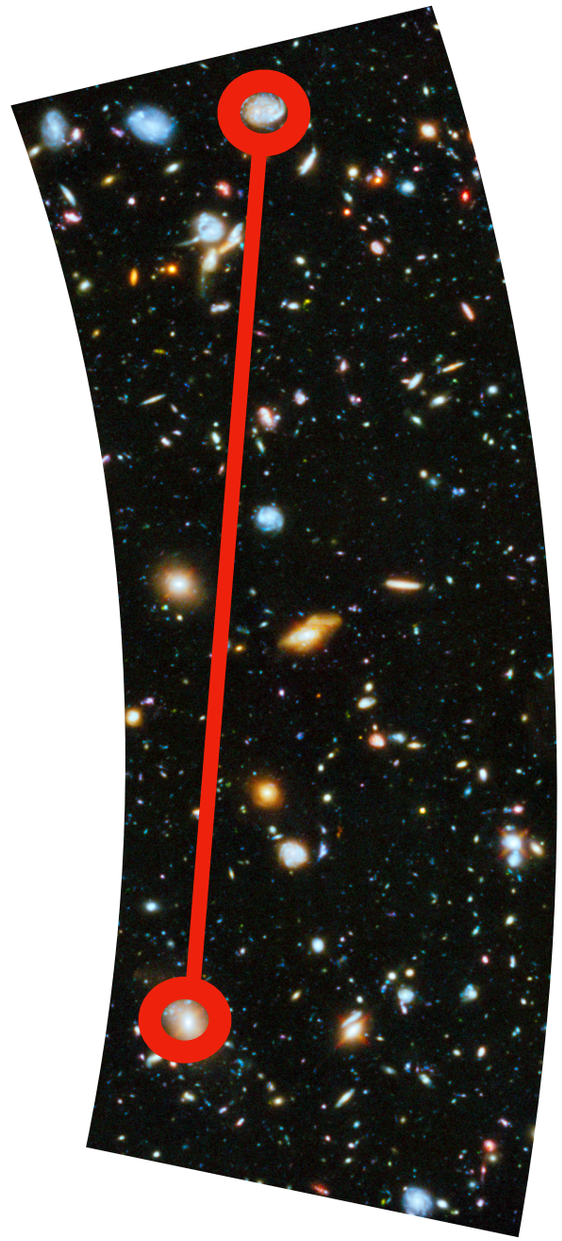
Cosmology from galaxy surveys

Compute the probability of finding two galaxies in volume elements dV_1 and dV_2

$$dP = dV_1 dV_2 \langle n_g(x_1) n_g(x_2) \rangle$$

$$= dV_1 dV_2 \, \bar{n}_g^2 \left[1 + \langle \delta_g(x_1) \delta_g(x_2) \rangle \right]$$

↑ Excess probability



Assumption: statistical homogeneity and isotropy

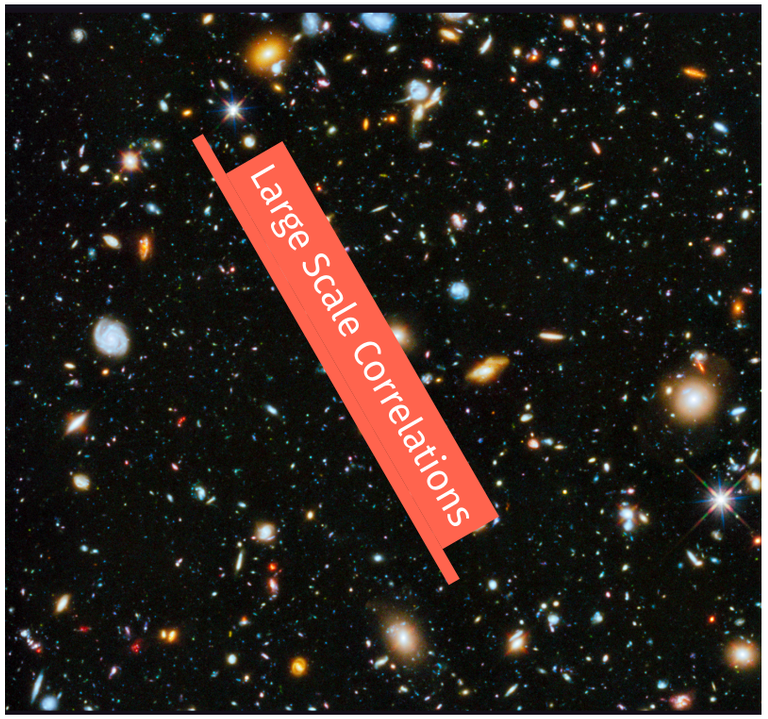
$$\xi_g(|\vec{x}_1 - \vec{x}_2|) \equiv \langle \delta_g(\vec{x}_1) \delta_g(\vec{x}_2) \rangle$$

Two-point function only depends on the distance $r = |\vec{x}_1 - \vec{x}_2|$ between two points

Cosmology from galaxy surveys

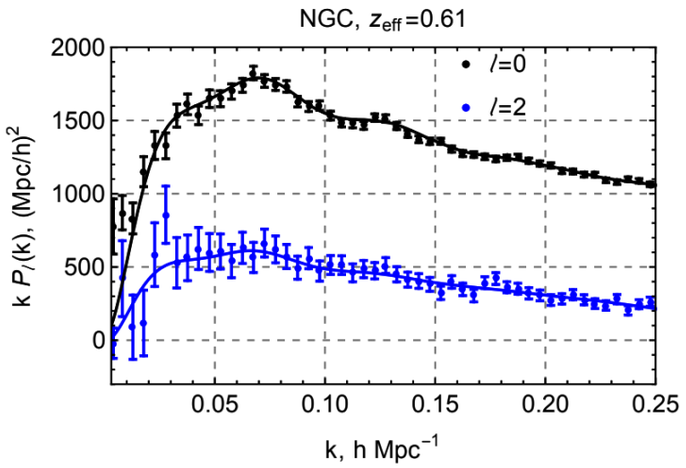
"Conventional" methodology

Observation: 10^7 galaxies



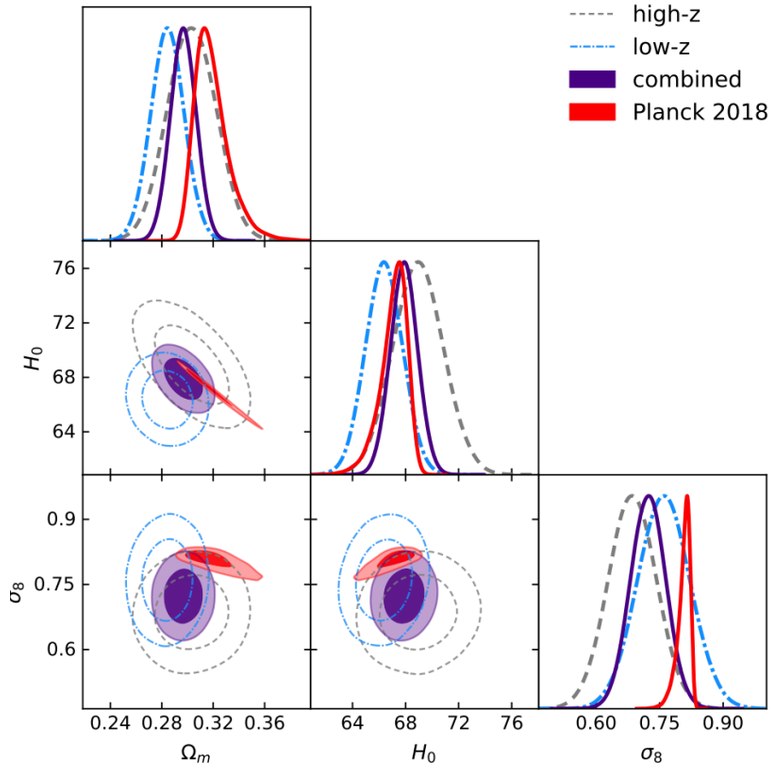
(Discrete points in a 3D volume)

Method: perturbative models



Measure correlation of galaxy pairs and ***compare to theoretical model*** based on matter dynamics, galaxy distribution, etc.

Goal: parameter inference



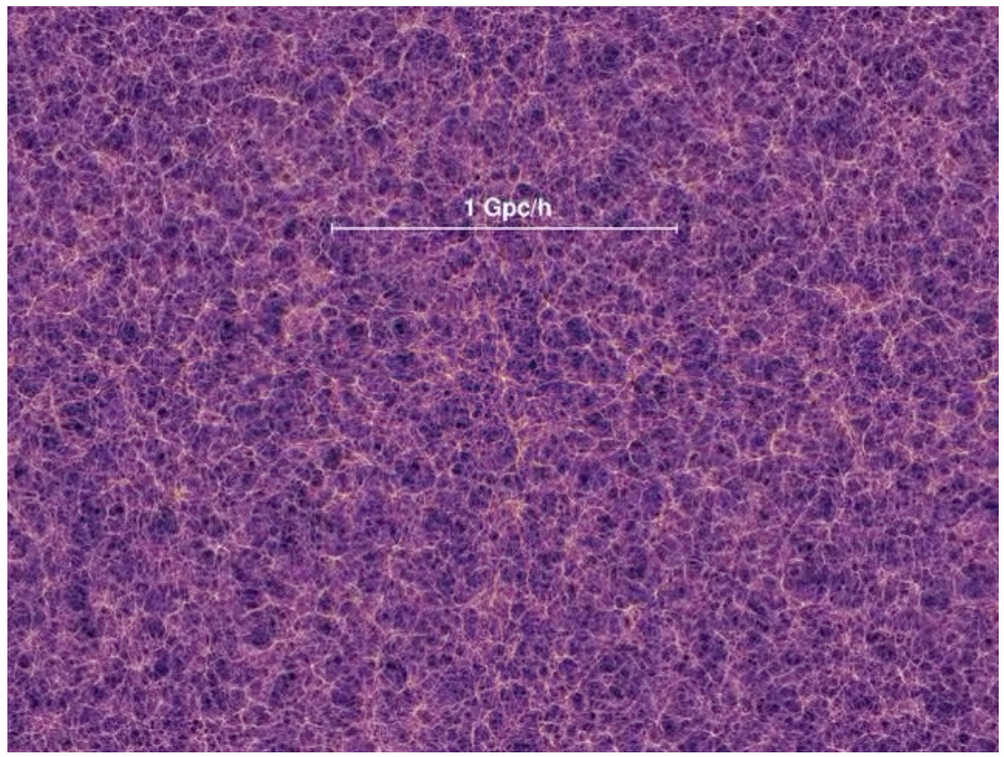
Cosmological parameters:
amount of matter in the Universe,
expansion rate,
amplitude of matter fluctuations,
...

Extreme compression: is it the right one?

Cosmology from galaxy surveys

Extreme compression: is it the right one? NO!

1) A non-Gaussian field is not well described only with two-point statistics

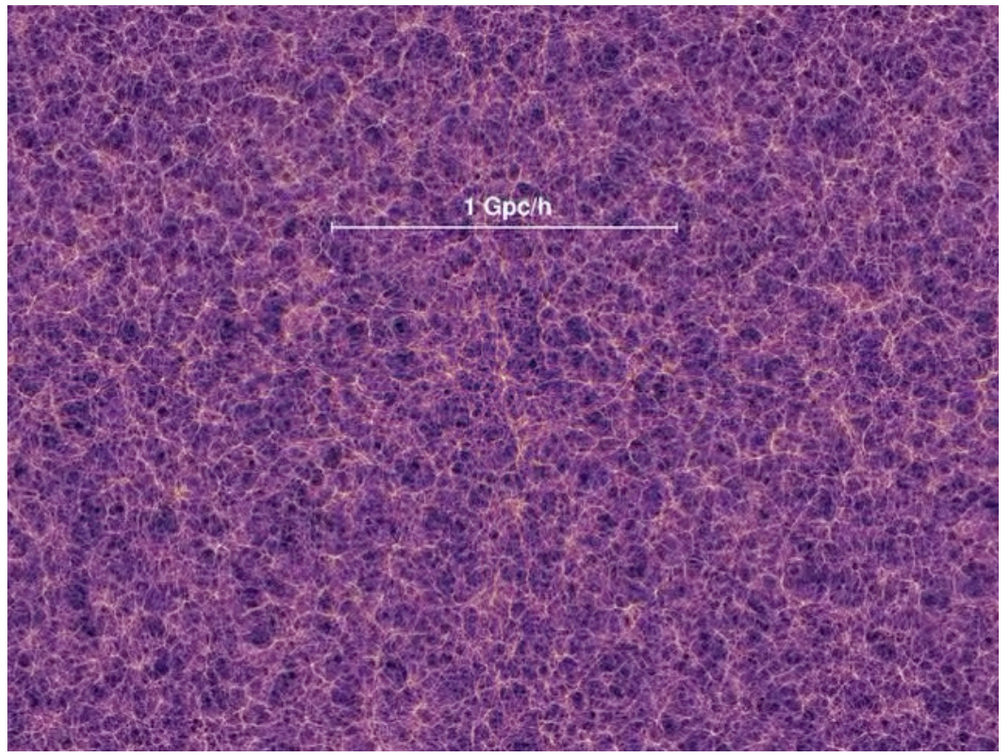


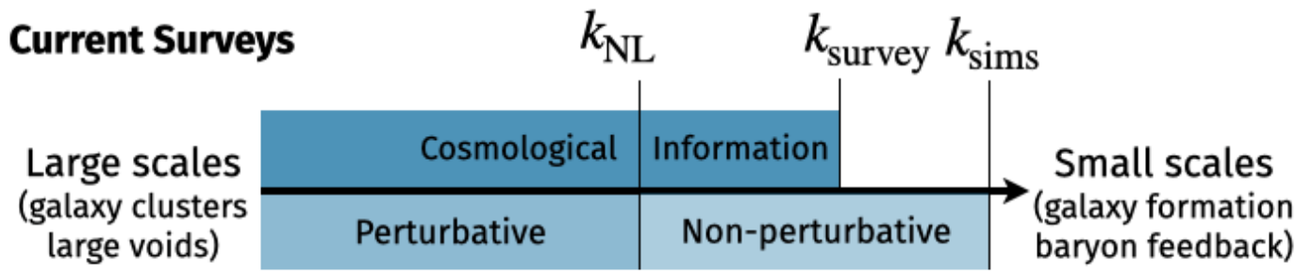
Cosmology from galaxy surveys

Extreme compression: is it the right one? NO!

1) A non-Gaussian field is not well described only with two-point statistics

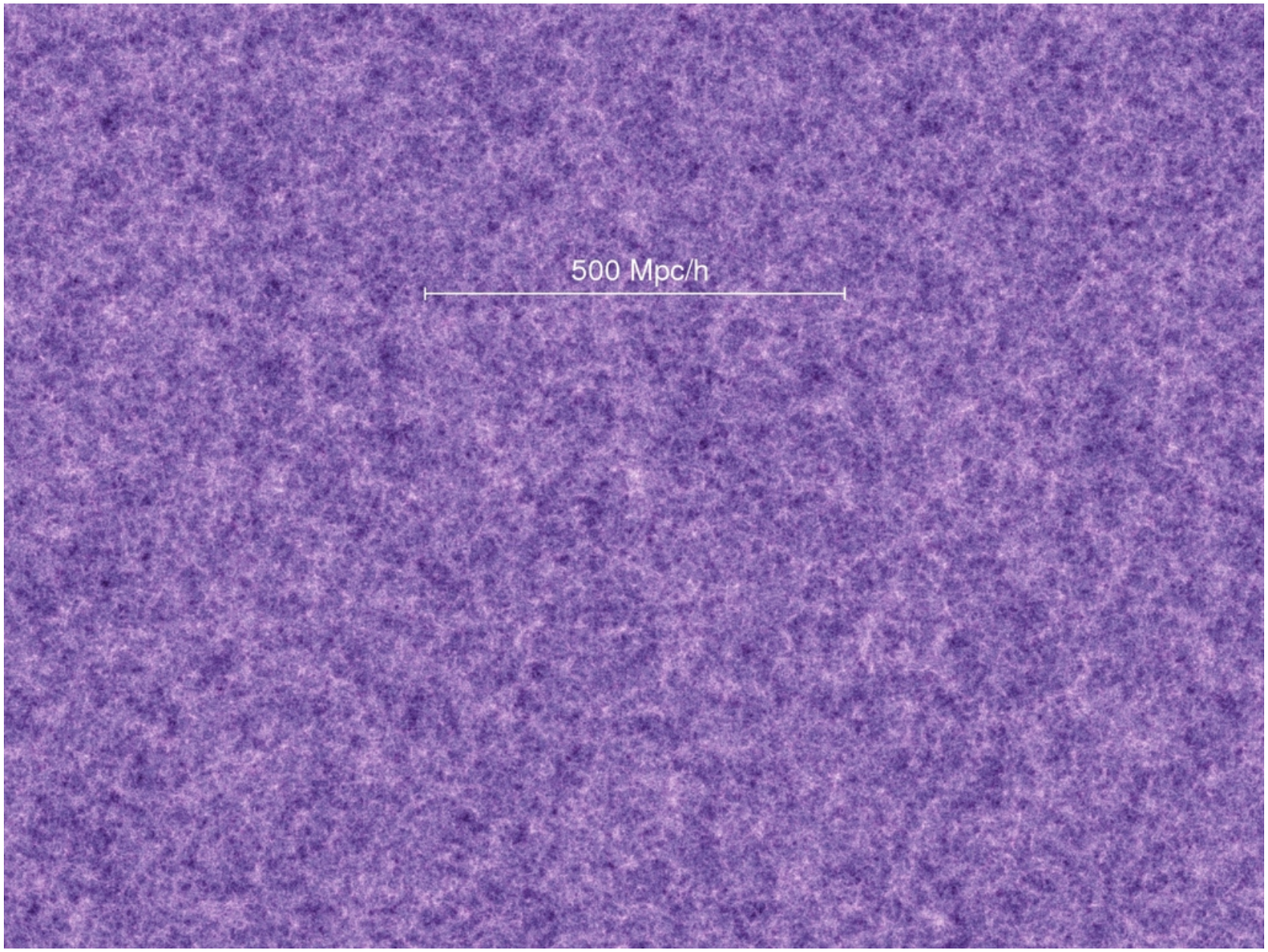
2) Information loss: observational data will improve faster than theoretical model





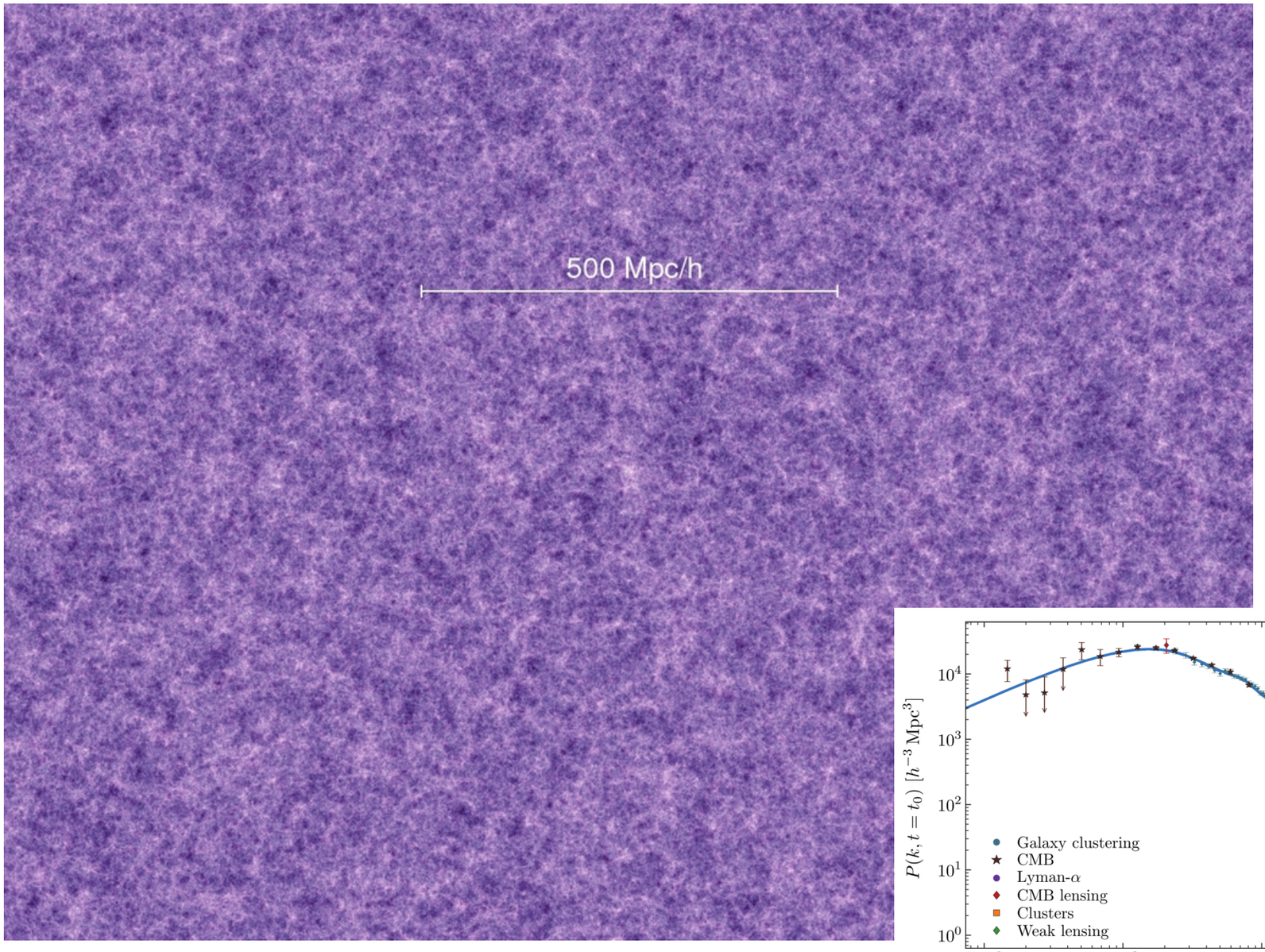
Initial conditions: Gaussian random field

t_0



Initial conditions: Gaussian random field

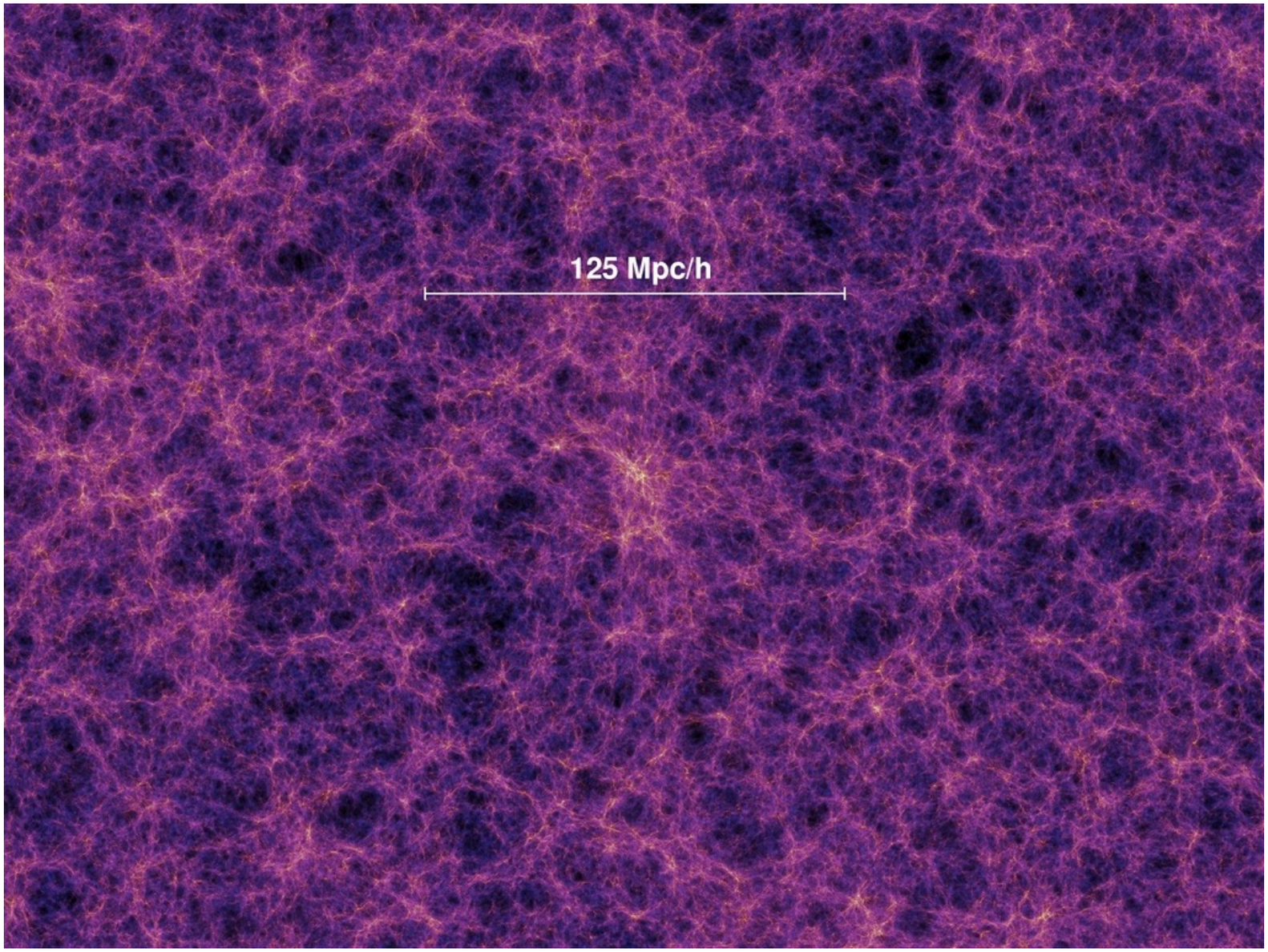
t_0



Power spectrum fully determines GRF: contains interesting parameters

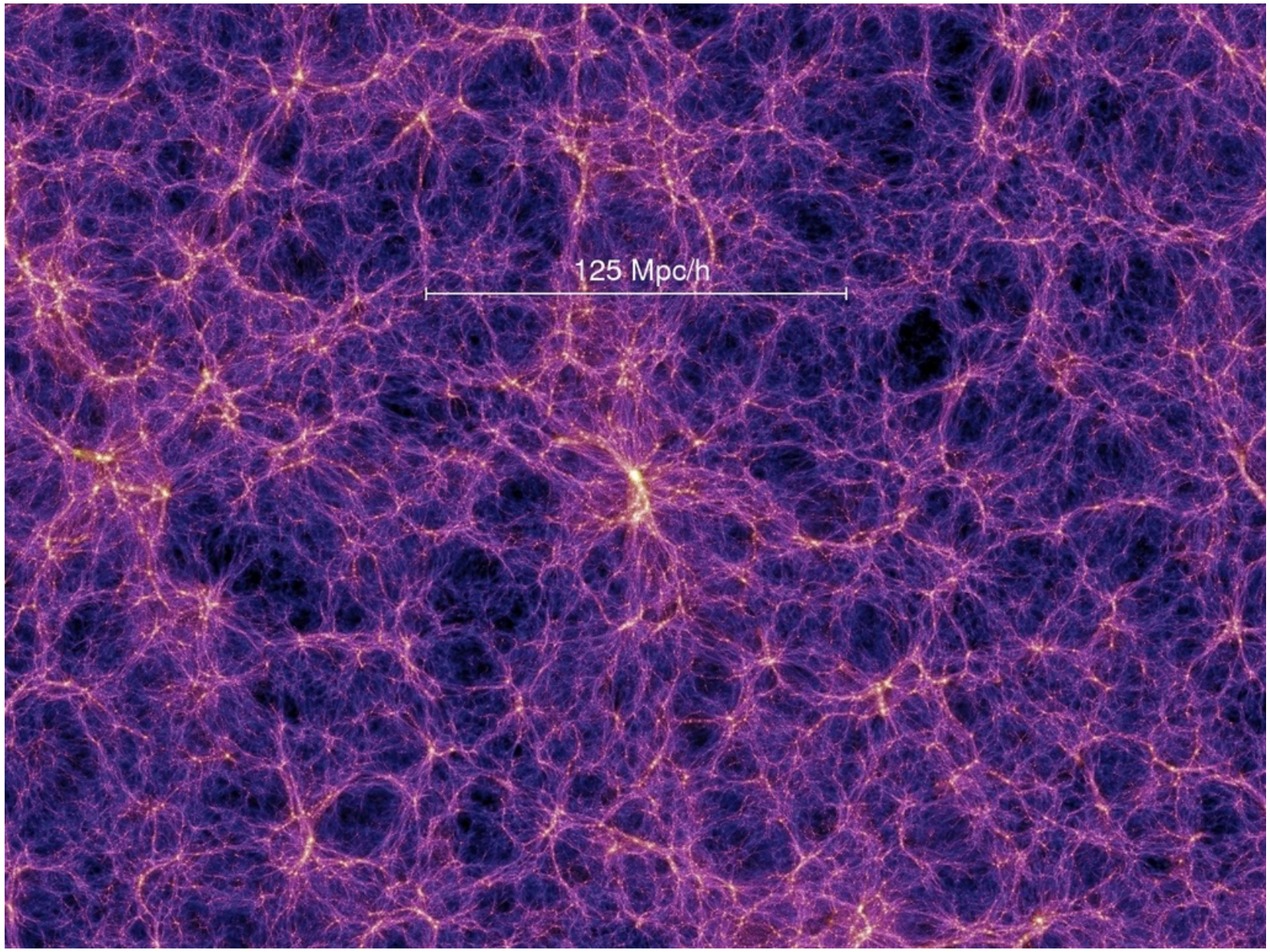
Evolution (gravity): field becomes gradually non-Gaussian

t_1



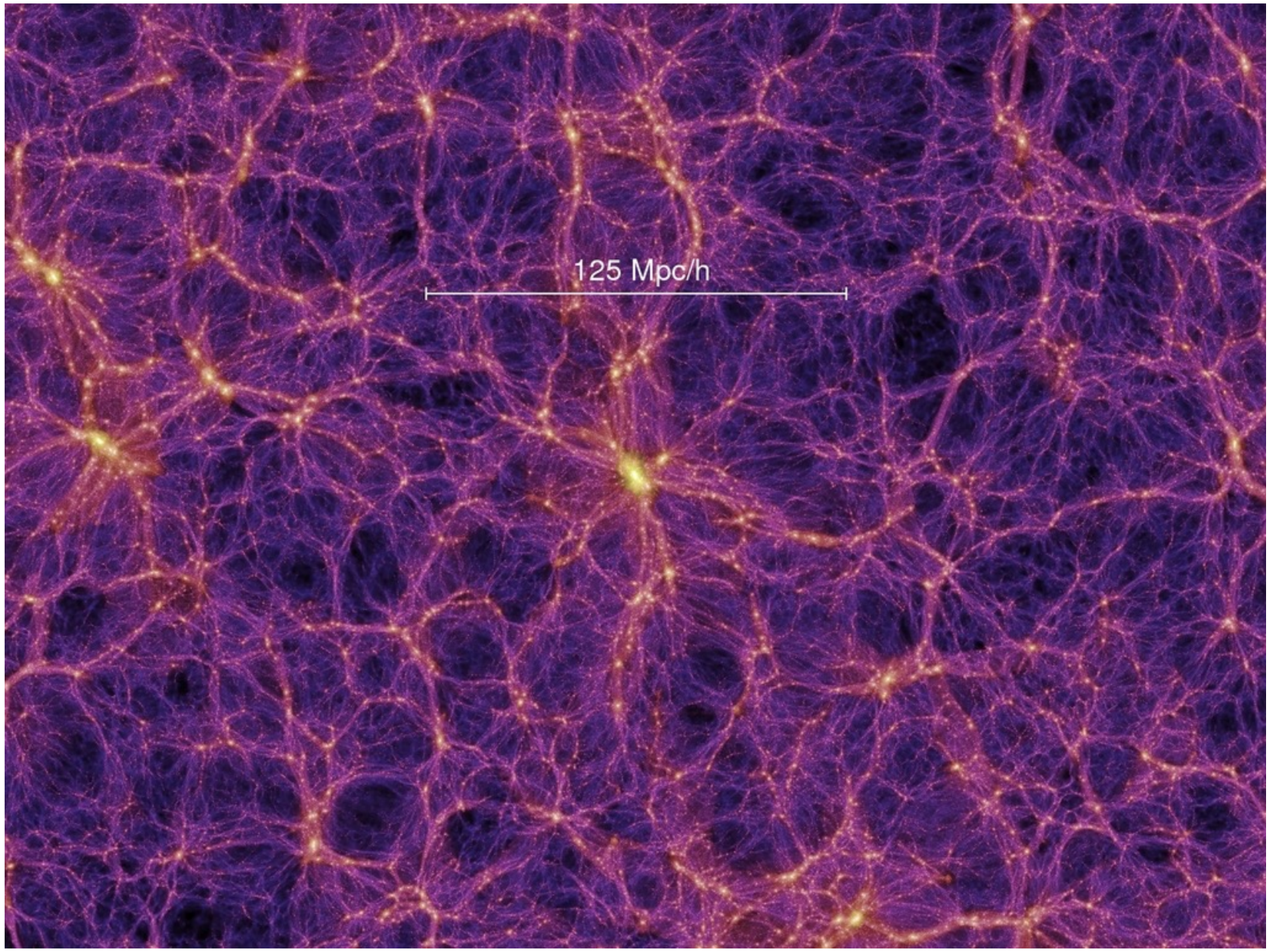
Evolution (gravity): field becomes gradually non-Gaussian

t_2



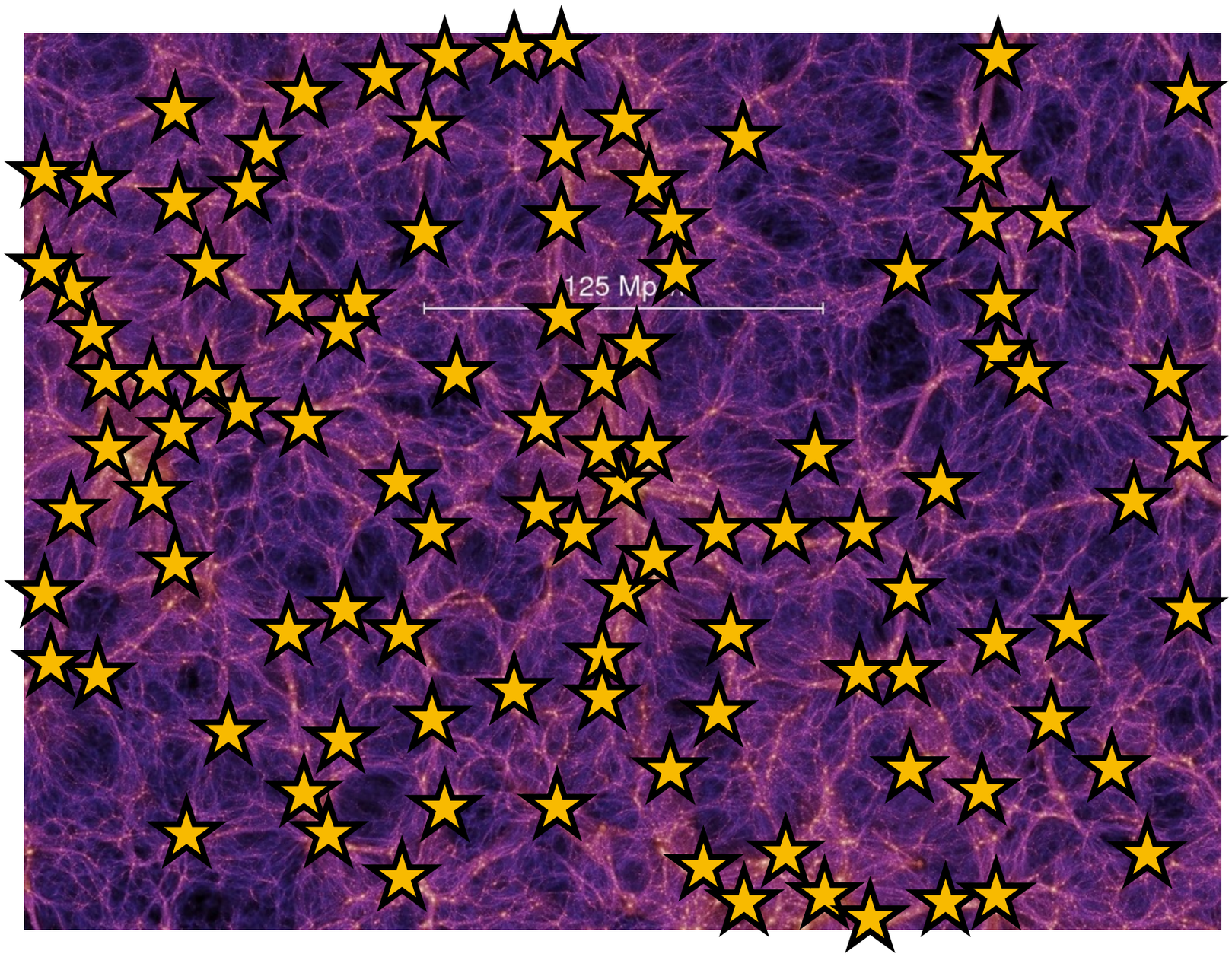
Evolution (gravity): field becomes gradually non-Gaussian

t_f



Evolution (gravity): field becomes gradually non-Gaussian

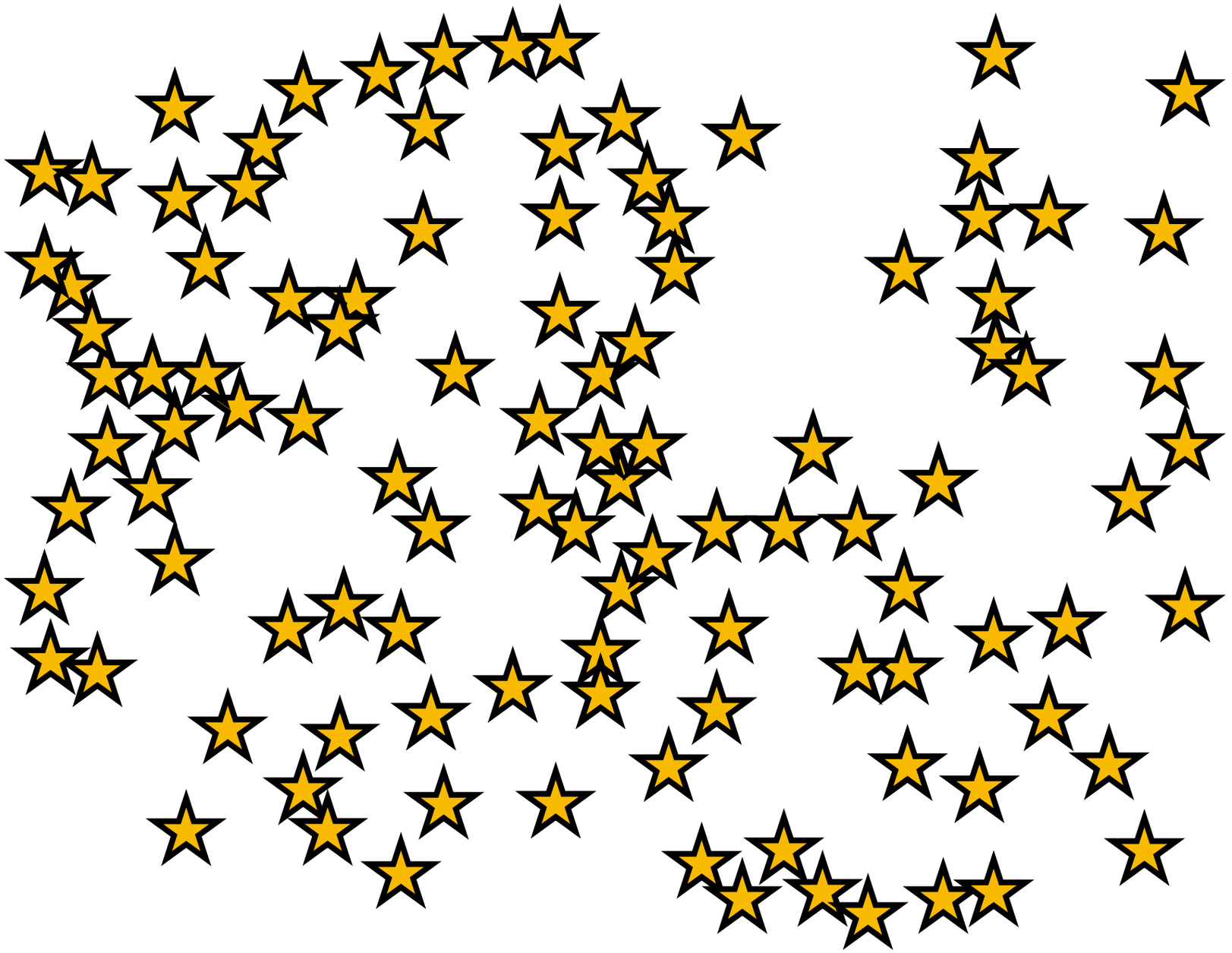
t_f



We only observe galaxies: a discrete (biased) sample of the matter field

Evolution (gravity): field becomes gradually non-Gaussian

t_f



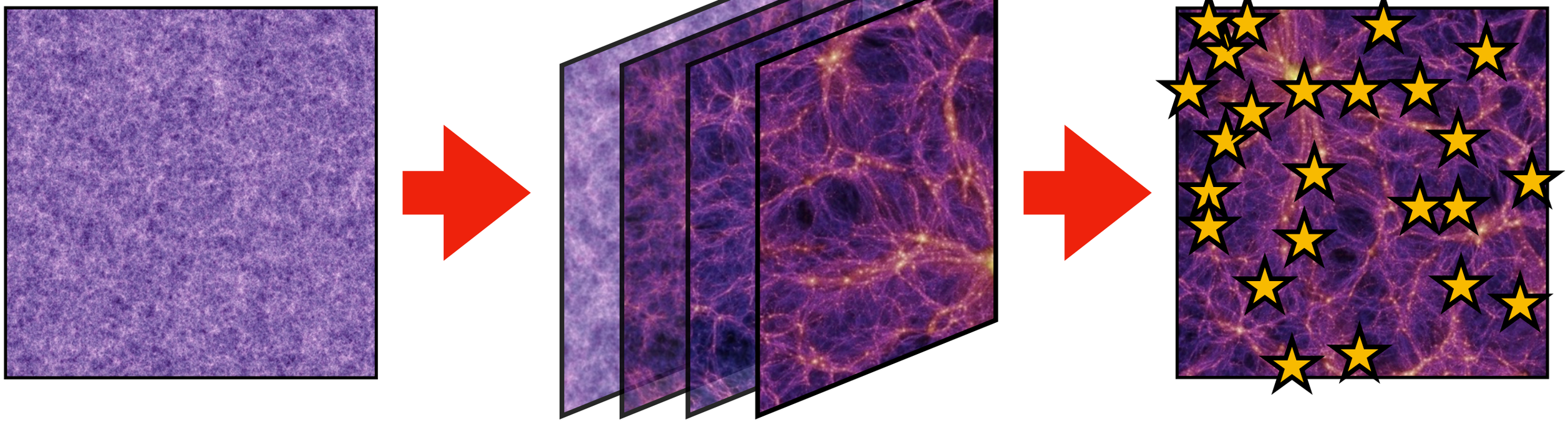
We only observe galaxies: a discrete (biased) sample of this gaussian field

Summary: cosmic evolution

Initial conditions
t_0

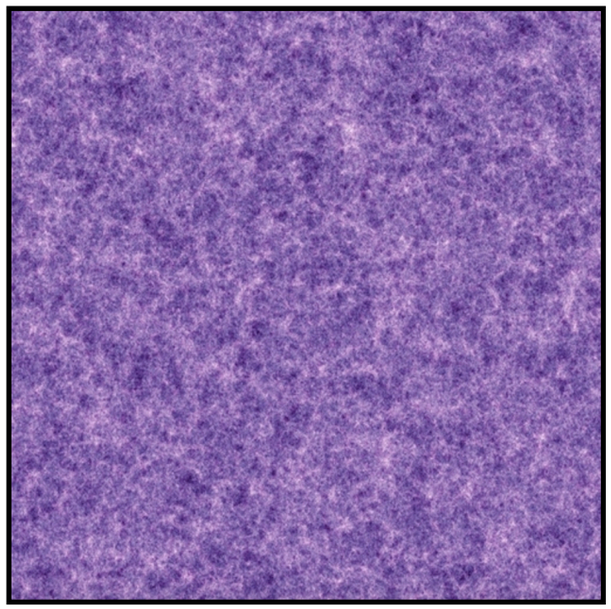
Evolution

Late-time
t_f

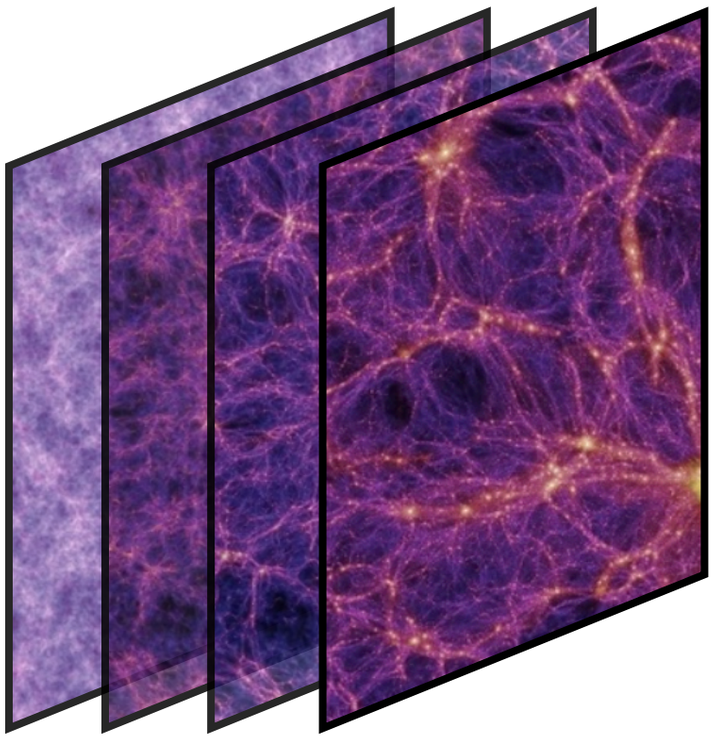


Summary: inference

Initial conditions
t_0



Evolution

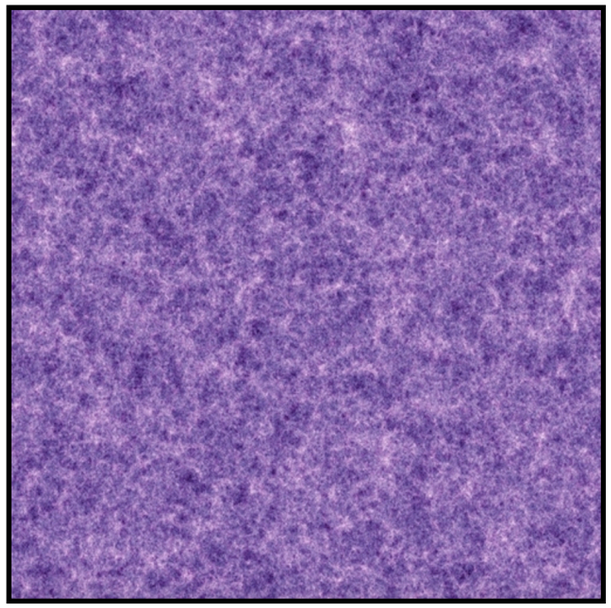


Late-time
t_f

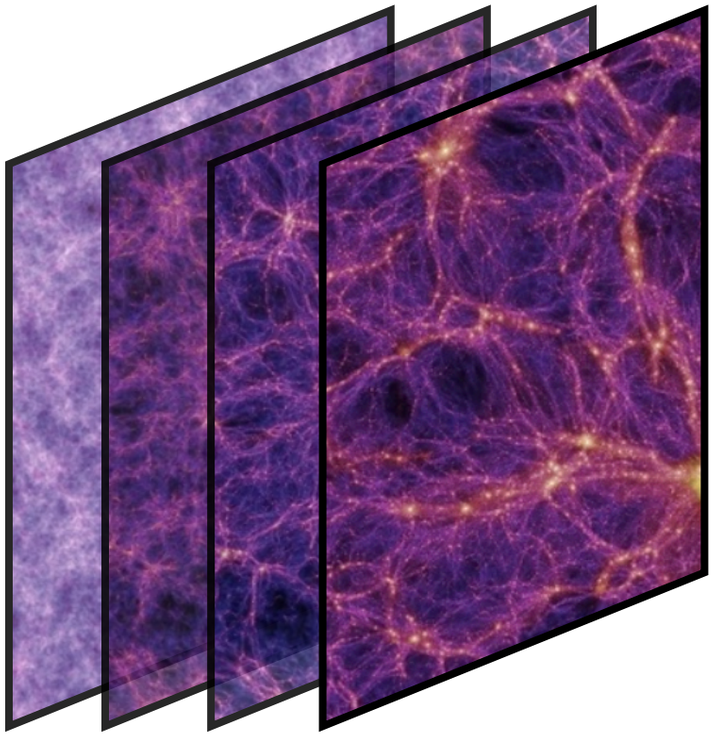
"Conventional strategy": compute excess probability of finding two (three) galaxies at a given distance (configuration) and compare to model

Summary: inference

Initial conditions
t_0



Evolution



Late-time
t_f

Why not directly using numerical simulations?

Flagship 2 simulations

U.Zurich (Stadel et al.)

- 2 different Nbody simulations:
  - **WIDE**: 10^9 Msun resolution (4.1 trillion particles, 3600 Mpc/h box)
    - x2 improvement in mass res wrt FS1
  - **DEEP**: 10^8 Msun resolution (0.9 trillion particles, 1000 Mpc/h box)

WIDE run compute time 800.000 node hours

- some compromises taken (no merger-trees)
- GR effects (metric), radiation and mass. neutrino linear pert. per timestep
- **all-sky 3D light cone up to z=3**, healpix DM counts, Rockstar halos
- data volume (~1 PB): 0.6 PB (LC only) + 5 snapshots (112 TB each)
- Status: completed

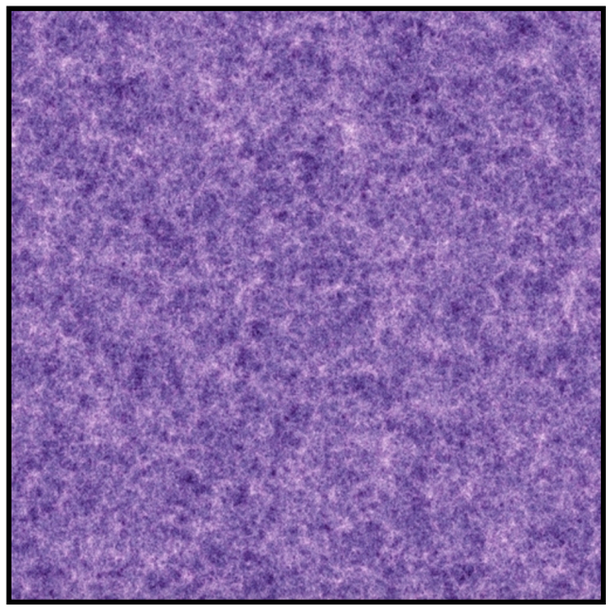
⇨ **Input for FS2.0 WIDE mock** I will be talking about…

DEEP run estimated compute time 200.000 node hours

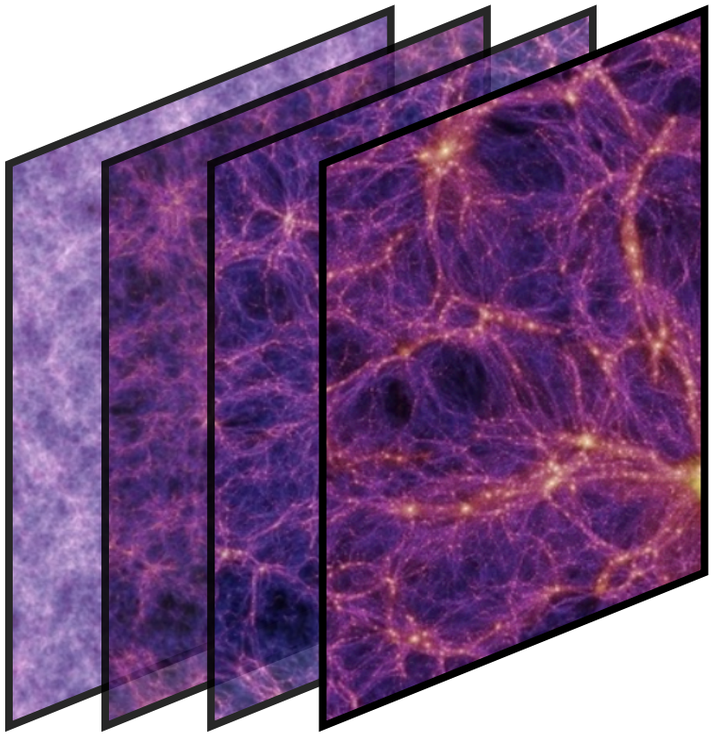
- **lightcone outputs: "pencil beams" of 40 sq.deg. each up to z=10**
- same outputs (healpix, halo catalogs) as in WIDE
- data volume : 0.2 PB (LC only) + snapshots (25 TB each)
- **Status: simulation developed up to z=10**

Summary: inference

Initial conditions
t_0



Evolution



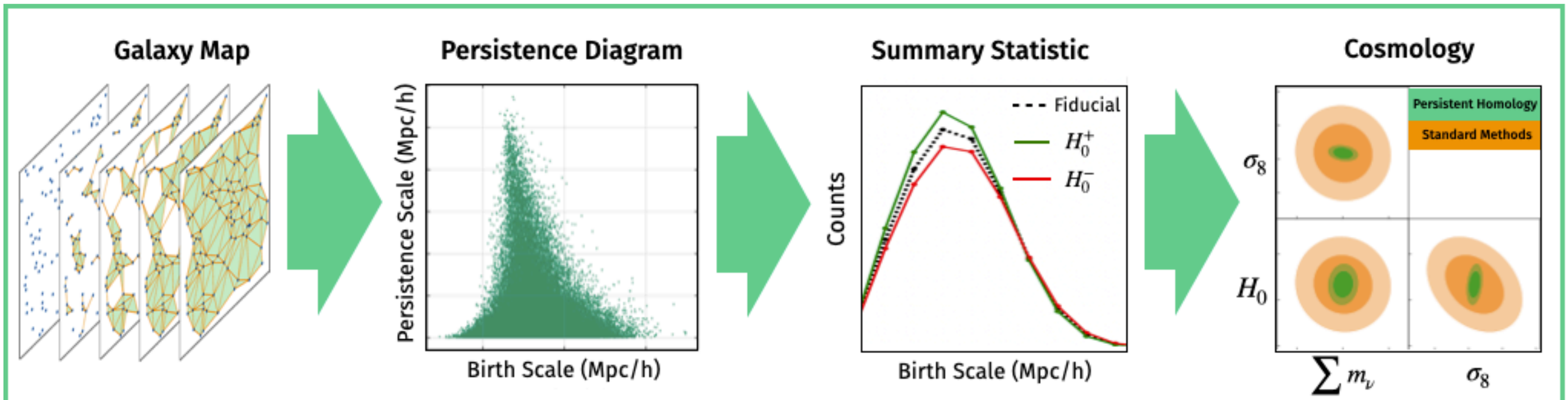
Late-time
t_f

New strategy: describe set of discrete points (galaxies) using ***persistent homology***

Idea: persistent homology provides a hardcoded but interpretable transformation of the galaxy map (e.g. might help a ML architecture to learn better and faster)

Plan of the talk

- Introduction to Persistent Homology
- From Galaxy Maps to Persistence Diagrams
- From Persistence Diagrams to Summary Statistics
- From Summary Statistics to Cosmology



Introduction to Persistent Homology

Topology and Homology

topologically equivalent

$=$

not topologically equivalent

$\neq$

Introduction to Persistent Homology

Topology and Homology

topologically equivalent

$=$

not topologically equivalent

$\neq$

$H_i(X)$

i[th] homology group of X

$= \left\{ \begin{matrix} \text{i-dim} \\ \text{holes} \end{matrix} \right\}$

Betti number

$b_i = \text{rank}(H_i)$

1 hole — 1 hole — 3 holes

$b_1(\text{mug}) = b_1(\text{donut}) = 1$

$b_1(\text{pretzel}) = 3$

$b_0 = 1$ # connected components

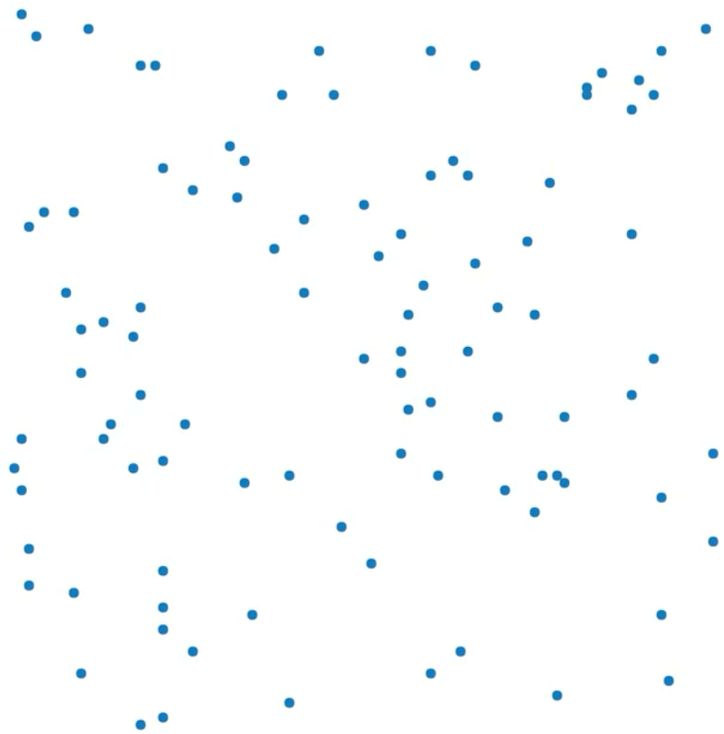
$b_2 = 0$ # voids

Introduction to Persistent Homology

Changing homology across scales

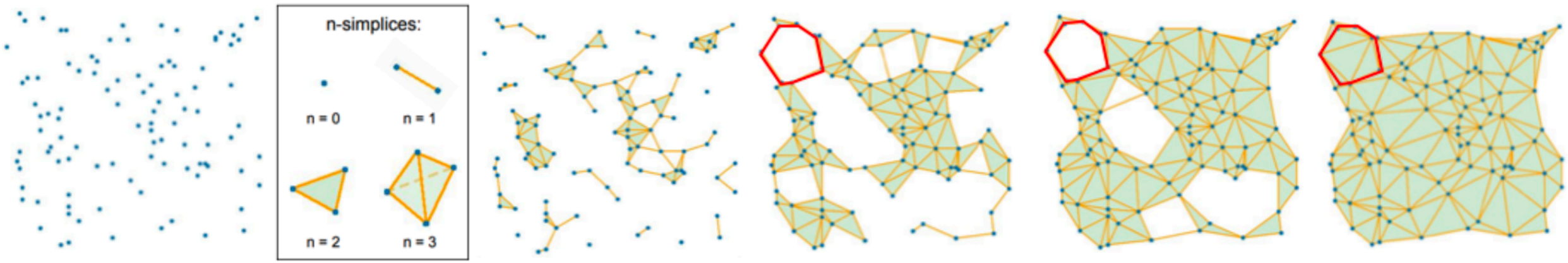
Length scale parameter

imagine a ball of radius ν, when balls touch simplices are added to the complex



Introduction to Persistent Homology

Goal: track ***persistent*** features



(Ambitious) goal: find the ***underlying manifold*** of sampled data

Goal: find the ***topological features*** of sampled data

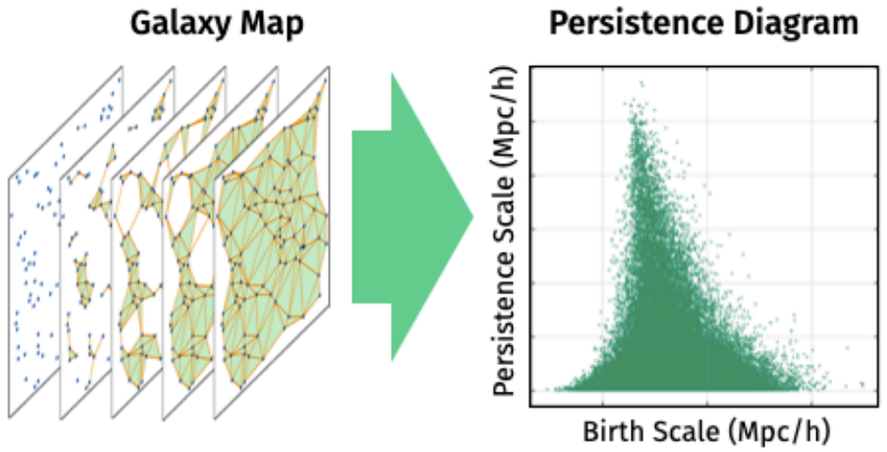
Topological features are connected to observable aspects of the cosmic web

0-dim holes: galaxy clusters

1-dim holes: filaments

2-dim holes: voids

From Galaxy Maps to Persistence Diagrams



From Galaxy Maps to Persistence Diagrams

The Sancho dataset

CMASS-like mock galaxy catalogs at $z = 0.5$

- varying 6 cosmological parameters and 5 HOD parameters
- 15K catalogs fiducial cosmology, 15K for each parameter, 250K+ total

Quijote suite

The fiducial cosmological parameter values are

Amount of matter	$\Omega_m = 0.3175$	$\Omega_b = 0.049$	Amount of baryons
Expansion Rate	$h = 0.6711$	$n_s = 0.9624$	Spectral index
Amplitude of fluctuations	$\sigma_8 = 0.834$	$M_\nu = 0.0\text{eV}$	Mass of neutrinos

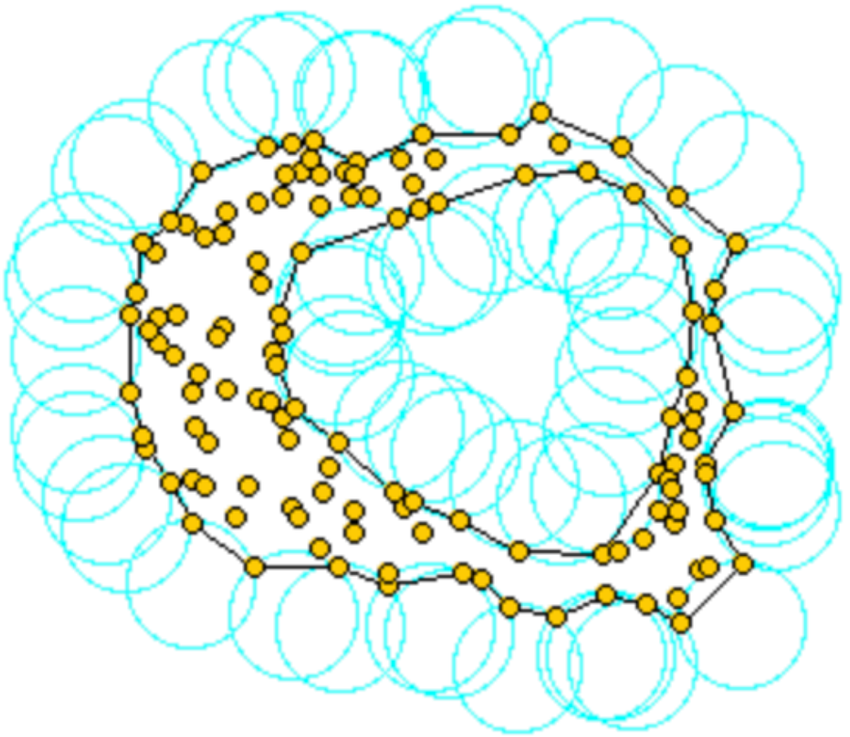
With steps:

$$\{\Delta\Omega_m, \Delta\Omega_b, \Delta h, \Delta n_s, \Delta\sigma_8\} = \{0.01, 0.002, 0.02, 0.02, 0.015\}$$

From Galaxy Maps to Persistence Diagrams

Let us now compute persistent homology on (simulated) galaxies

We use the alpha-filtration (alpha complex in Gudhi) + DTM function (see next slide)



Allows fast computation of large sets (300K points)

From Galaxy Maps to Persistence Diagrams

Let us now compute persistent homology on (simulated) galaxies

Introduce a Distance-To-Measure function

$$DTM(x) = \left(\frac{1}{k} \sum_{x_i \in N_k(x)} |x - x_i|^p \right)^{1/p}$$

k : # of nearest neighbours

$N_k(x)$: the set of k-nearest neighbours of x

p : a mixing parameter (e.g. $p = 2$)

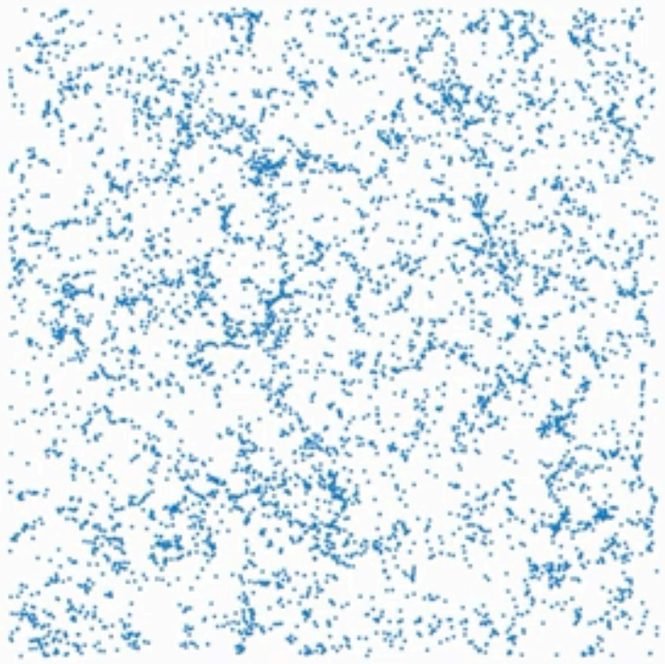
To modify the scale ν at which the radius r of the balls starts growing:

$$r_x(\nu) = (\nu^q - DTM(x)^q)^{1/q} \qquad (q = 2)$$

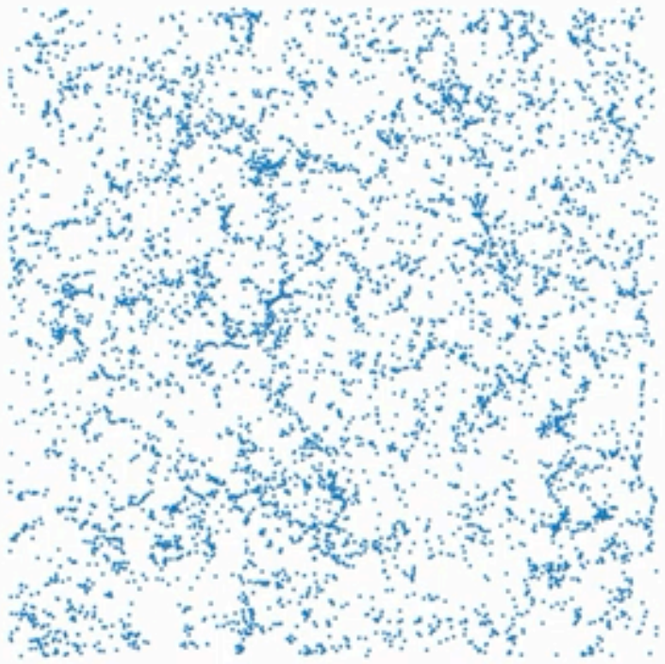
In a galaxy map, there is no clear sense of outliers, but the DTM function allows to *give weight to features at different scales*

From Galaxy Maps to Persistence Diagrams

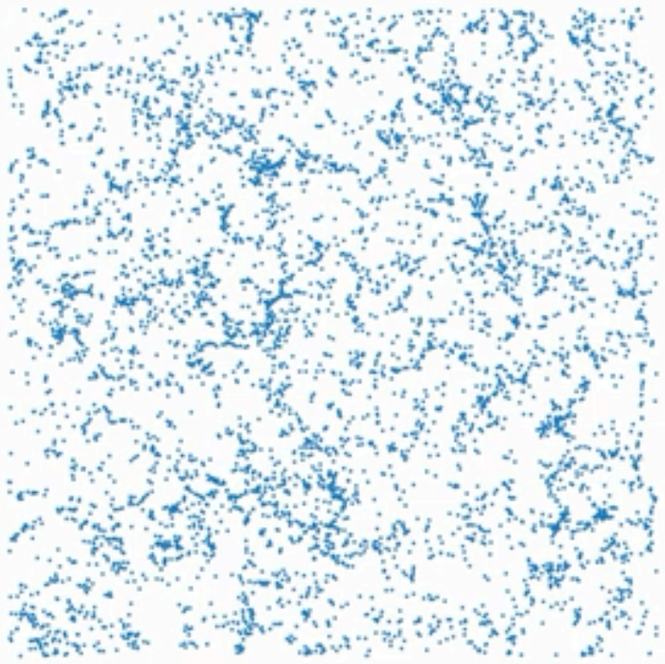
Let us now compute persistent homology on (simulated) galaxies



$k = 1$



$k = 15$

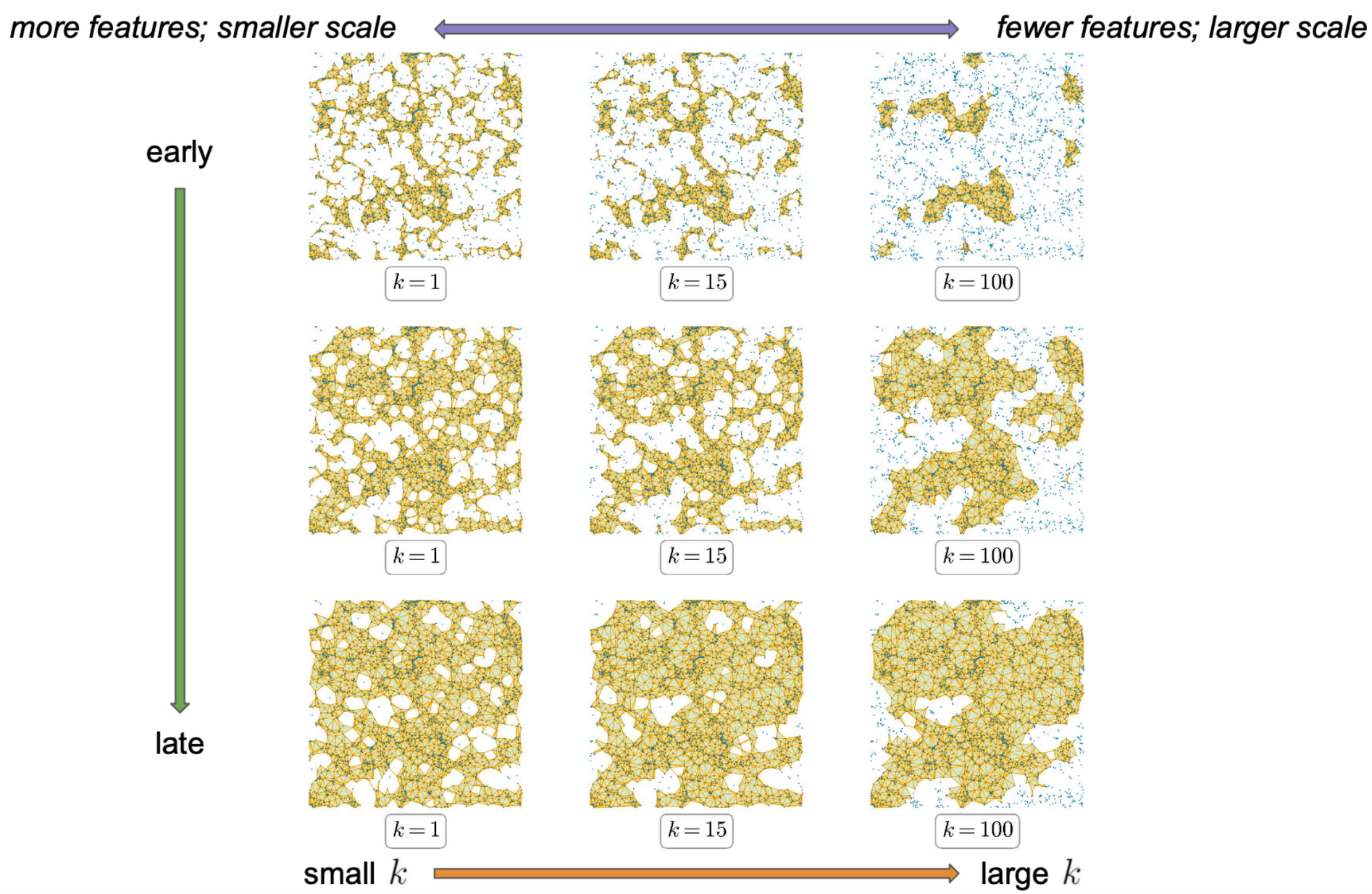


$k = 100$

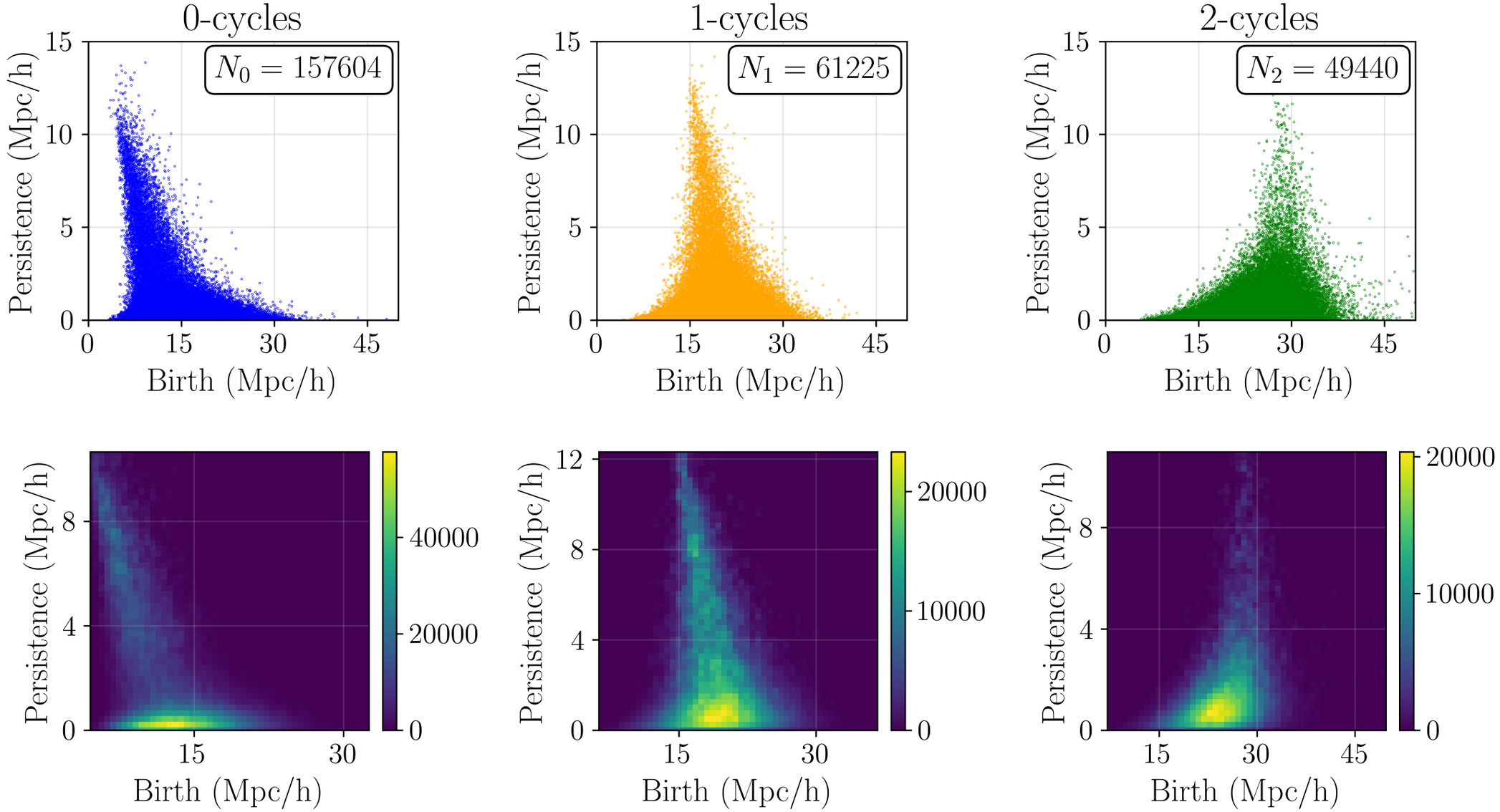
e.g.: varying the number of nearest neighbours of the DTM function $DTM(x) = \left(\frac{1}{k} \sum_{x_i \in N_k(x)} |x - x_i|^p \right)^{1/p}$

From Galaxy Maps to Persistence Diagrams

Compute persistence diagrams at varying k

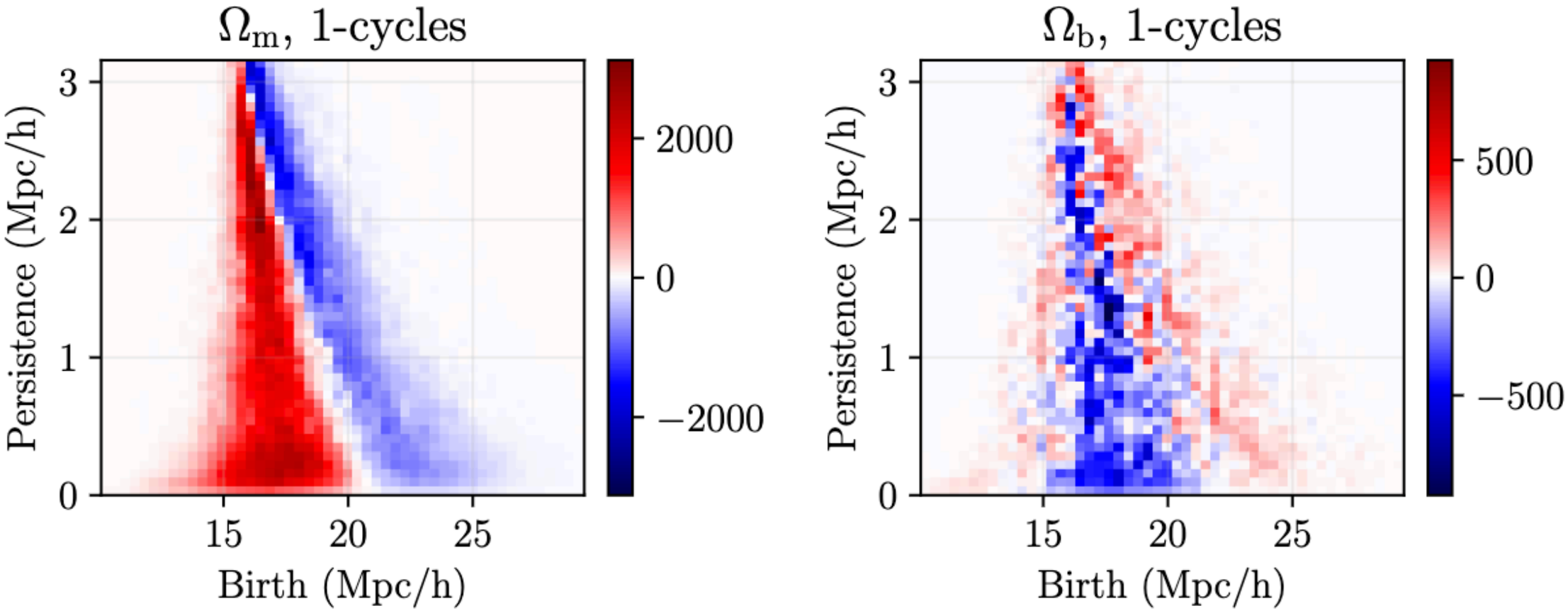


From Galaxy Maps to Persistence Diagrams



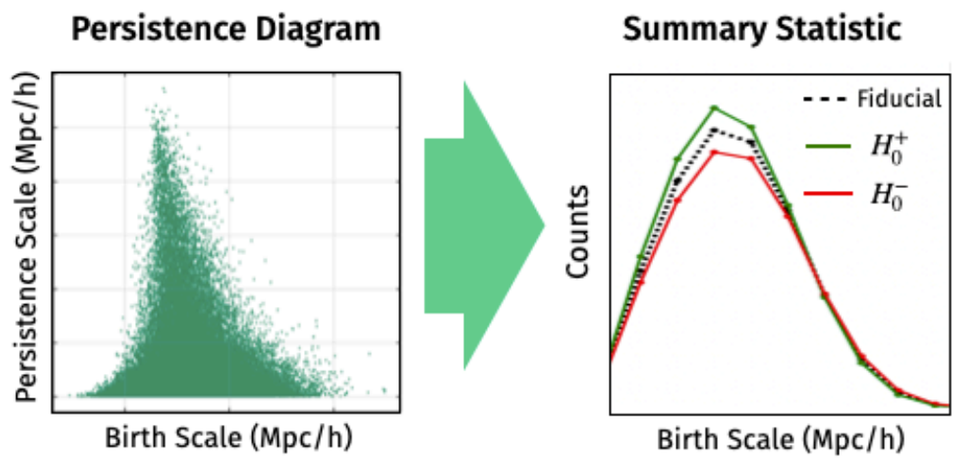
Persistence diagrams and images of a single fiducial box from Sancho

From Galaxy Maps to Persistence Diagrams



Difference between Sancho boxes with varying Ω_m and Ω_b

From Persistence Diagrams to Summary Statistics



From Persistence Diagrams to Summary Statistics

Summary statistic should be:

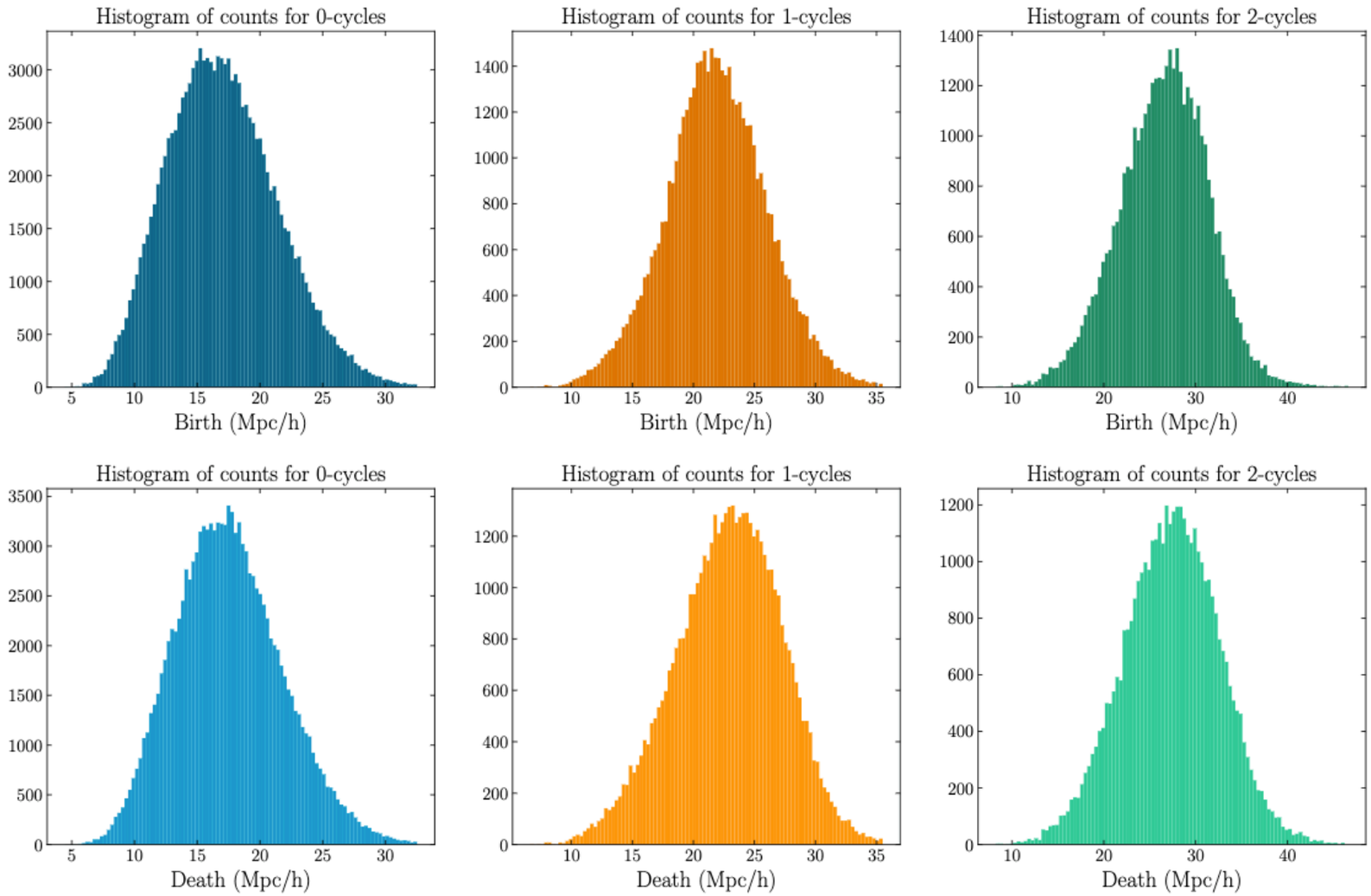
- computationally tractable (vector)
- Informative

This optimisation procedure is non-trivial and currently under study in our context

Viswanathan, Cole, MB (in progress)

For now, we choose a simple (and sub-optimal) approach: ***histograms of counts***

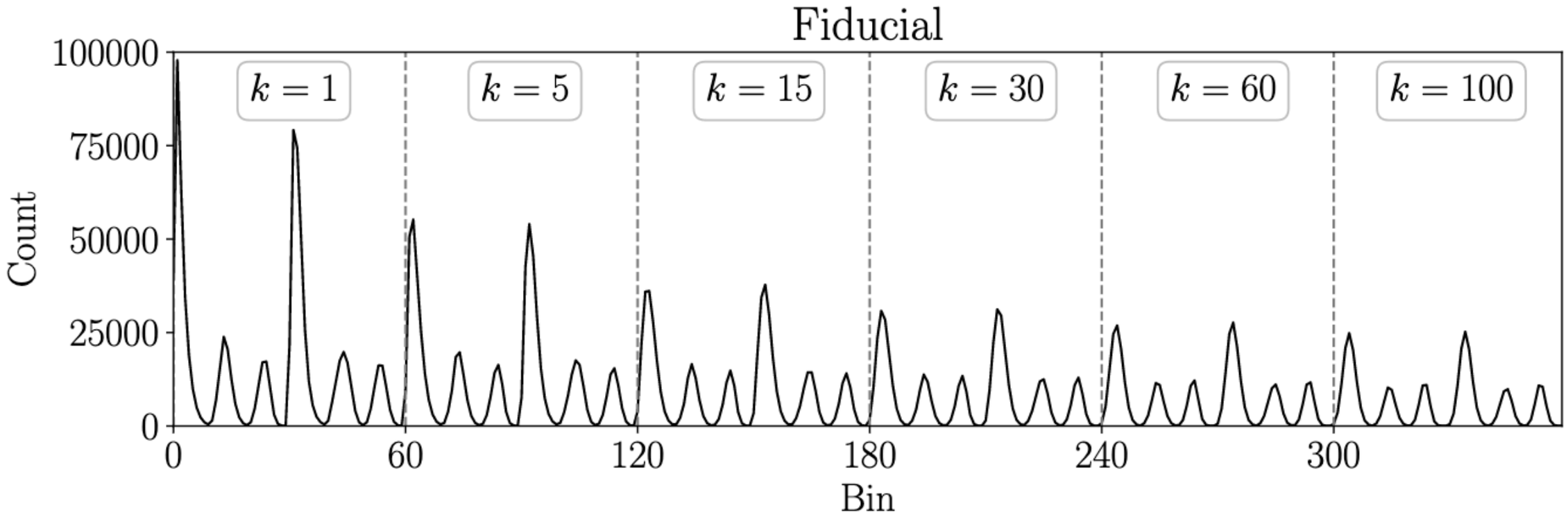
From Persistence Diagrams to Summary Statistics



- Divide each persistence diagram into N bins
- Count births and deaths
- Downsample histogram if needed for covariance inversion/convergence

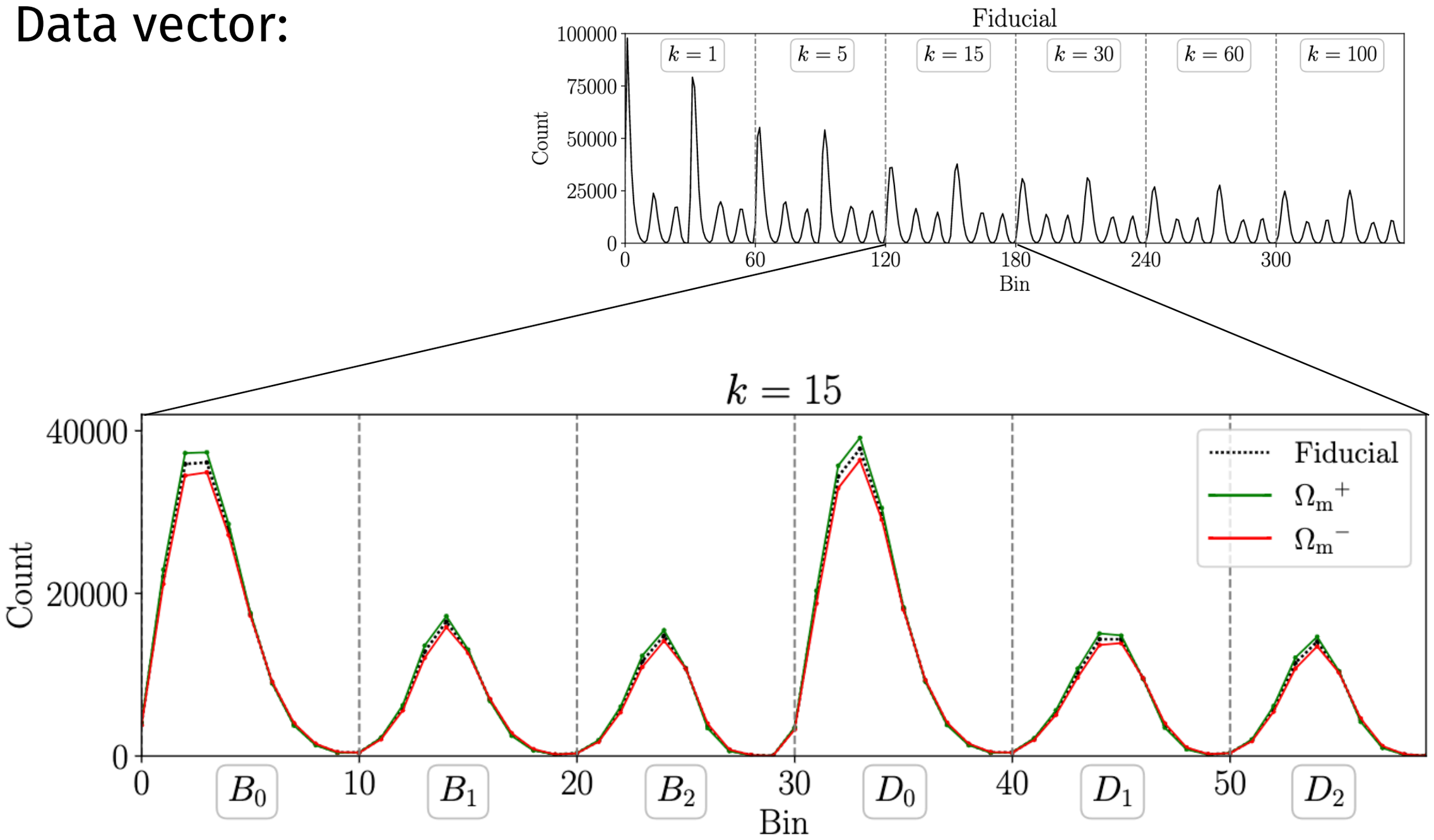
From Persistence Diagrams to Summary Statistics

Data vector:



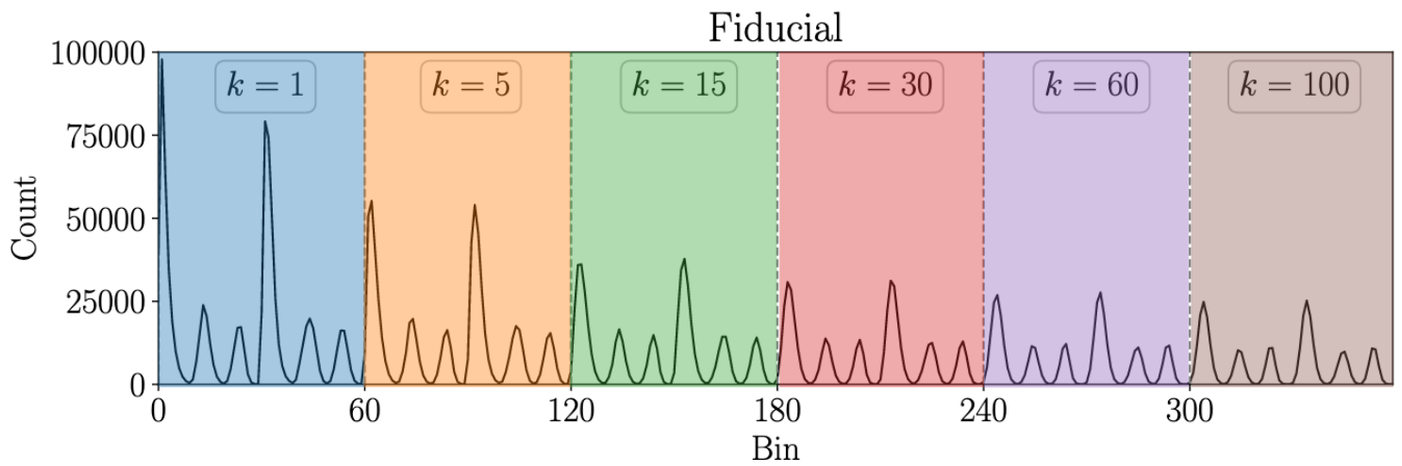
From Persistence Diagrams to Summary Statistics

Data vector:

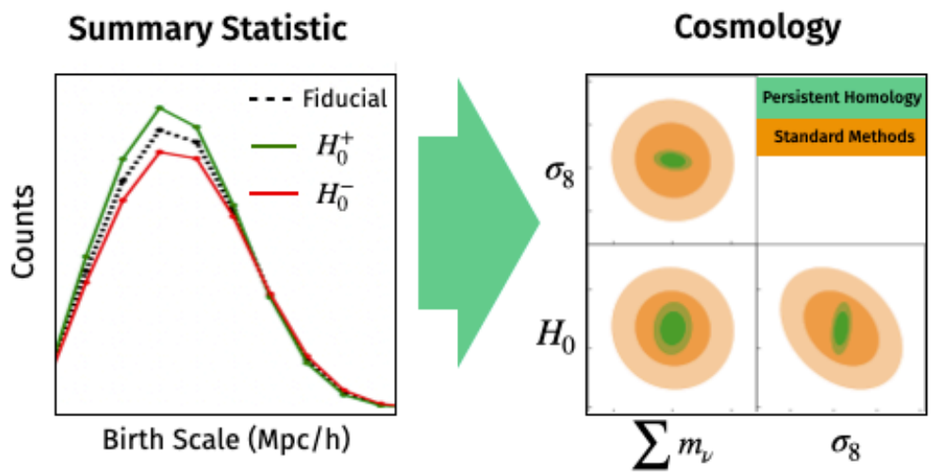


From Persistence Diagrams to Summary Statistics

Data vector:



From Summary Statistics to Cosmology



From Summary Statistics to Cosmology

All nice and clean but do we actually *learn* something new on the parameters?

To quantify this we use the Fisher Information Matrix

$$F_{ij} = \left(\partial_i D\right)^T C^{-1} \left(\partial_j D\right) + \frac{1}{2}\mathrm{Tr}\left[\left(C^{-1}C_{,i}\right)\left(C^{-1}C_{,j}\right)\right]$$

the inverse of the Fisher information is a lower bound
on the variance of any unbiased estimator of *parameter i,j*
(Cramer-Rao bound)

From Summary Statistics to Cosmology

Information content through Fisher Matrix

$$F_{ij} = \left(\partial_i D\right)^T C^{-1} \left(\partial_j D\right) + \frac{1}{2}\mathrm{Tr}\left[\left(C^{-1}C_{,i}\right)\left(C^{-1}C_{,j}\right)\right]$$

Data Vector: counts of connected components, loops, voids for births and deaths

From Summary Statistics to Cosmology

Information content through Fisher Matrix

$$F_{ij} = \left(\partial_i D\right)^T C^{-1} \left(\partial_j D\right) + \frac{1}{2}\mathrm{Tr}\left[\left(C^{-1}C_{,i}\right)\left(C^{-1}C_{,j}\right)\right]$$

Data Vector: counts of connected components, loops, voids for births and deaths

Numerical derivative $\dfrac{D(\theta_i^+) - D(\theta_i^-)}{\theta_i^+ - \theta_i^-}$

From Summary Statistics to Cosmology

Information content through Fisher Matrix

$$F_{ij} = \left(\partial_i D\right)^T C^{-1} \left(\partial_j D\right) + \frac{1}{2}\mathrm{Tr}\left[\left(C^{-1}C_{,i}\right)\left(C^{-1}C_{,j}\right)\right]$$

Data Vector: counts of connected components, loops, voids for births and deaths

Numerical derivative $\dfrac{D(\theta_i^+) - D(\theta_i^-)}{\theta_i^+ - \theta_i^-}$

Covariance $C = \dfrac{1}{N_{sims}-1}\sum_k^{N_{sims}} (D_k - \bar{D})(D_k - \bar{D})^T$

From Summary Statistics to Cosmology

Information content through Fisher Matrix

$$F_{ij} = \left(\partial_i D\right)^T C^{-1} \left(\partial_j D\right) + \frac{1}{2}\mathrm{Tr}\left[\left(C^{-1}C_{,i}\right)\left(C^{-1}C_{,j}\right)\right]$$

Data Vector: counts of connected components, loops, voids for births and deaths

Numerical derivative $\dfrac{D(\theta_i^+) - D(\theta_i^-)}{\theta_i^+ - \theta_i^-}$

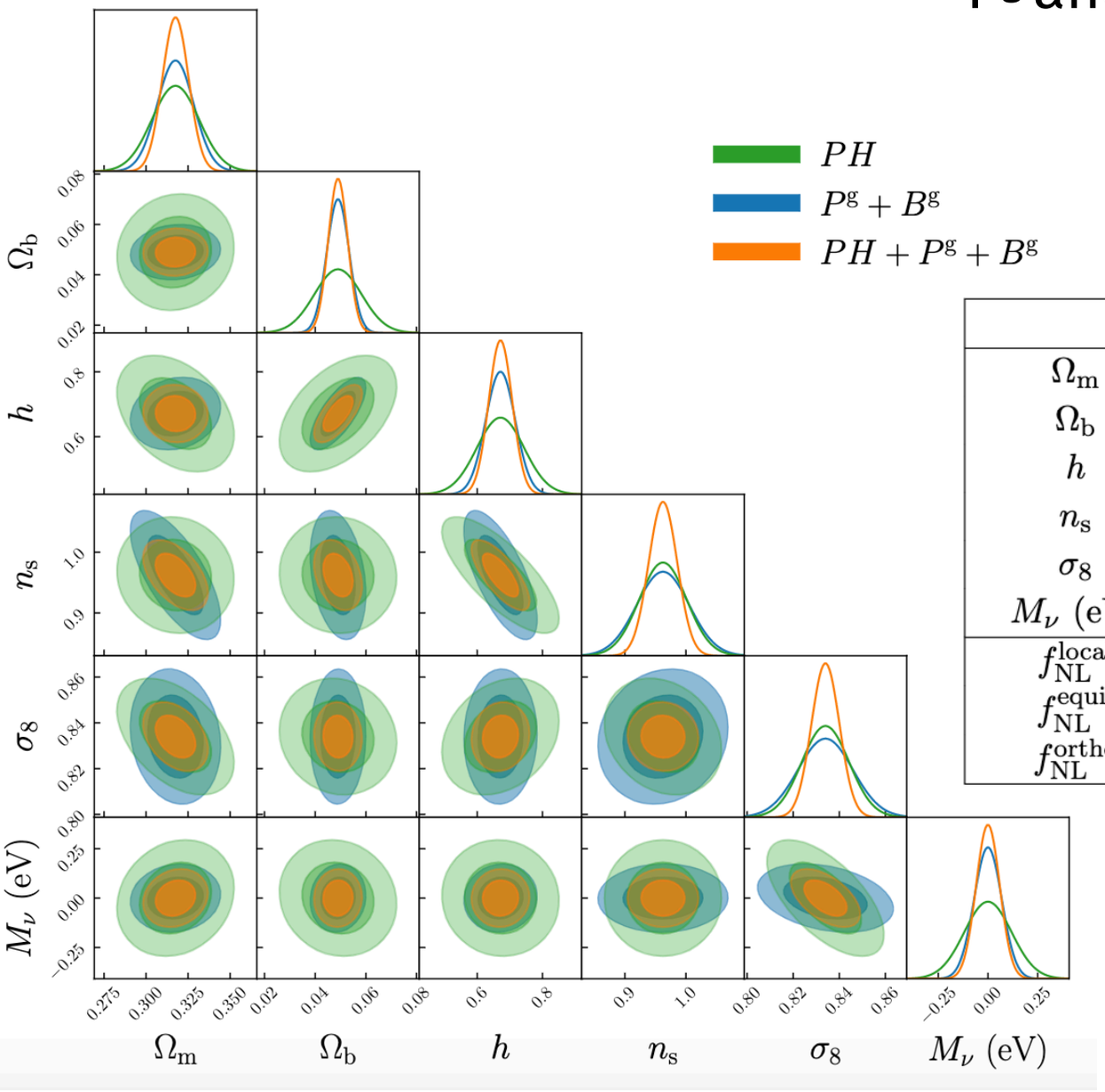
Covariance $C = \dfrac{1}{N_{sims} - 1} \sum_k^{N_{sims}} (D_k - \bar{D})(D_k - \bar{D})^T$

Covariance Derivative: neglected

From Summary Statistics to Cosmology

(Preliminary, conservative) Results

PH: Persistent Homology
P^g and B^g: Power Spectrum + Bispectrum



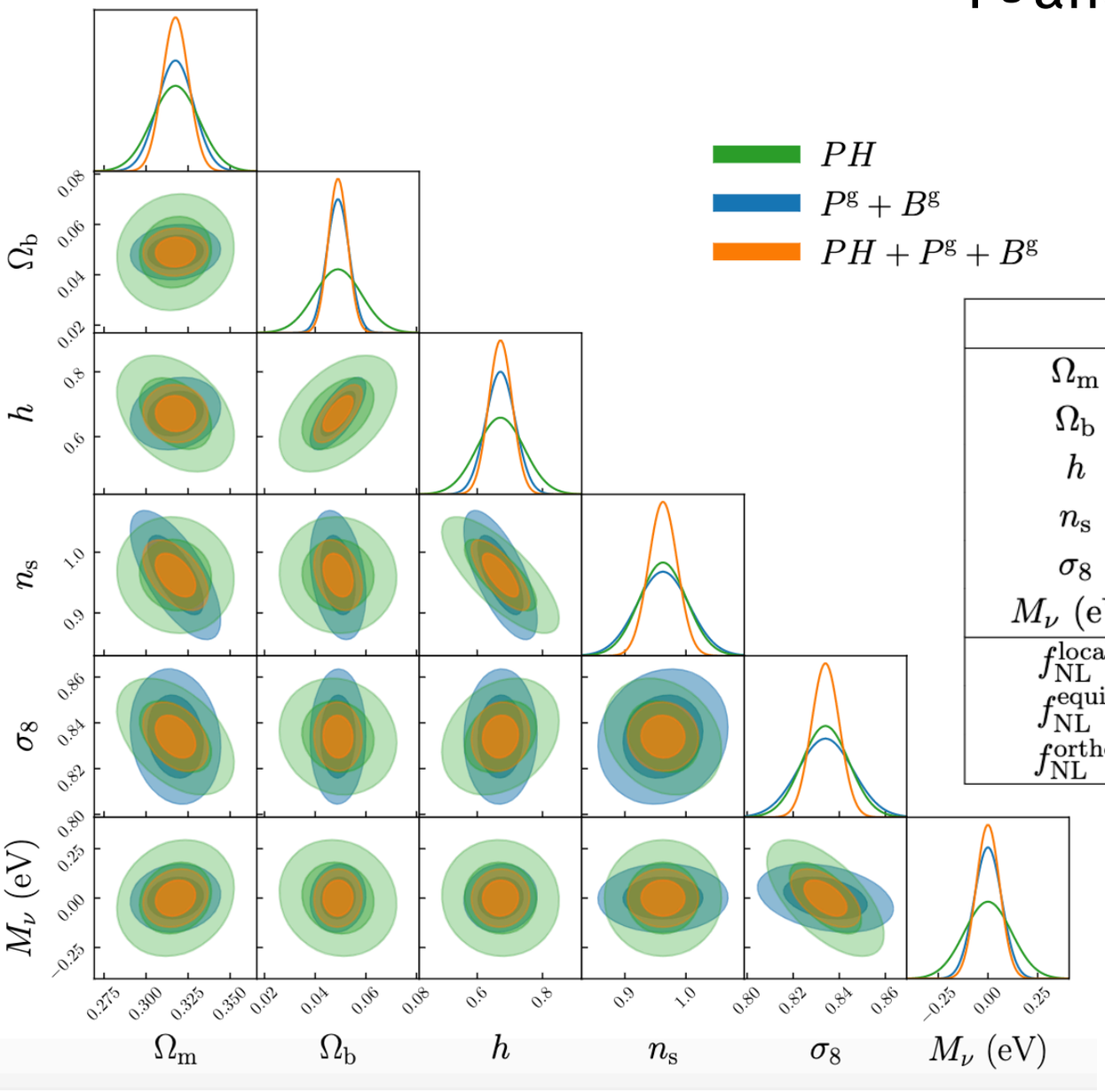
	PH	$P^g + B^g$	$PH + P^g + B^g$
Ω_m	0.014 (0.009)	0.011 (0.007)	0.008 (0.005)
Ω_b	0.009 (0.006)	0.004 (0.003)	0.004 (0.002)
h	0.072 (0.047)	0.045 (0.029)	0.036 (0.023)
n_s	0.039 (0.028)	0.043 (0.028)	0.023 (0.015)
σ_8	0.010 (0.007)	0.012 (0.008)	0.006 (0.004)
M_ν (eV)	0.119 (0.045)	0.070 (0.045)	0.059 (0.038)
f_{NL}^{local}	141.0 (91.0)	59.6 (38.5)	53.0 (34.2)
f_{NL}^{equil}	216.2 (139.5)	239.9 (154.9)	149.2 (96.3)
f_{NL}^{ortho}	82.4 (53.2)	66.4 (42.9)	50.2 (32.4)

Yip, MB, Cole, Calles, Shiu (to appear)

From Summary Statistics to Cosmology

(Preliminary, conservative) Results

PH: Persistent Homology
P^g and B^g: Power Spectrum + Bispectrum

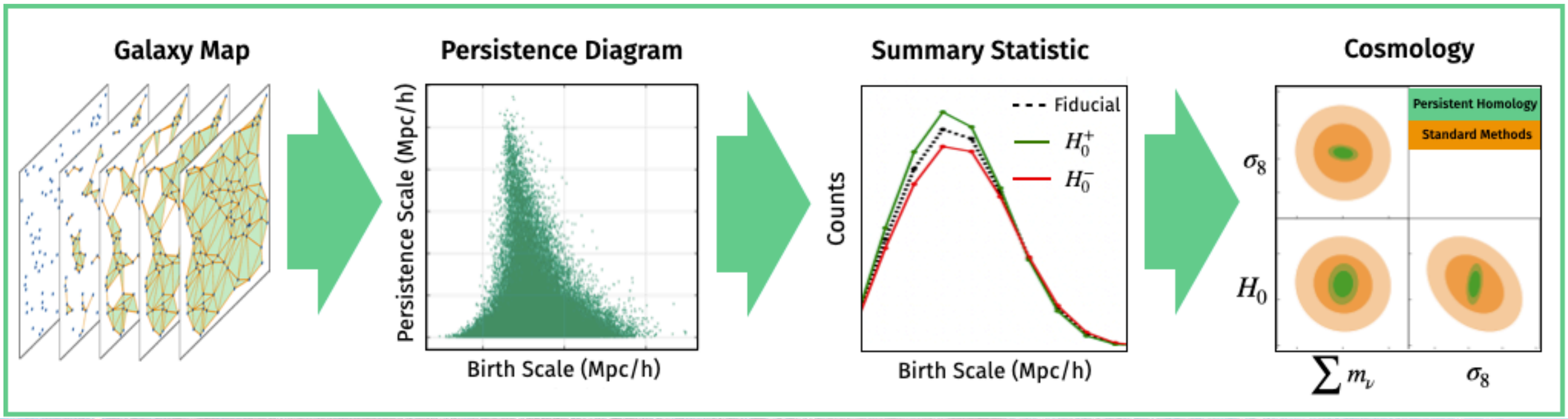


	PH	$P^g + B^g$	$PH + P^g + B^g$
Ω_m	0.014 (0.009)	0.011 (0.007)	0.008 (0.005)
Ω_b	0.009 (0.006)	0.004 (0.003)	0.004 (0.002)
h	0.072 (0.047)	0.045 (0.029)	0.036 (0.023)
n_s	0.039 (0.028)	0.043 (0.028)	0.023 (0.015)
σ_8	0.010 (0.007)	0.012 (0.008)	0.006 (0.004)
M_ν (eV)	0.119 (0.045)	0.070 (0.045)	0.059 (0.038)
f_{NL}^{local}	141.0 (91.0)	59.6 (38.5)	53.0 (34.2)
f_{NL}^{equil}	216.2 (139.5)	239.9 (154.9)	149.2 (96.3)
f_{NL}^{ortho}	82.4 (53.2)	66.4 (42.9)	50.2 (32.4)

Yip, MB, Cole, Calles, Shiu (to appear)

Checkpoint Remarks

- Persistent Homology as a way to describe the Large Scale Structures of the Universe
- Hybrid approach: data-driven, but interpretable
- Preliminary study shows there is complementary cosmological information



Things we are trying

- Generative models using information from persistent homology for fast simulations
- Maximise information on underlying parameters using persistent homology (for best constraints)
- Transform discrete set of points with properties A to another one with properties B using information from persistent homology (for fast simulations)

Things we are trying

- **Generative models using information from persistent homology for fast simulations**

- Maximise information on underlying parameters using persistent homology (for best constraints)

- Transform discrete set of points with properties A to another one with properties B using information from persistent homology (for fast simulations)

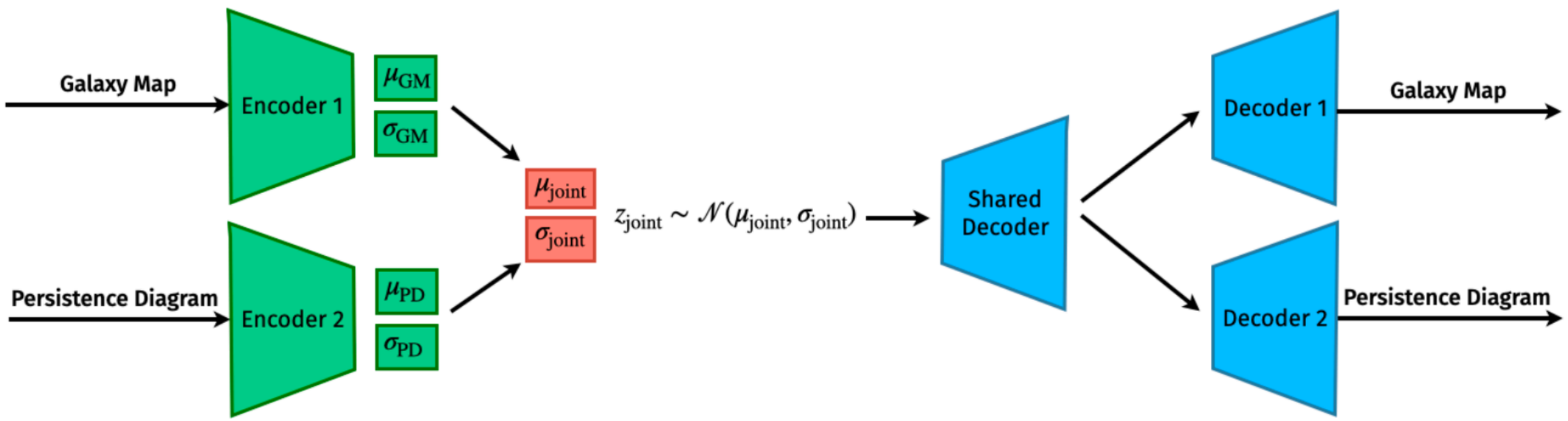
Generative models

Goal: avoid running the full numerical sim (computationally expensive)

Objectives: inference, build training sets

Idea: persistent homology provides a transformation of the galaxy map which might help a ML architecture to learn better and faster

Methodology: Variational Auto-encoders

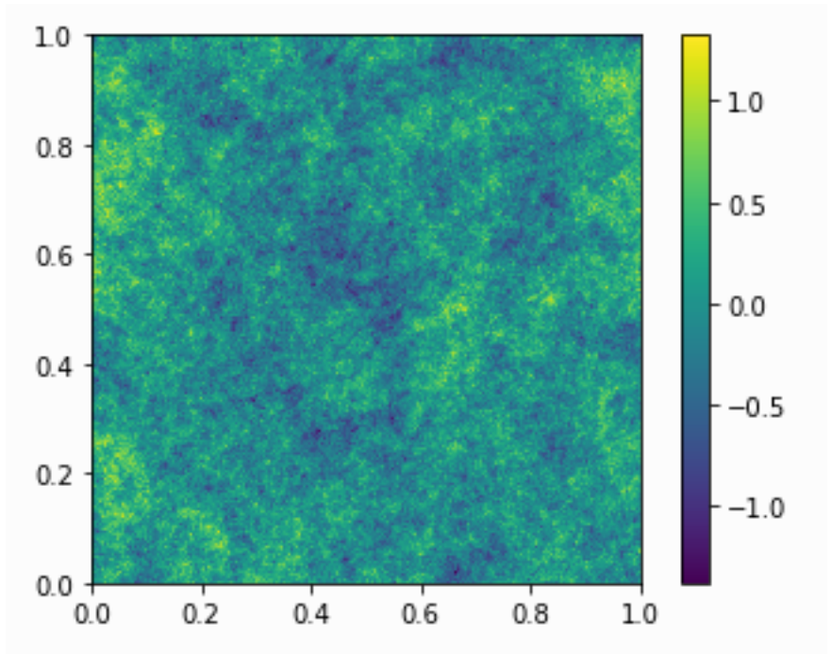


Generative models

Start simple with Gaussian Random Fields with a given power spectrum

GRF generated with $P(k) = Ak^{-B}$

(In the figure A=0.1, B=2)



Useful features:

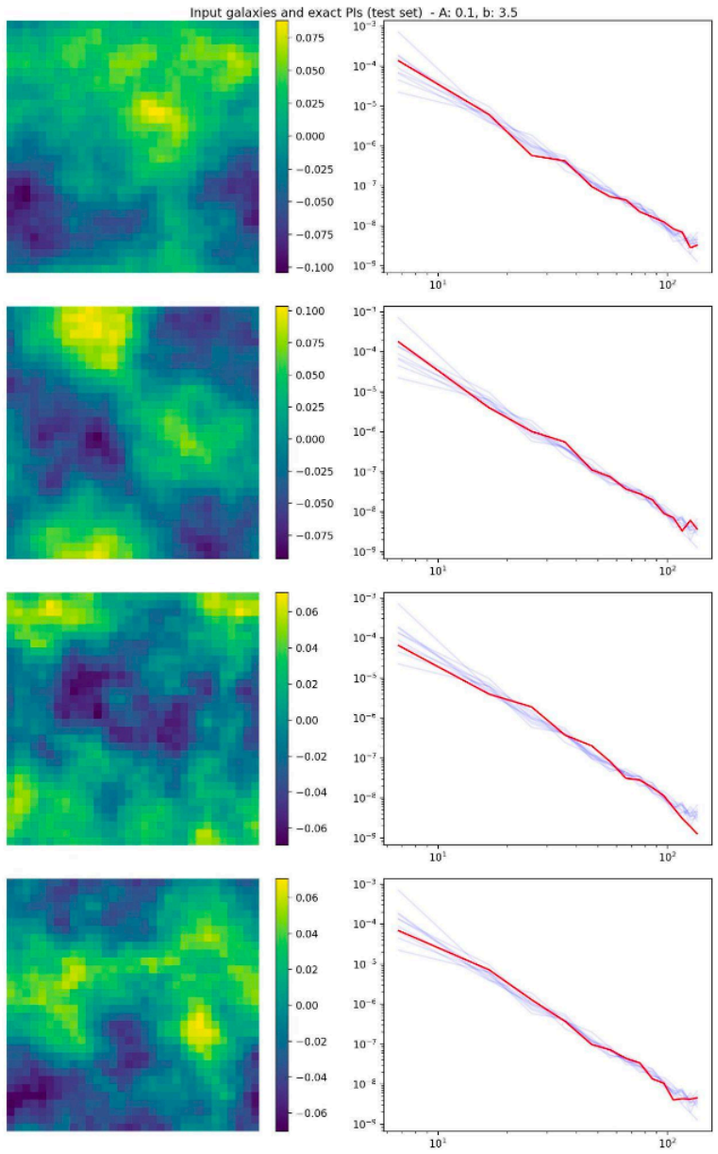
1) we know the ground truth (power spectrum specifies fully the field)

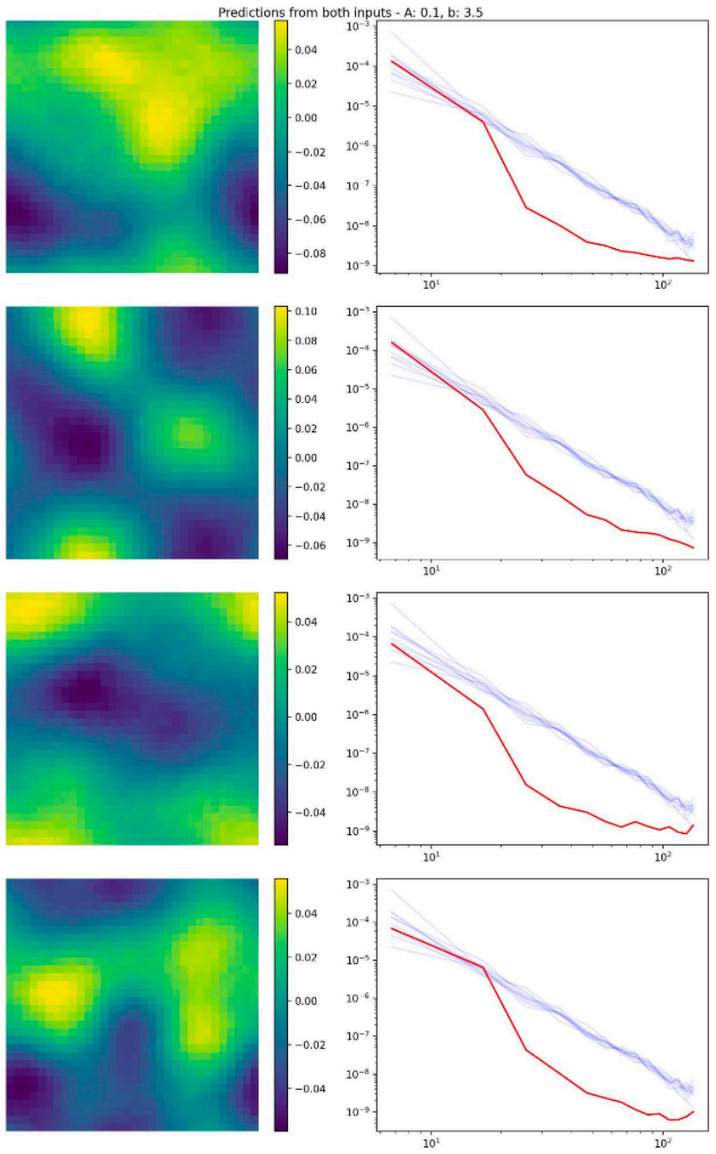
2) Easy to generate

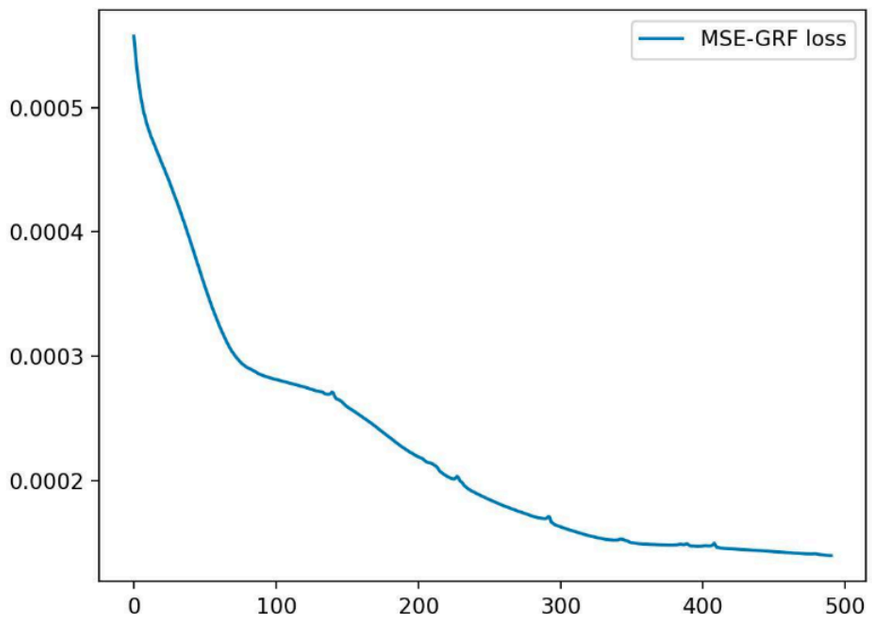
3) Input of cosmological simulations: evolution makes field non-Gaussian, but good starting point

Generative models

AE: loss: MSE on x-space, A = 0.1, b =3.5

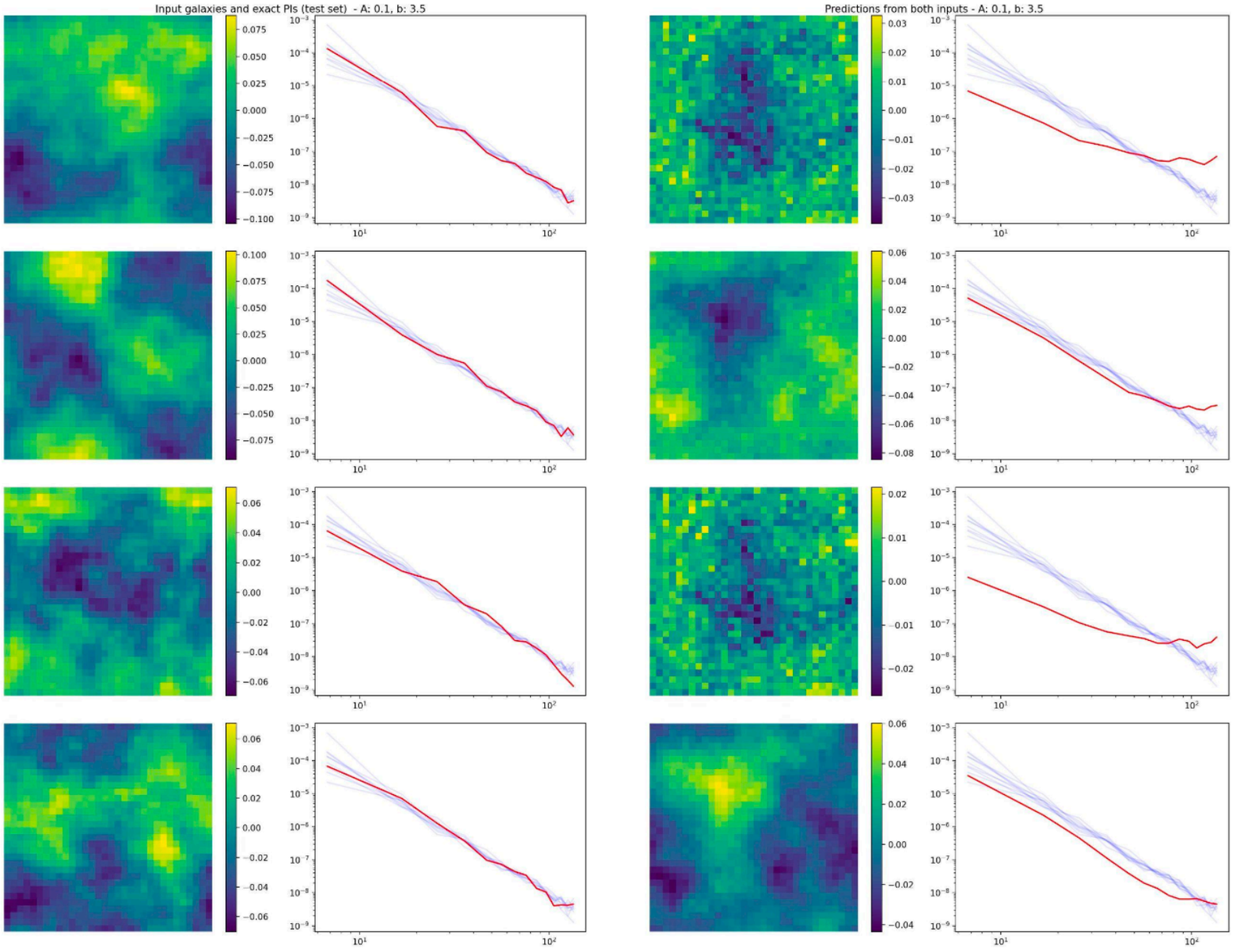


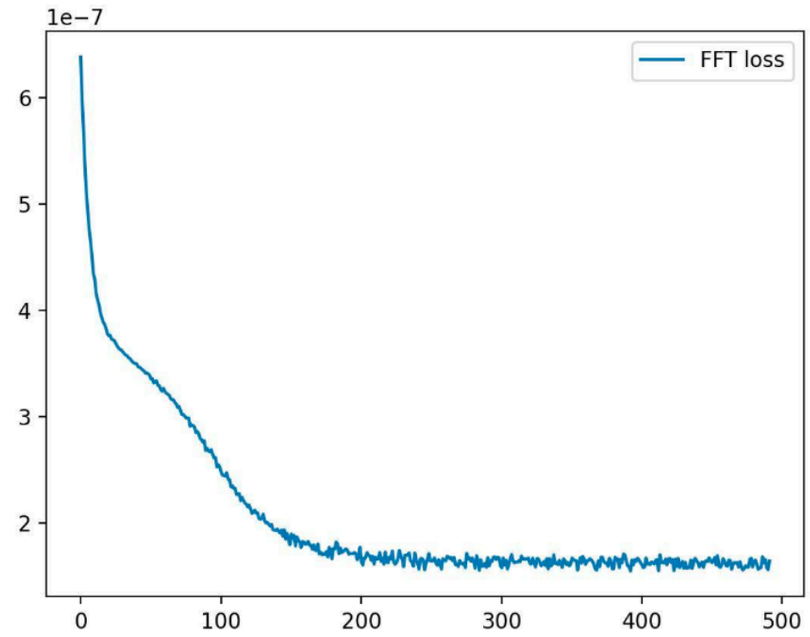




Generative models

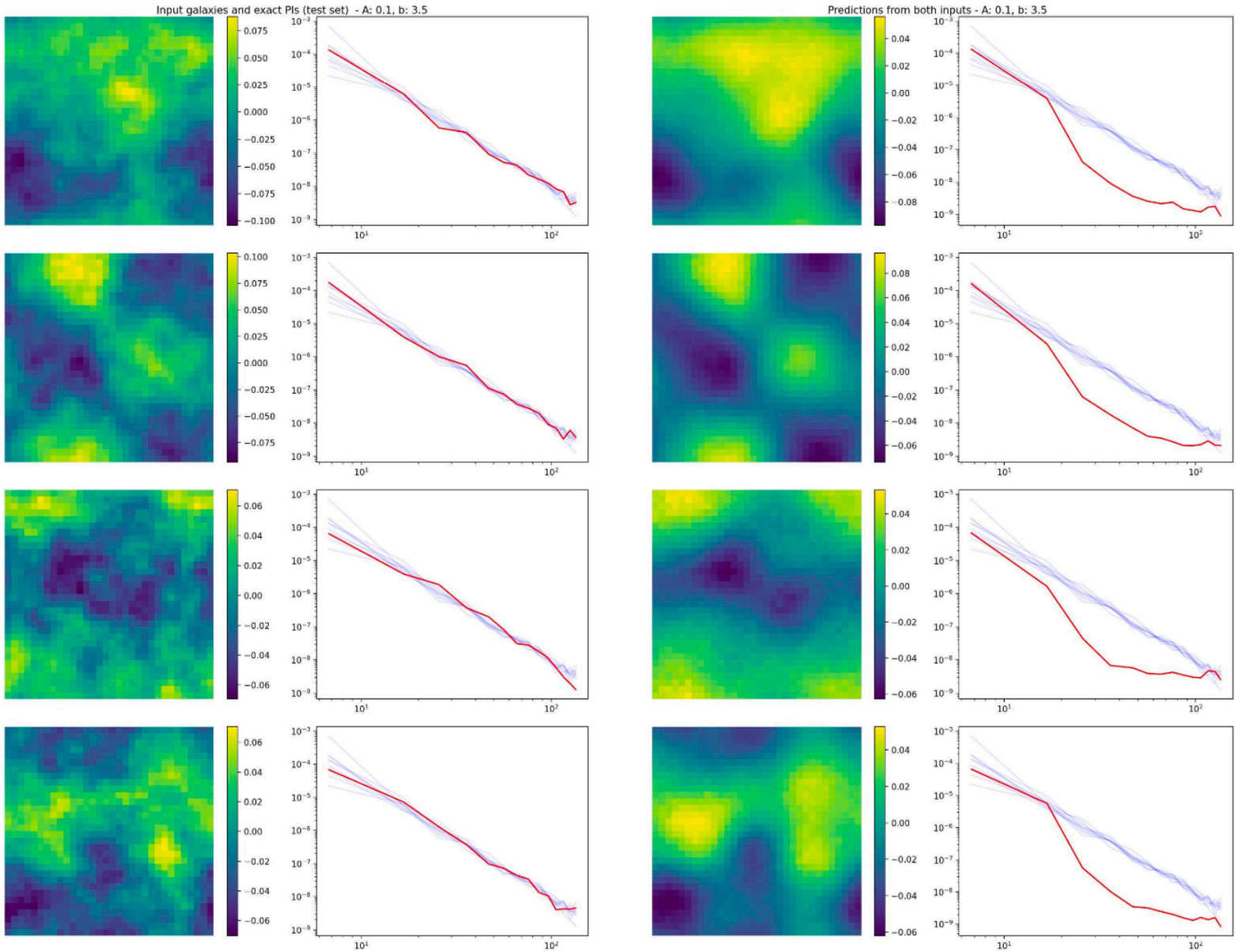
AE: loss: MSE on k-space, A = 0.1, b =3.5

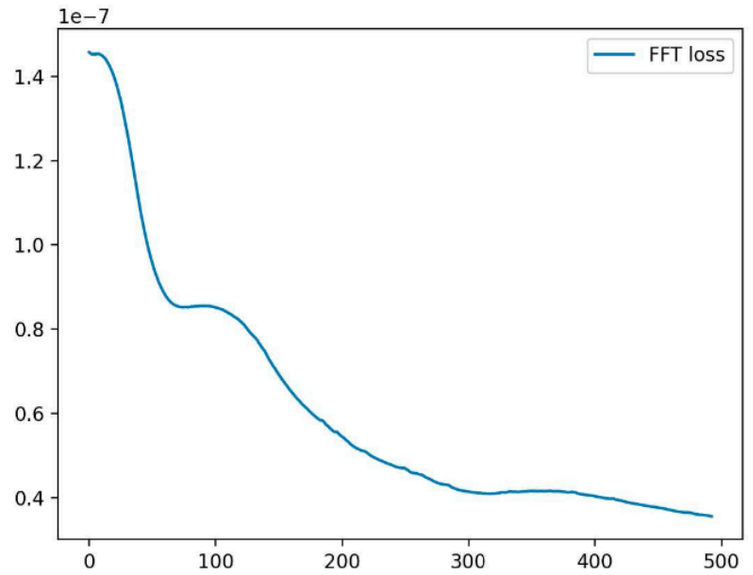


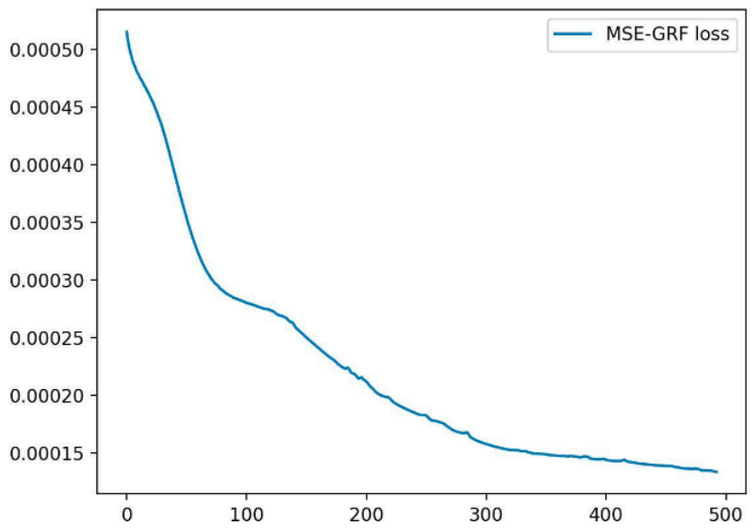


Generative models

AE: loss: MSE on both x- and k-space, A = 0.1, b =3.5, loss factor = 1.0

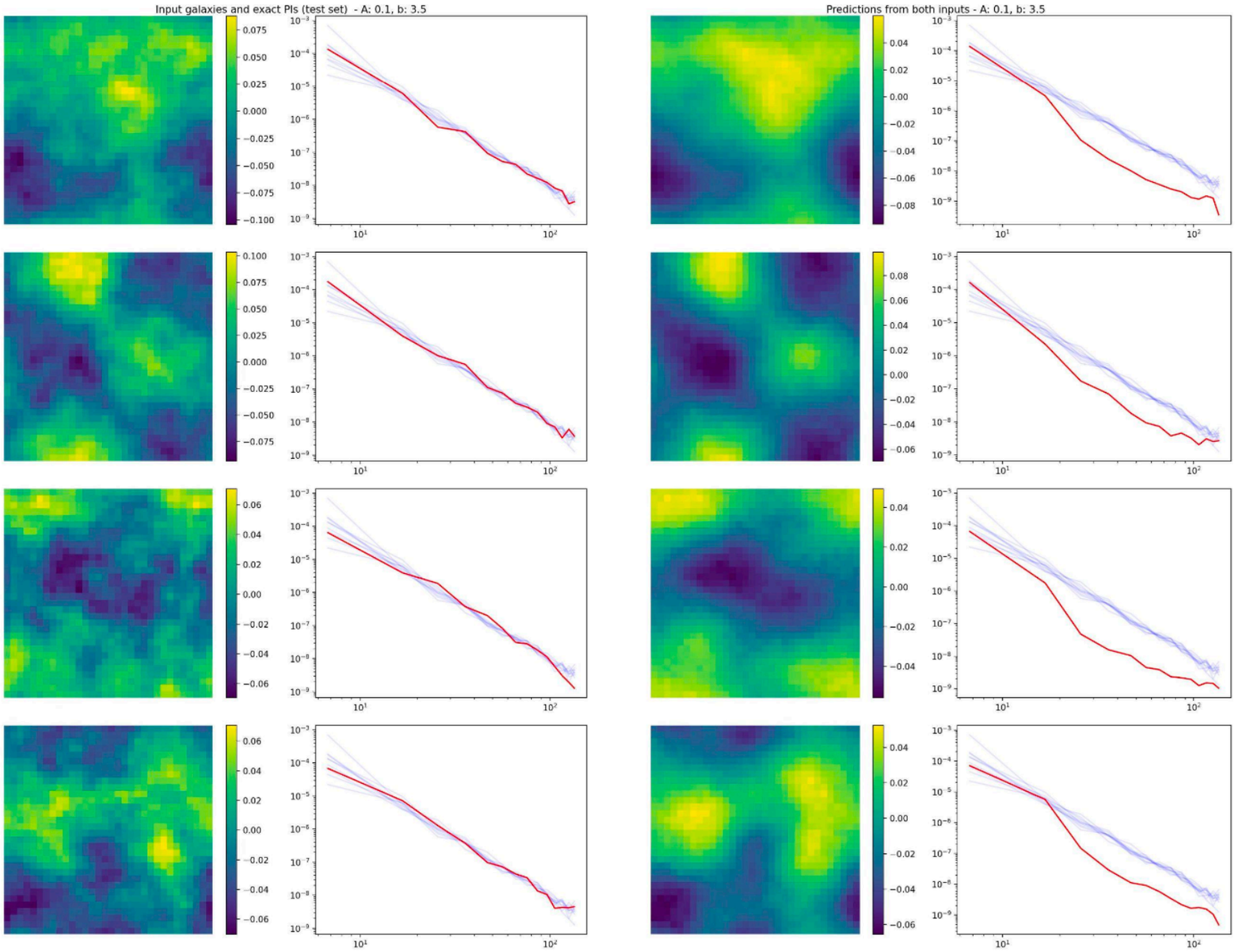


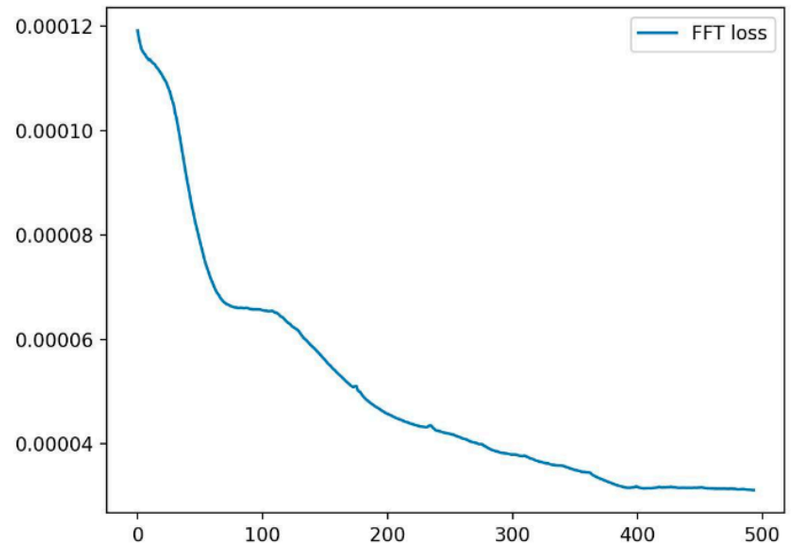


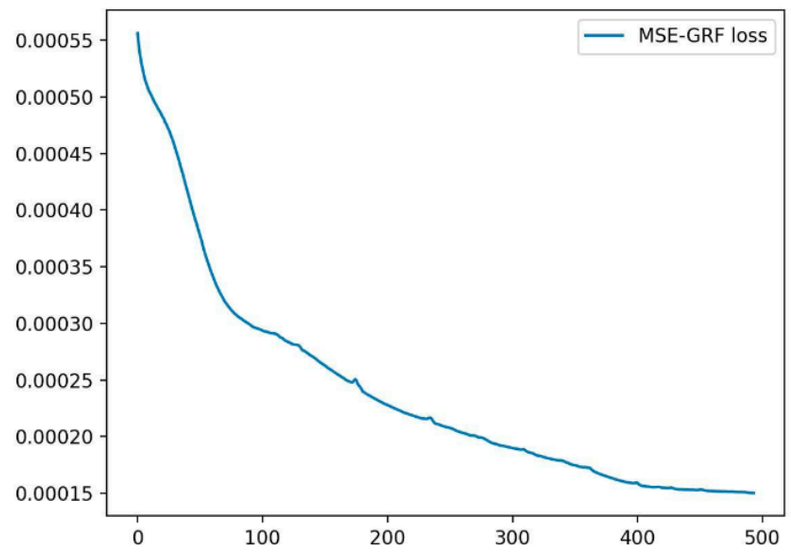


Generative models

AE: loss: MSE on both x- and k-space, A = 0.1, b =3.5, loss factor = 10^3

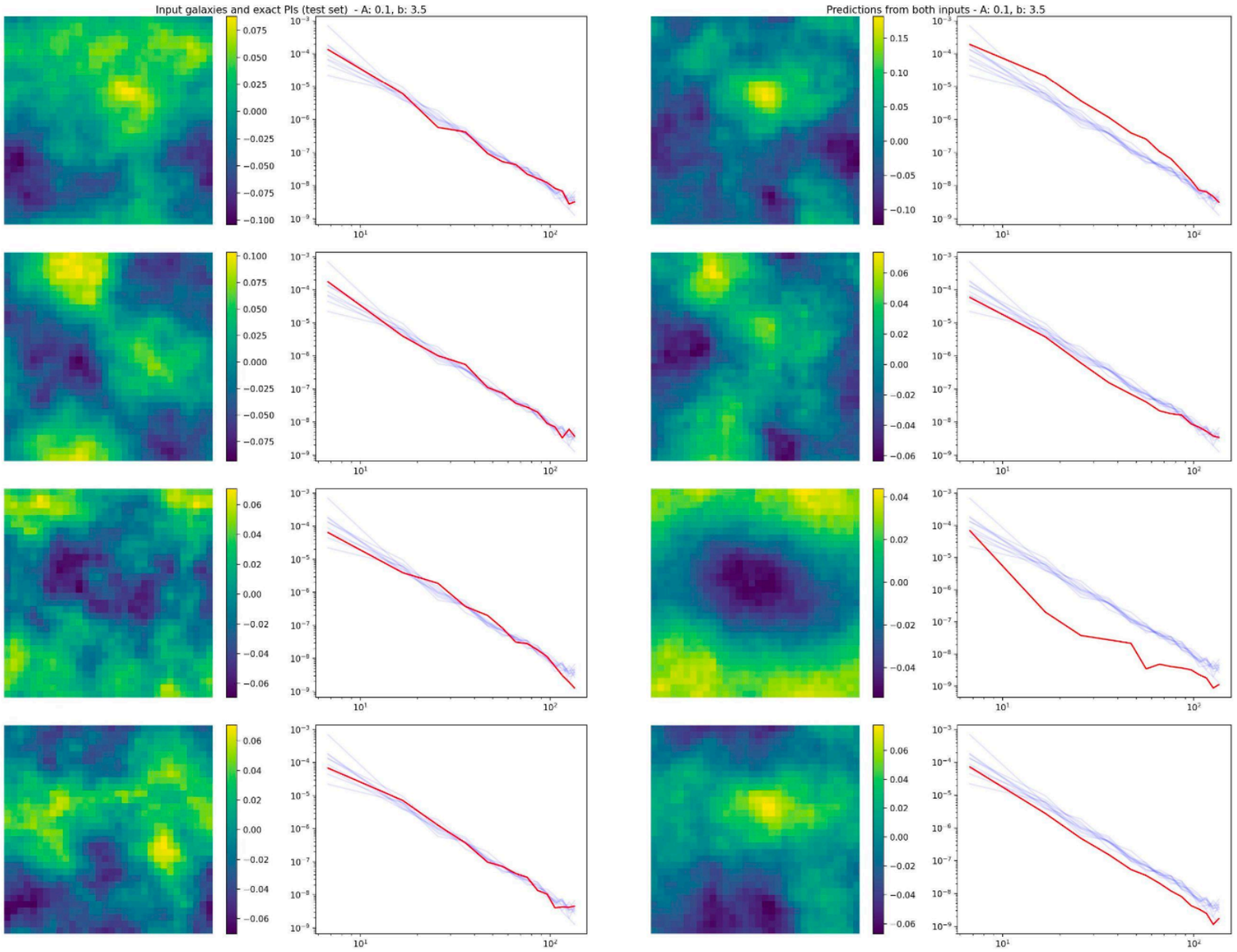


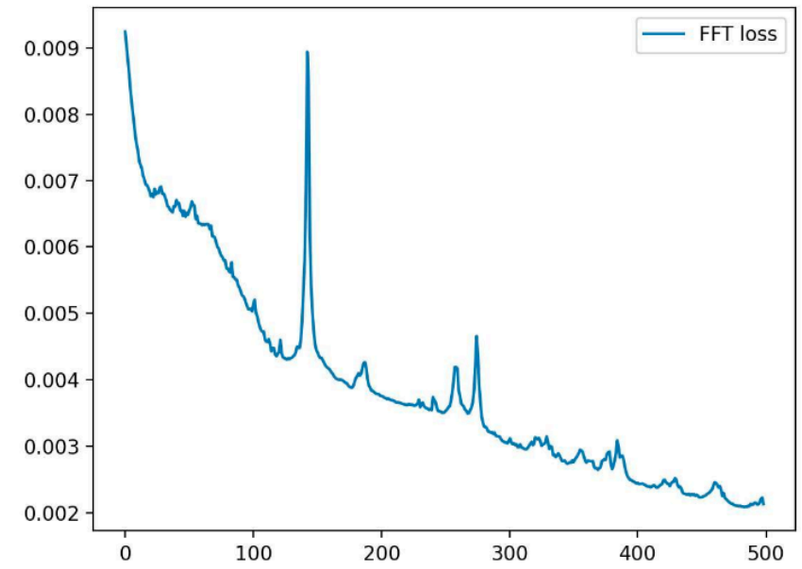


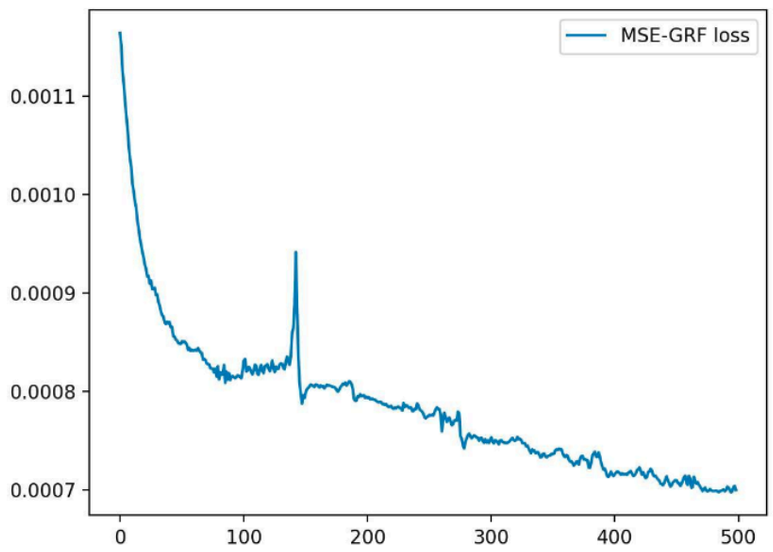


Generative models

AE: loss: MSE on both x- and k-space, A = 0.1, b =3.5, loss factor = 10^5

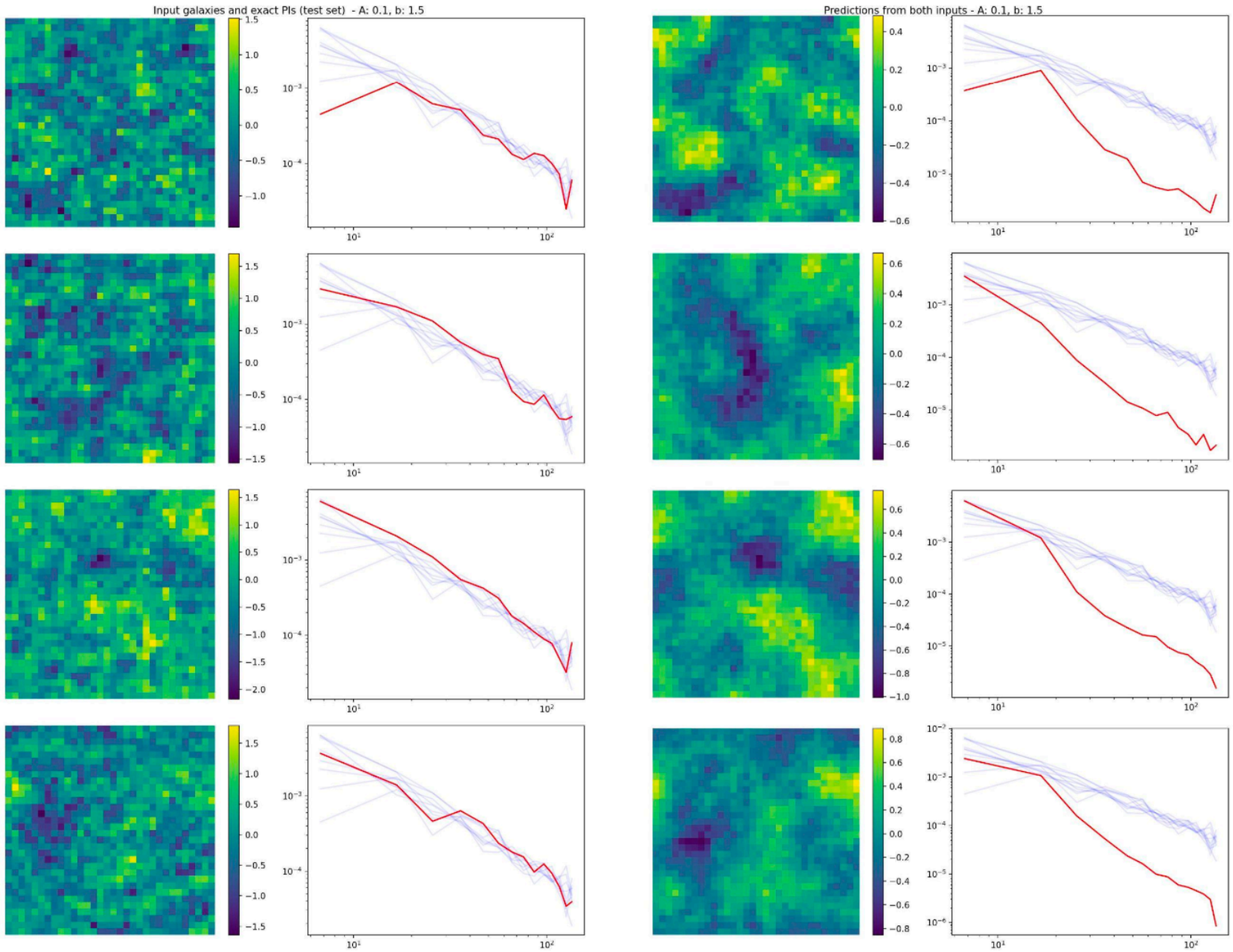


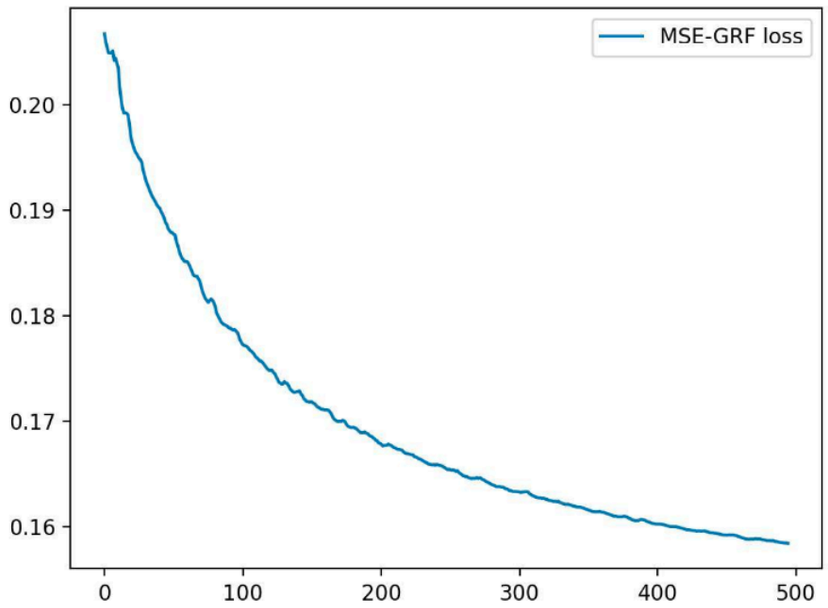




Generative models

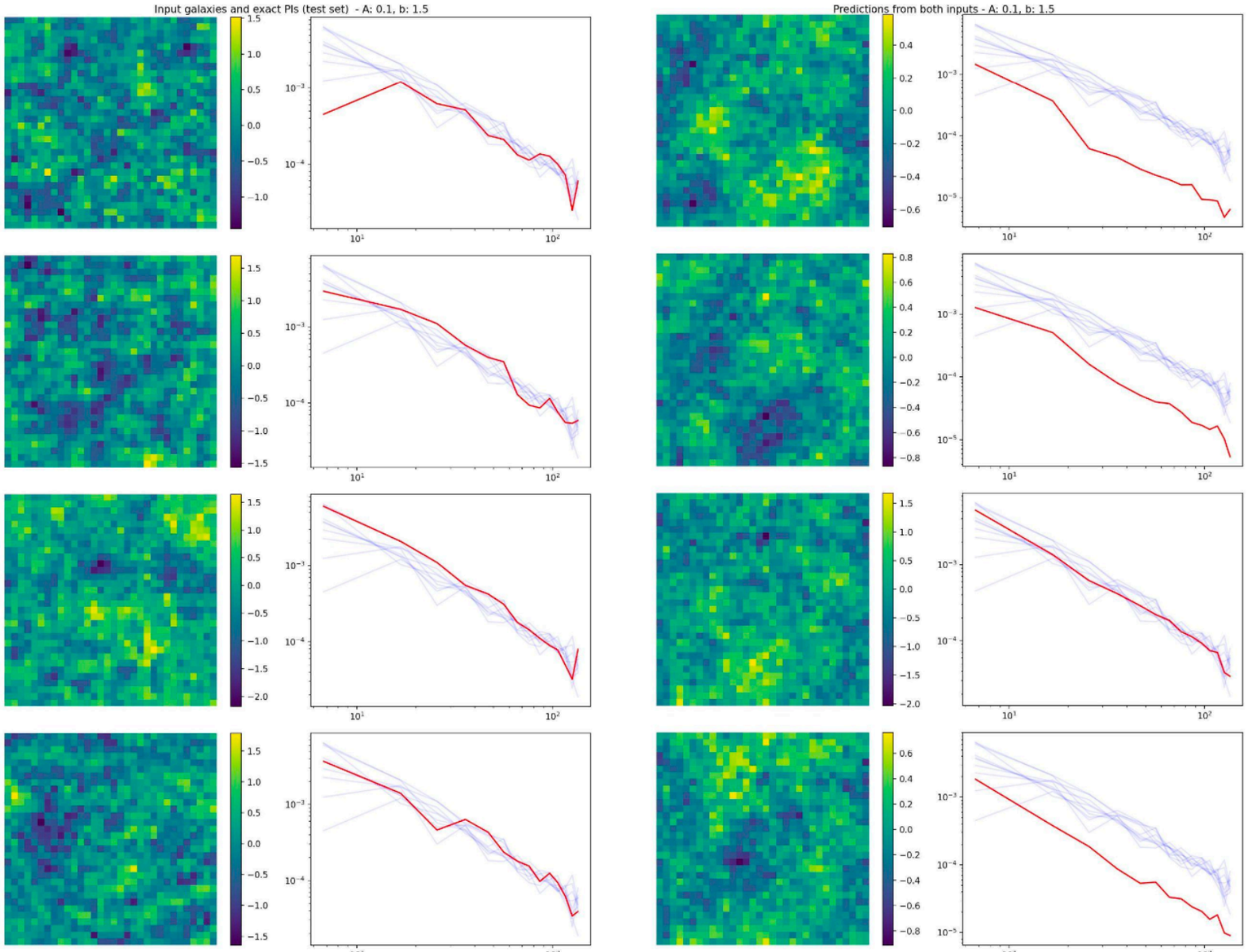
AE: loss: MSE on x-space, A = 0.1, b =1.5

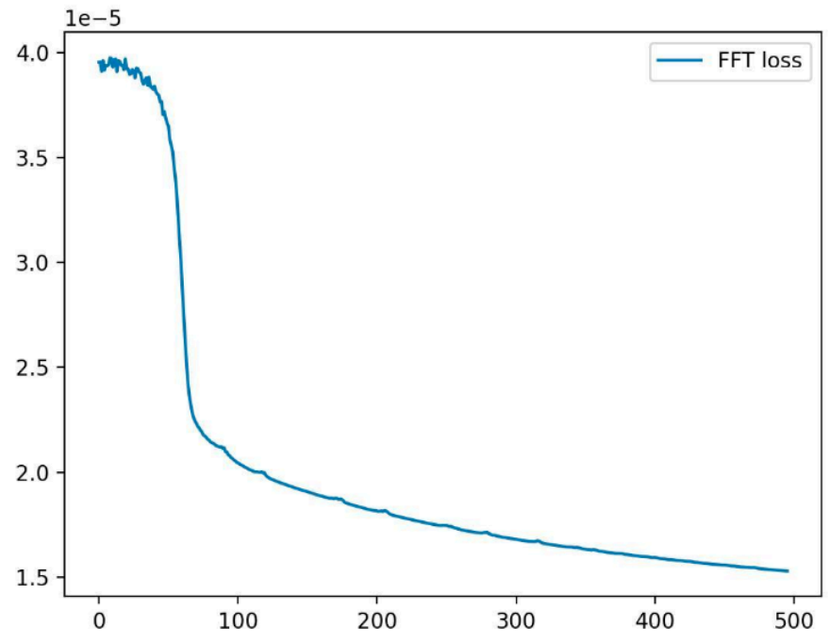




Generative models

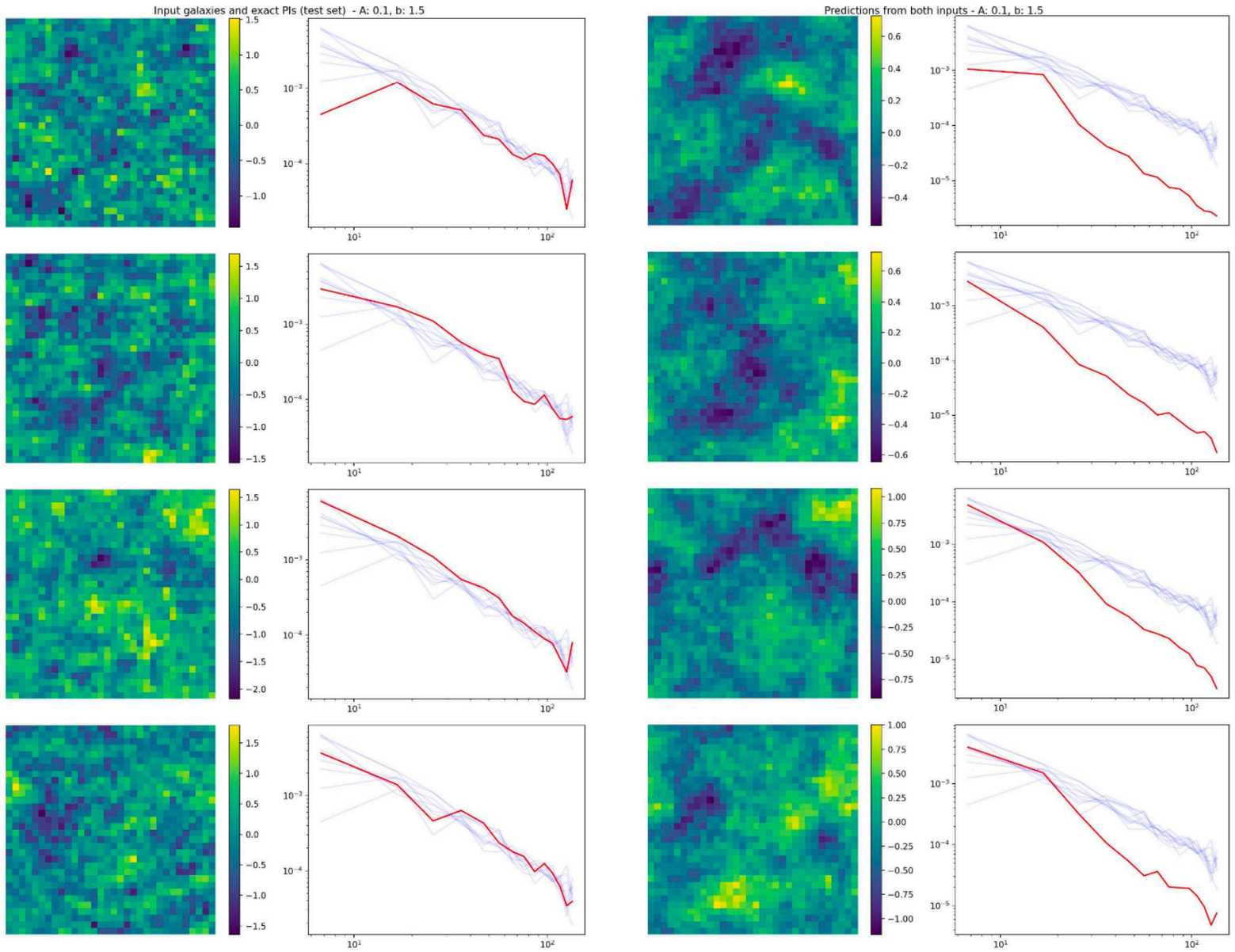
AE: loss: MSE on k-space, A = 0.1, b =1.5

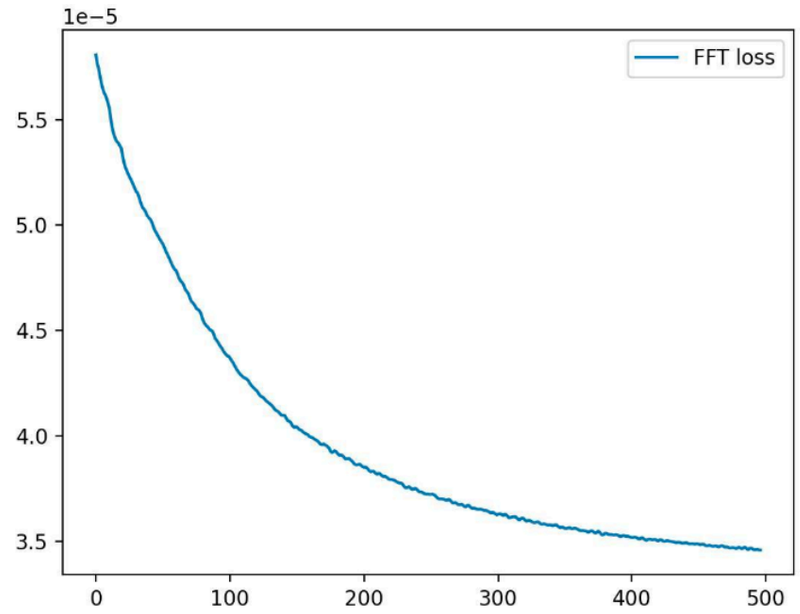


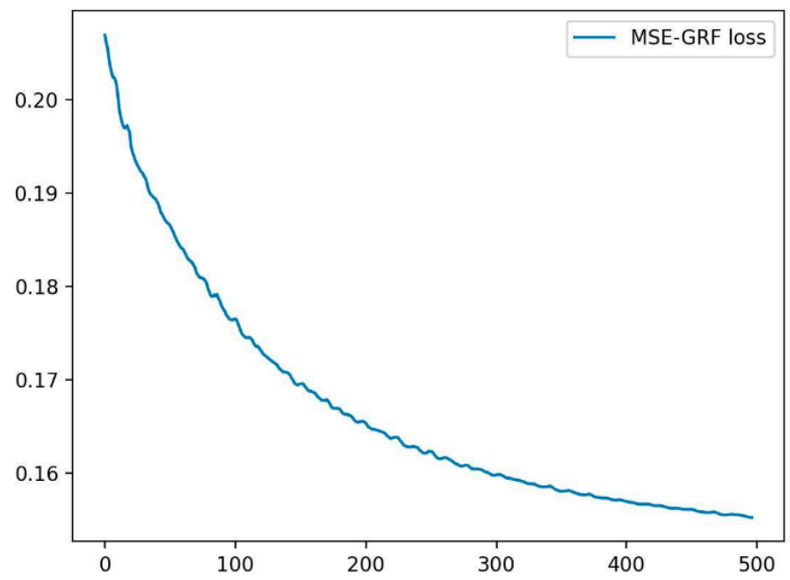


Generative models

AE: loss: MSE on both x- and k-space, A = 0.1, b =1.5, loss factor = 1.0

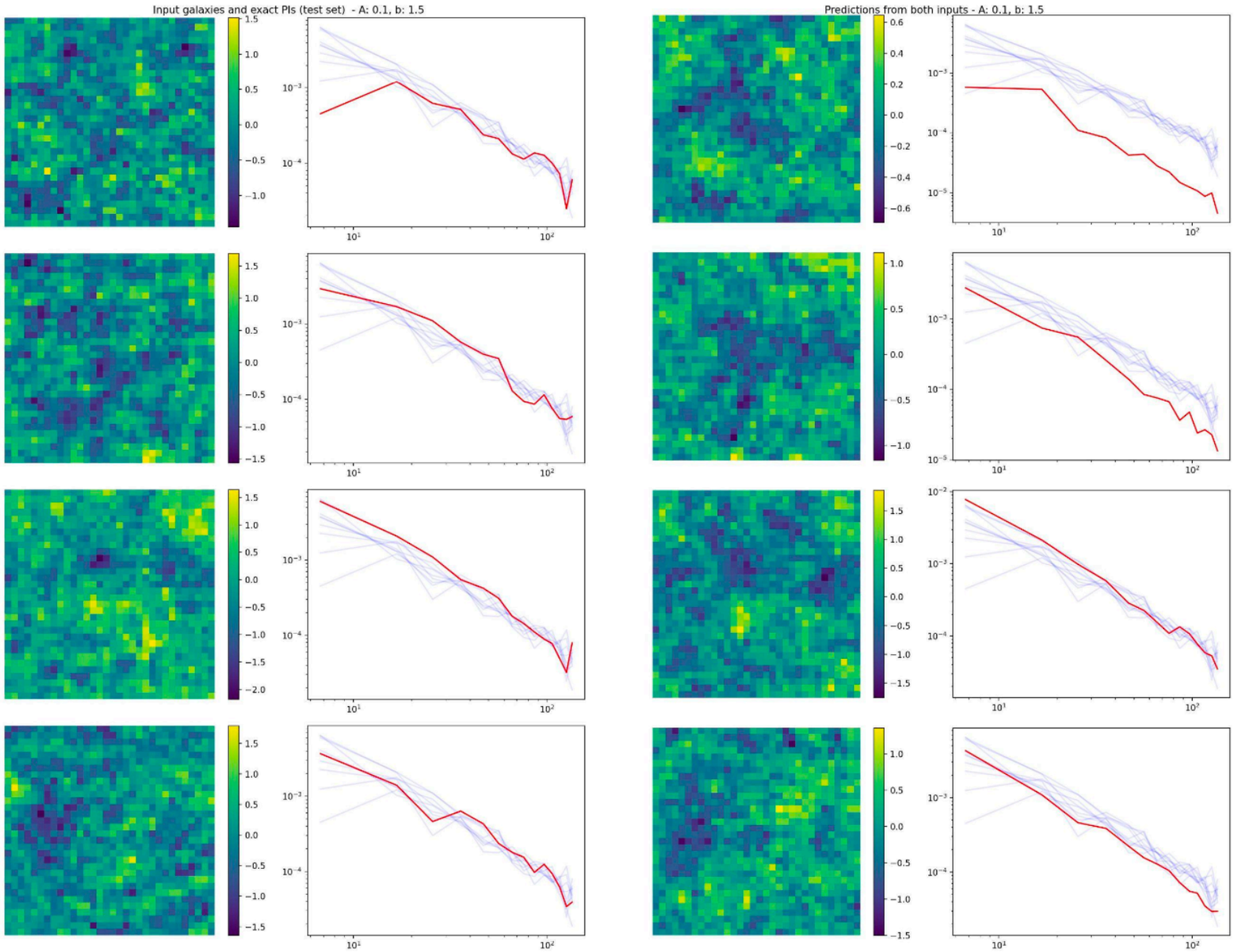


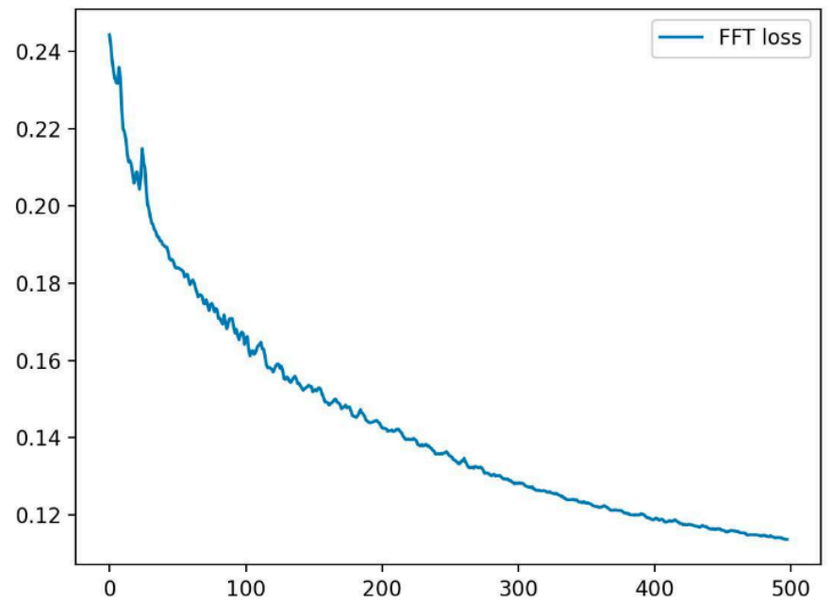


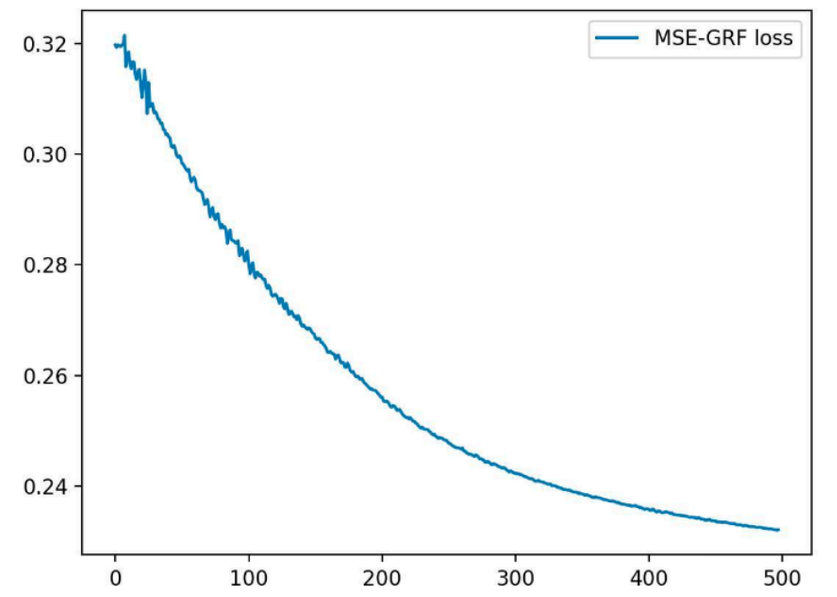


Generative models

AE: loss: MSE on both x- and k-space, A = 0.1, b =1.5, loss factor = 10^4

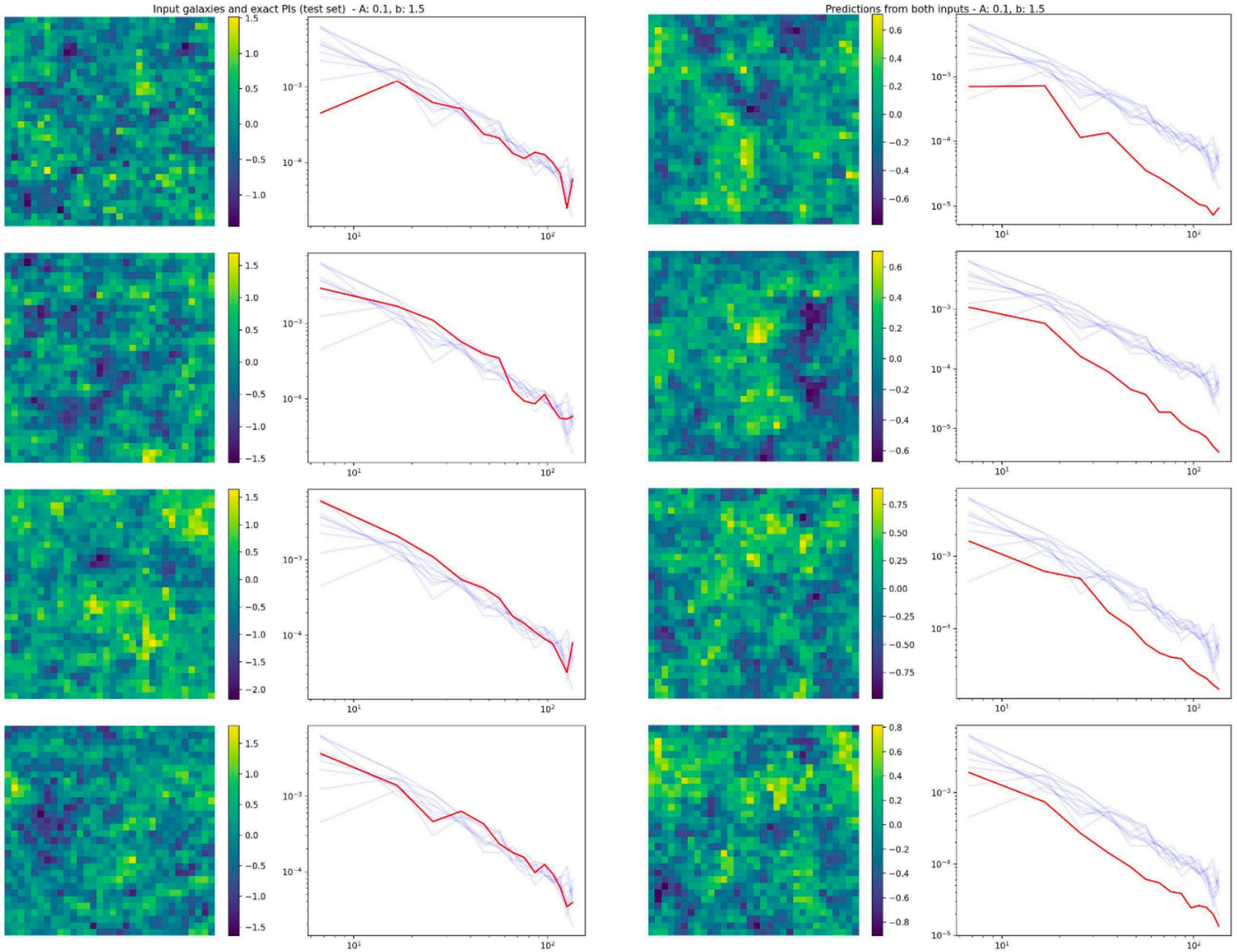


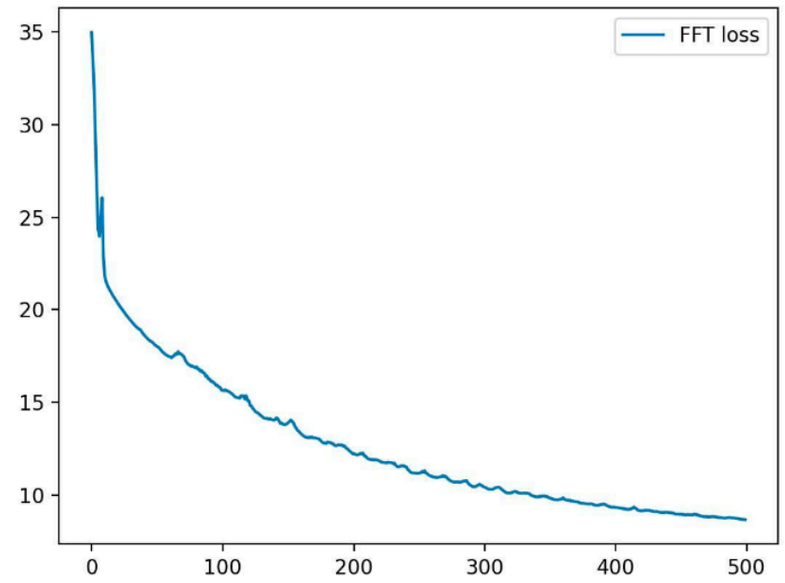


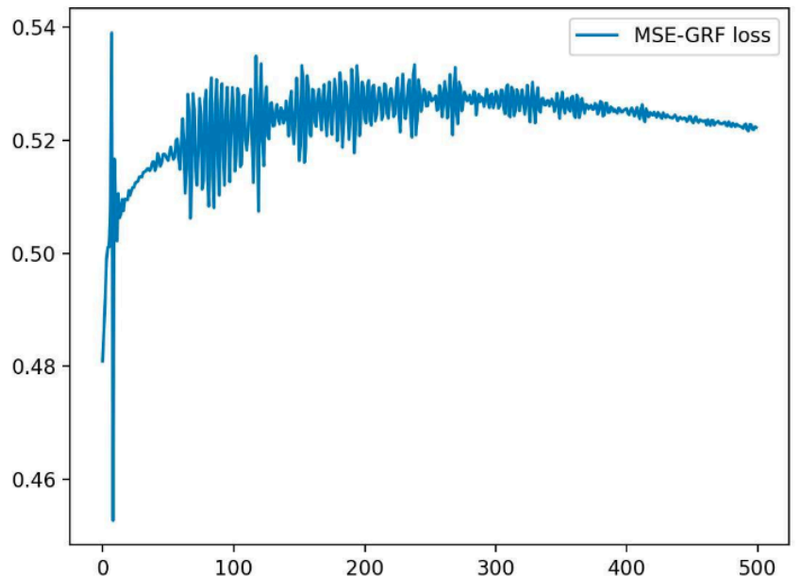


Generative models

AE: loss: MSE on both x- and k-space, A = 0.1, b =1.5, loss factor = 10^5







Persistent Homology of Cosmic Structures

Matteo Biagetti

DataShape Seminar