Dark Quest. I. Fast and Accurate Emulation of Halo Clustering Statistics and Its Application to Galaxy Clustering

based on arXiv:1811.09504; ApJ accepted + updates

TAKAHIRO NISHIMICHI (YITP, KYOTO)

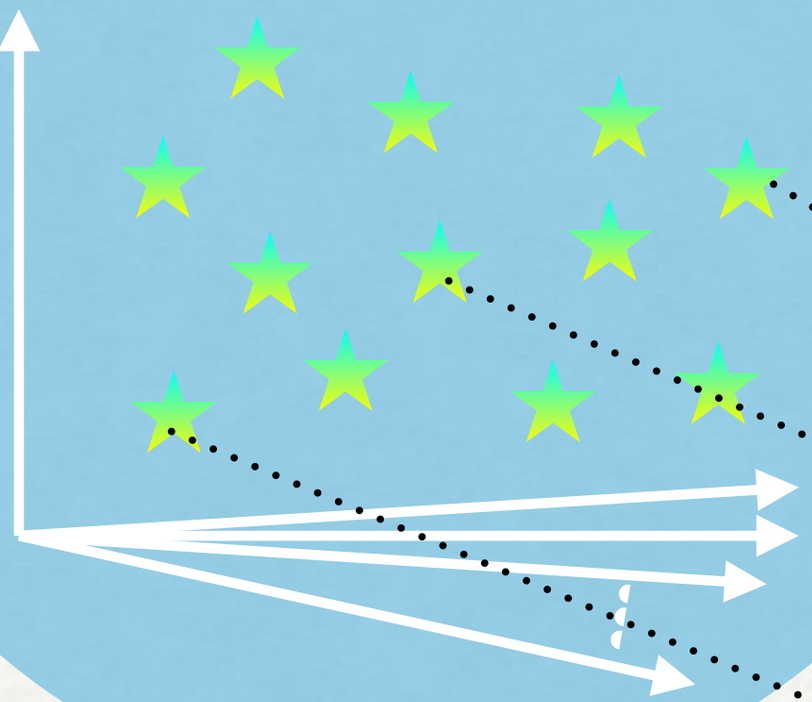
In collaboration with M. Takada, R. Takahashi, K. Osato,
M. Shirasaki, T. Oogi, H. Miyatake, M. Oguri, R. Murata,
Y. Kobayashi, N. Yoshida (HSC weak lensing WG)

15th Rencontre du Vietnam "Cosmology" @ ICISE, Quy Nhon, Vietnam, 16th Aug. 2019

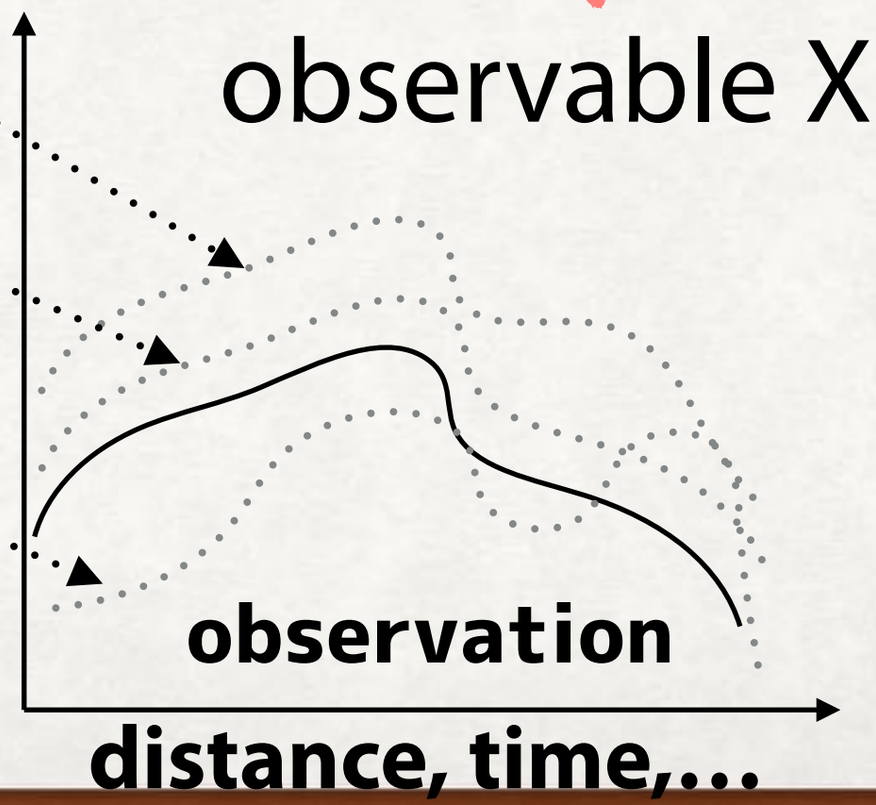
THE GAME

- Efficiently cover the space
- Controlled error

Multi-D cosmological parameter / model space

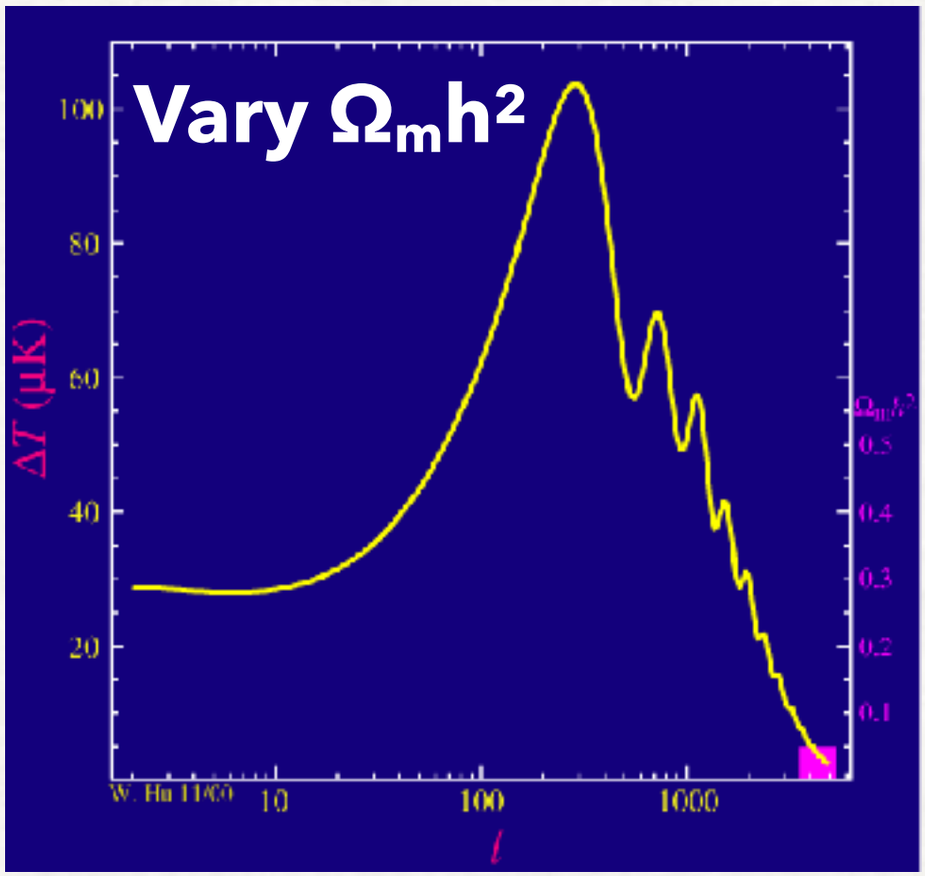


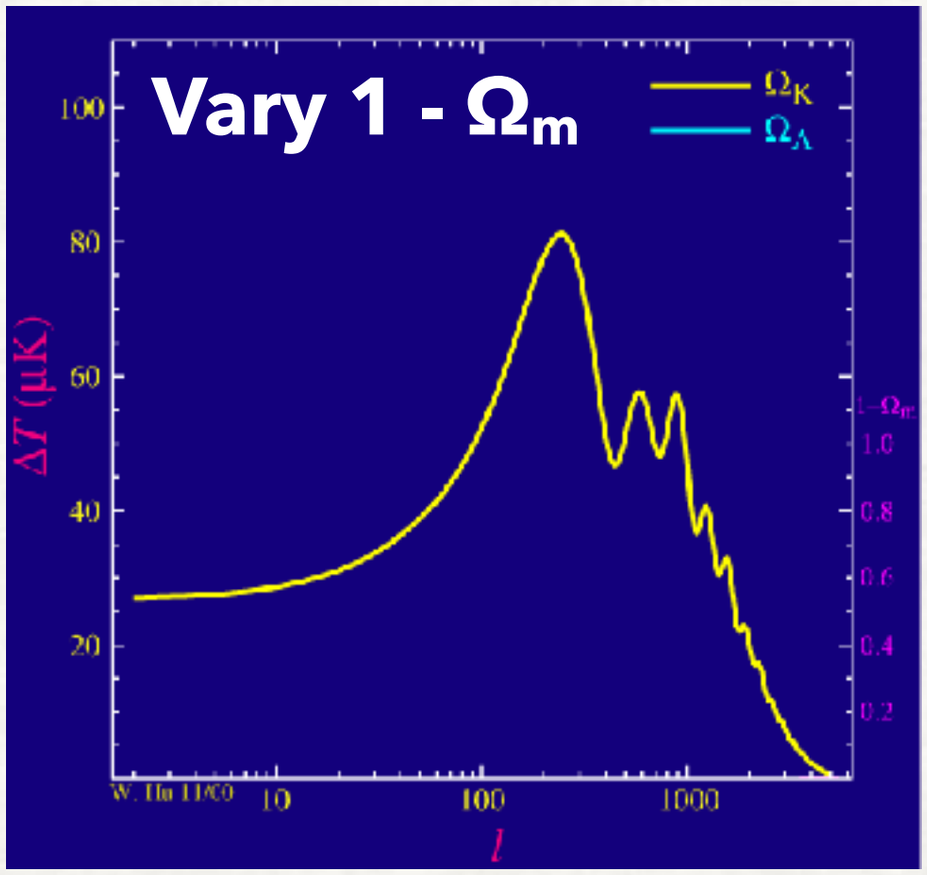
Forward modeling



LSS JUST LIKE CMB…?

- This is quite well established in case of CMB analyses.
  - Thanks to the smallness of fluctuations -> linear Boltzmann
  - Typically Markov-Chain Monte Carlo works comfortably
- Can we do the same for **large scale structure**?

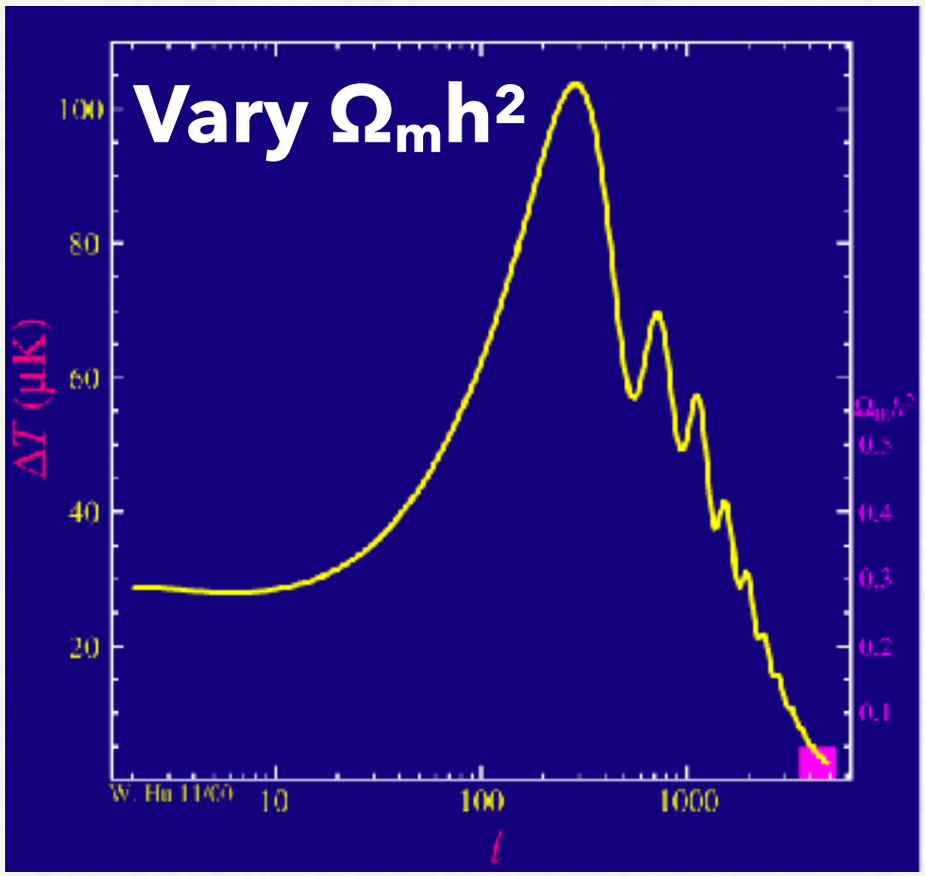


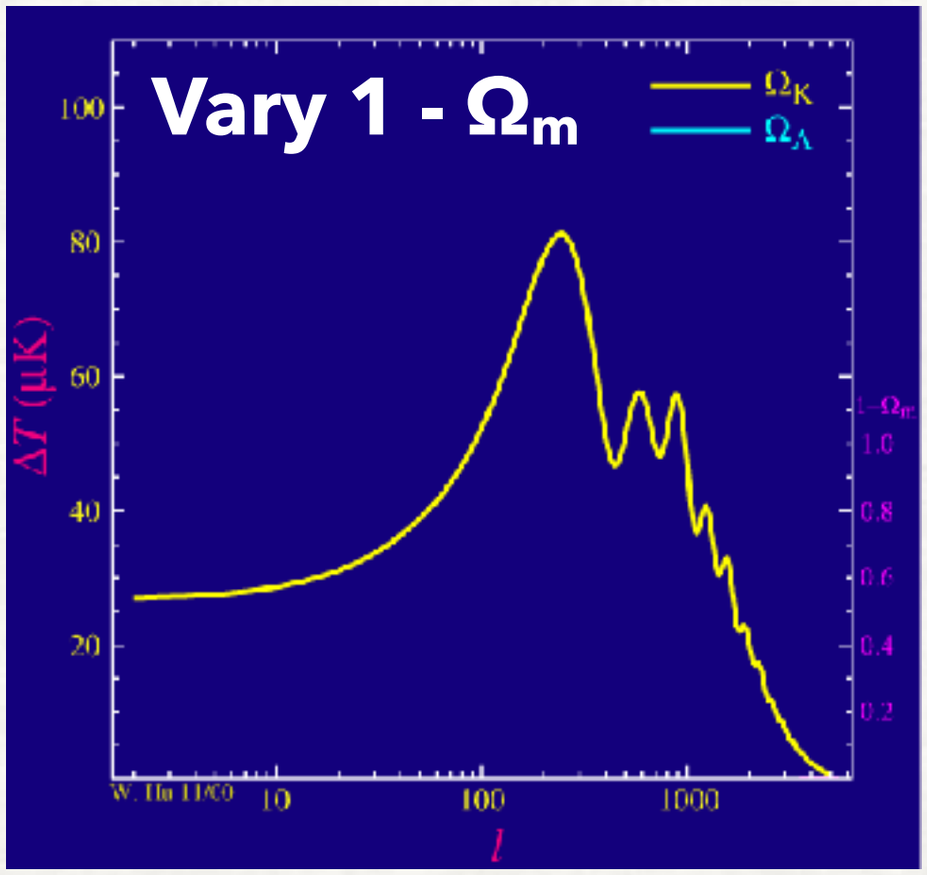


http://background.uchicago.edu/~whu © Wayne Hu

LSS JUST LIKE CMB…?

- This is quite well established in case of CMB analyses.
  - Thanks to the smallness of fluctuations -> linear Boltzmann
  - Typically Markov-Chain Monte Carlo works comfortably
- Can we do the same for **large scale structure**?

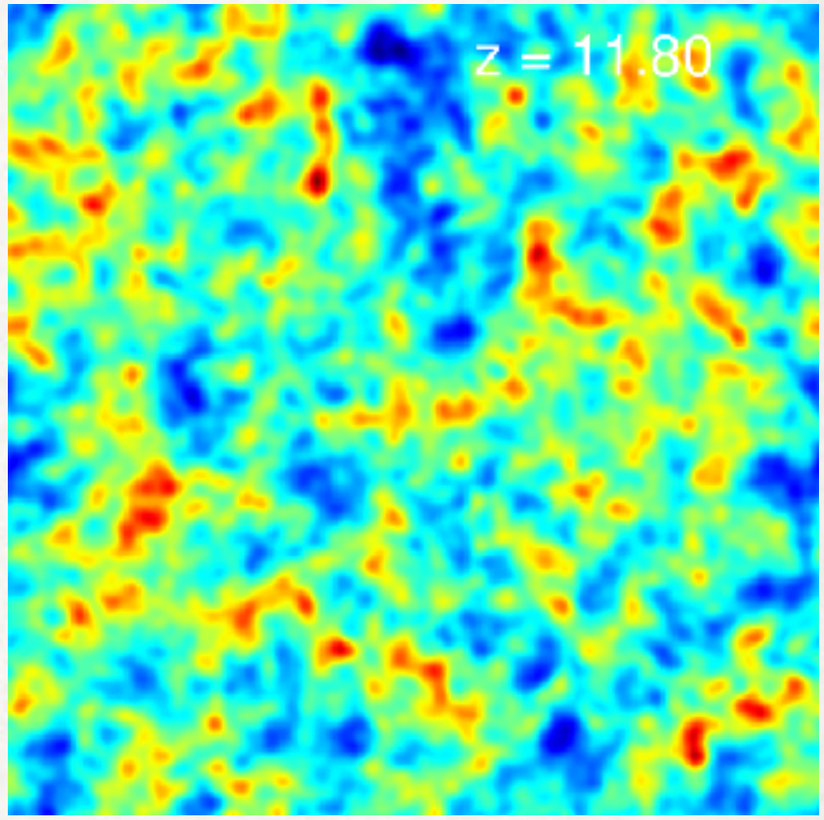


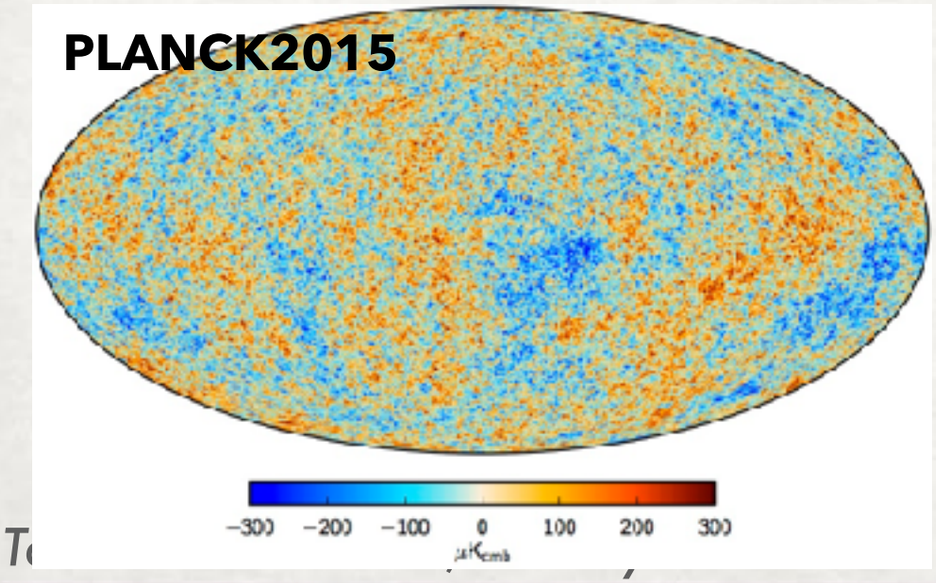


http://background.uchicago.edu/~whu © Wayne Hu

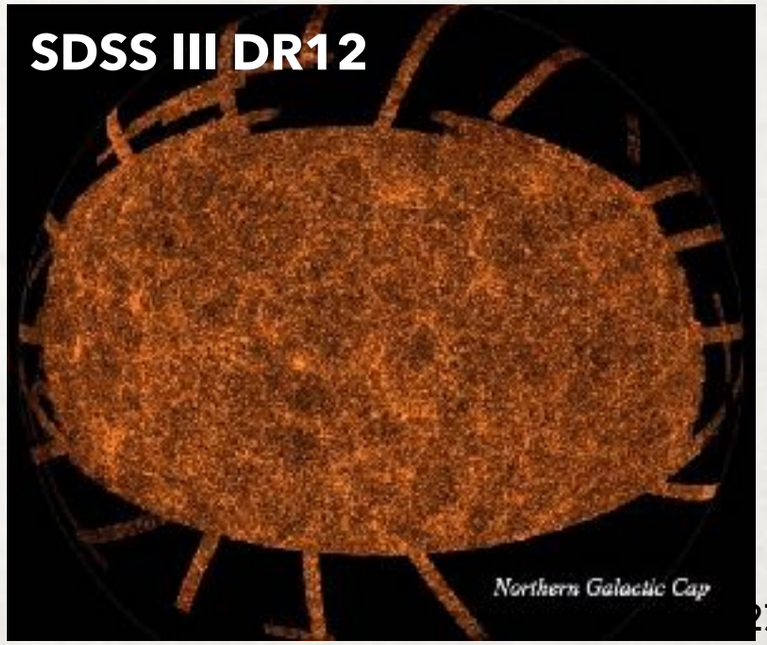
LARGE SCALE STRUCTURE

- Complementarities to CMB
  - Dark energy dominates the nearby universe
    - Equation of state?
  - Gravity is the driver of structure growth
    - Test of GR?
- **Understanding nonlinearity is the key**



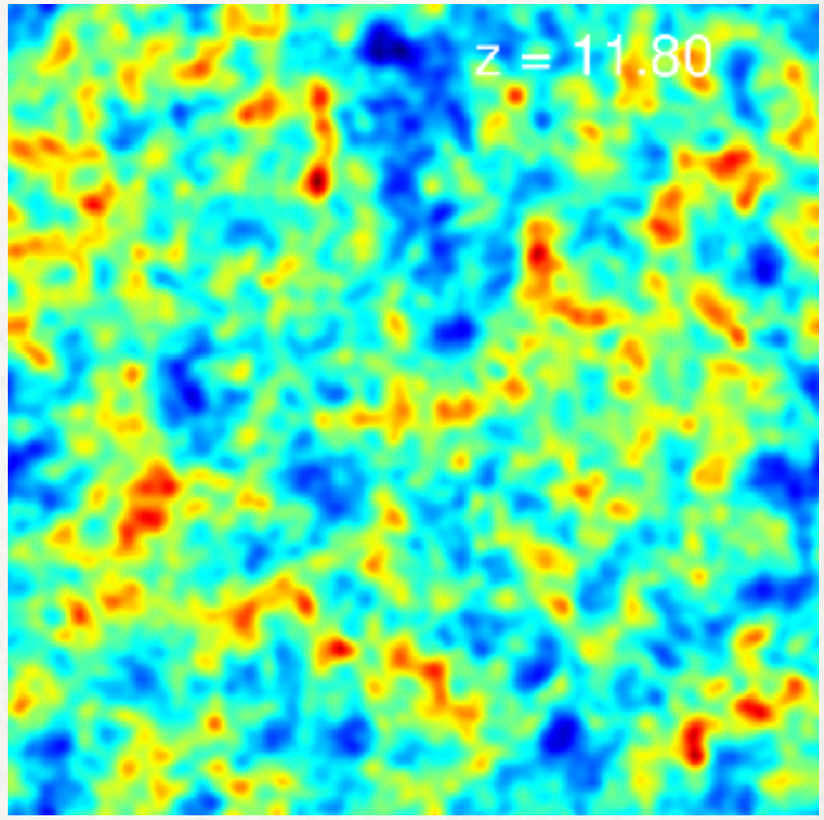


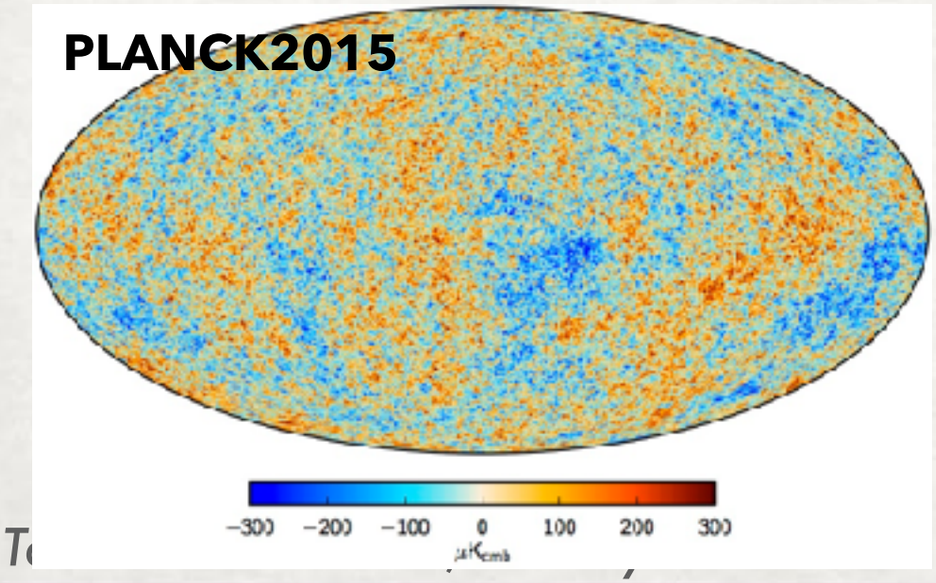
gravity !



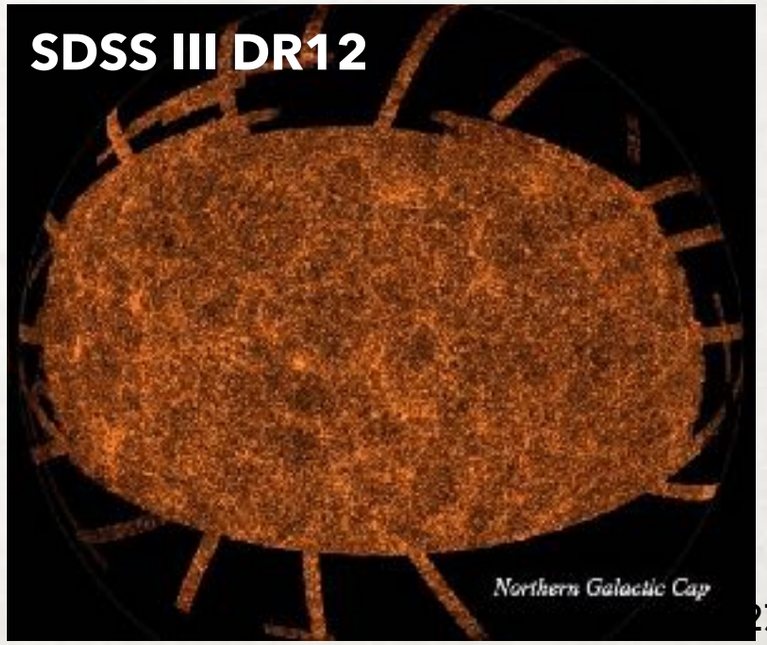
LARGE SCALE STRUCTURE

- Complementarities to CMB
  - Dark energy dominates the nearby universe
    - Equation of state?
  - Gravity is the driver of structure growth
    - Test of GR?
- **Understanding nonlinearity is the key**



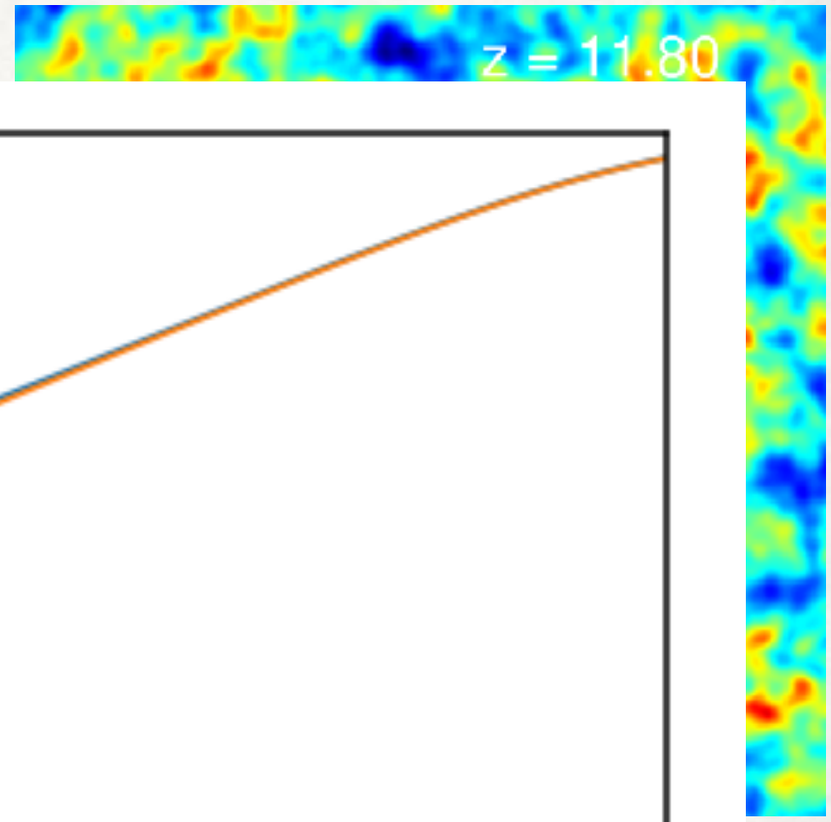


gravity !



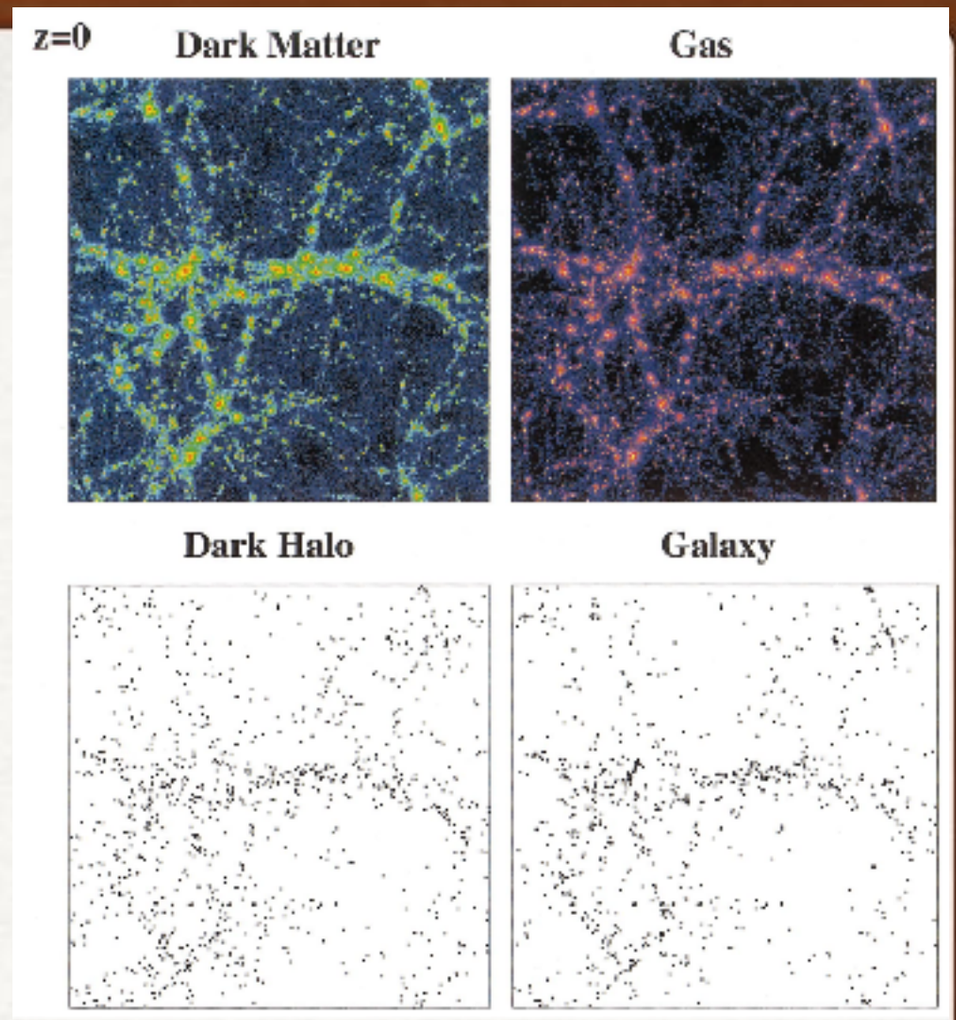
LARGE SCALE STRUCTURE

- Same phase as initial conditions in CMB



BIAS IS ANNOYING

- Luminous things are "tracers" of the underlying matter field (Kaiser '84)
- No first-principle analytical approach available (but hydro sims)
- Have to introduce many (really many!!) nuisance parameters?



Yoshikawa+'01

The Galaxy Power Spectrum and Bispectrum in Redshift Space

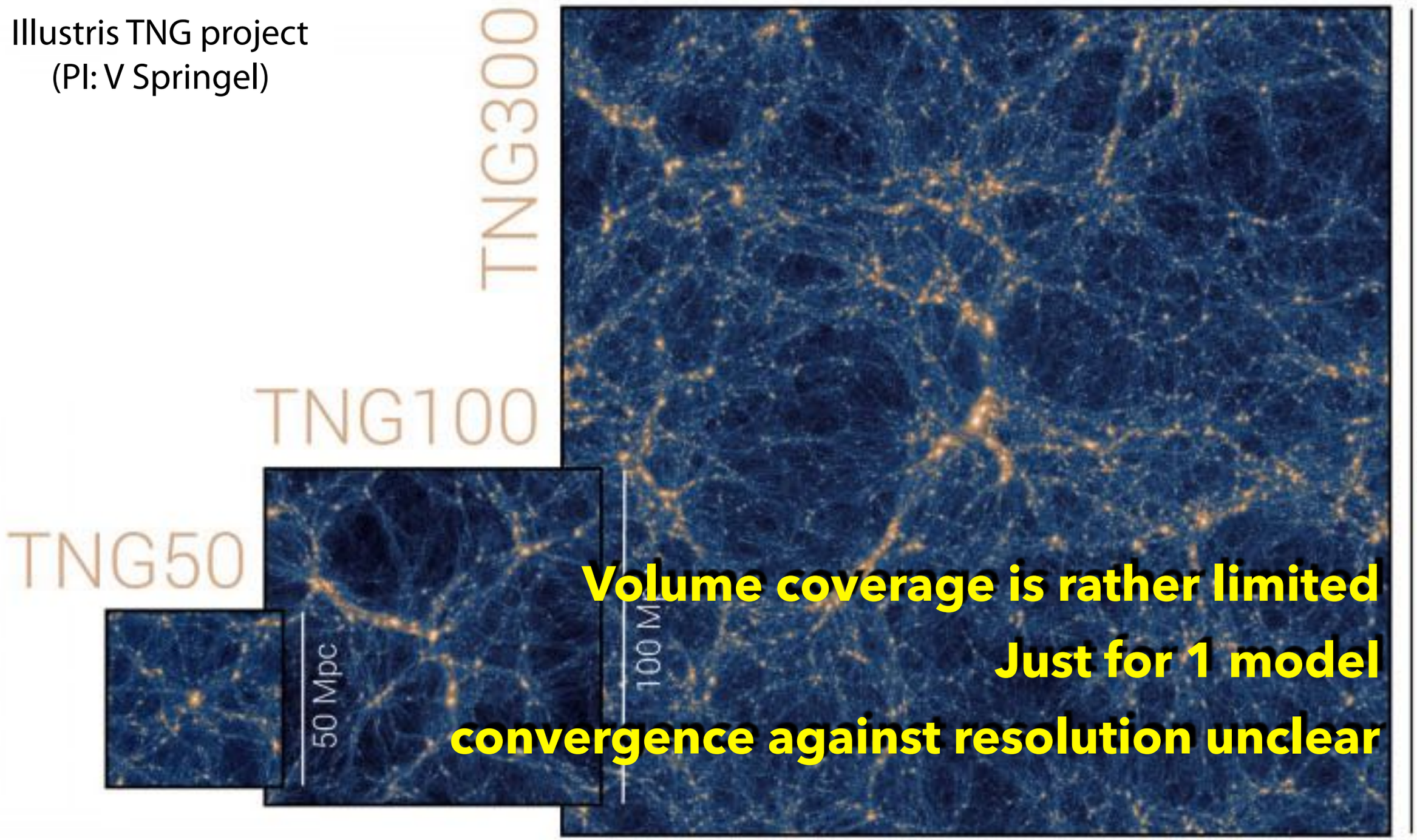
Vincent Desjacques, Donghui Jeong, Fabian Schmidt

(Submitted on 11 Jun 2018)

We present the complete expression for the next-to-leading (1-loop) order galaxy power spectrum and the leading-o space in the general bias expansion, or equivalently the effective field theory of biased tracers. We consistently includ selection effects. These are degenerate with many, but not all, of the redshift-space distortion contributions, and hav before. Moreover, we show that, in the framework of effective field theory, a consistent bias expansion in redshift spa contributions. Physical arguments about the tracer sample considered and its observational selection have to be used contributions. In summary, the next-to-leading order galaxy power spectrum and leading-order galaxy bispectrum i described by 22 parameters, which reduces to 11 parameters if selection effects can be neglected. All contributions t in terms of 28 independent loop integrals.

BIAS IS DIFFICULT

Illustris TNG project
(PI: V Springel)



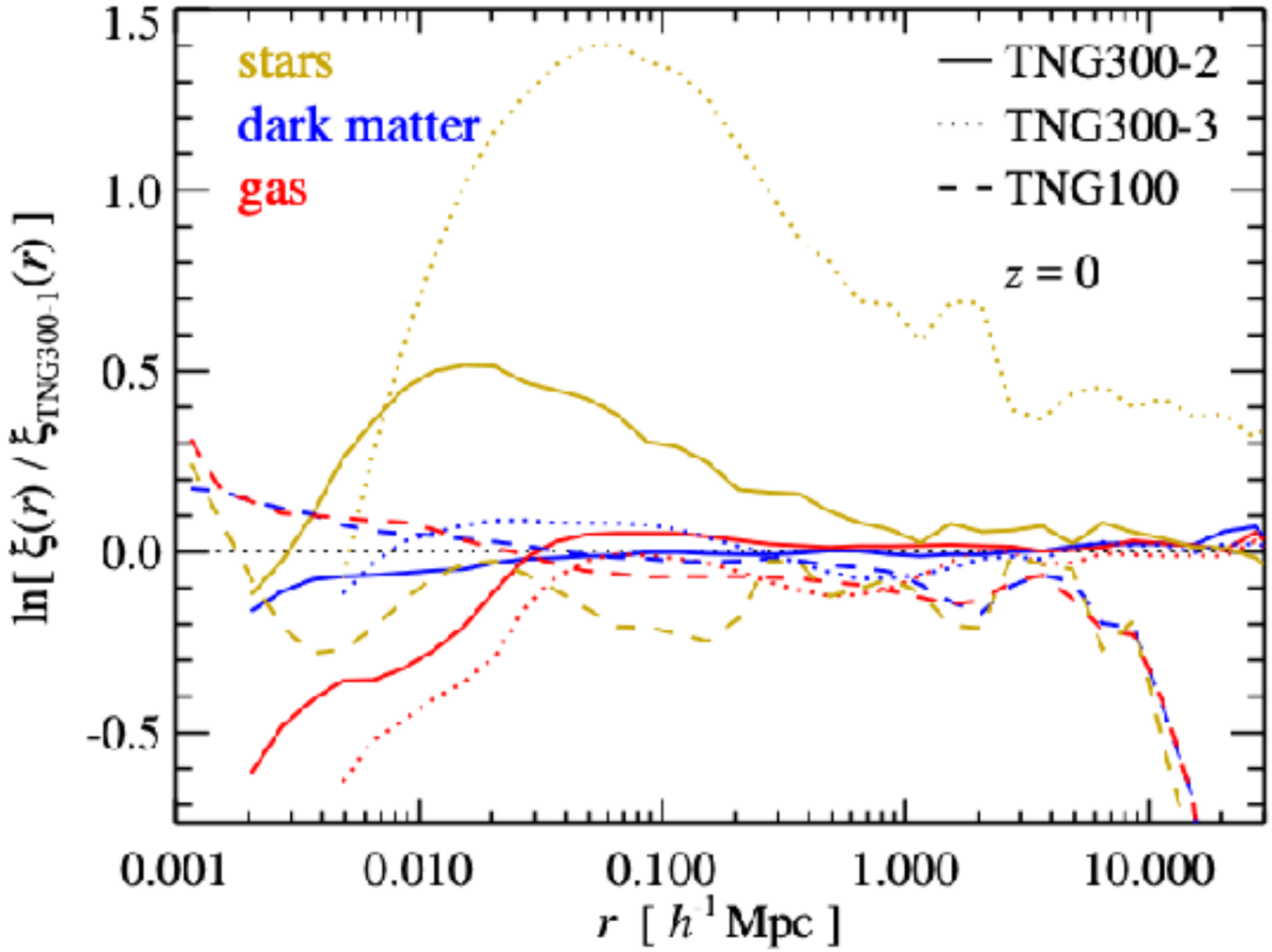
Volume coverage is rather limited

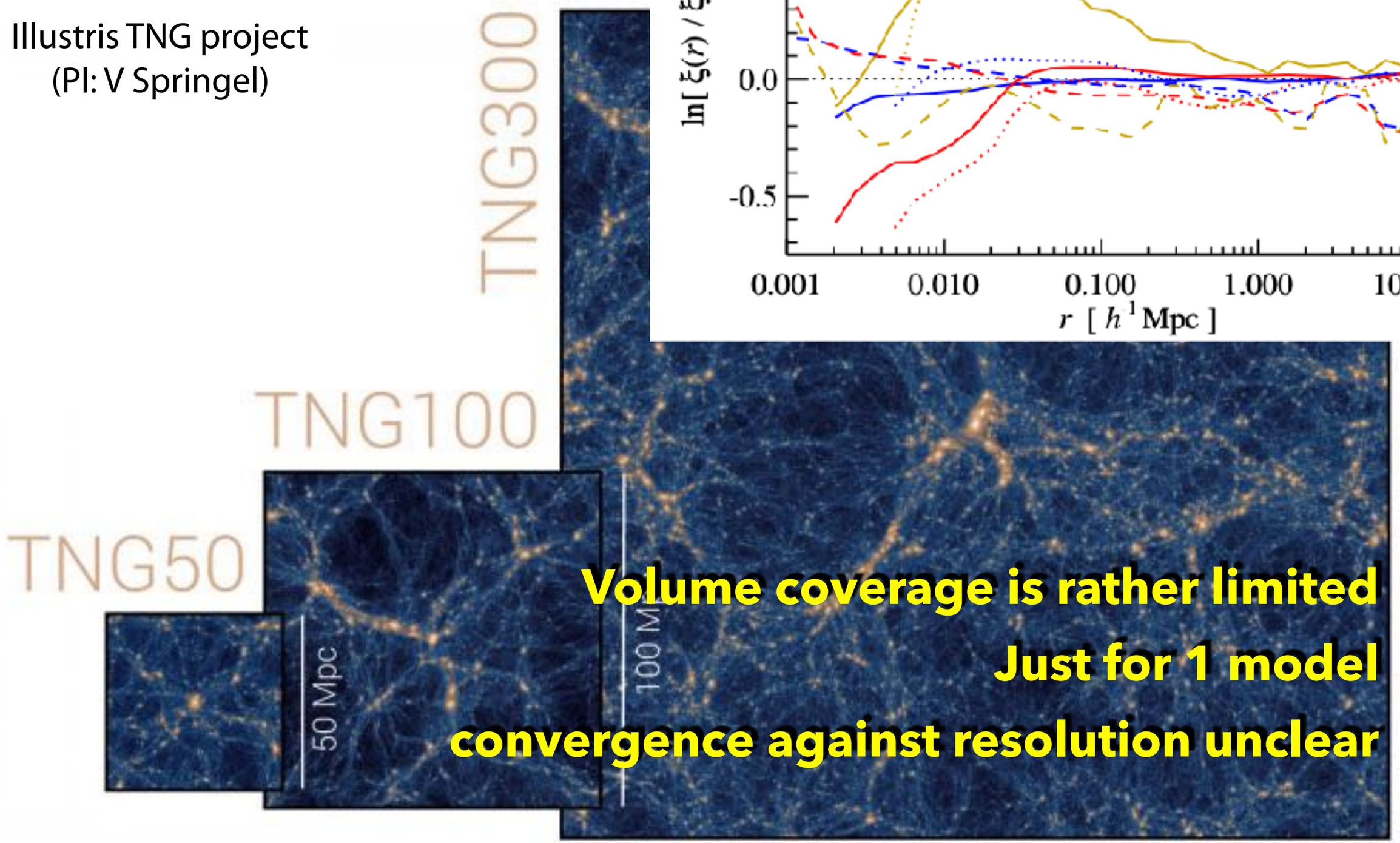
Just for 1 model

convergence against resolution unclear

BIAS IS DIFFICULT

Illustris TNG project
(PI: V Springel)





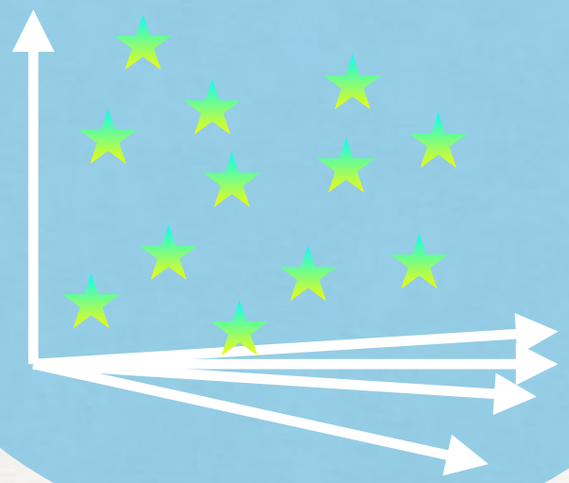
Volume coverage is rather limited

Just for 1 model

convergence against resolution unclear

HOW TO DO IT?

Cosmological parameters



Ω_m, Ω_b, H_0, A_s, n_s, w, …

If you are a big fun of halo approach …

HOD parameters

N(M): M_{min}, M_1, α, …

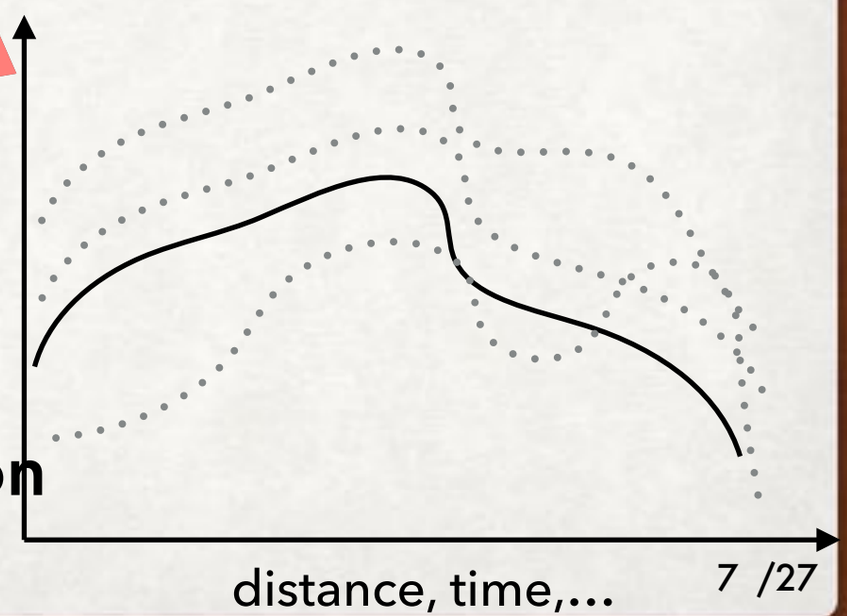
Halos

Statistics of halos based on N-body and Machine Learning

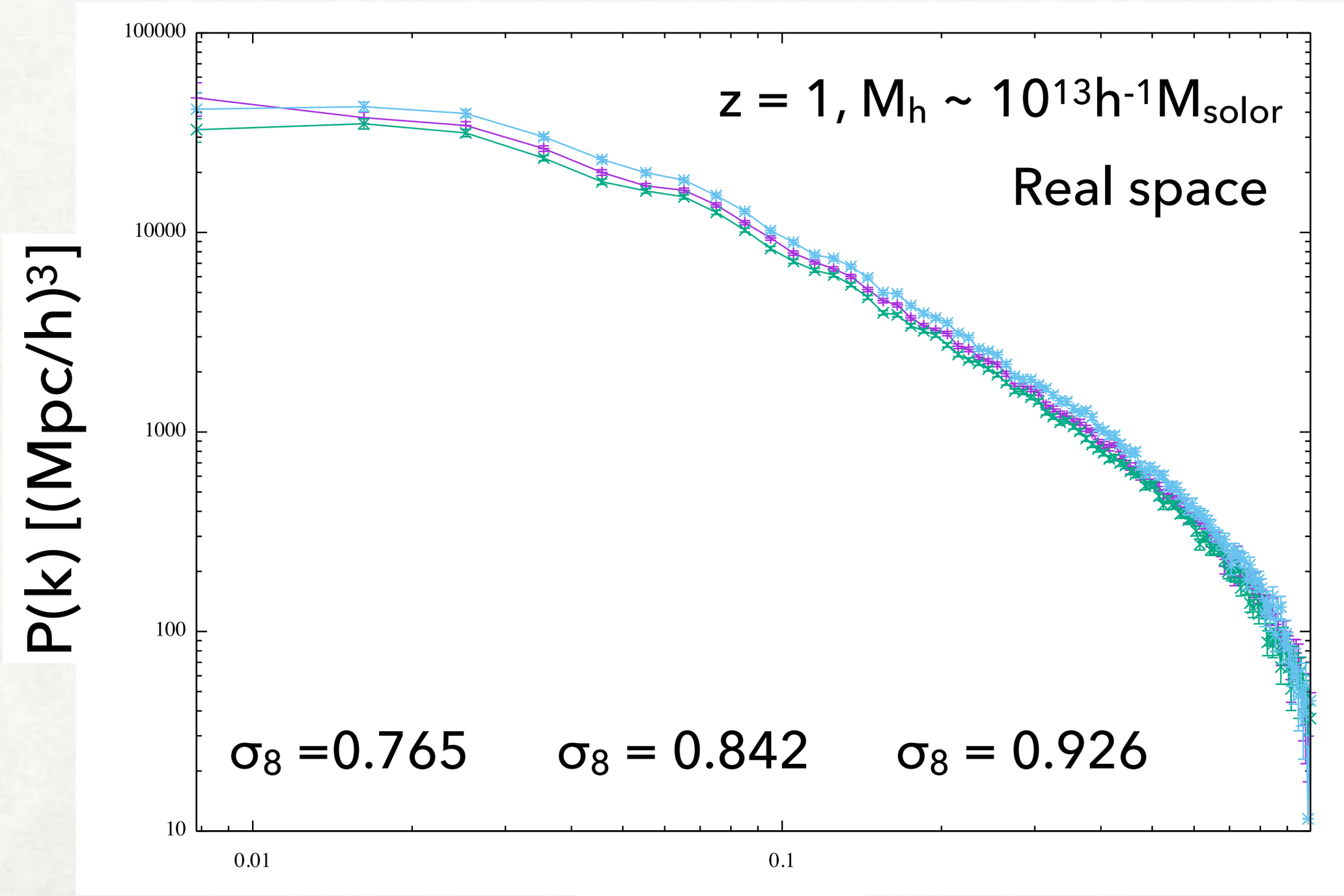
Analytical calculation

Can be updated if you want, e.g., off-centering, incompleteness, or more complex parameterization

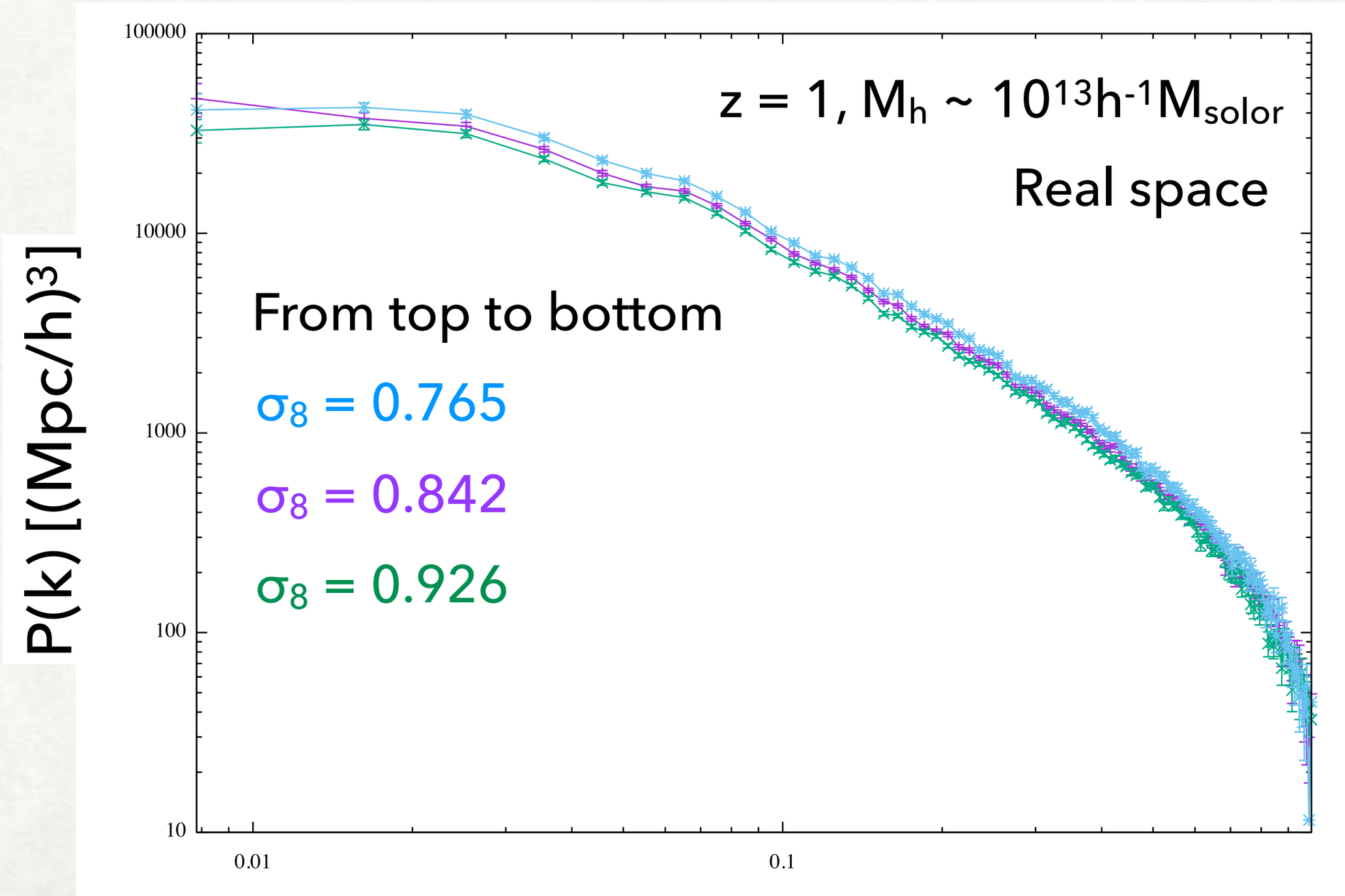
Galaxy observables



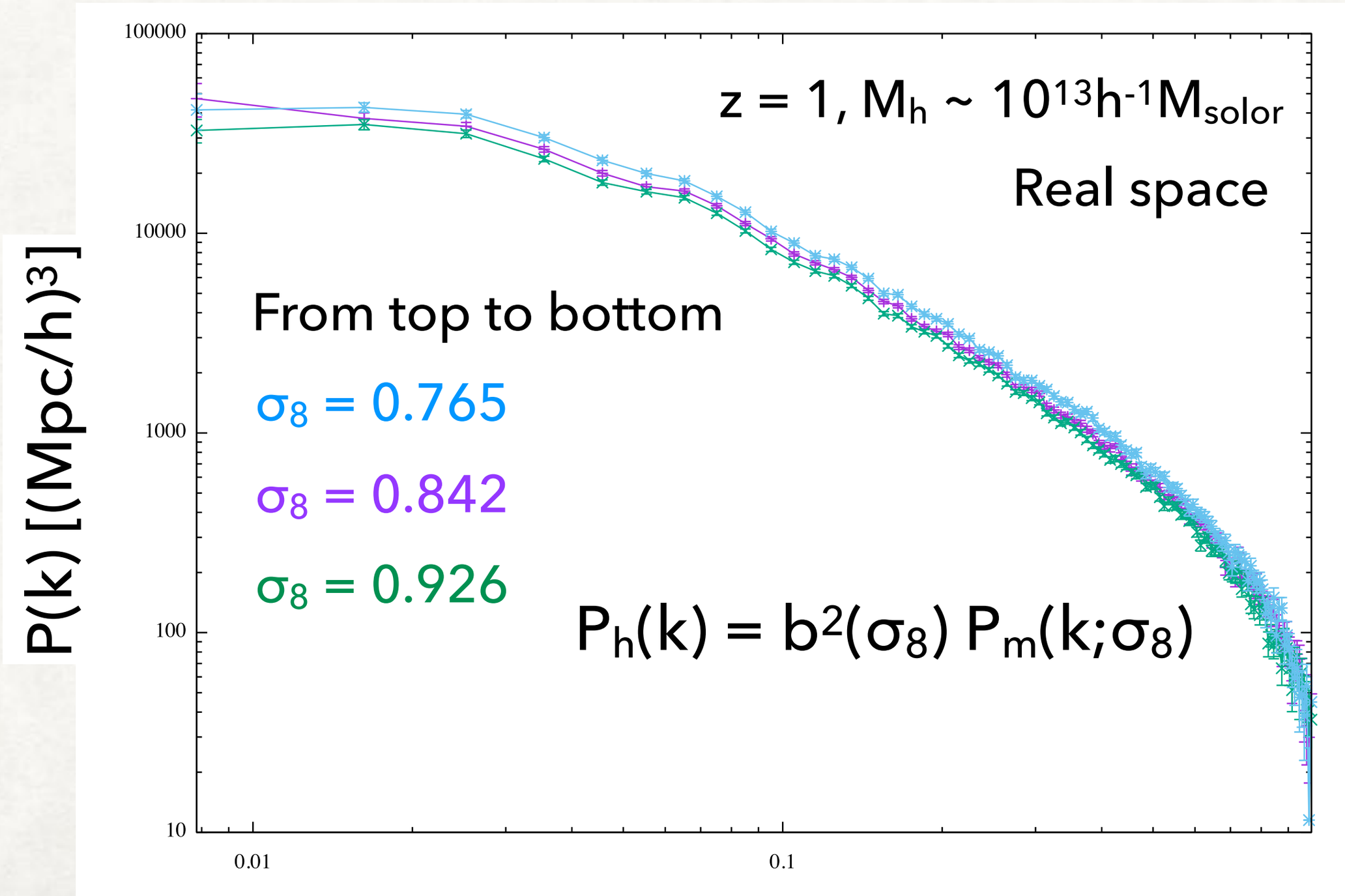
BIAS IS STILL TRICKY
(EVEN AT THE LEVEL OF HALOS AND AT LINEAR SCALE)



BIAS IS STILL TRICKY
(EVEN AT THE LEVEL OF HALOS AND AT LINEAR SCALE)



BIAS IS STILL TRICKY
(EVEN AT THE LEVEL OF HALOS AND AT LINEAR SCALE)

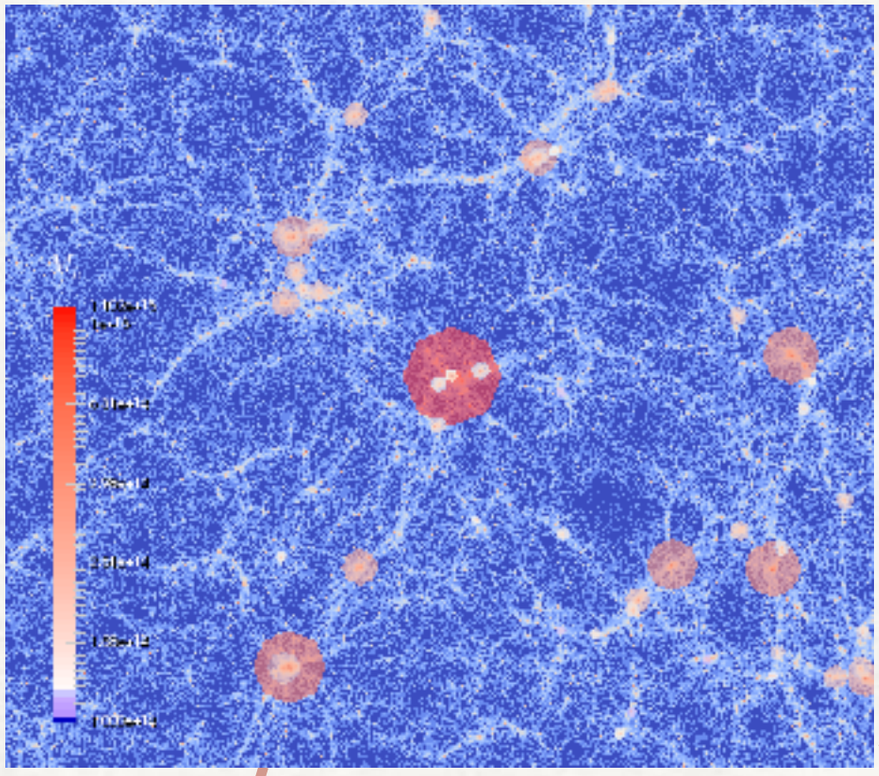


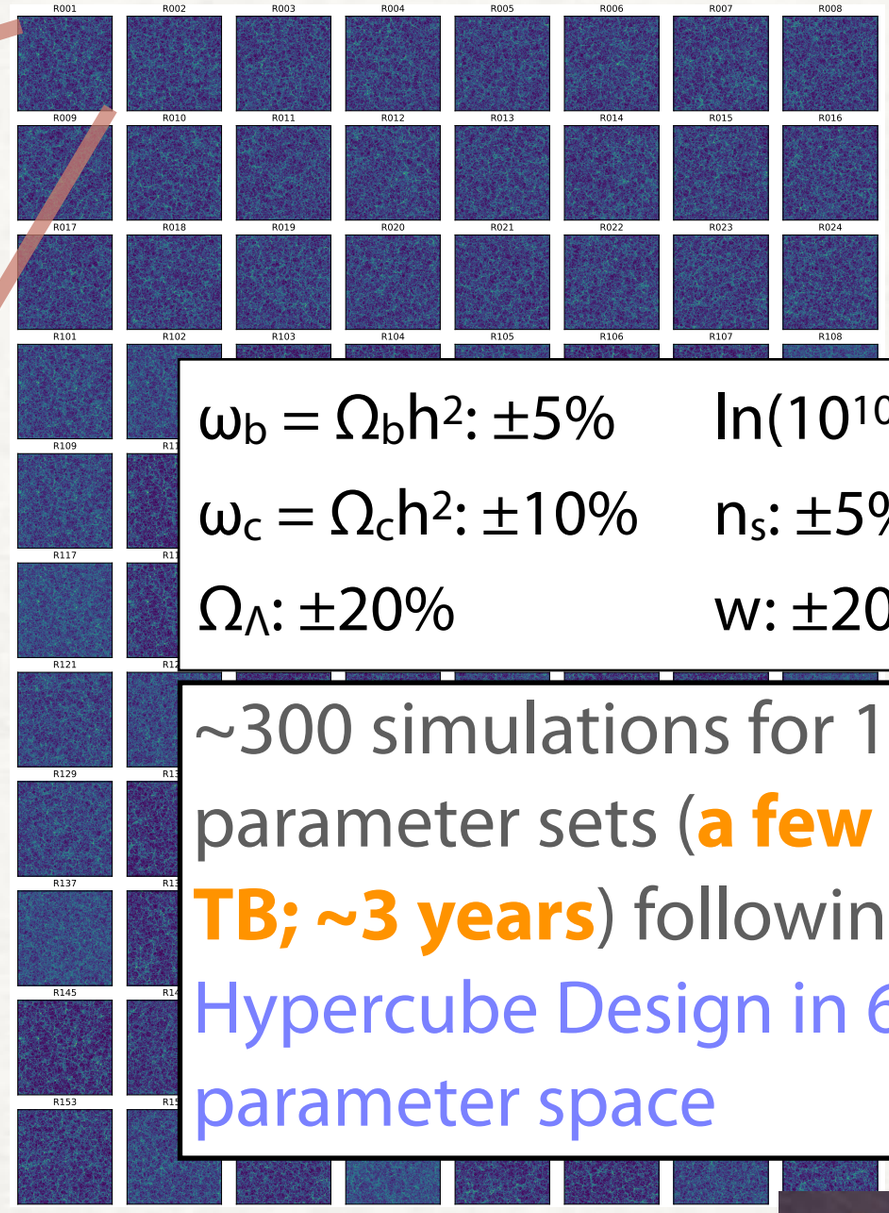
DARK QUEST PROJECT

arXiv:1811.09504

COSMIC STRUCTURE SIM ENSEMBLE IN 6D PARAM SPACE

Zoom of a 2048^3-body simu





$\omega_b = \Omega_b h^2$: ±5%

$\omega_c = \Omega_c h^2$: ±10%

Ω_Λ: ±20%

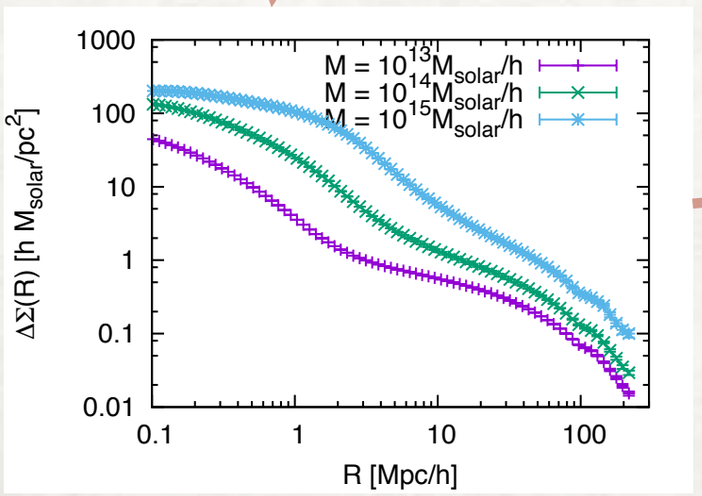
$\ln(10^{10} A_s)$: ±20%

n_s: ±5%

w: ±20%

~300 simulations for 100 parameter sets (**a few hundred TB; ~3 years**) following a Latin Hypercube Design in 6D parameter space

Extract halo statistics



Machine learning

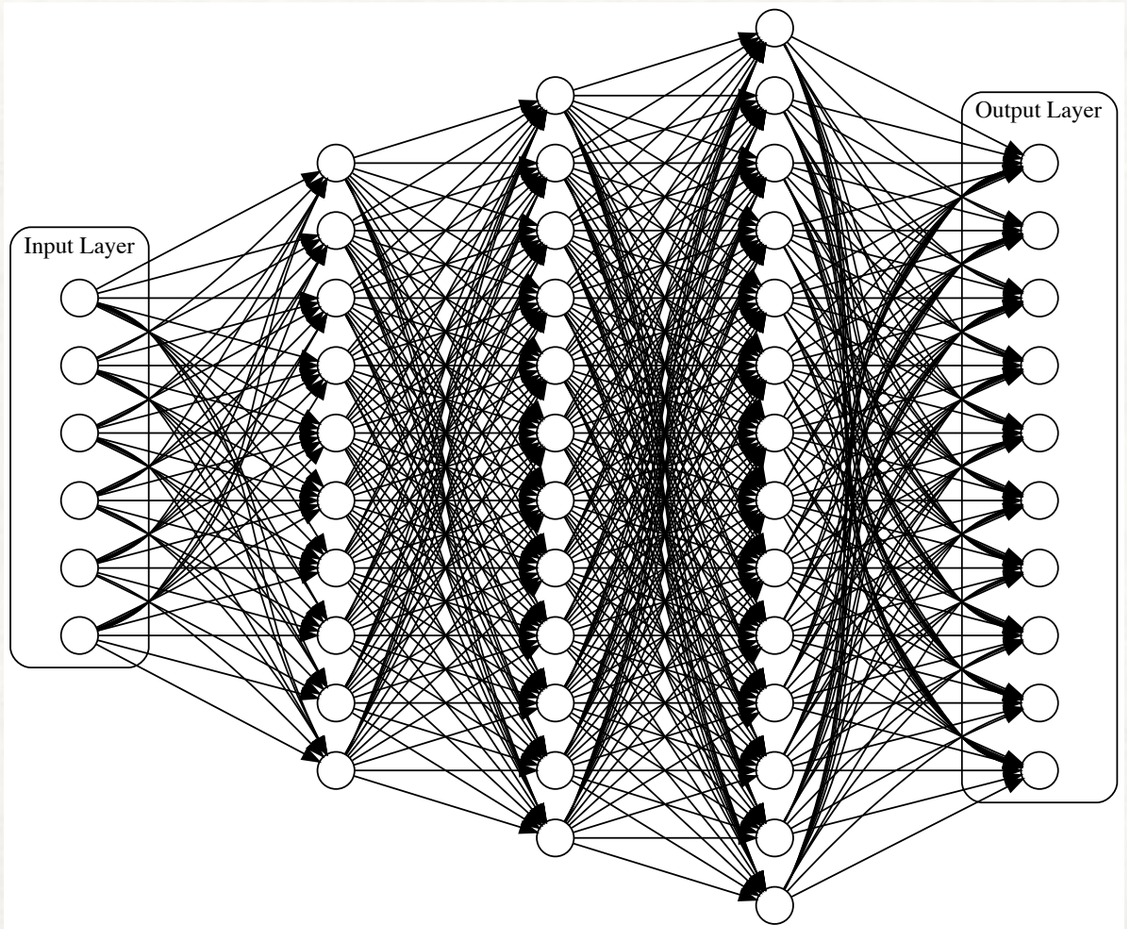
Prediction

~ 100 mili sec / one model

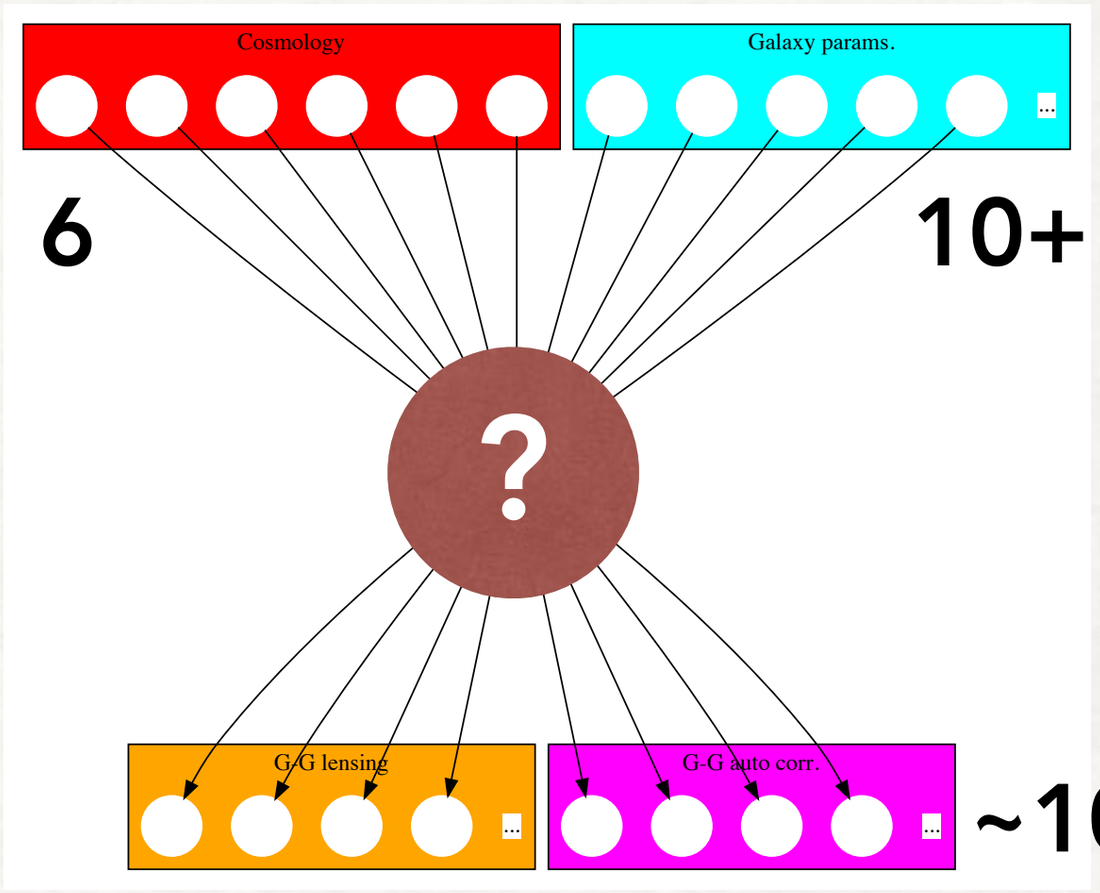
with accuracy (~3%)

HOW DO WE SET UP THE ML ARCHITECTURE?

Input Hidden layers Output



Cosmology Galaxy params.



Lensing Clustering

- Deep architectures are popular
  - Availability of large ensemble of data
  - Flexibility of the function
  - Tricks to avoid overfitting, i.e., regularization

HOW DO WE SET UP THE ML ARCHITECTURE?

Cosmology

6

GPR

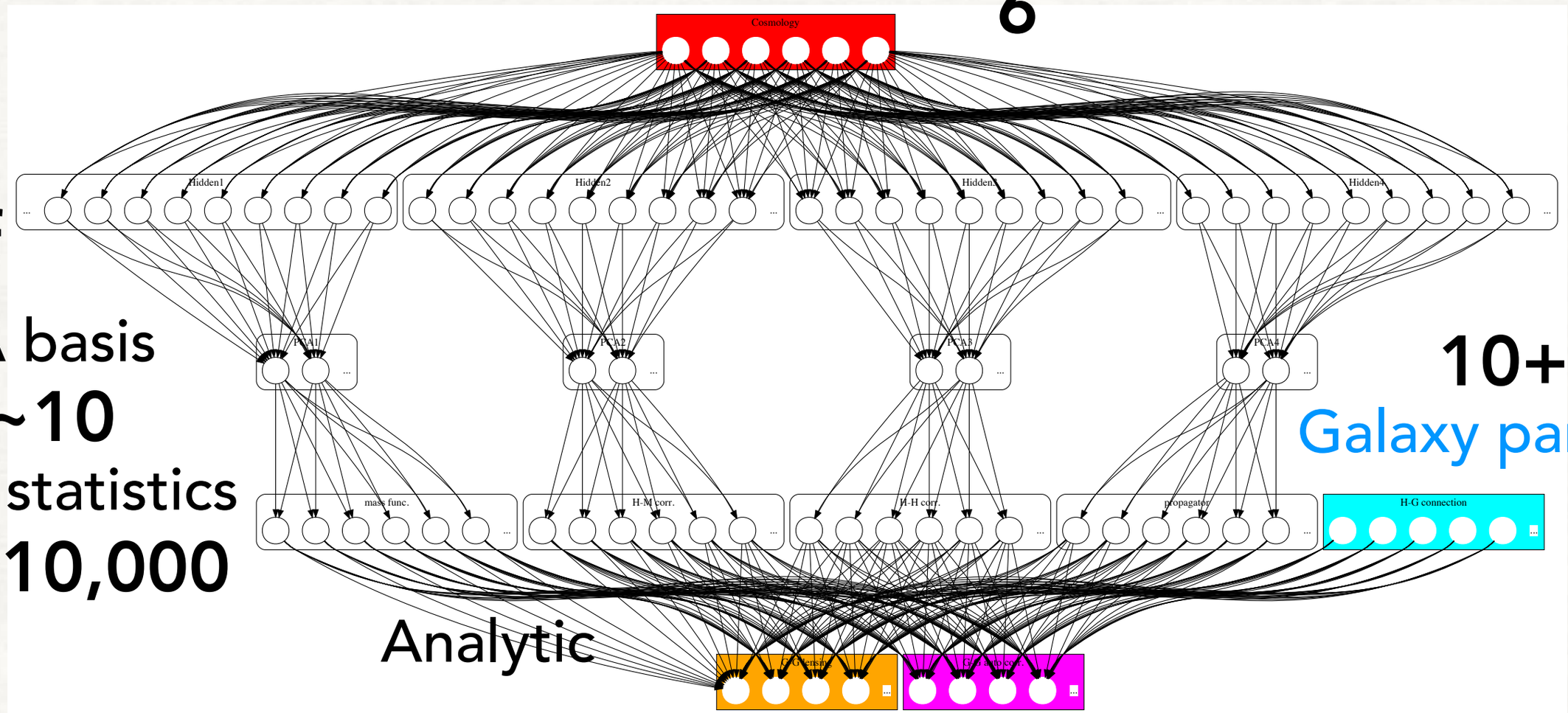
Inf

PCA basis

~10

Halo statistics

~10,000



10+

Galaxy params.

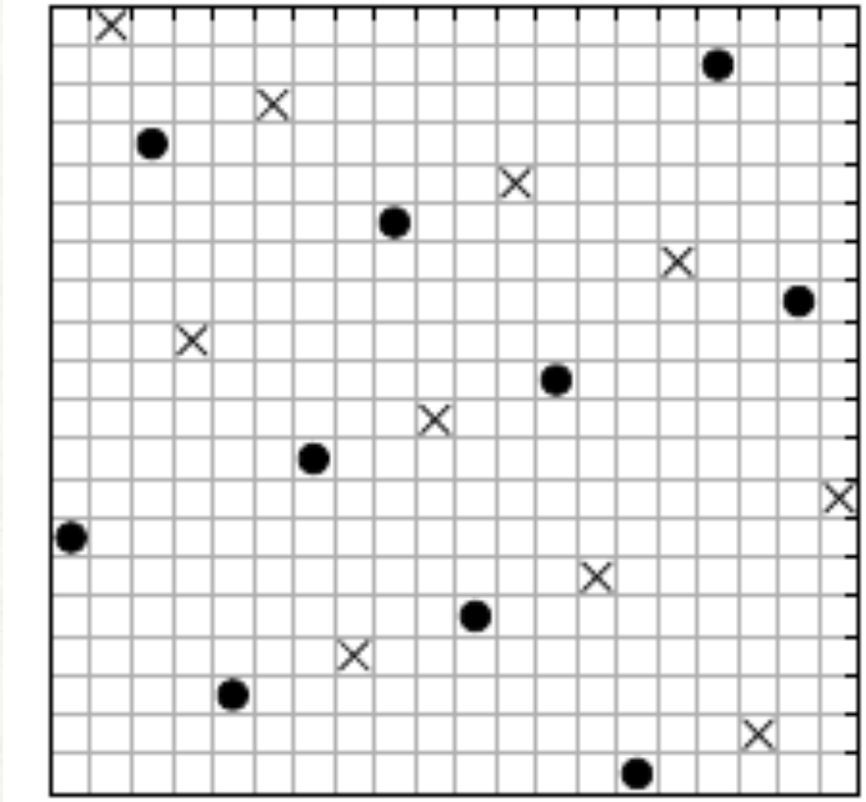
Analytic

Lensing Clustering ~100

- Deep architectures are popular
  - Availability of large ensemble of data -> **not enough!**
  - Flexibility of the function
  - Tricks to avoid overfitting, i.e., **regularization**

SLICED LATIN HYPERCUBE DESIGN

- Latin Hypercube Designs
  - one and only one in every row and column
  - Good projection properties
- Hierarchical LHD (Ba, Myers, Brenneman '15)
  - Each symbol forms an LHD
  - All the points together form an LHD
  - Space filling property is ensured by minimizing



$$\Phi(\mathbf{X}_N) = \frac{1}{2}\left(\phi_{\text{all}} + \frac{1}{m}\sum_{t-1}^{m}\phi_t\right),$$

$$\phi(\mathbf{X}_N) = \left(\frac{2}{N(N-1)}\sum_{i\neq j}\frac{1}{d^r(\mathbf{x}^{(i)},\mathbf{x}^{(j)})}\right)^{1/r}$$

Cost function for all the samples within the *t*-th slice with some large r

- Useful
  - for a stringent cross-validation test
  - by splitting the sample into training and validation sets

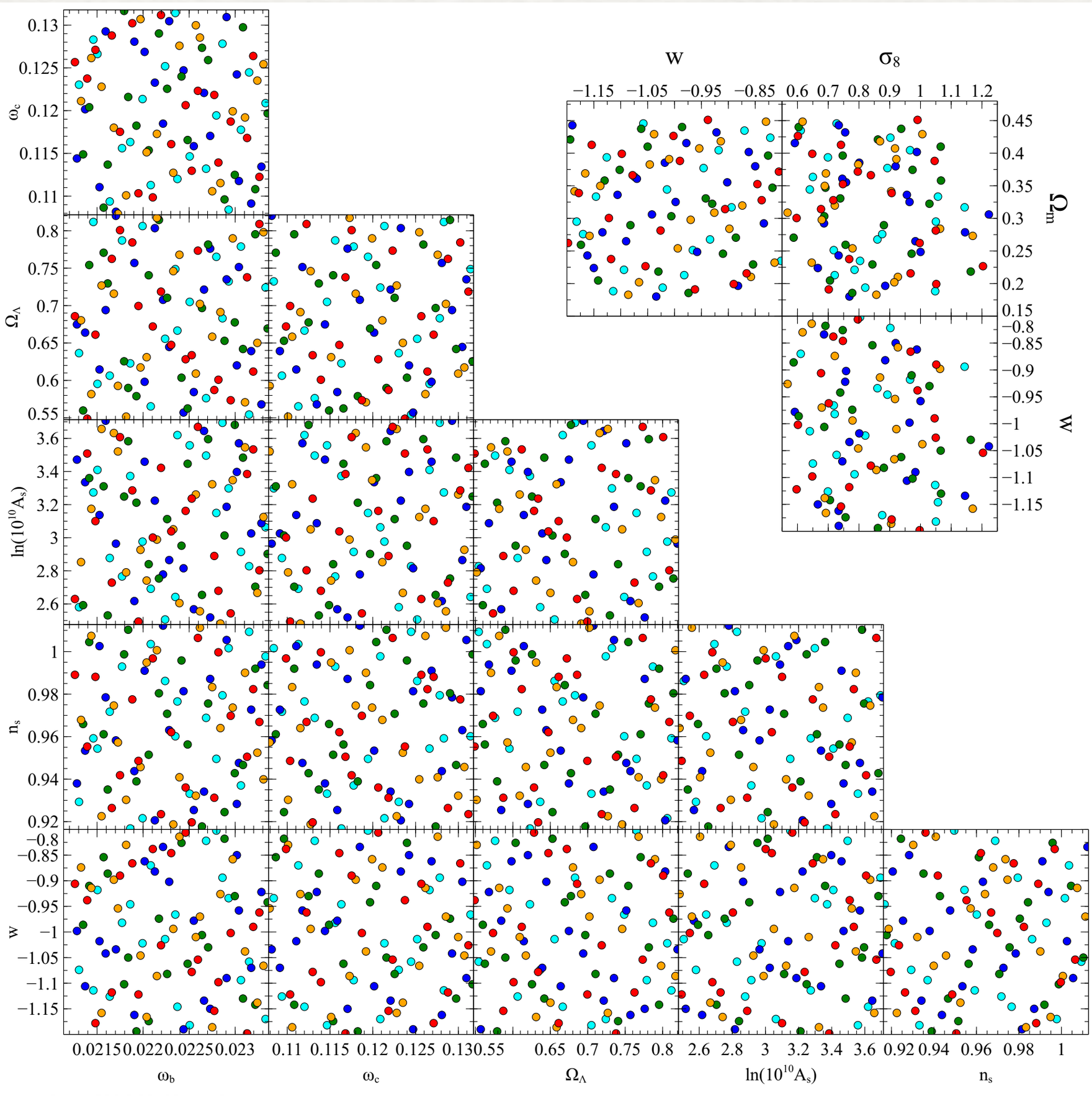
SLIC

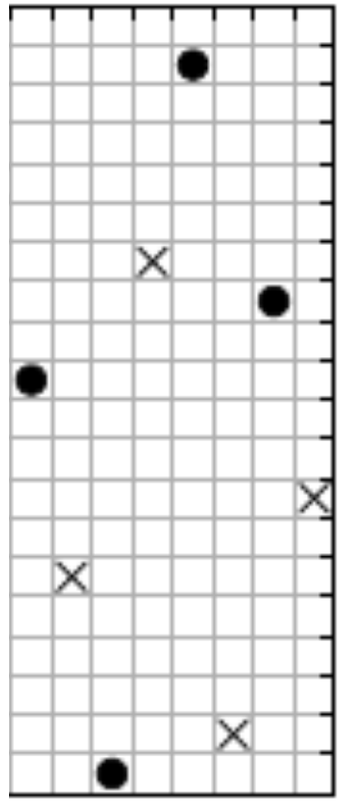
- Latin
  - one
  - Go
- Hiera
  - Eac
  - All
- Spa

min

Cost fu

- Use
  - fo
  - b





$$\left(\frac{1}{(i)}, \mathbf{x}^{(j)} \right)^{1/r}$$

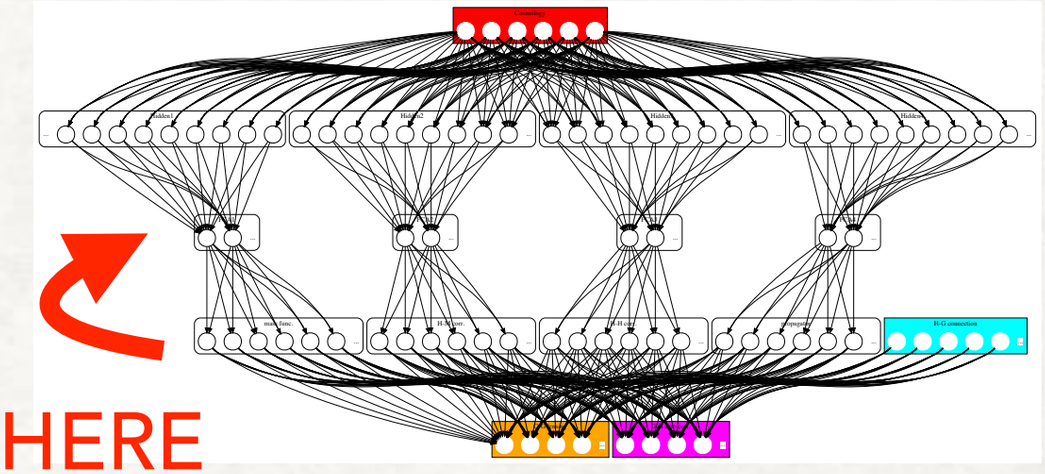
me large r

DIMENSION REDUCTION

- Mass function

  fitting $f(\sigma) = A[\sigma^{-a} + b] \exp\left[-\frac{c}{\sigma^2}\right]$

  - Sheth-Tormen type functional form
  - b and c from Tinker
  - (A, a) at 21 redshifts = 42 component vector -> 6 PCs
- Halo-matter cross correlation
  - (r-bin, n-bin, z-bin) = (66, 13, 21)
  - 18,018 components -> 5 PCs

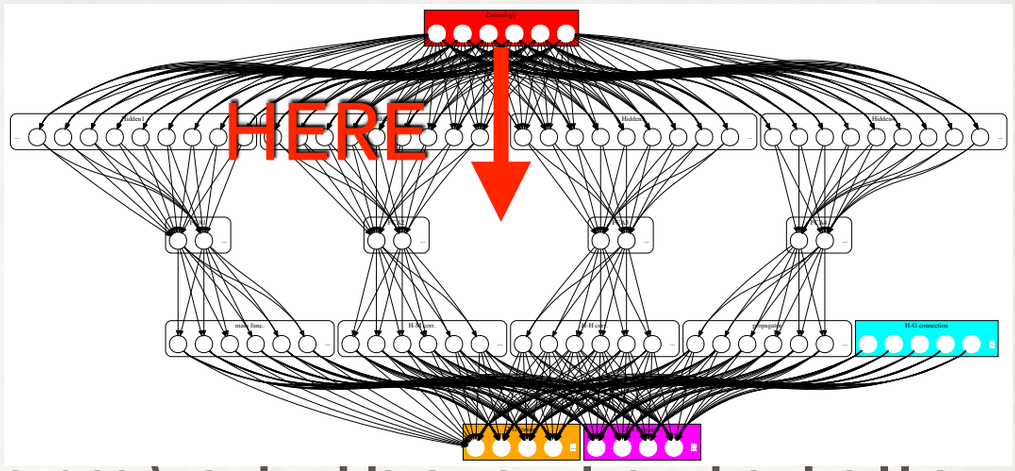


HERE

- Halo auto correlation
  - (r-bin, n_1-bin, n_2-bin, z-bin) = (21, 8, 8, 21)
  - 28,224 components -> 8 PCs
- Propagator

  fitting $G(k) = [c_0 + c_2 k^2 + c_4 k^4] \exp\left[-k^2 \sigma_d^2 / 2\right]$
  - 3 parameters at 21 redshifts = 63 components -> 3PCs

GAUSSIAN PROCESS REGRESSION

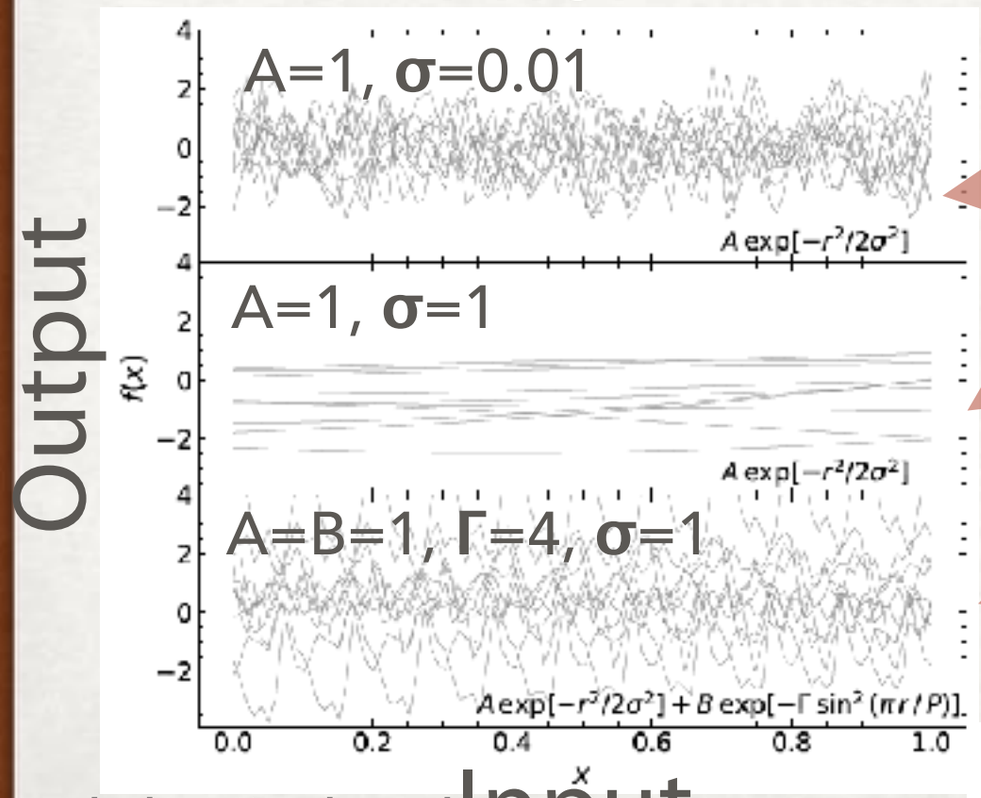


- Take the limit of infinite # of nodes
- Regularize the coefficients (i.e., L2 norm) s.t. the outputs follow i.i.d. -> multivariate Gaussian (c.f., central limit theorem)
  - Specified by the correlation (kernel) function
  - works as a **non-parametric regressor or classifier**

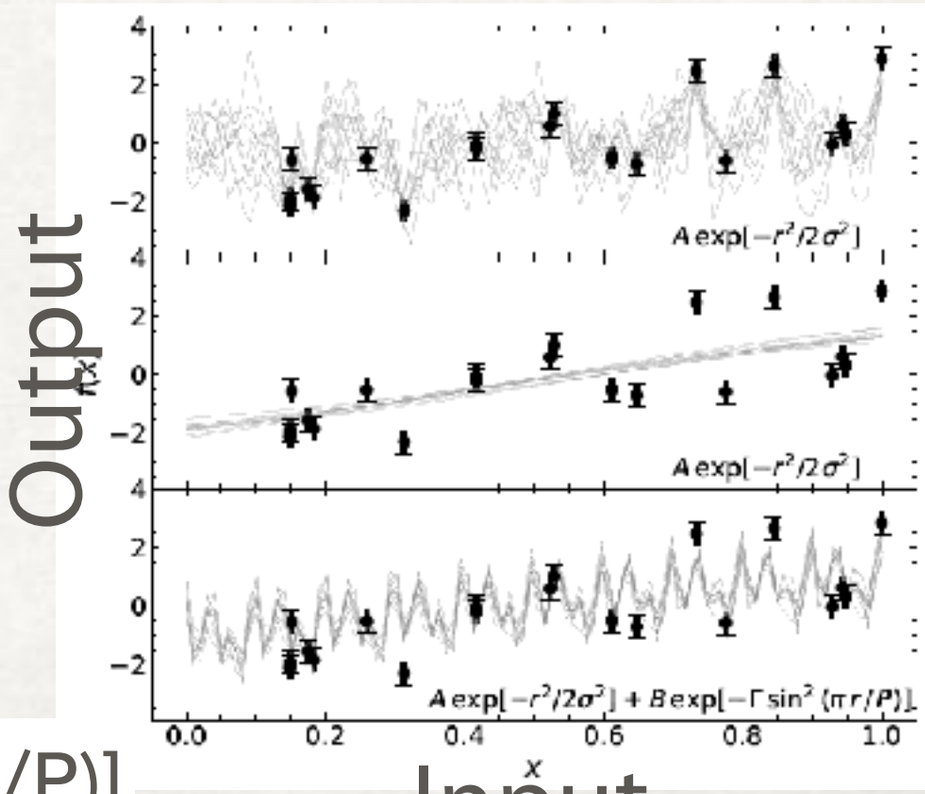
Prior

Posterior

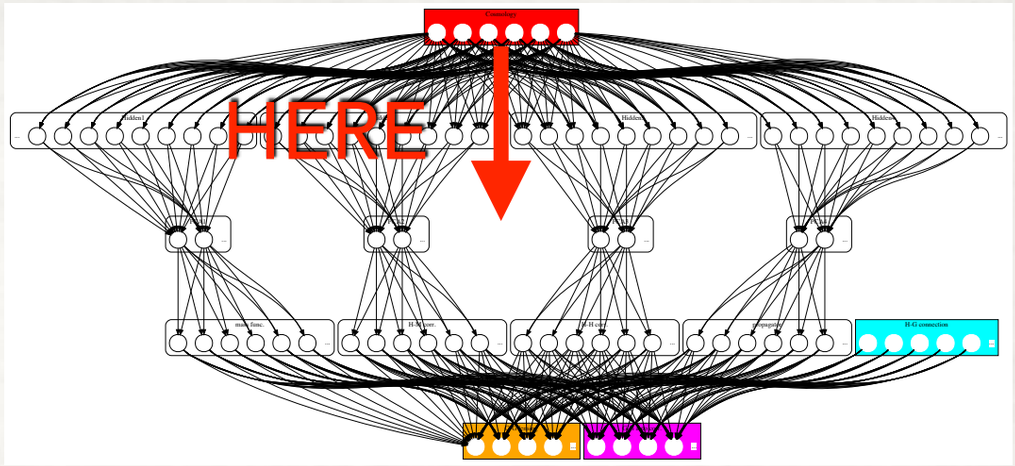
A exp[-r²/2σ²]



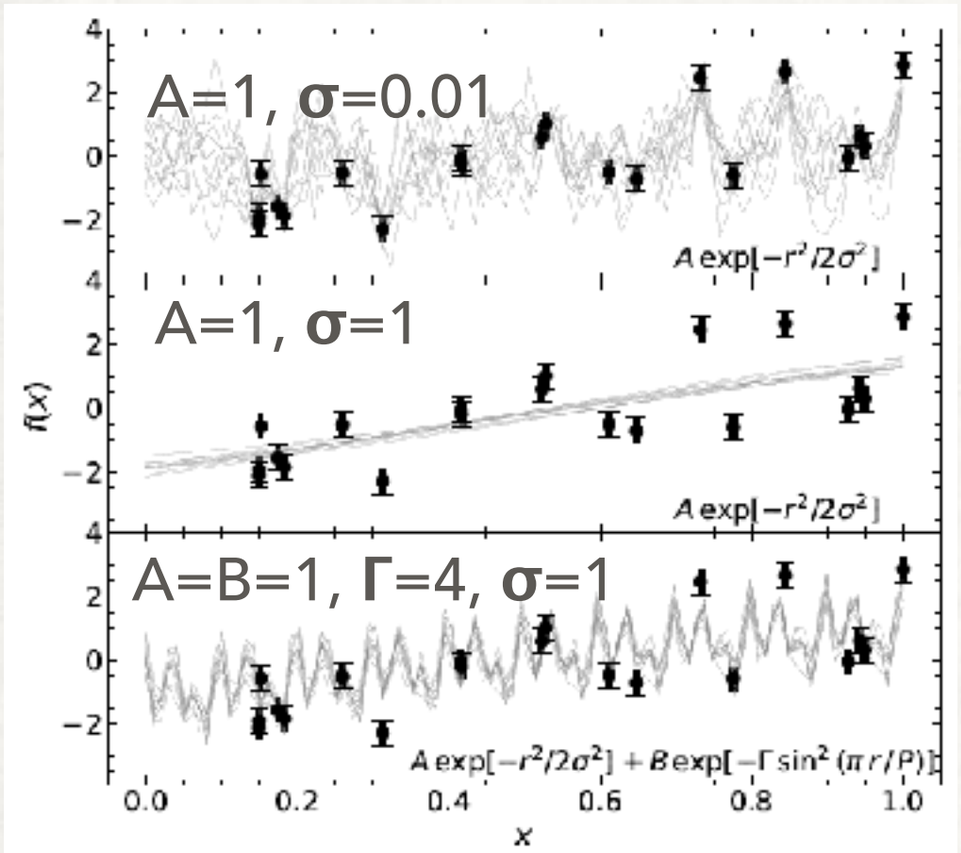
+B exp[-Γsin2(πr/P)]



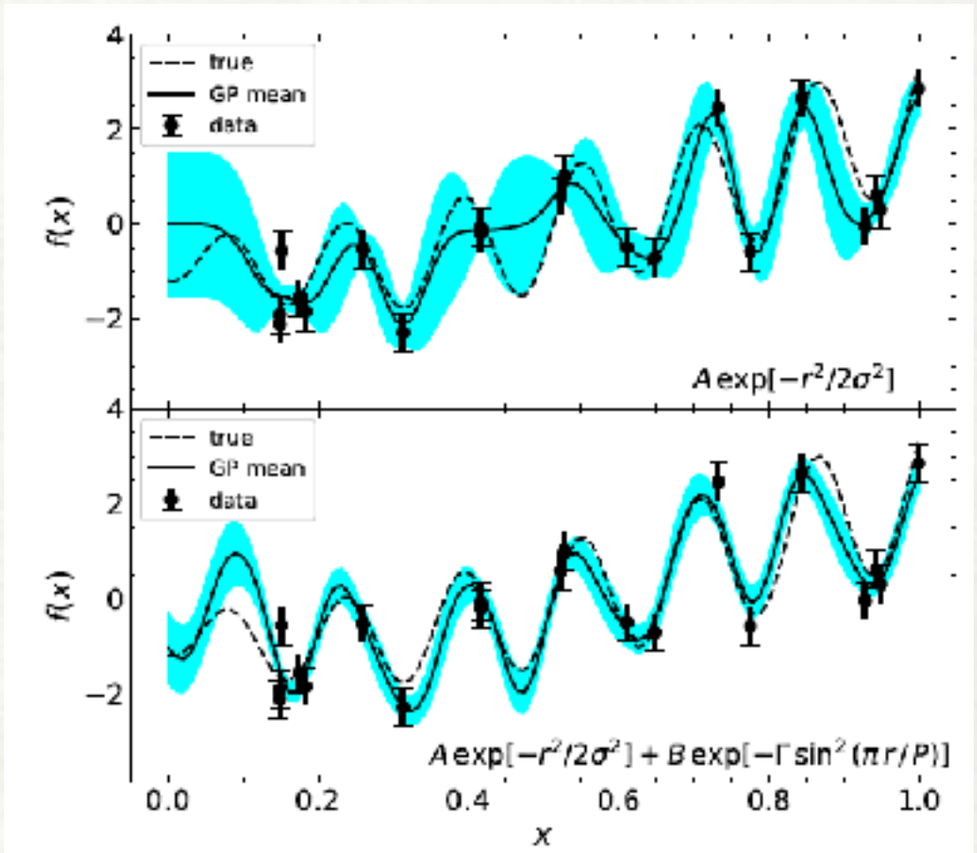
GAUSSIAN PROCESS REGRESSION



Posterior (before optimization)



Posterior (after optimization)

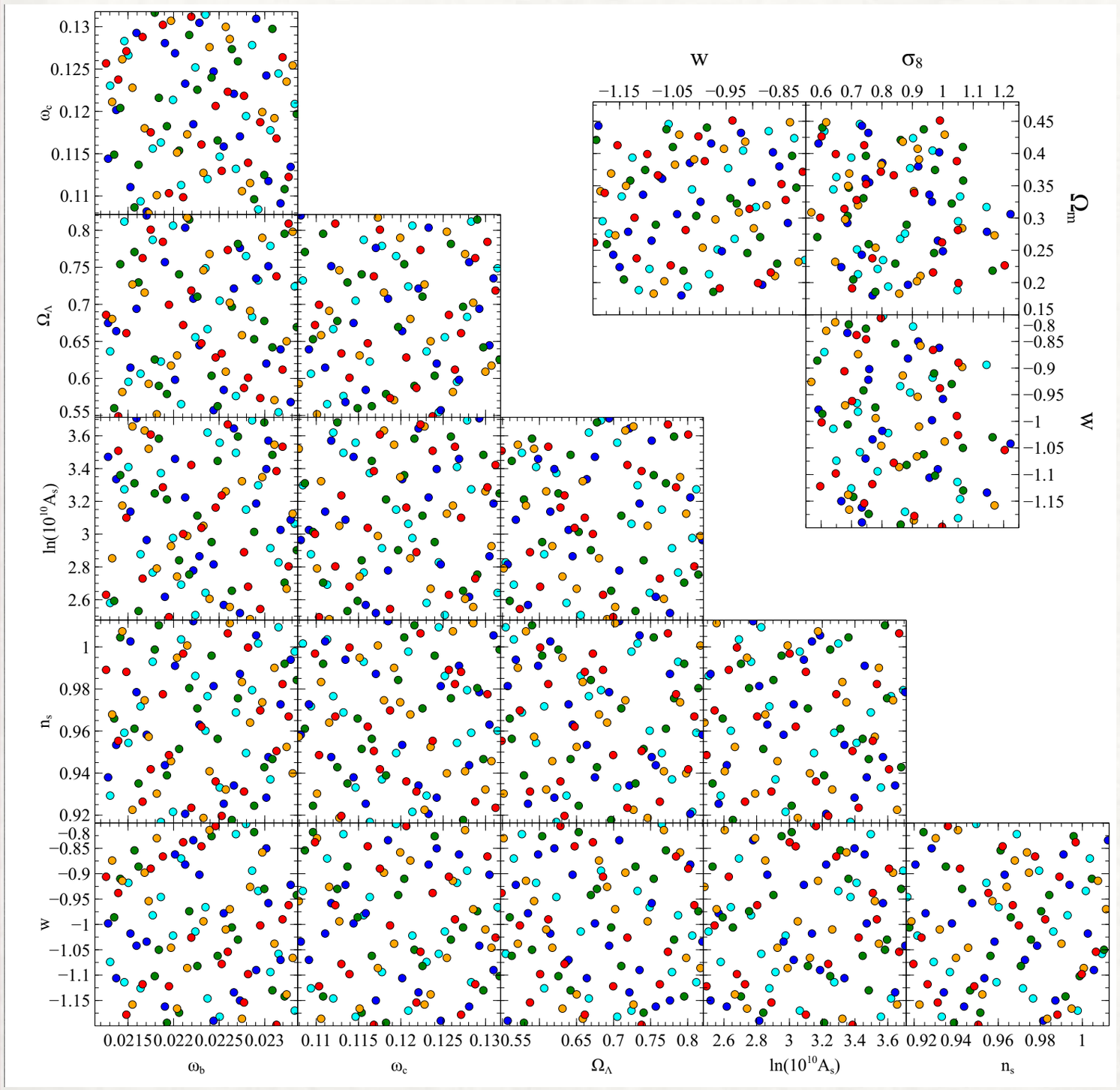


- All the quantities are analytic thanks to Gaussianity
- "Hyperparameters" can be ``trained"

$$P(t_{N+1}\,|\,\mathbf{t}_N) \propto \exp\left[-\frac{1}{2}\left[\mathbf{t}_N\ t_{N+1}\right]\mathbf{C}_{N+1}^{-1}\begin{bmatrix}\mathbf{t}_N\\ t_{N+1}\end{bmatrix}\right].$$

$$\begin{aligned}\hat{t}_{N+1} &= \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{t}_N\\ \sigma^2_{\hat{t}_{N+1}} &= \kappa - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}.\end{aligned}$$

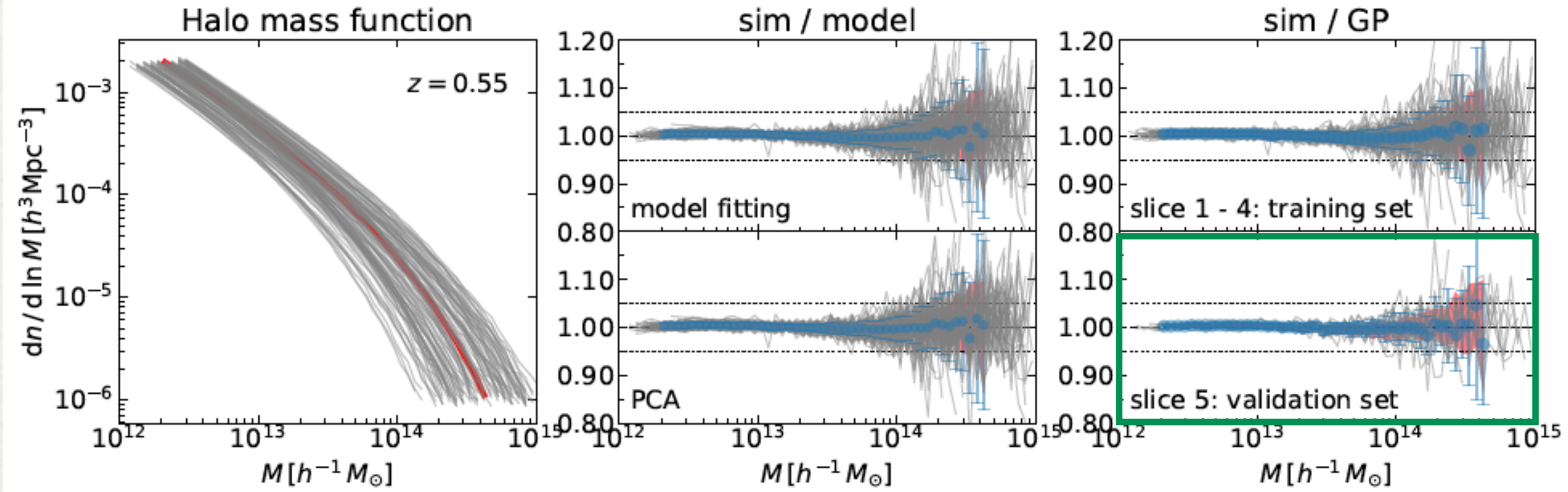
CROSS VALIDATION STUDIES



- Train the network using 80 points in 4 different colors.
- Validate the prediction of the network using the remaining 20 points in cyan.

ACCURACY: HALO MASS FUNCTION

Example plot at z = 0.55



Spread in HMF among the 100 models

Gaussian Process Regression

Upper: Model fitting w/ Sheth-Tormen type function (2 free parameters)

Training set

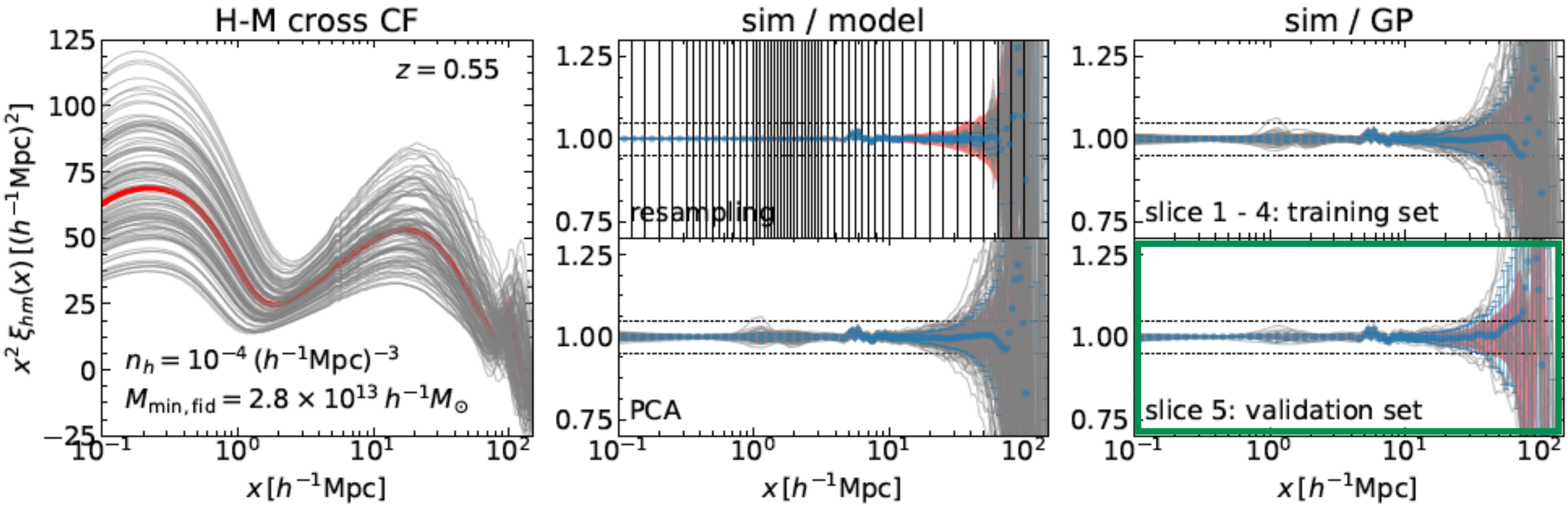
Lower: Compress the 42 (=2 x 21 redshifts) coefficients into 6 PCs

Validation set

Red shades: scatter of 28 fiducial runs

ACCURACY: CROSS CORRELATION

Example plot at z = 0.55 for a halo sample with 10^{-4} $(h^{-1}\mathrm{Mpc})^{-3}$



Spread in xi_hm among the 100 models

Gaussian Process Regression

Upper: Sample fewer number of data points using cubic spline interpolation

Training set

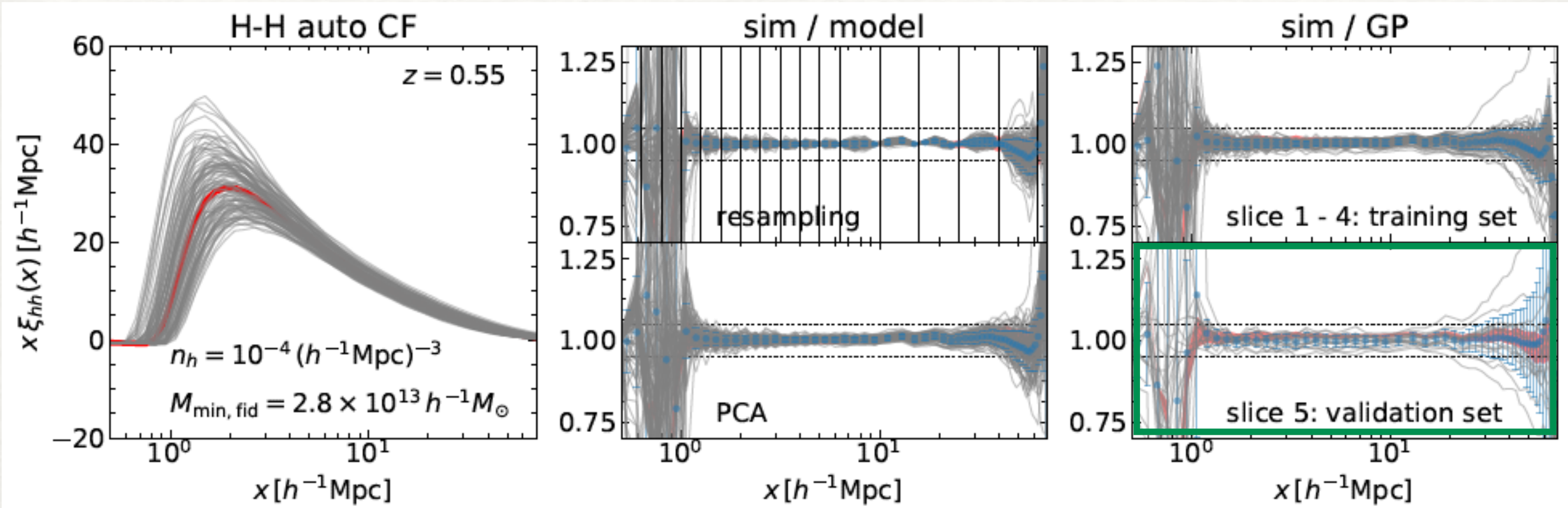
Lower: Compress the 18,018 (= 66x13x21 radial x number density x redshift bins) coefficients into 5 PCs

Validation set

Red shades: scatter of 28 fiducial runs

ACCURACY: AUTO CORRELATION

Example plot at z = 0.55 for a halo sample with 10^{-4} $(h^{-1}\mathrm{Mpc})^{-3}$



Spread in xi_hh among the 100 models

Gaussian Process Regression

Upper: Sample fewer number of data points using cubic spline interpolation

Training set

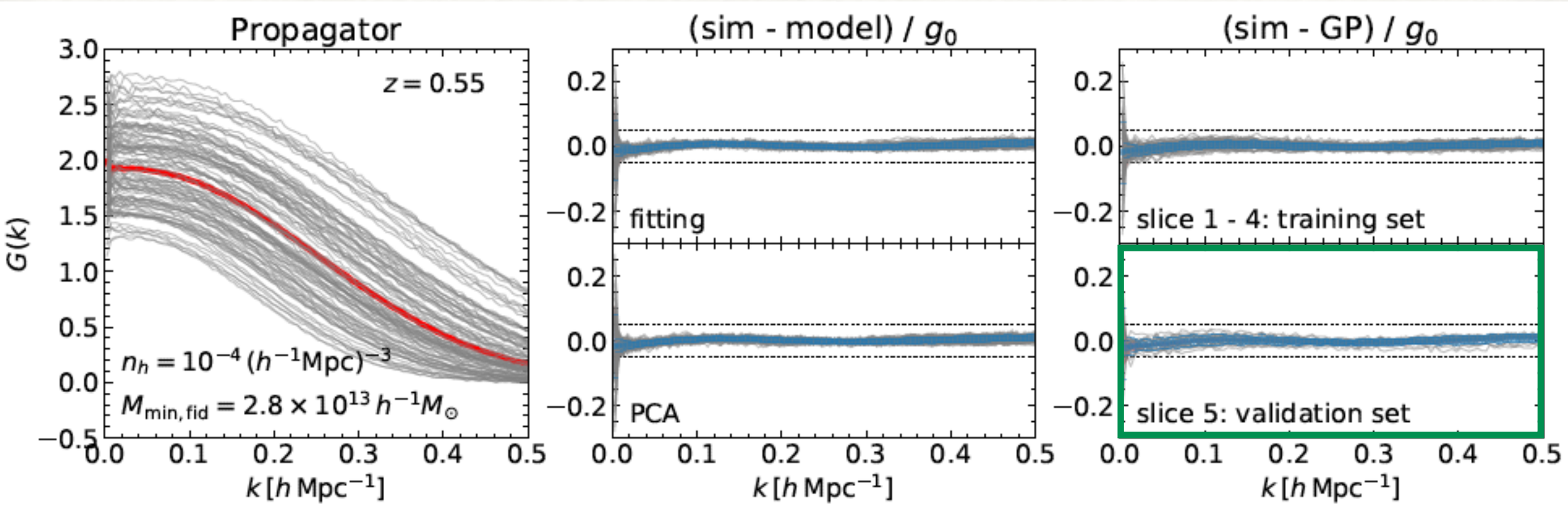
Lower: Compress the 28,224 (= 21x8x8x21 radial x number density 1 x number density 2 x redshift bins) coefficients into 8 PCs

Validation set

Red shades: scatter of 14 fiducial runs

ACCURACY: PROPAGATOR

Example plot at z = 0.55 for a halo sample with 10^{-4} $(h^{-1}\mathrm{Mpc})^{-3}$



Spread in G(k) among the 100 models

Gaussian Process Regression

Upper: Model fitting with Gaussian + corrections (3 free params)

Training set

Lower: Compress the 63 (=3 x 21 redshifts) coefficients into PCs

Validation set

Red shades: scatter of 14 fiducial runs

DARK EMULATOR: WHAT IT CAN DO

OVERVIEW

```
In [18]: import darkemu

In [19]: emu = darkemu.base_class()

initialize cosmo_class
Initialize xilin emulator
initialize xinl emulator
Initialize pklin emulator
initialize propagator emulator
Initialize sigma_d emulator
initialize cross-correlation emulator
initialize auto-correlation emulator
Initialize hmf emulator
Initialize sigmaM emulator
```

$$(\omega_{\rm b}, \omega_{\rm c}, \Omega_{\rm de}, \ln(10^{10}A_{\rm s}), n_{\rm s}, w)$$

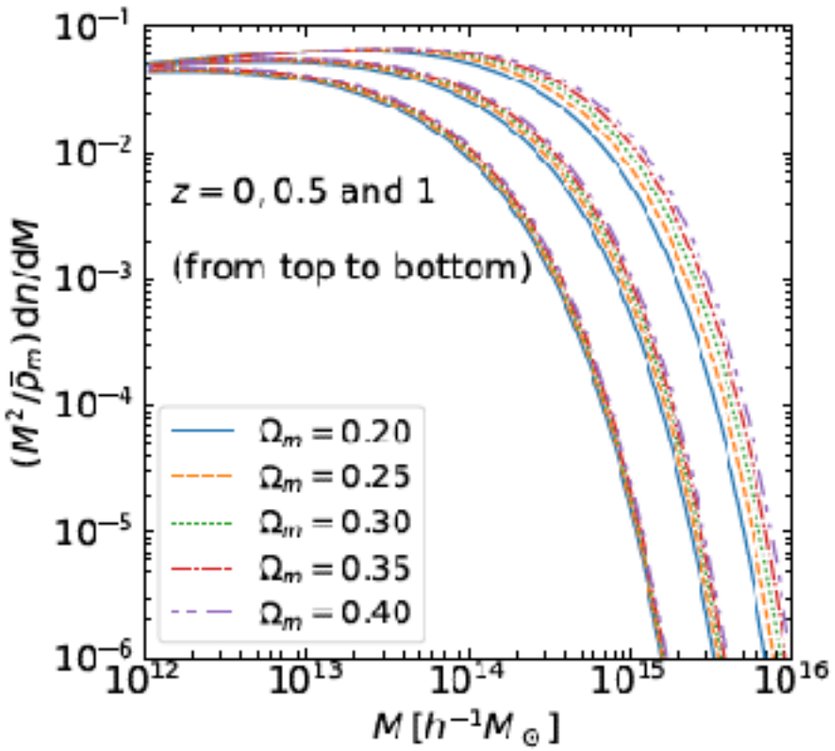
```
In [14]: cparam = np.array([0.02225,0.1198,0.6844,3.094,0.9645,-1.])
         emu.set_cosmology(cparam)
```

```
emu.get_nhalo(massbins[ii],massbins[ii+1],1.,z)
```

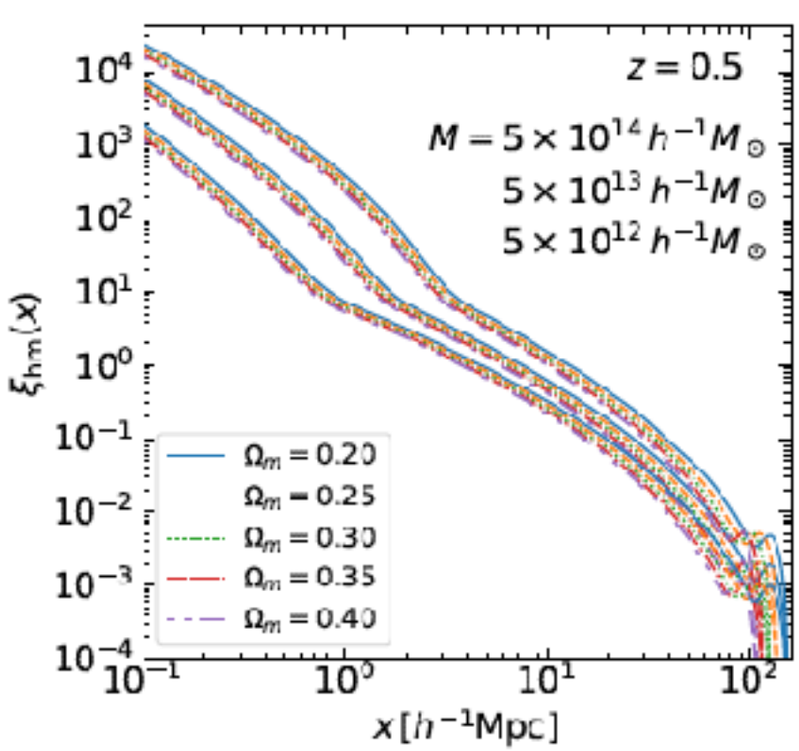
```
emu.get_xicross_mass(rs,Mh,z)
```

```
emu.get_xiauto_mass(rs,Mh,Mh,z)
```

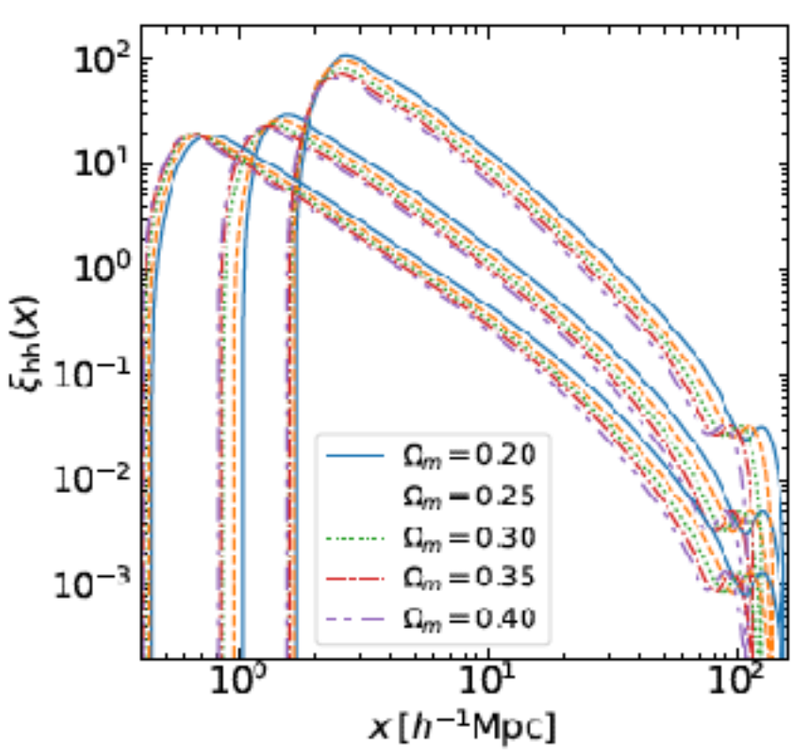
Halo mass function



Halo-Matter Cross CF



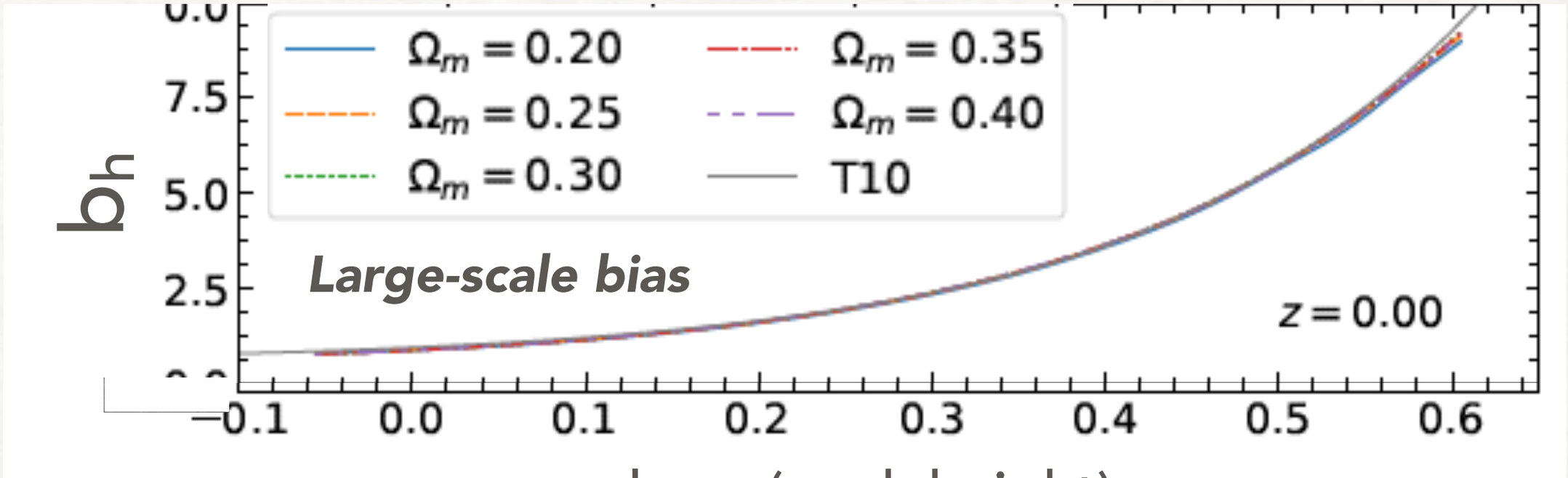
Halo-Halo Auto CF



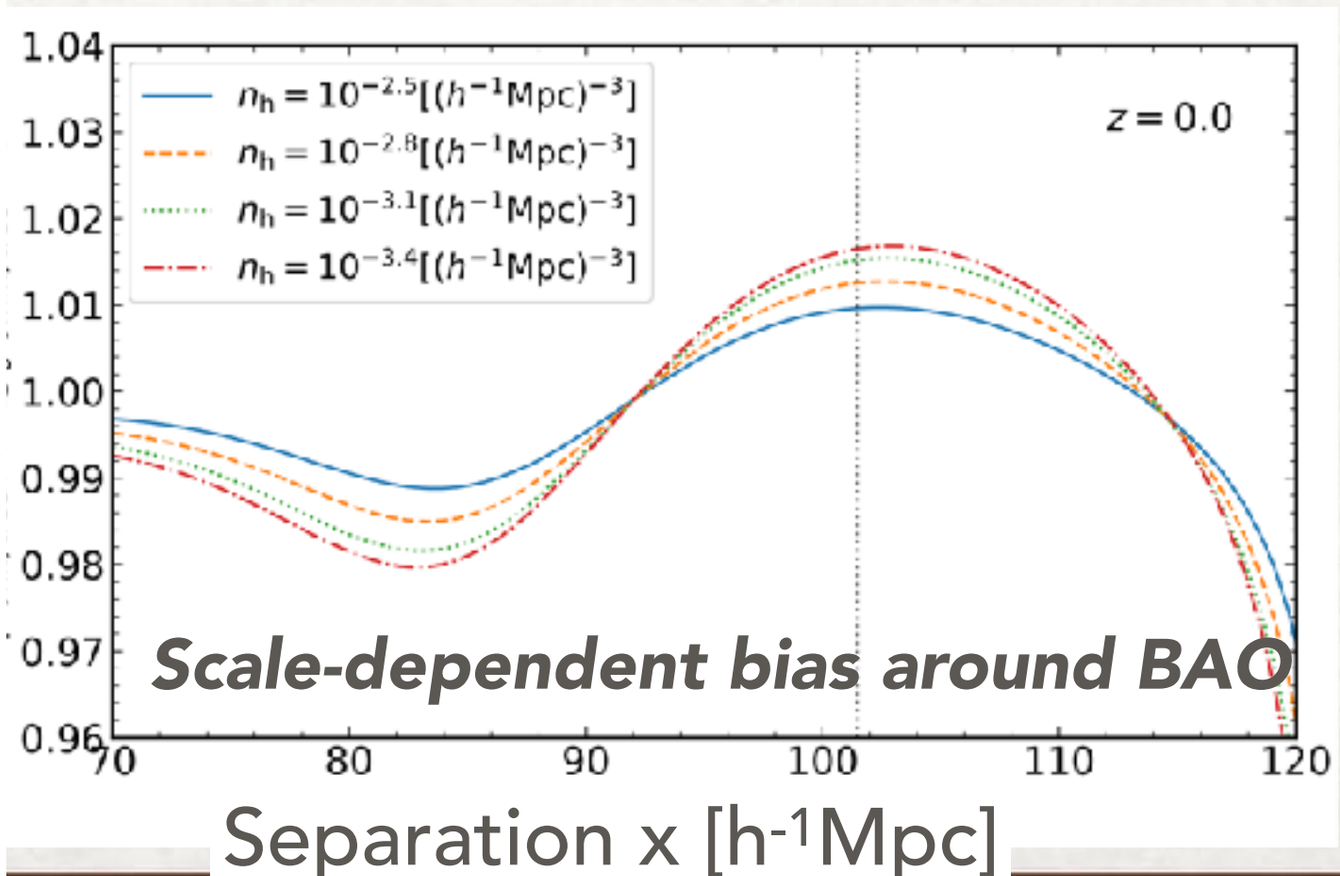
1 curve ~ 100 mili secs on a typical laptop computer

DARK EMULATOR: WHAT IT CAN DO

LARGE SCALES



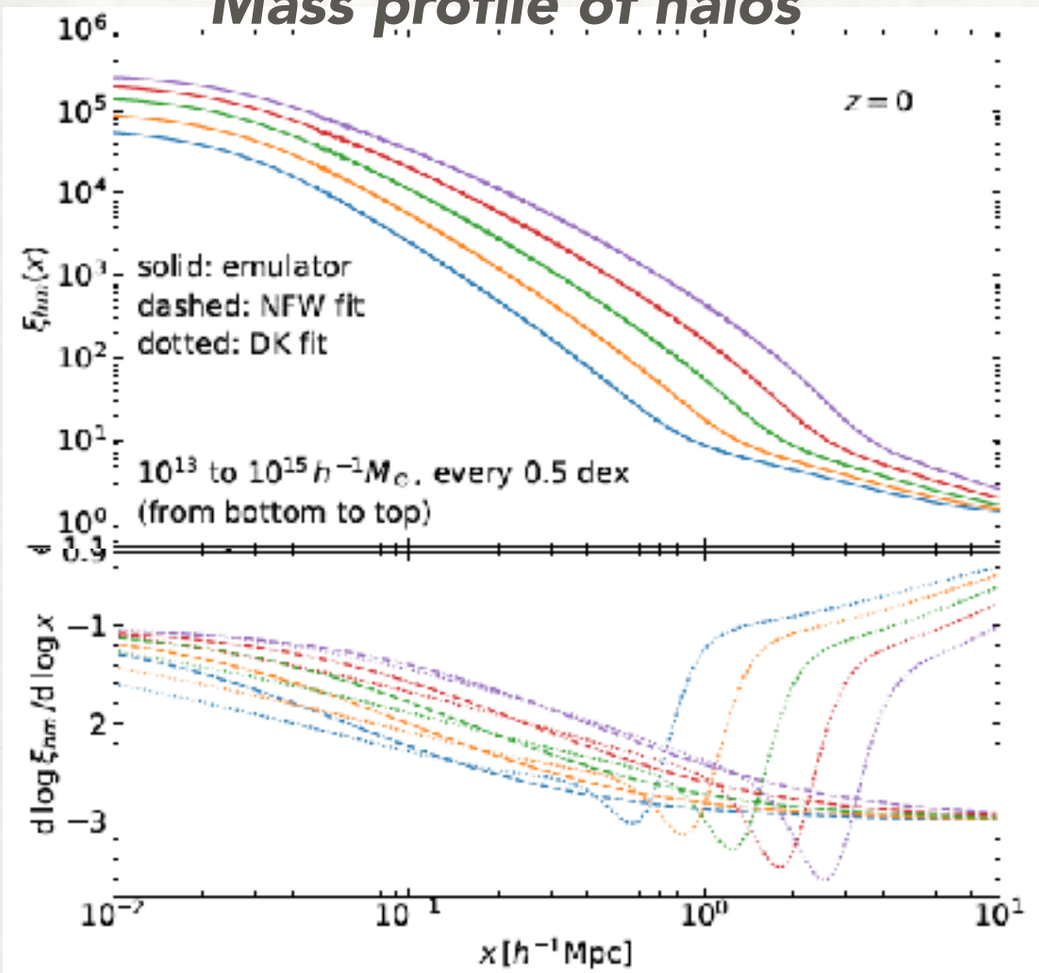
$$\frac{1}{b_{\rm h}}\sqrt{\frac{\xi_{\rm hh}(x)}{\xi_{\rm mm}(x)}}$$



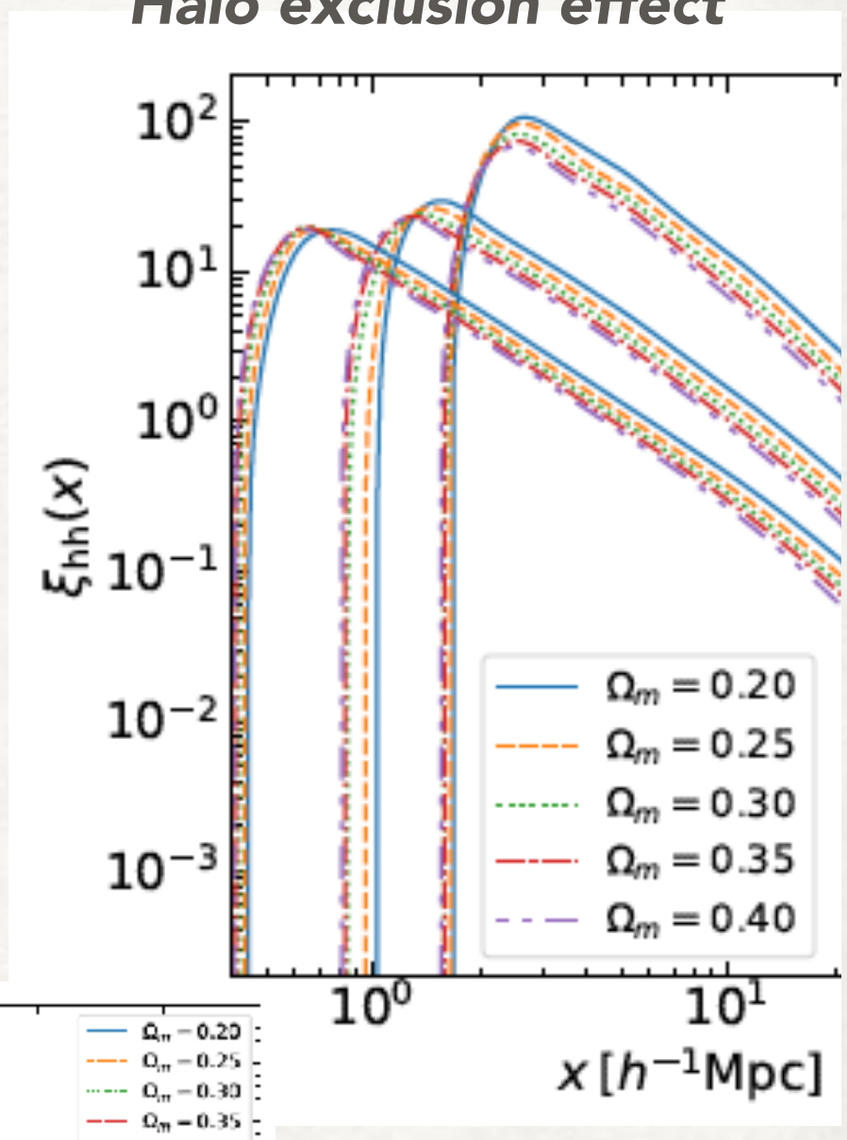
DARK EMULATOR: WHAT IT CAN DO

SMALL SCALES

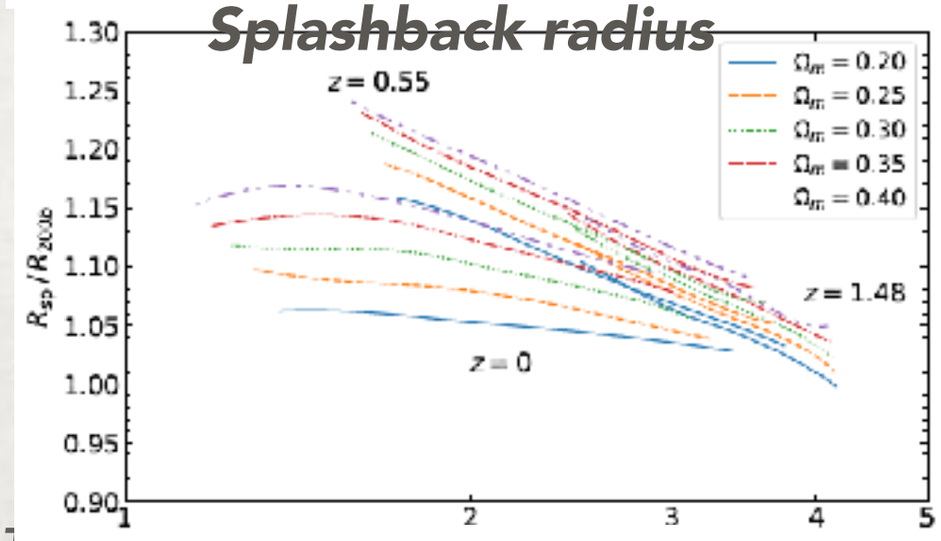
Mass profile of halos



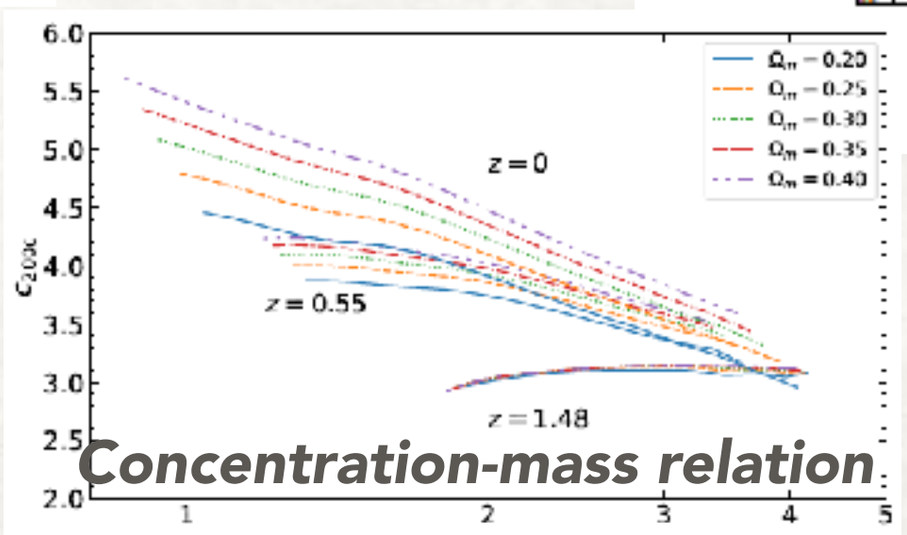
Halo exclusion effect



Splashback radius



Concentration-mass relation



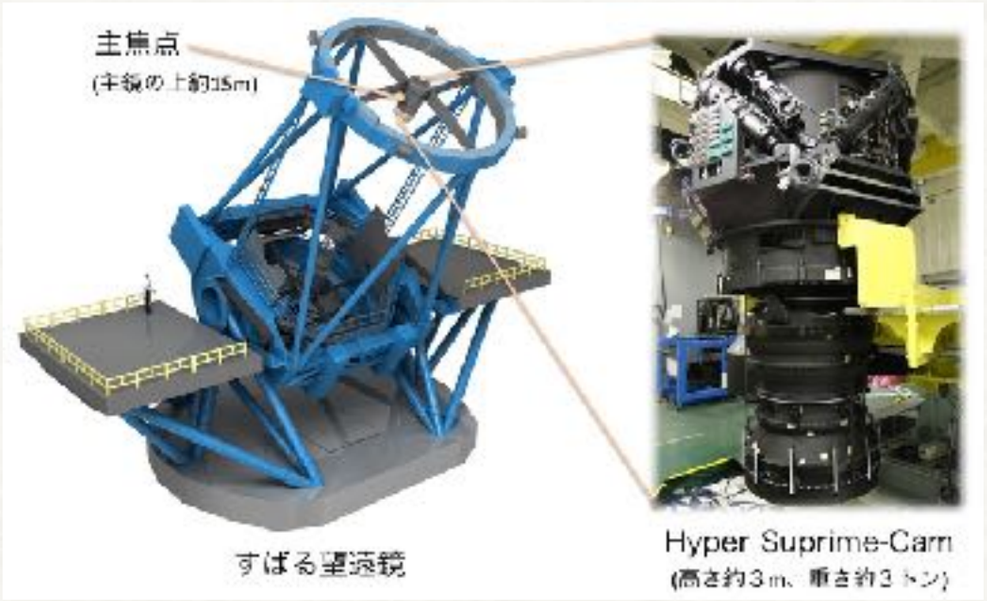
peak height

peak height

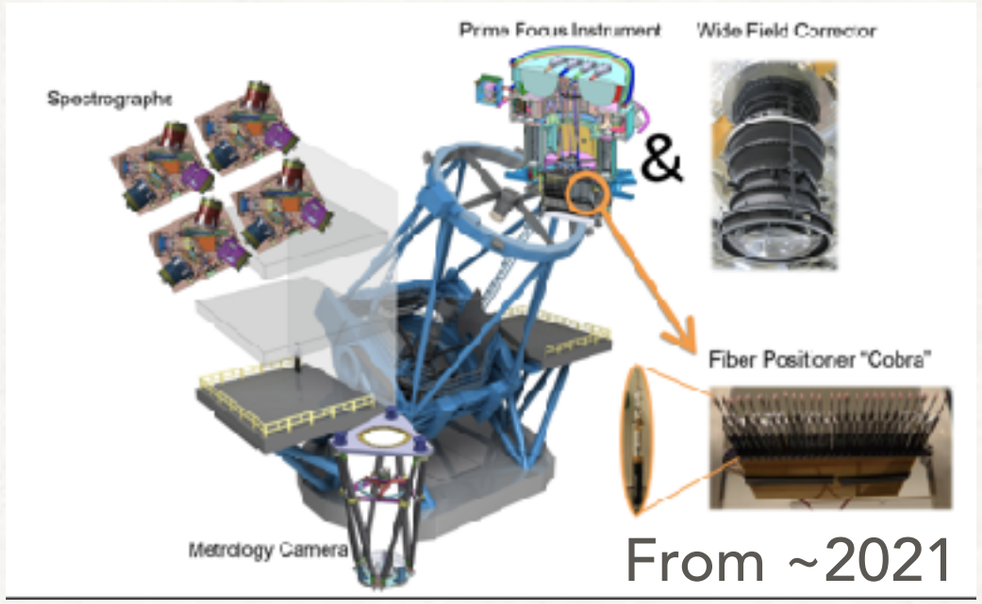
BREAK THE DEGENERACY

Subaru Measurement of Images and Redshifts (SuMIRe) Project

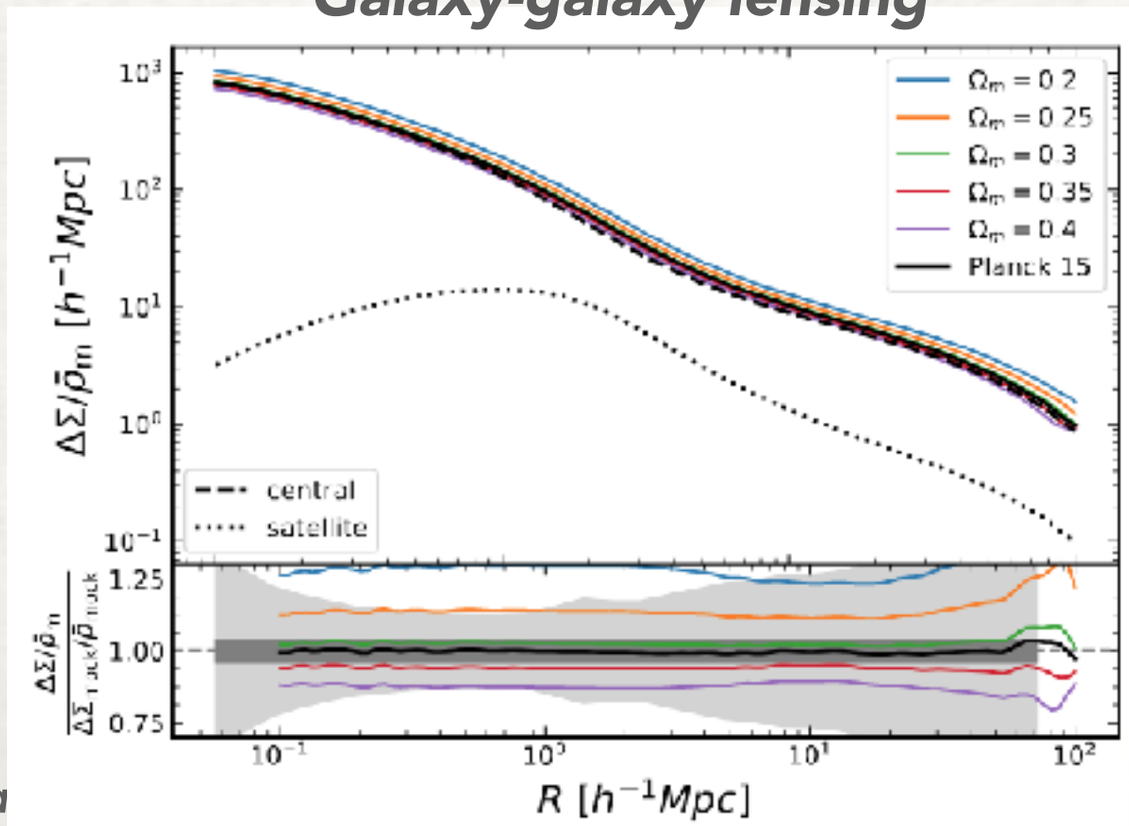
Hyper Suprime Cam (HSC)



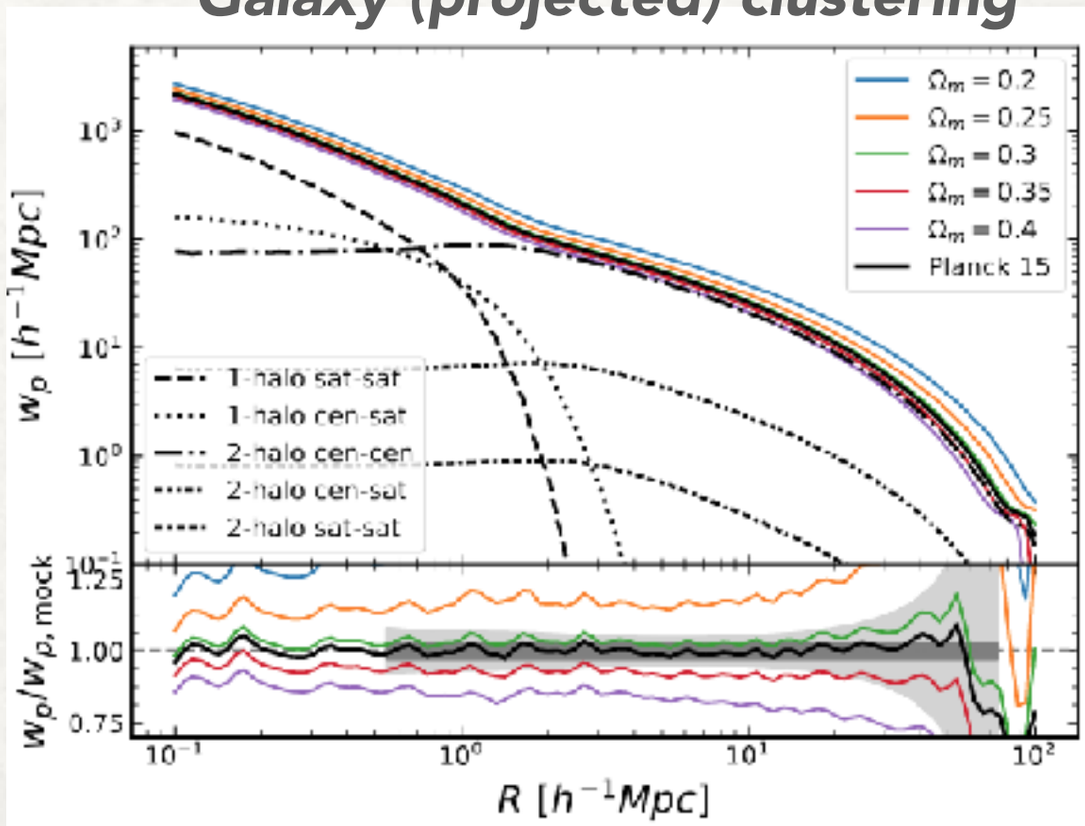
Prime Focus Spectrograph (PFS)



Galaxy-galaxy lensing



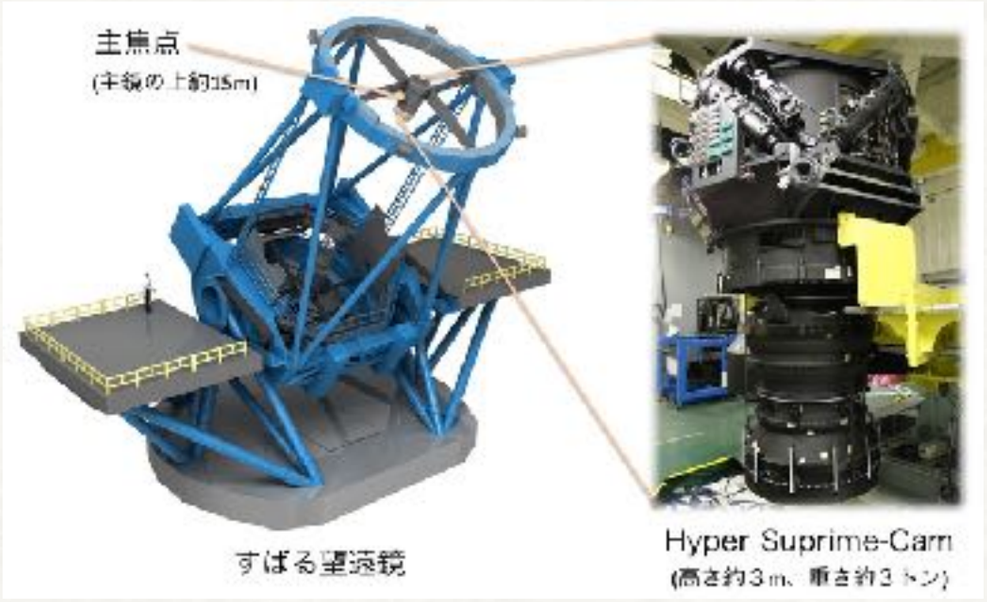
Galaxy (projected) clustering



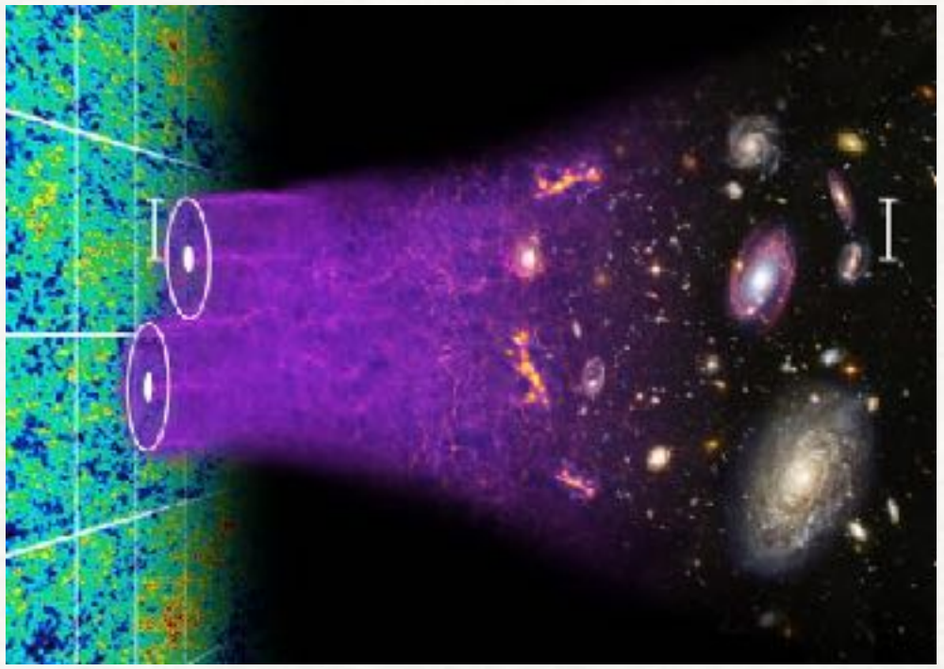
BREAK THE DEGENERACY

Subaru Measurement of Images and Redshifts (SuMIRe) Project

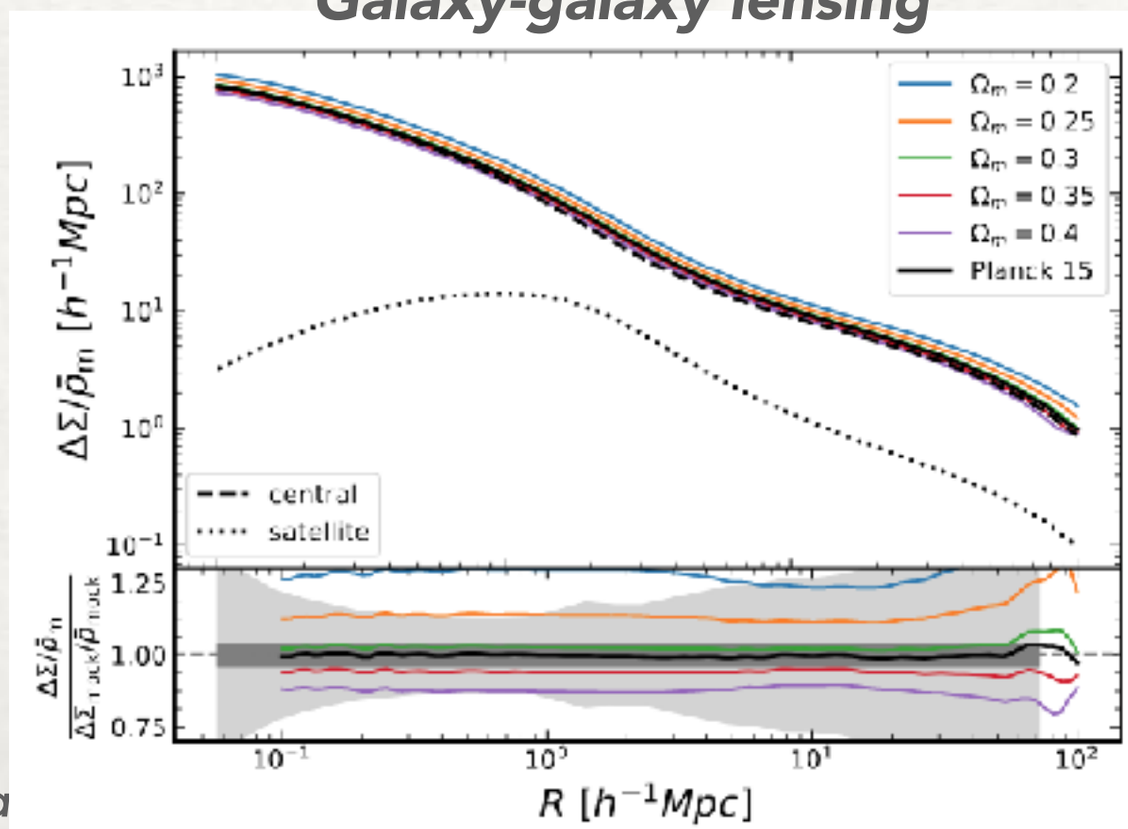
Hyper Suprime Cam (HSC)



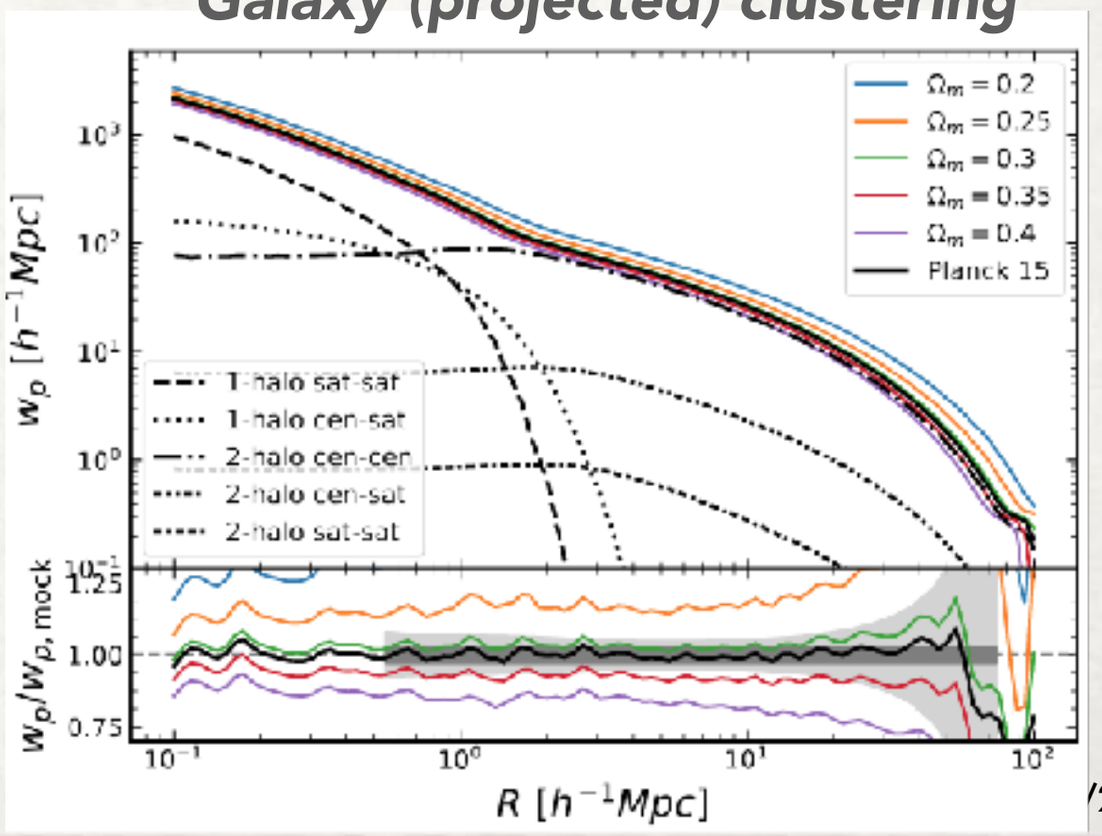
SDSS BOSS (CMASS + LOWZ)



Galaxy-galaxy lensing



Galaxy (projected) clustering



BREAK THE DEGENERACY

Subaru Measurement of Images and Redshifts (SuMIRe) Project

Hyper Suprime Cam (HSC)

SDSS BOSS (CMASS + LOWZ)

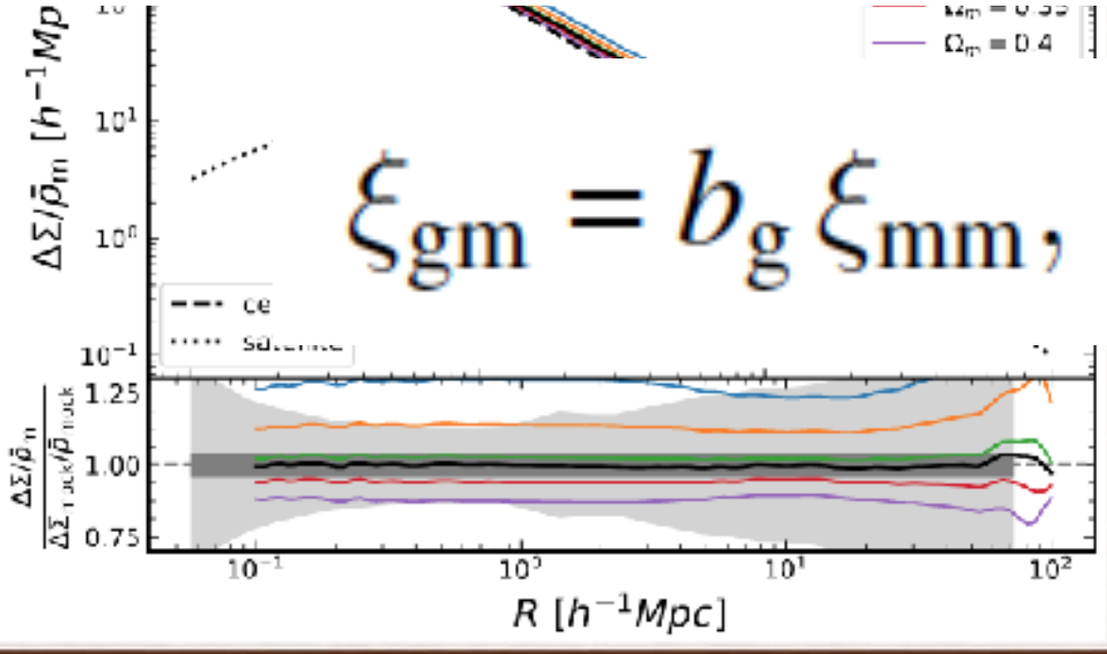
$$\Delta\Sigma_{\rm g}(R) = \bar{\Sigma}_{\rm g}(<R) - \Sigma_{\rm g}(R),$$

$$\bar{\Sigma}_{\rm g}(<R) = \frac{2}{R^2}\int_0^R \Sigma_{\rm g}(y)\mathrm{d}y,$$

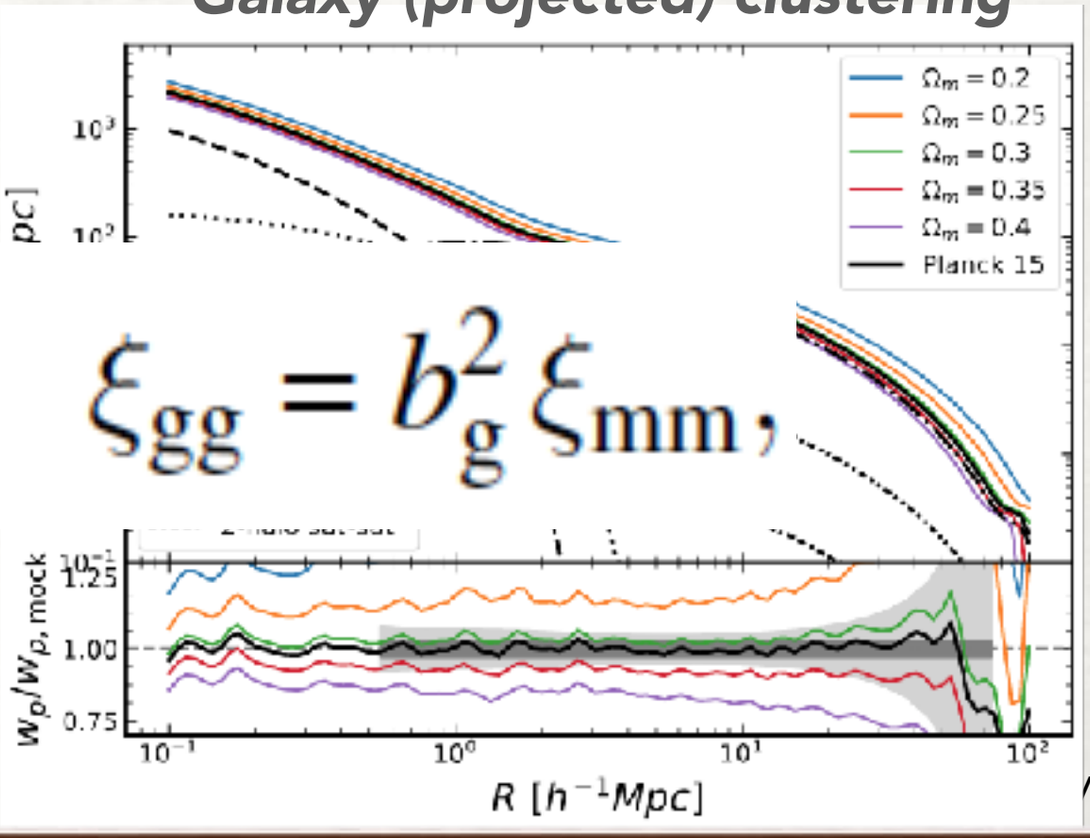
$$\Sigma_{\rm g}(R) = \bar{\rho}_{\rm m0}\int \xi_{\rm gm}(R,\pi)\,\mathrm{d}\pi,$$

$$w_{\rm gg}(R) = 2\int_0^{\pi_{\rm max}} \xi_{\rm gg}(R,\pi)\,\mathrm{d}\pi,$$

Galaxy (projected) clustering

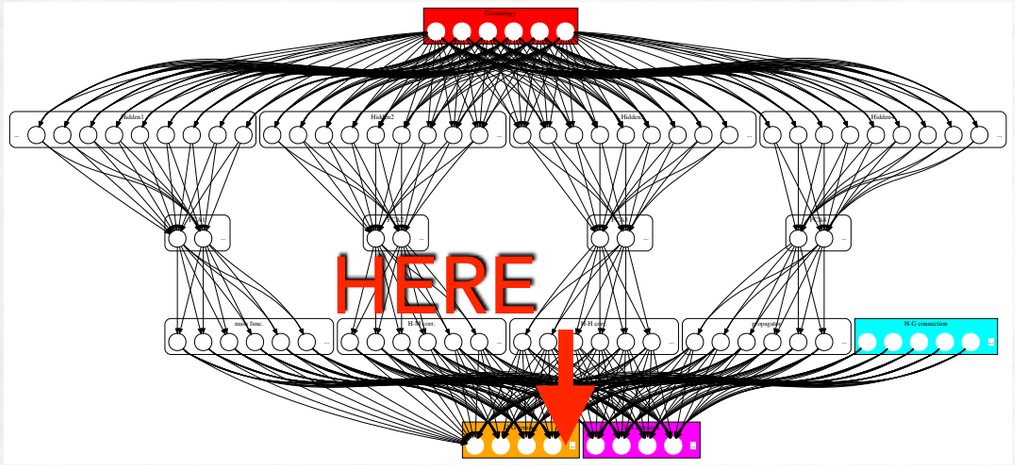


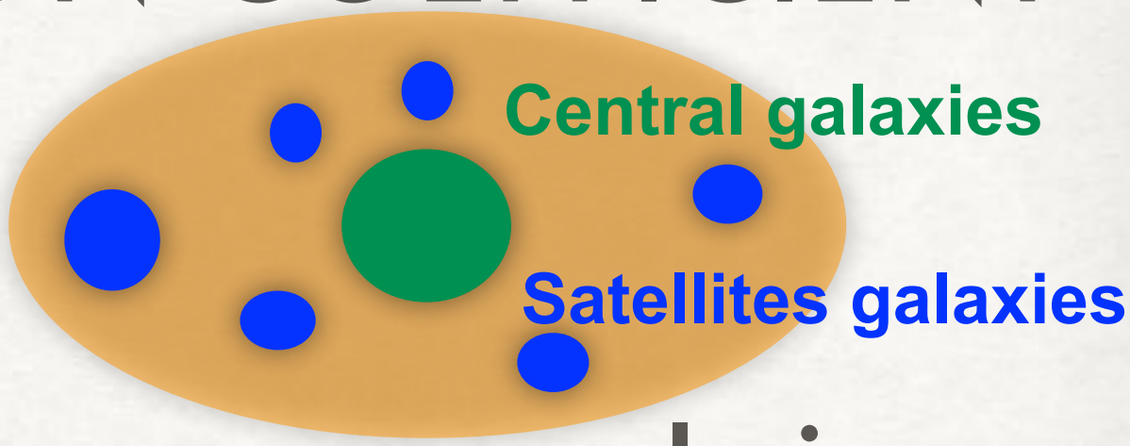
$$\xi_{\rm gm} = b_{\rm g}\,\xi_{\rm mm},$$



$$\xi_{\rm gg} = b_{\rm g}^2\,\xi_{\rm mm},$$

CROSS-CORRELATION COEFFICIENT





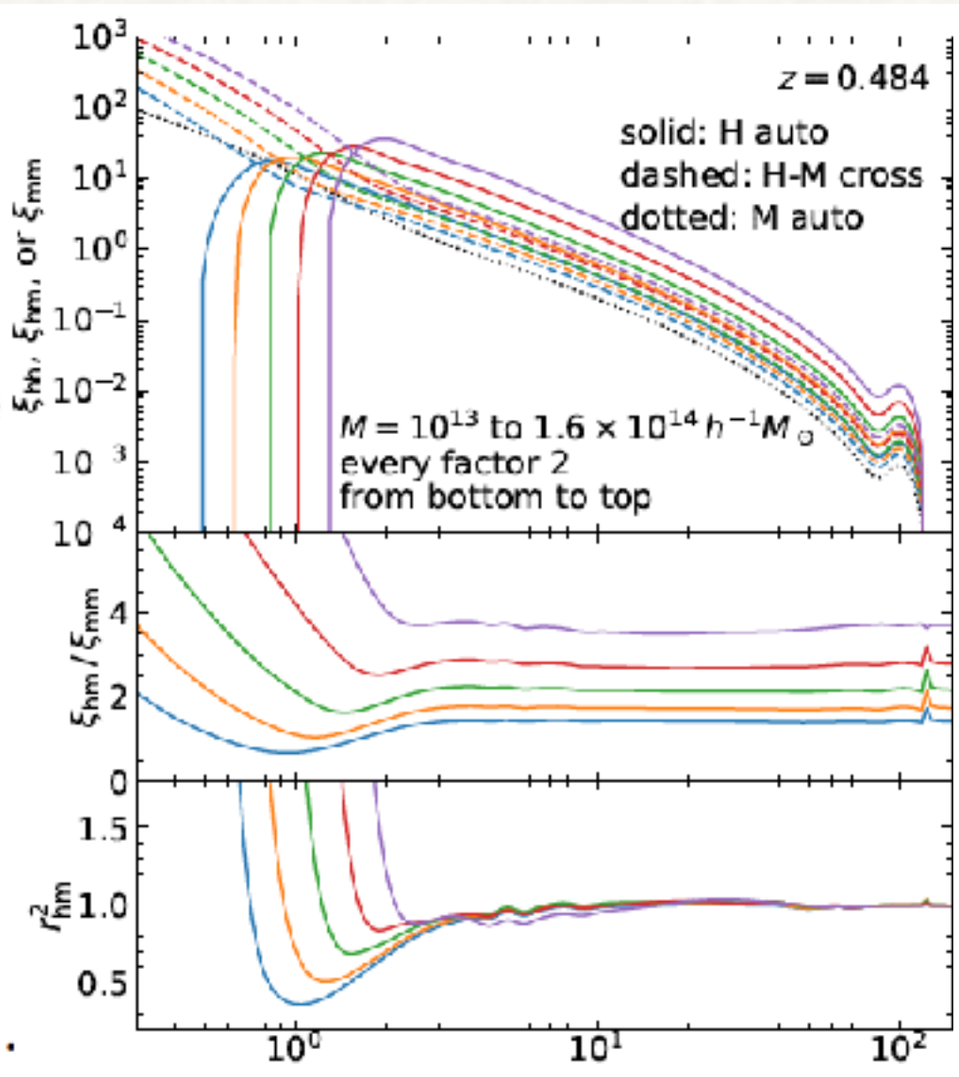
Halos galaxies

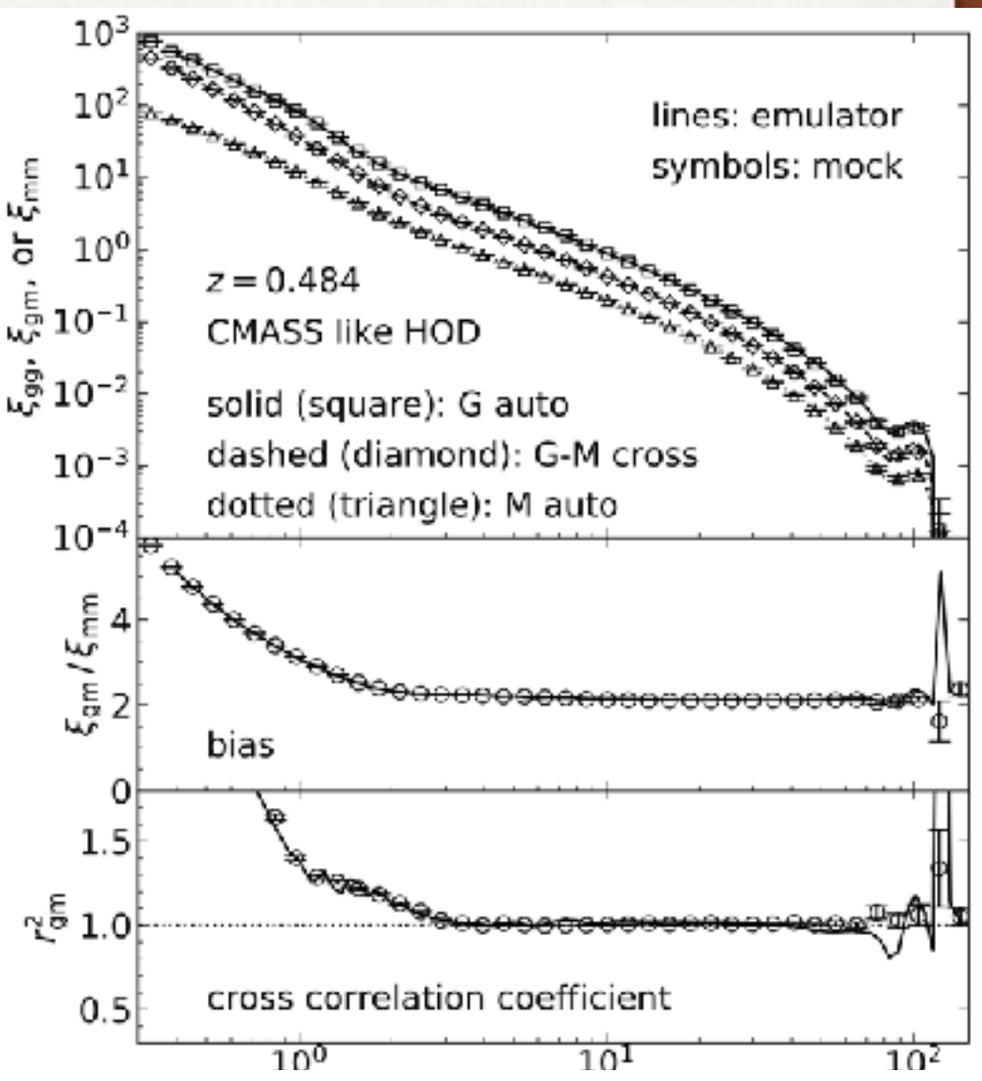
Correlation functions (matter & tracers)

Bias

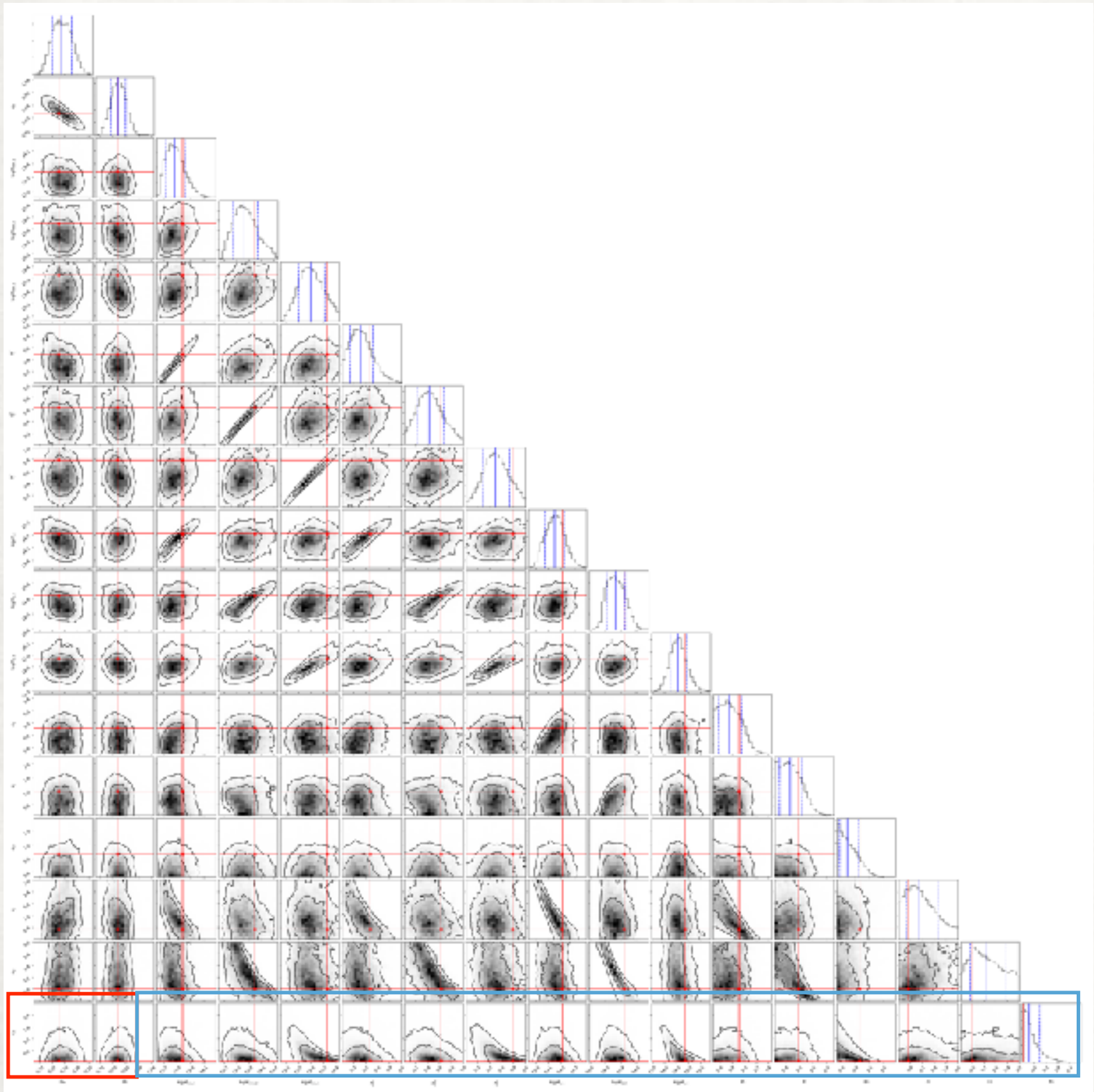
Correl. coeff.

$$r_{\rm gm}^2(r) \equiv \frac{\left[\xi_{\rm gm}(r)\right]^2}{\xi_{\rm gg}(r)\,\xi_{\rm mm}(r)}.$$





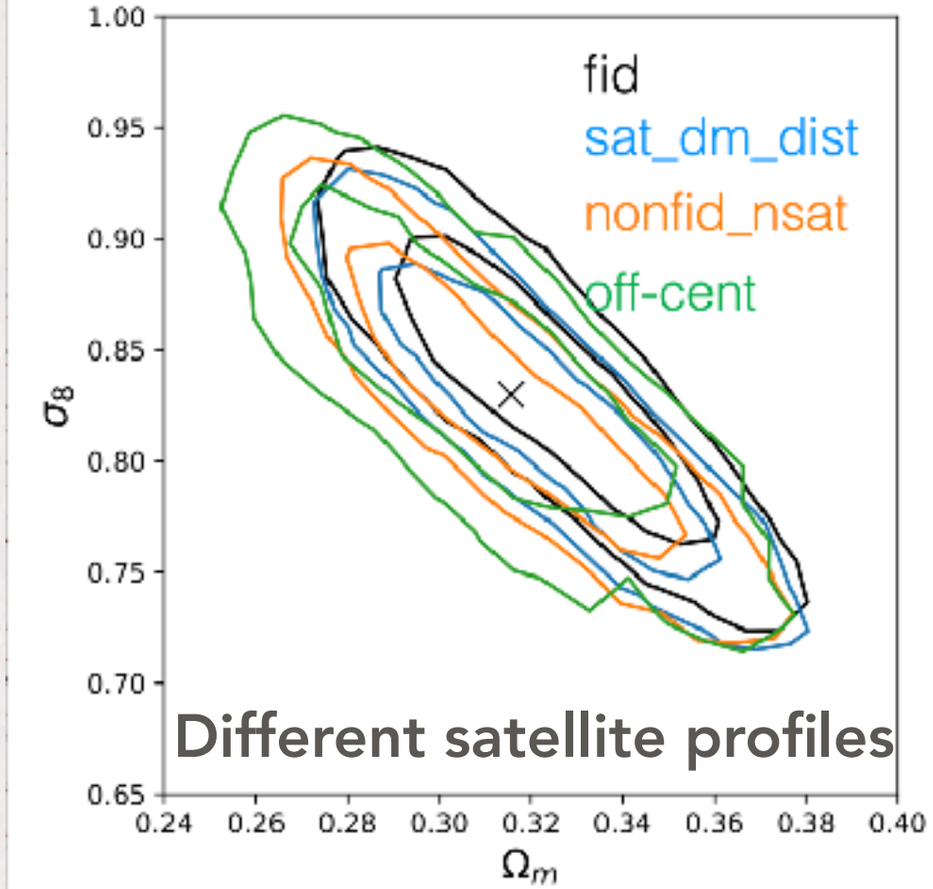
MOCK UNIVERSE VS EMULATOR



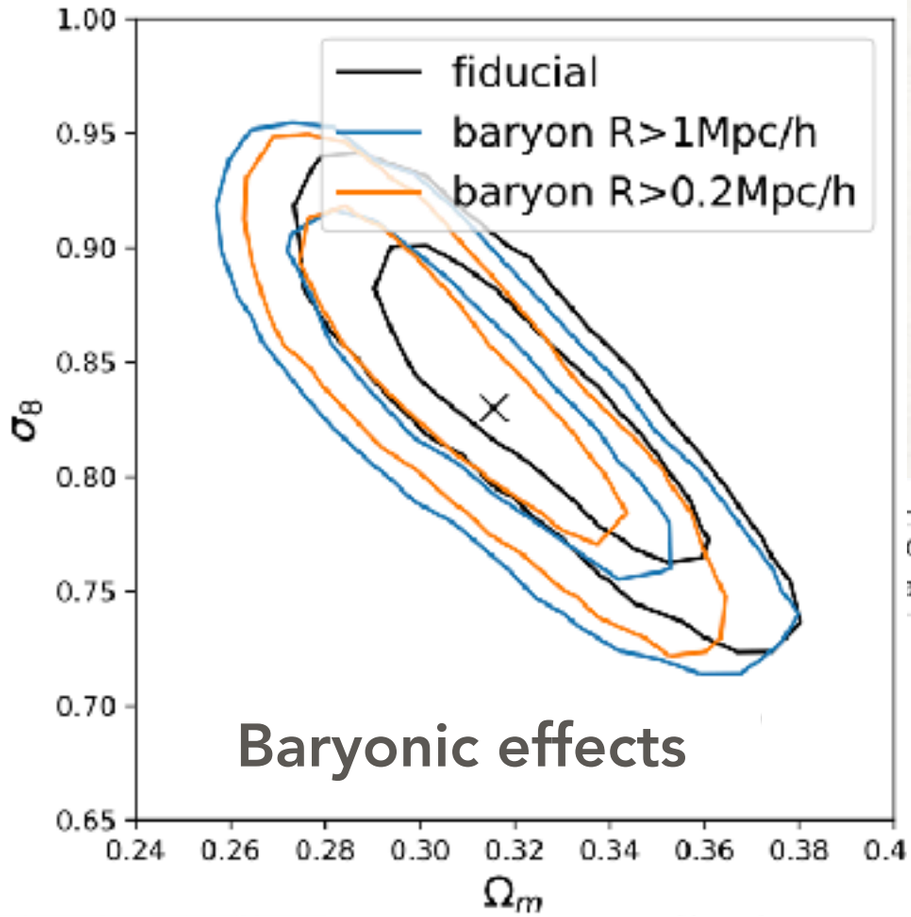
COSMO

GALAXY (HOD + extra)

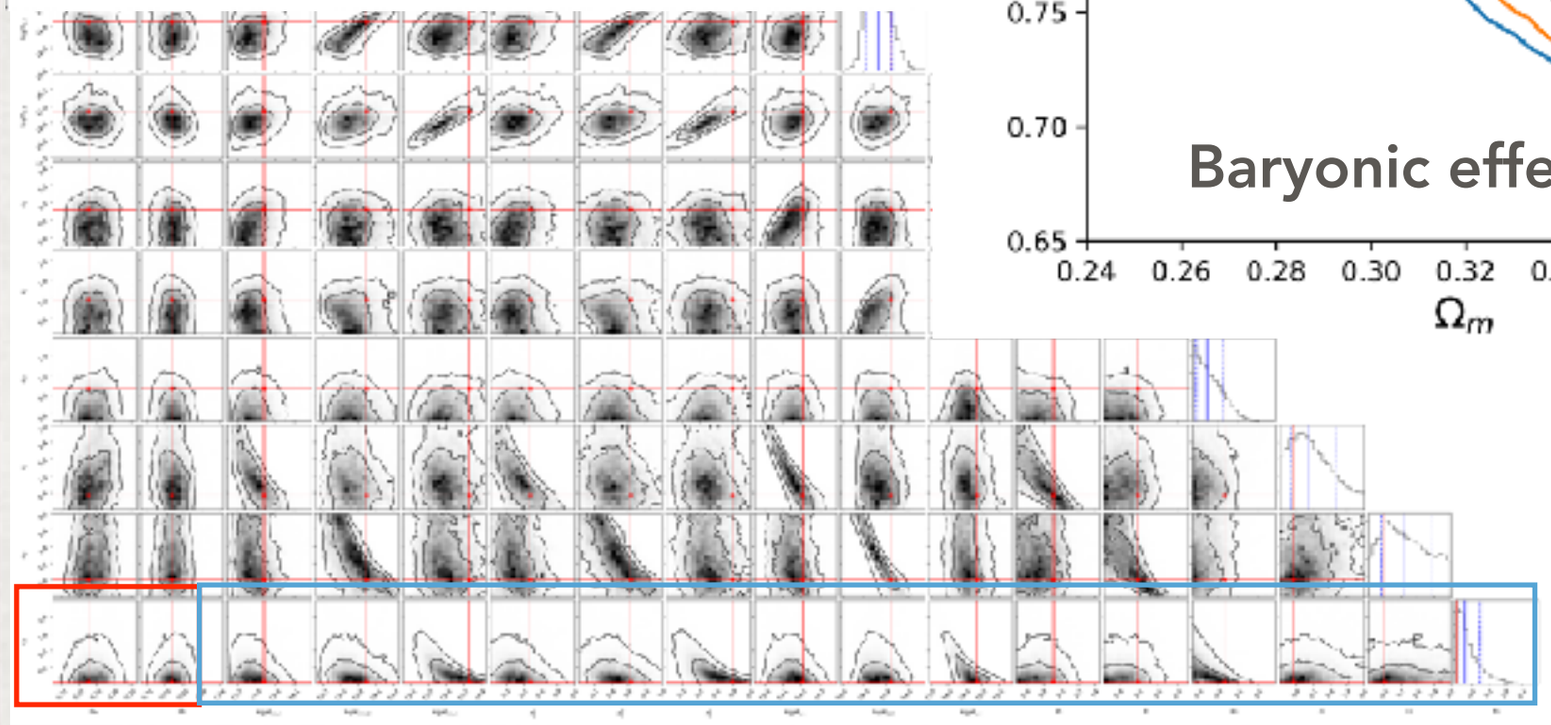
MOCK UNIVERSE VS EMULATOR



Different satellite profiles

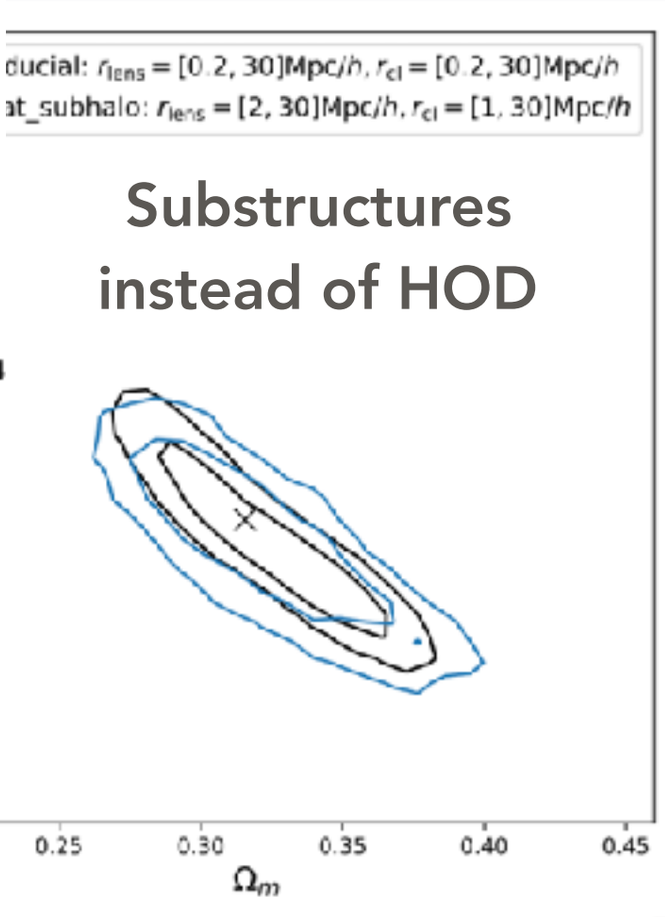


Baryonic effects



COSMO

GALAXY (HOD + extra)



Substructures instead of HOD

SUMMARY

- Modeling halo clustering signal based on simulations
  - **Dark Quest** simulation suite with Latin Hypercube Design
  - Nonparametric regression based on Gaussian Process
  - currently 2~3% accuracy and hopefully this gets better
  - **"Dark emulator" public after the HSC g-g lensing analysis**
- HSC (g-g lensing) + BOSS (w_p) analysis
  - **Break the degeneracy!**
  - **"Galaxy parameters" on small scales (g-g lensing)**
  - **Cosmology on large scale (g-g clustering)**
  - Mock challenge shows good results
- To follow
  - Redshift-space distortions
  - Expand the input parameter space