Fundamental Physics from the Large Scale Structure of the Universe: the next generation of analysis

Uroš Seljak
UC Berkeley/LBNL
February 21 2018

Motivation

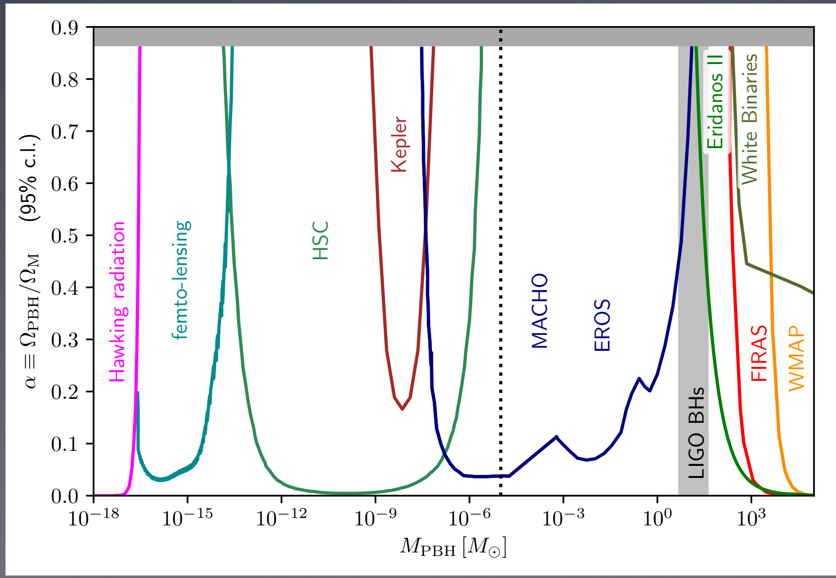
- LSS can probe fundamental physics:
- Growth of structure and BAO constrain dark energy and modified gravity
- Growth of structure, shape of power spectrum and lensing versus galaxy clustering constrains neutrino mass and number of effective neutrino families
- Inflation parameters (slope, running) from power spectrum shape
- Features in power spectrum constrain some models of inflation
- Primordial non-gaussianity probes inflation models and alternatives
- LSS can probe the nature of dark matter (warm or fuzzy dark matter, PBH as dark matter)

Cosmology meets statistics

- Main theme of the talk: using the best possible statistical analysis can greatly enhance statistical power of large scale structure (LSS)
- Example 1: PBH DM constraints from SN1A statistical analysis (Zumalacarregui & US 2017)
- Example 2: towards "optimal" analysis of LSS (US et al 2017)
- Example 3: combining (CMB) lensing and LSS (Schmittfull & US 2017)

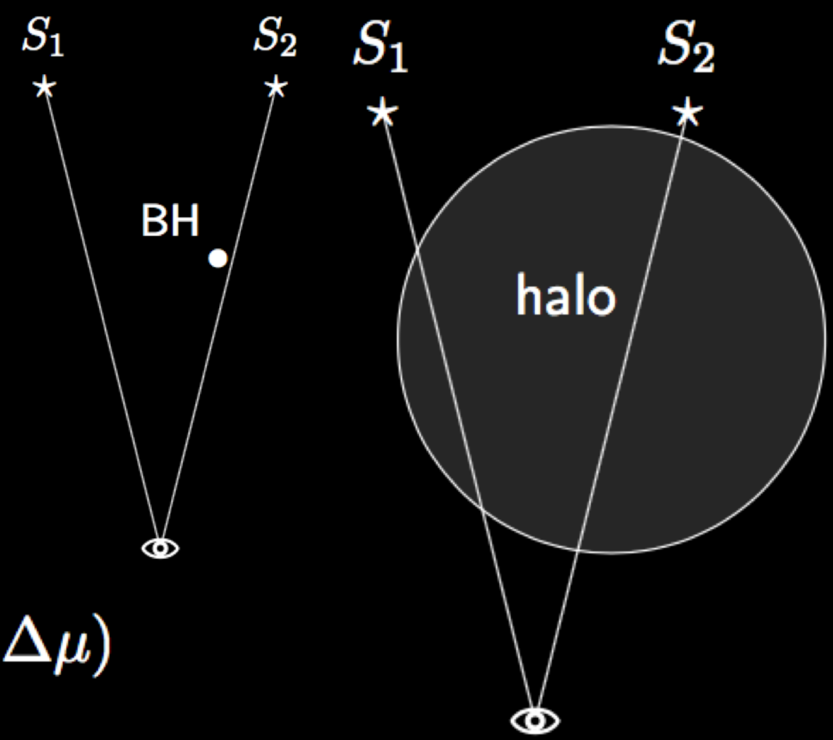
LIGO black holes as dark matter?

- Intriguing: not completely ruled out? (Bird etal, Sasaki etal)
- LIGO rate gives very tight constraint (Misao's talk)
- Clustering of BH and wider BH mass distribution has been invoked to counter some of these limits (Garcia-Bellido etal)



- Perhaps we do not need another method to rule this model out, but I will present a new method that is very clean
- Basic idea: lensing changes SN1A as a standard candle

Lensing by compact objects



Magnifiaction and distance

$$D(z, \mu) = \frac{\bar{D}(z)}{\sqrt{1 + \Delta\mu}}$$

SNe distance modulus

$$m = 5\log_{10}(\bar{D}_L) + 25 - 2.5\log_{10}(1 + \Delta\mu)$$

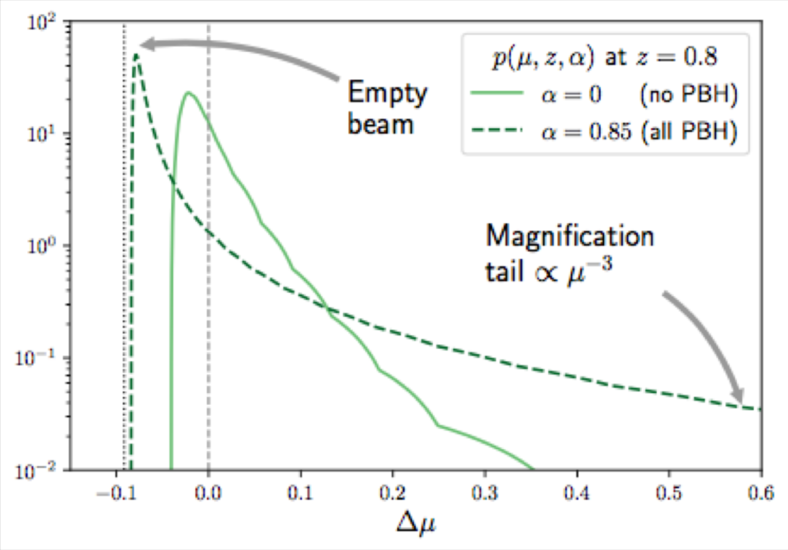
PBH signatures (magnification PDF)

- Most lines of sight: $\sim$ empty universe

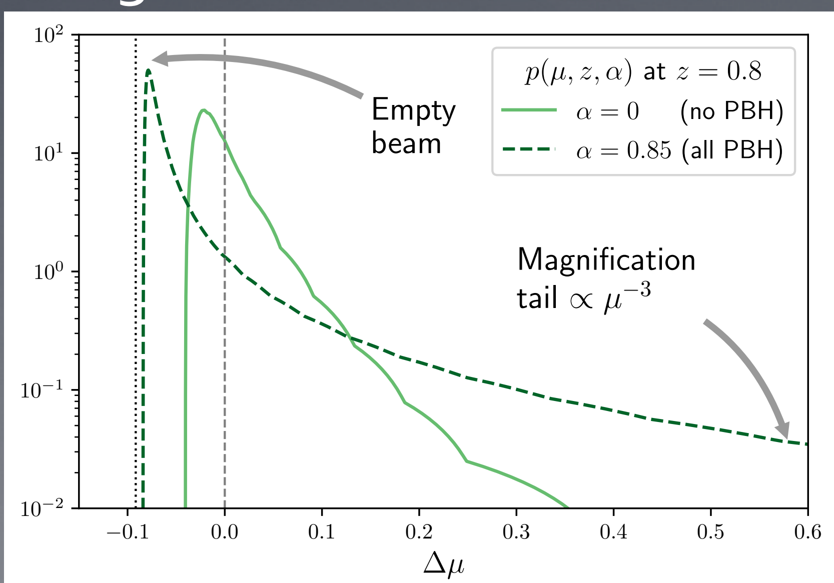
$$\Delta\mu = \left(\bar{D}/D_e\right)^2 - 1 < 0$$

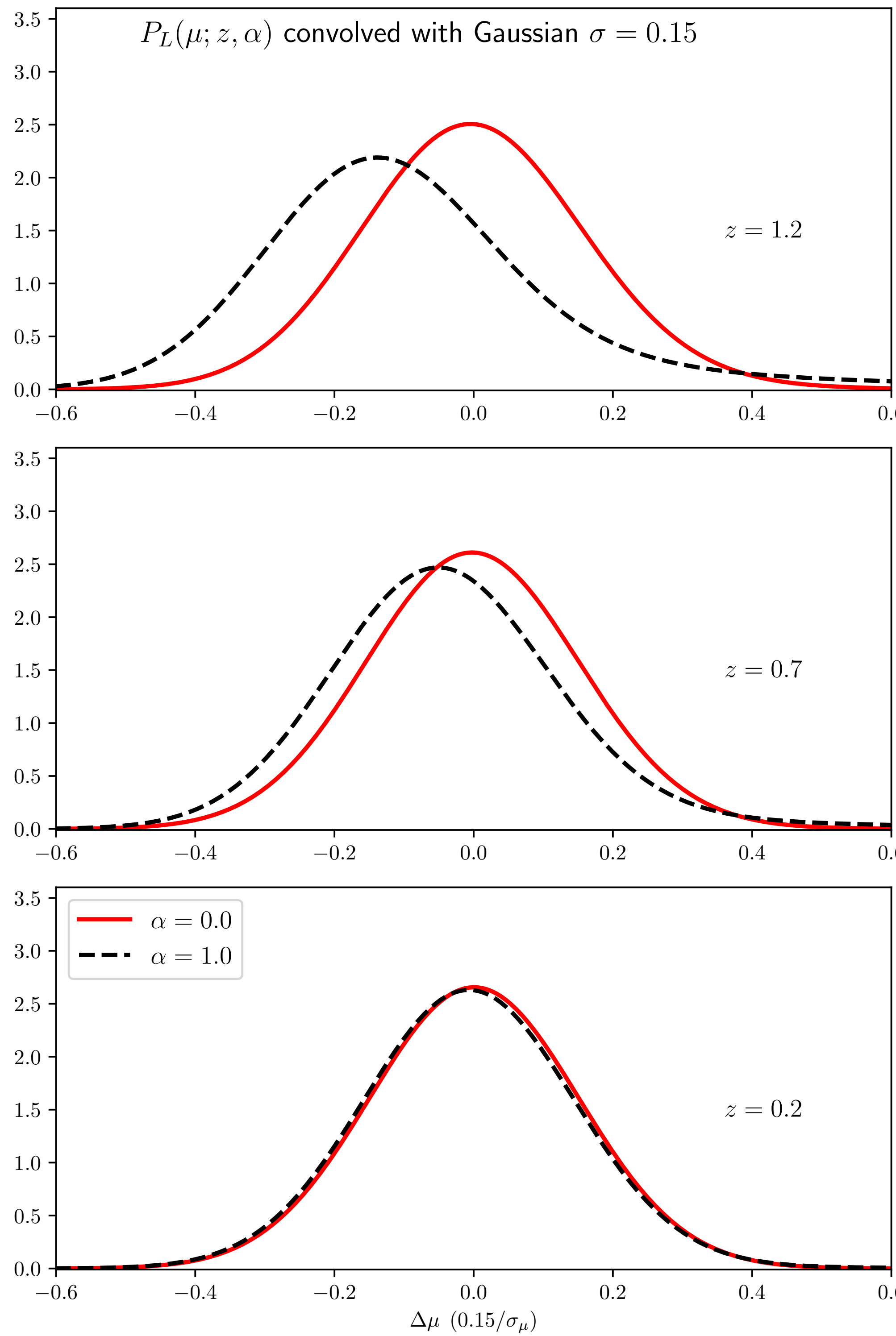
- Few lines of sight: high magnification

$$P(\Delta\mu \gg 1) \sim \Delta\mu^{-3}$$



PBH-only (Rauch '91) convolved with LSS (Seljak & Holz '99)

- Strong redshift dependence
- Signatures: look for redshift dependent change of PDF
- Look for peak PDF away from correct cosmology prediction (degenerate with cosmology, use external priors)
- Look for (absence of) high magnification events
- 



Statistical analysis

- Information is in full (redshift dependent) PDF: we can adopt likelihood analysis with marginalization over all nuisance parameters (noise, LSS, SN stretch and color...): hierarchical Bayesian analysis with analytic marginalizations

- Analytic marginalization: numerical integrals of convolutions of PBH signal with noise, with LSS, with intrinsic SN PDF, all redshift dependent

$$P_L(\mu; z, \alpha) = \int_0^{\frac{\mu}{1-\alpha}} d\mu' P_{LSS}(\mu', z) P_C[\mu - \mu'(1-\alpha), \alpha\mu']$$

$$L_i(\vec{\theta}, \alpha) = \int d\mu P_L(\mu; , z_i, \alpha) P_{SNe}(m_i, \sigma_i, z_i, \mu, \vec{\theta})$$

- SNe are independent, so likelihood is $L=\Pi_i L_i$

SNe population

$$m_{th} = 5\log_{10}\left(\bar{D}_L(z)\right) - 2.5\log_{10}(1+\Delta\mu) + 25$$

$$L_i(\vec{\theta},\alpha) = \int d\mu P_L(\Delta\mu;,z_i,\alpha)P_{SNe}(m_i,\sigma_i,z_i,m_{th},\vec{\theta})$$

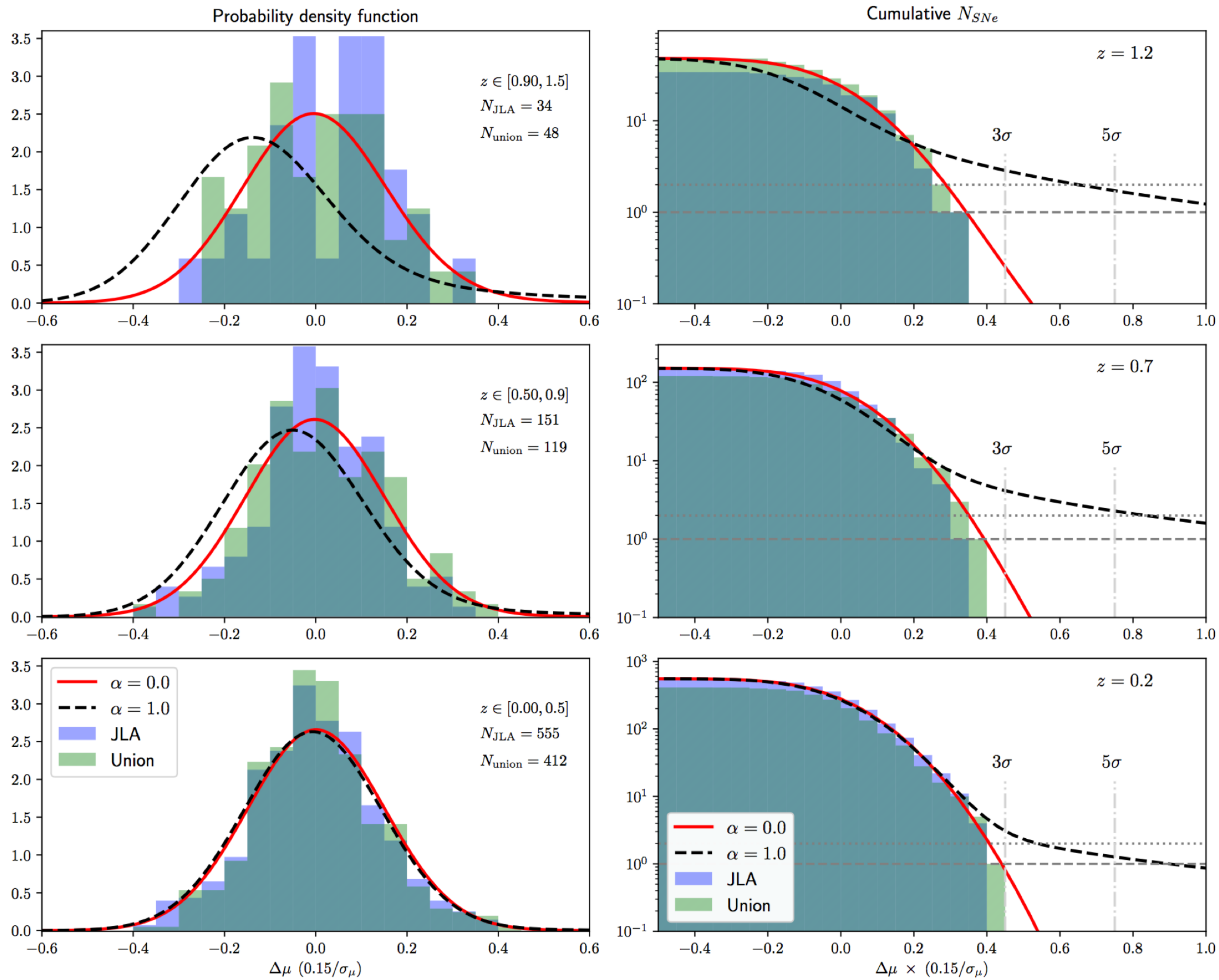
SNe distribution

$$x_i = \frac{1}{\sigma_i}\left(m_{ob,i} - m_{th}(z_i,\Delta\mu) - \bar{m}\right)$$

$$\sigma_i^2 = \sigma_{ob,i}^2 + k_2 - \sigma_L^2(z)$$

$$P_{SNe}(x) = \mathcal{N}\left(1+\mathrm{erf}\left(\frac{k_3 x}{\sqrt{2}}\right)\right)\exp\left(-\frac{1}{2}|x|^{2-k_4}\right)$$

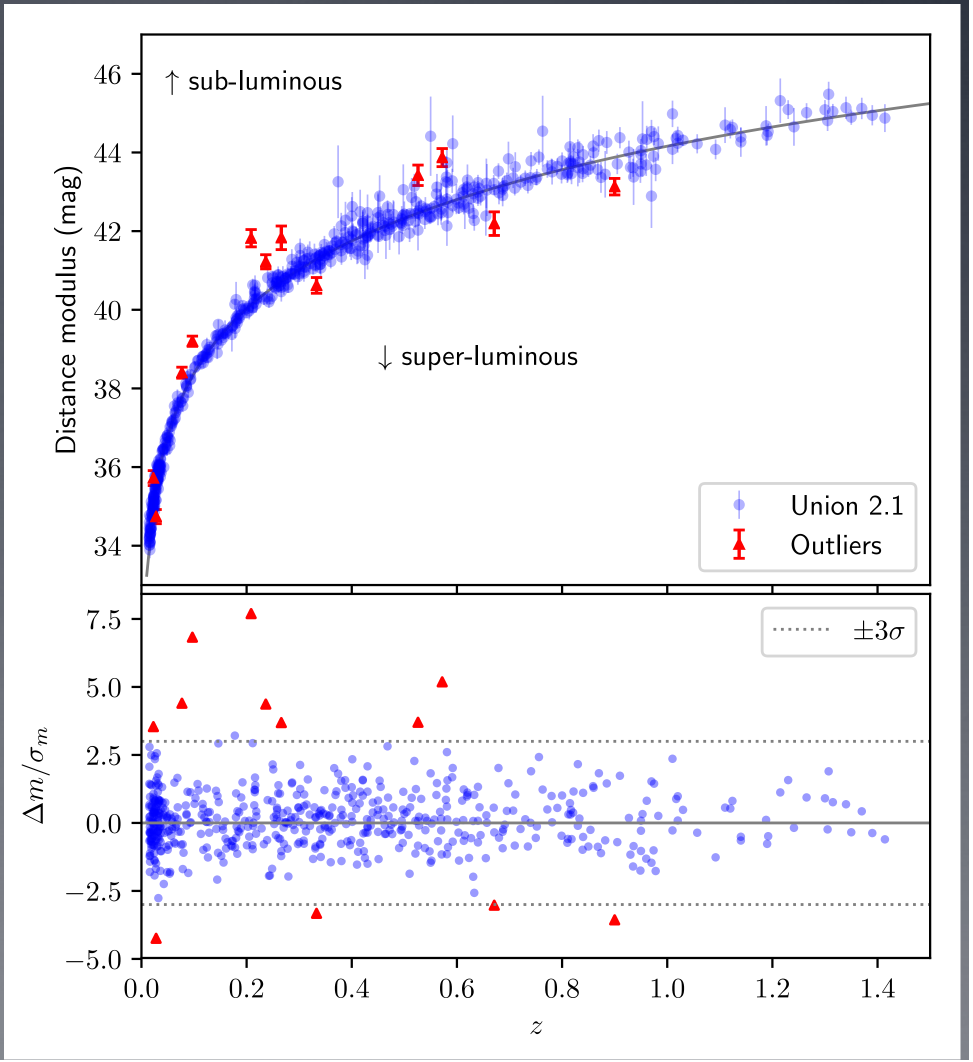
$\theta \supset$ mean ($\bar{m}$), intrinsic scatter (k_2), skewness (k_3), kurtosis ($k4$), standardization $m_{ob,i} = m^{\star}_{B,i} - (M_B - \alpha X_{1,i} + \beta C_i)$



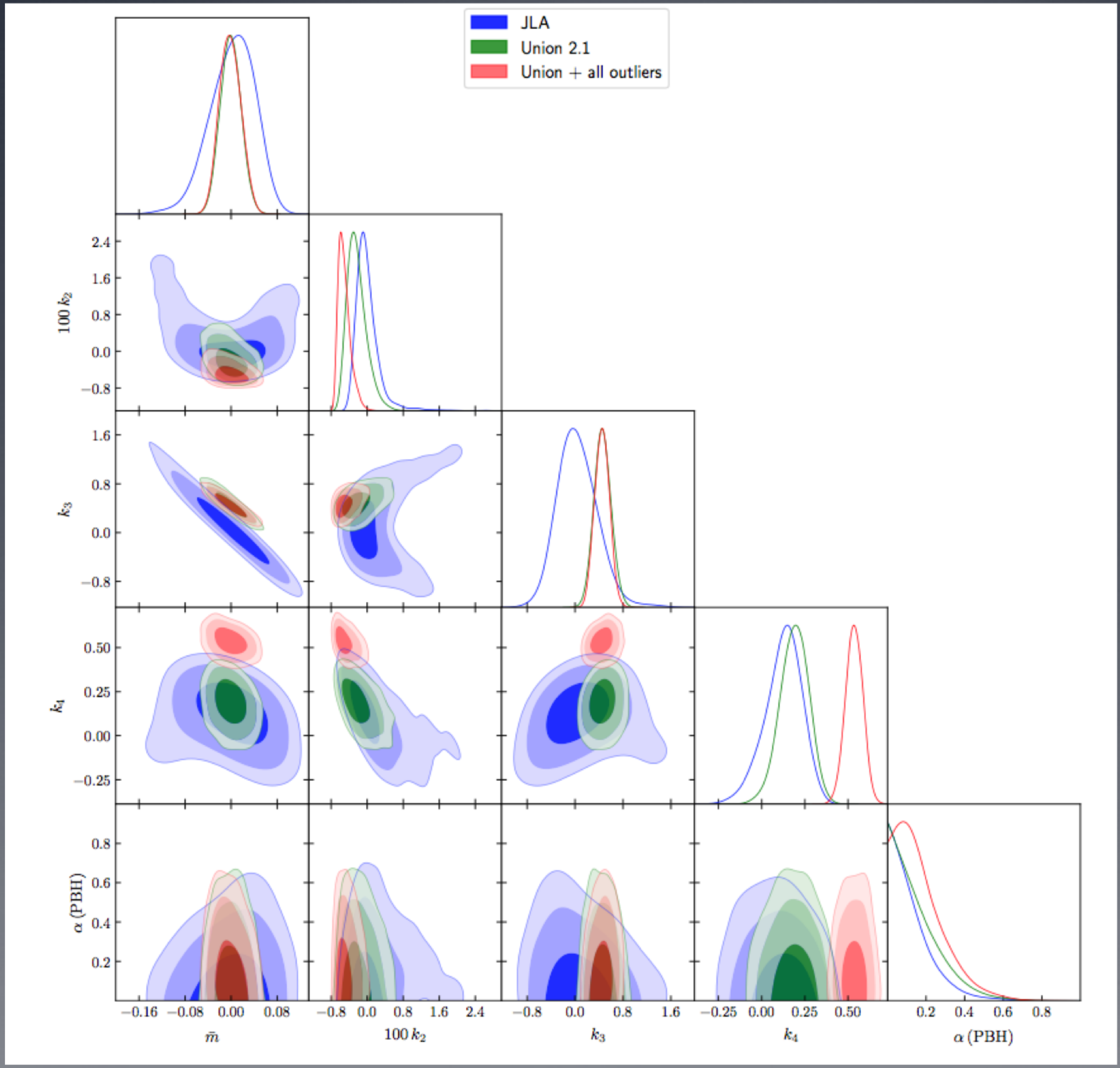
Probability density function
Cumulative N_{SNe}
$z \in [0.90, 1.5]$
$N_{\mathrm{JLA}} = 34$
$N_{\mathrm{union}} = 48$
$z = 1.2$
$z \in [0.50, 0.9]$
$N_{\mathrm{JLA}} = 151$
$N_{\mathrm{union}} = 119$
$z = 0.7$
$z \in [0.00, 0.5]$
$N_{\mathrm{JLA}} = 555$
$N_{\mathrm{union}} = 412$
$z = 0.2$
3σ
5σ
$\alpha = 0.0$
$\alpha = 1.0$
JLA
Union
$\Delta\mu\ (0.15/\sigma_\mu)$
$\Delta\mu \times (0.15/\sigma_\mu)$

Outliers

- SN folks remove outliers
- Most outliers are fainter, not brighter
- No evidence of outliers getting more frequent with redshift

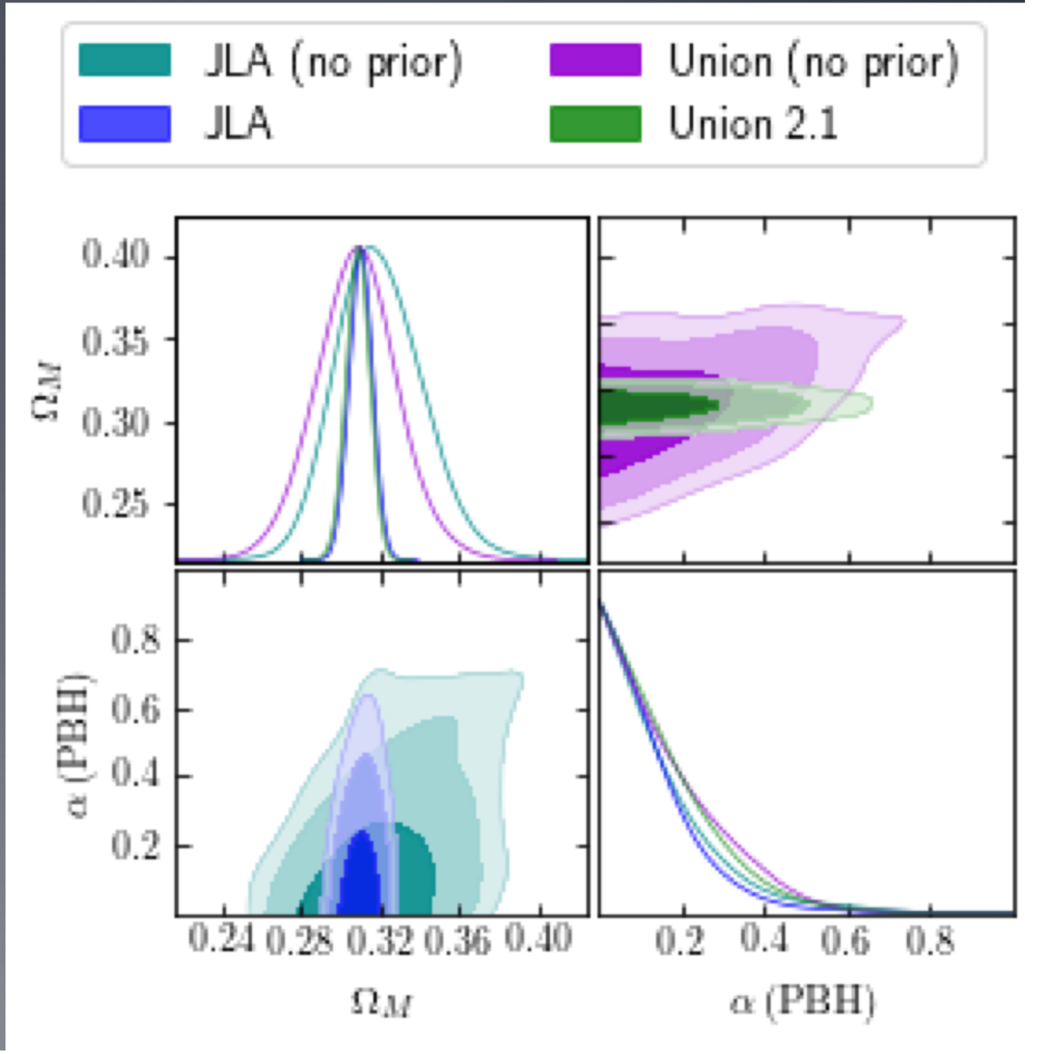


Results



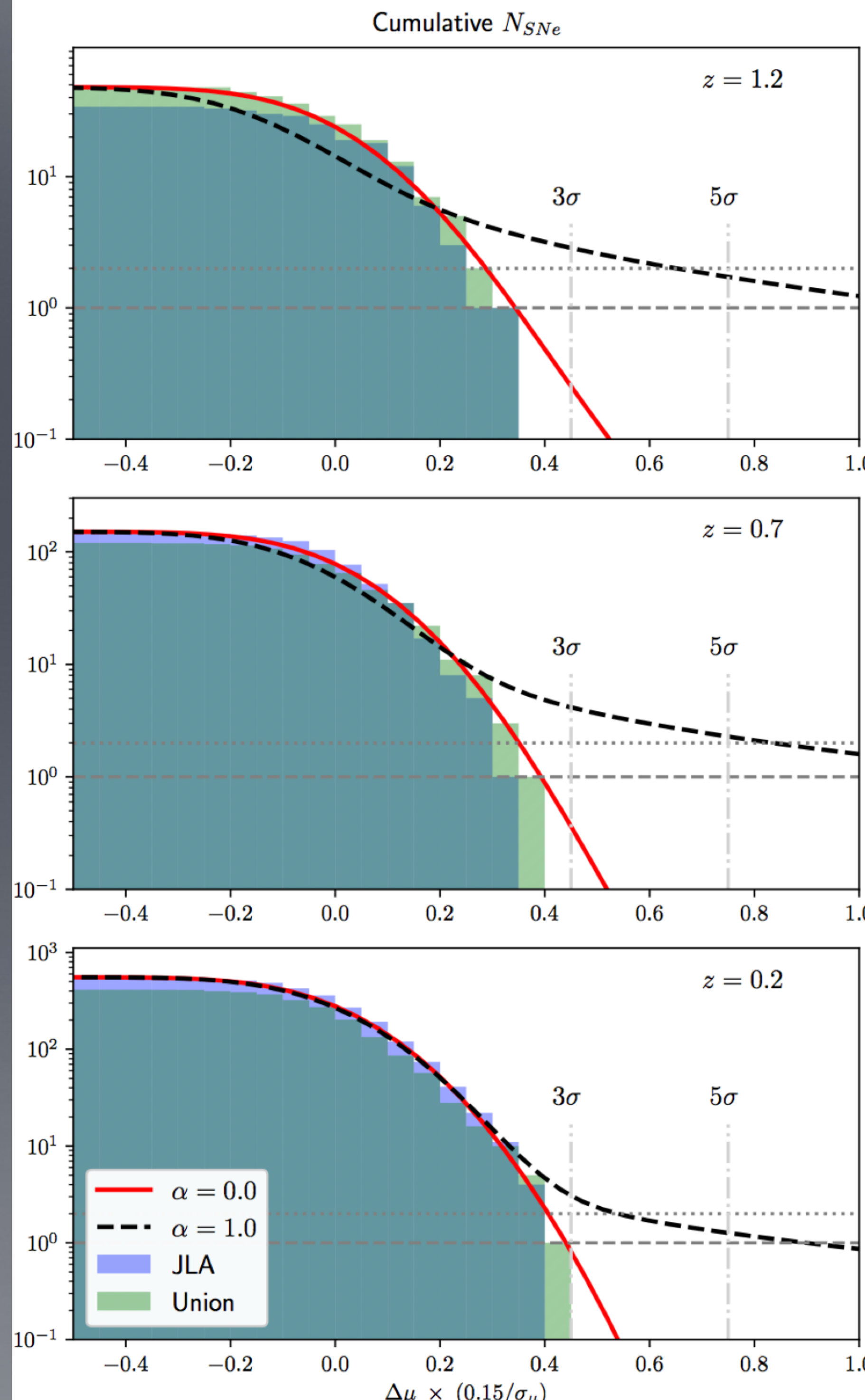
Matter density prior

- We use Planck CMB+SDSS BAO
- We can also relax this prior: not very important
- Suggests tails of pdf more important than peak (since peak PDF is correlated with SN cosmology)



Absence of strongly magnified events

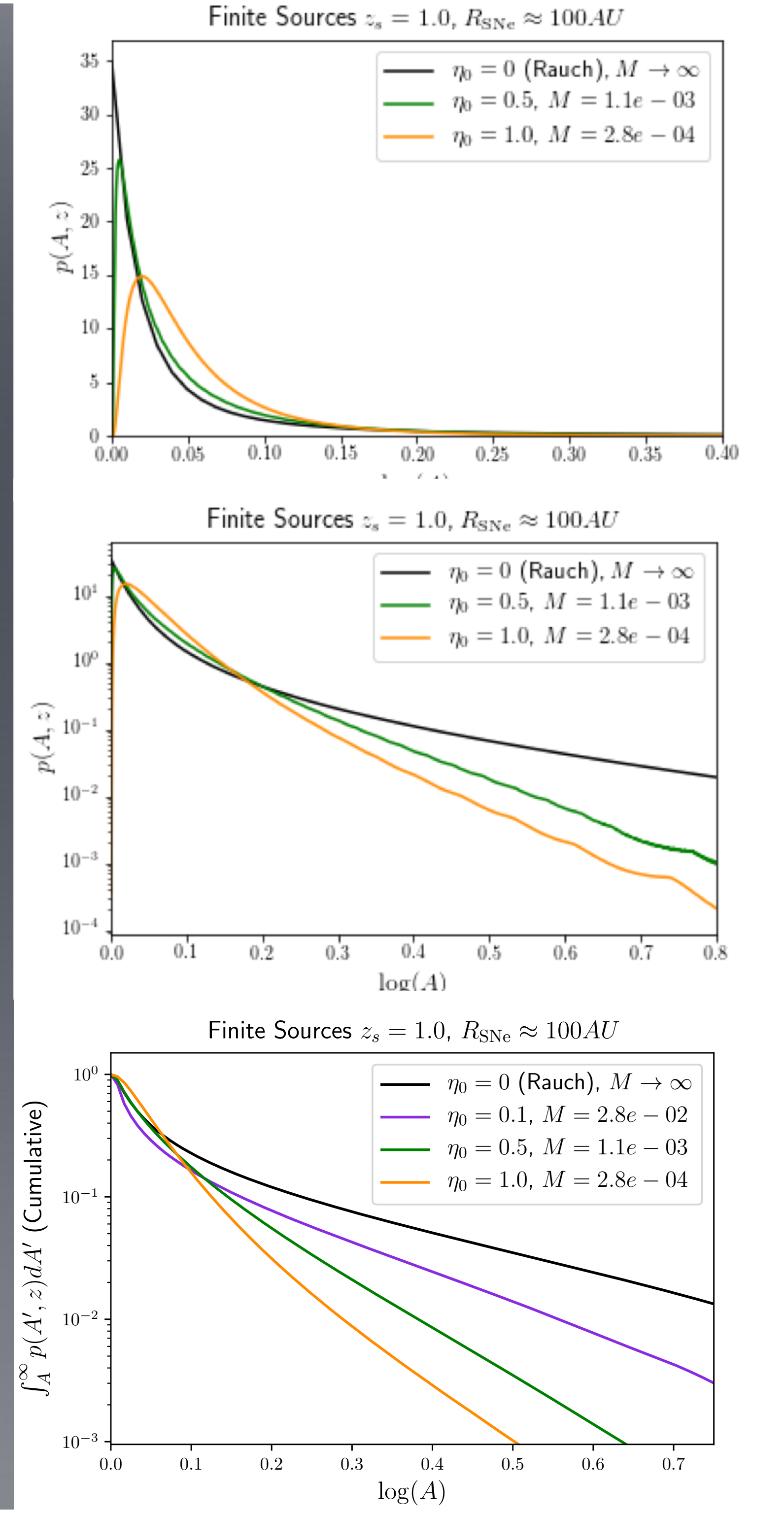
- For $\alpha=1$ there are 10 events predicted with $\Delta\mu>0.4$, none are observed
- P(0,mean=10)= exp(-10) =exp(-20/2), so this is about 4.5 sigma significance
- Absence of events is very constraining!
- Very robust prediction



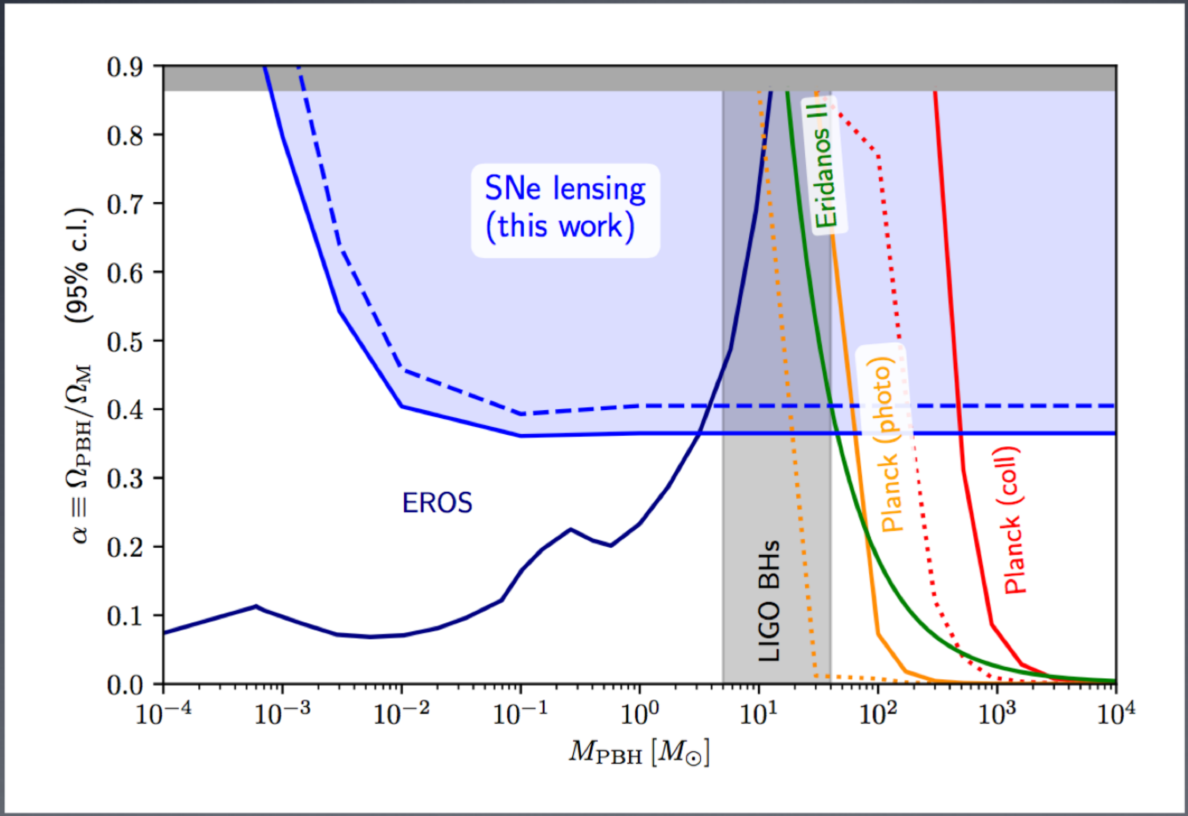
Finite size of SN

- The constraints do not have an upper mass limit and are independent of mass distribution if Einstein radius larger than SN1A size
- Lower mass limit determined by the intrinsic size of SN1A photosphere relative to mean Einstein radius of PBH averaged along the line of sight
- These effects kick in first at high magnifications (smallest impact parameter) and last for low magnification lines of sight (far from PBH)
- SN1A: 20 days peak, 10,000km/s expansion, size of order 1.5×10^{10}km, Einstein radius of order $10^{-3} M_{sun}$
- We did a detailed analysis following Pei 1998

- These finite size effects start to matter around $10^{-2}M_{sun}$ and below
- The predicted number of highly magnified events is suppressed
- There are no useful constraints for PBH mass below $10^{-4}M_{sun}$, whose Einstein radius is small compared to SN1A size



PBH mass range and constraints

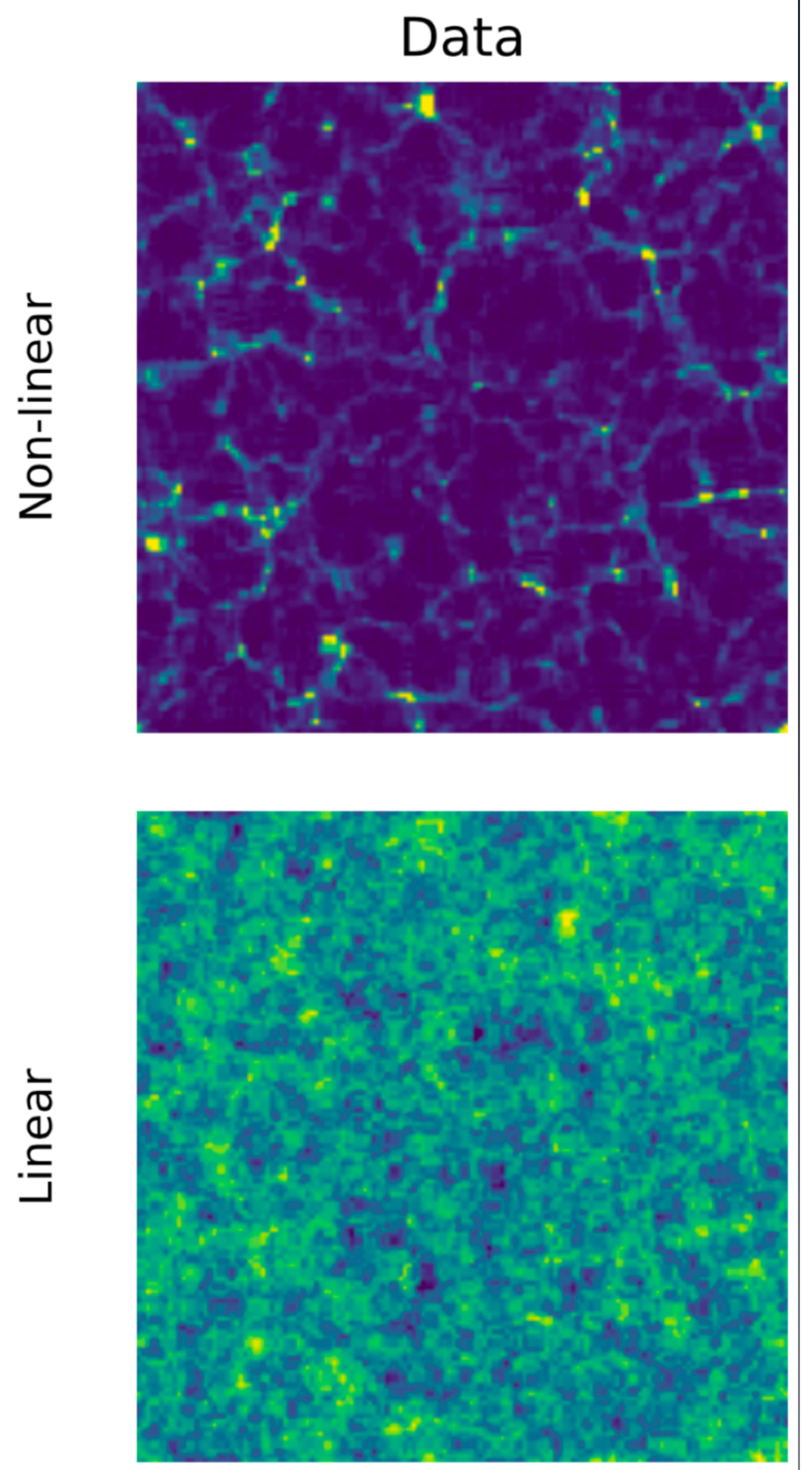


These constraints are on total DM fraction, independent of the mass distribution

The constraints state that $\alpha<0.4$ in PBG with $M>10^{-2}M_{sun}$

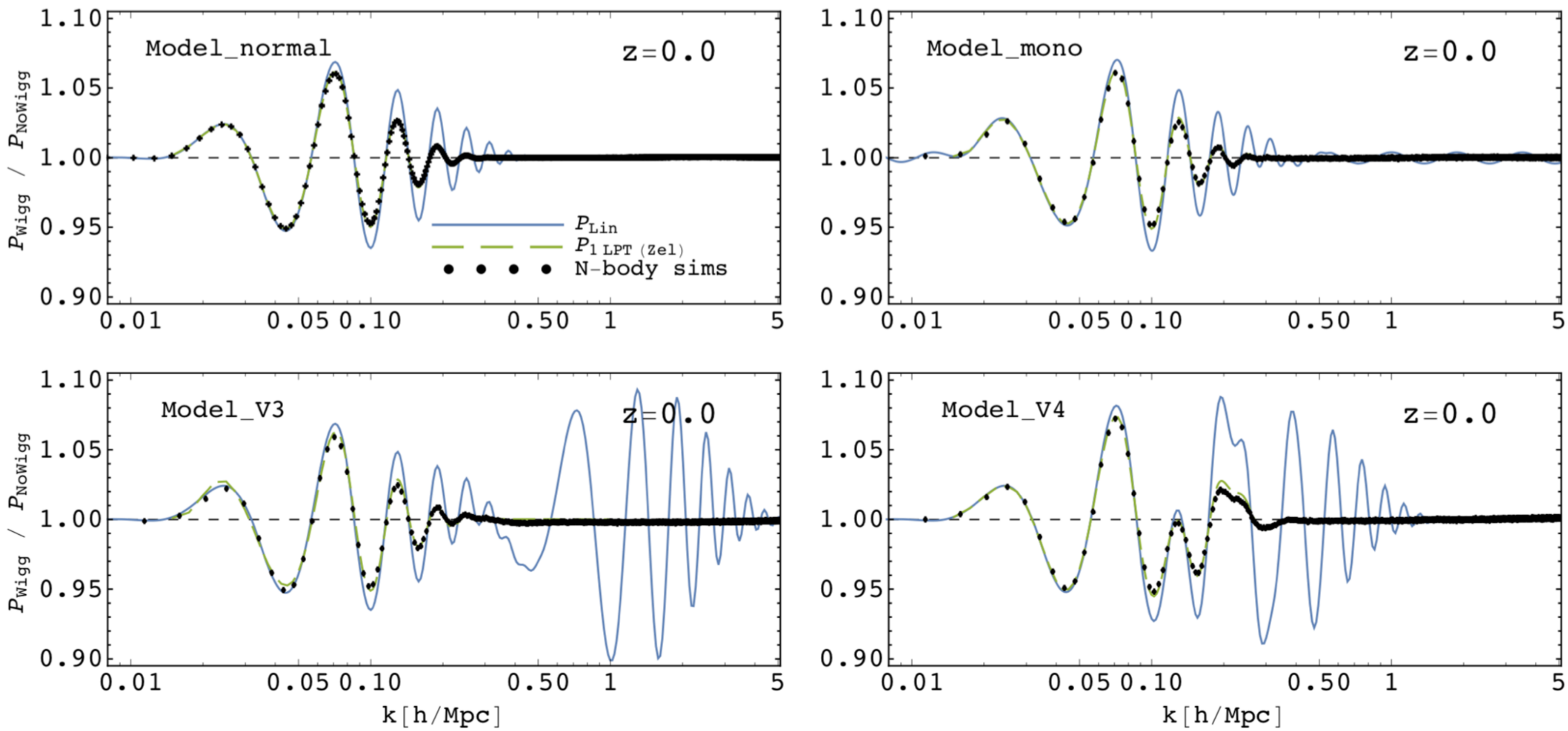
Statistical analysis of LSS

- LSS is very non-Gaussian: nonlinear evolution of structure
- How do we analyze such data?
- Workhorse: 2-point function (power spectrum or correlation function)
- Also need its covariance matrix (4-point function)
- Beyond power spectrum: higher order correlations (3-point etc)
- Peaks counting (clusters, or just density peaks)
- Voids, nonlinear transforms (clipping, lognormal…)
- Topological measures (genus…)
- BAO reconstruction
- Unclear how to combine these different statistics (mock sims)
- Nongaussian: need covariance matrix
- Do we just keep trying with new statistics or is there are way to extract maximal information and prove it?
- Can we convert 3d NL into 3d linear?



Example: features in P(k)

- Features are destroyed in nonlinear P(k) at high k due to nonlinear evolution
- Can we recover it back from non-Gaussian information?



A new approach: statistical Bayesian analysis with full forward model

(w. Yu Feng, G. Aslanyan, C. Modi

- Map making in cosmology: unique in that prior is well defined
- Use data likelihood together with the prior and forward model to maximize posterior

posterior prior data likelihood

$$L(\boldsymbol{s}|\boldsymbol{d}) = (2\pi)^{-(M+N)/2}\det(\boldsymbol{S})^{-1/2}\det(\boldsymbol{N})^{-1/2}\exp\left(-\frac{1}{2}\left\{\boldsymbol{s}^{\dagger}\boldsymbol{S}^{-1}\boldsymbol{s} + [\boldsymbol{d}-\boldsymbol{F}(\boldsymbol{s})]^{\dagger}\boldsymbol{N}^{-1}[\boldsymbol{d}-\boldsymbol{F}(\boldsymbol{s})]\right\}\right)$$

- Prior: gaussian initial density Fourier modes s with linear power spectrum S, acts as a regularization
- Forward model F(s): take initial density modes and evolve them to final galaxy positions F(s). This is a full simulation
- Data d: anything can be used: galaxy position, magnitudes, colors, noisy spectroscopy
- Data likelihood: often gaussian, diagonal in real space, with noise variance N

Solving full nonlinear problem: how to make the best possible MAP?

- To get the minimum variance map: maximize a posterior (MAP), ie solve the optimization problem

$$L(\boldsymbol{s}|\boldsymbol{d}) = (2\pi)^{-(M+N)/2} \det(\boldsymbol{S})^{-1/2} \det(\boldsymbol{N})^{-1/2} \exp\left(-\frac{1}{2}\left\{\boldsymbol{s}^\dagger \boldsymbol{S}^{-1} \boldsymbol{s} + [\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})]^\dagger \boldsymbol{N}^{-1} [\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})]\right\}\right)$$

- How to predict data F(s) given initial modes s: run a simulation for each configuration of s
- Typical size: 3Gpc^3 volume, resolution of 1Mpc: 10^{10} cells
- How to find the maximum posterior for 10^{10} modes s? Curse of dimensionality: $V=2^N$, $N=10^{10}$, means we cannot search blindly

How to find MAP in 10^{10} parameter space?

- Maximize posterior=minimize cost function

$$\chi^2(\boldsymbol{s}) = \boldsymbol{s}^\dagger \boldsymbol{S}^{-1}\boldsymbol{s} + [\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})]^\dagger \boldsymbol{N}^{-1}[\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})],$$

$$\chi^2(\boldsymbol{s}) = \chi_0^2 + 2\boldsymbol{g}\boldsymbol{\Delta s} + \boldsymbol{\Delta s}\boldsymbol{D}\boldsymbol{\Delta s},$$

$$\boldsymbol{g} = \frac{1}{2}\frac{\partial \chi^2}{\partial \boldsymbol{s}} = \frac{\boldsymbol{s}_m}{S} - \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}[\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s}_m)]$$

$$R_{ij} = \frac{\partial F(\boldsymbol{s}_m)_i}{\partial s_j}$$

gradient

$$\boldsymbol{D} = \frac{1}{2}\frac{\partial \chi^2}{\partial \boldsymbol{s}\partial \boldsymbol{s}} = \boldsymbol{S}^{-1} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R} + \boldsymbol{F}''[\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s}_m)]$$

Hessian

$$\frac{\partial \chi^2(\boldsymbol{s})}{\partial \Delta \boldsymbol{s}} = 0,$$

$$\boldsymbol{\Delta s} = -\boldsymbol{D}^{-1}\boldsymbol{g}.$$

Newton's method

Need a gradient R_{ij}: derivative of a full simulated data wrt all initial modes s

Also need nonlinear model F(s): a full simulation

Need to compute fast F(s) and its gradient

Our approach: try any optimizer that exists. Currently we are doing Gauss Newton with trust region and conjugate gradient

Needs gradients

Gradient with backpropagation

- One needs analytic derivative of every final data point with respect to every initial mode!
- Chain rule, applied to kick and drift PM operators
- Matrix products, all time steps need to be stored for backward prop.

Yu Feng slide

The basics of AD start from the chain rule of differentiation, which claims the following:

If we have two functions $y = f(x)$ and $z = g(y) = g(f(x))$, then

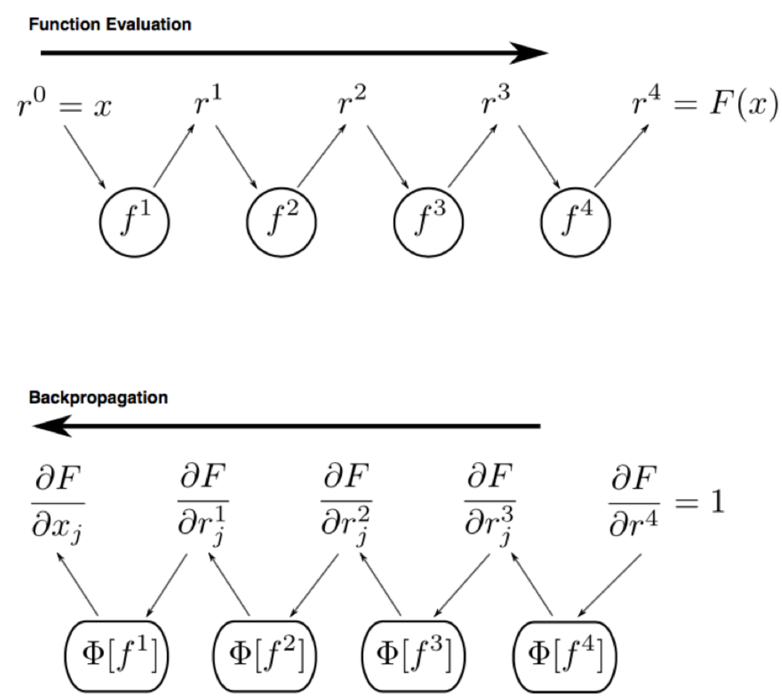
$$\frac{\partial z_j}{\partial x_i} = \sum_k \frac{\partial z_j}{\partial y_k}\frac{\partial y_k}{\partial x_i} = \frac{\partial z_j}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_i}.$$

We see that the chain rule converts a gradient of nested functions to a sequence of tensor products.

Let's now consider a scalar that comes from the nested evaluation of n functions,

$$F(x) := \left(f^1 \odot \cdots \odot f^n\right)(x) = f^n(f^{n-1}(\cdots(f^1(x))\cdots))).$$

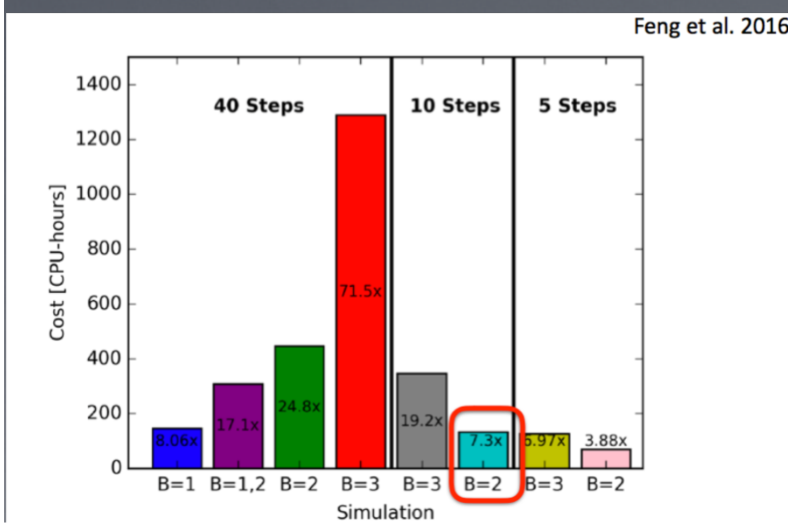
f^i maps to concepts in real-world problems:

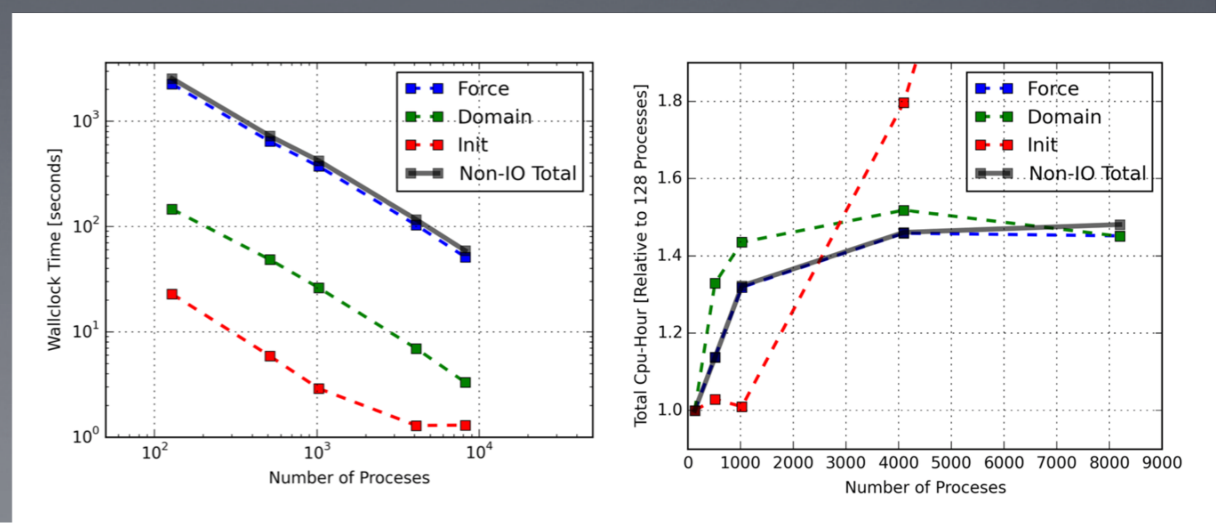


Fast forward model: FastPM

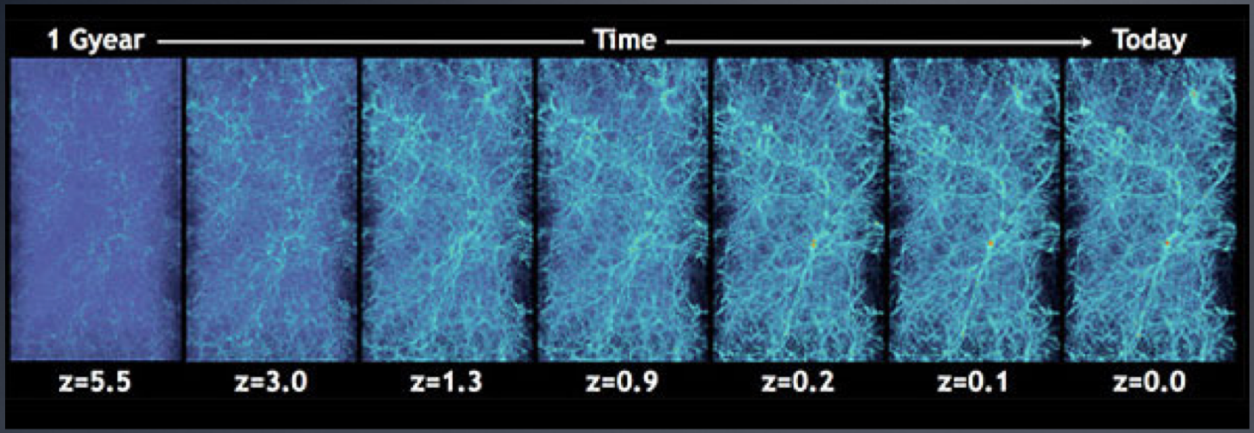
- Need a fast simulation that predicts the galaxy data sufficiently well
- FastPM: PM which enforces correct evolution on large scales even with few time steps (typical simulation 1000+ steps)
- Kick-Drift scheme is exact on Zeldovich (different from usual PM)
- Strong scaling tested to 10^4 cores (pencil FFTs)
- 5-10 time steps already give very good results: 100 CPU hours (minutes of real time) for 10^{10} particles

Feng etal 2016



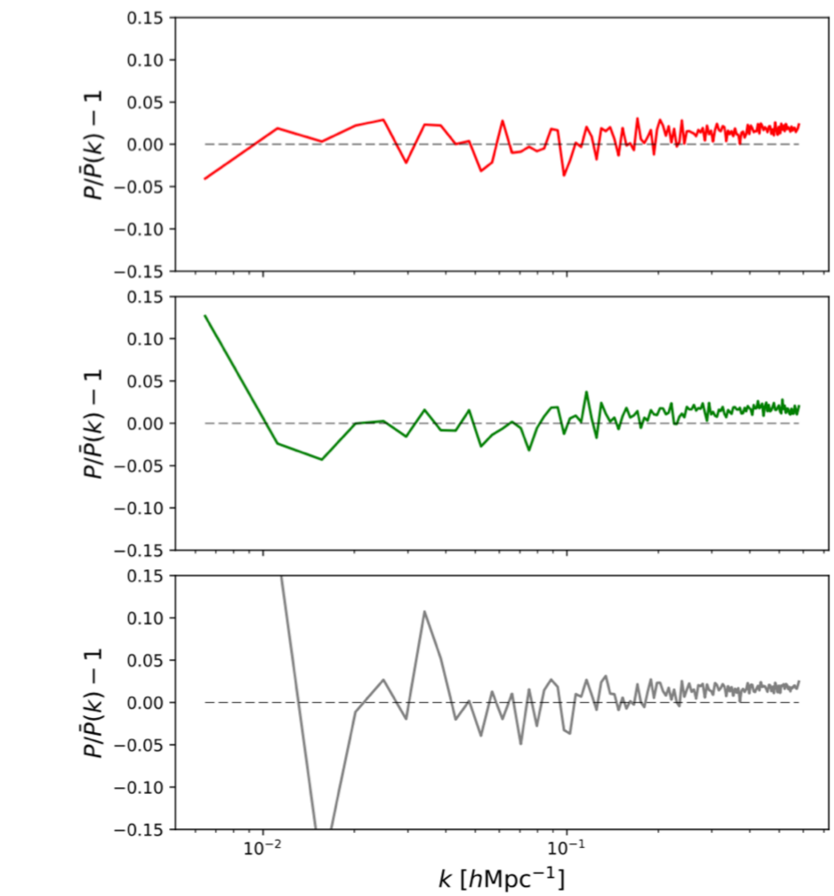


FastPM performance on halos

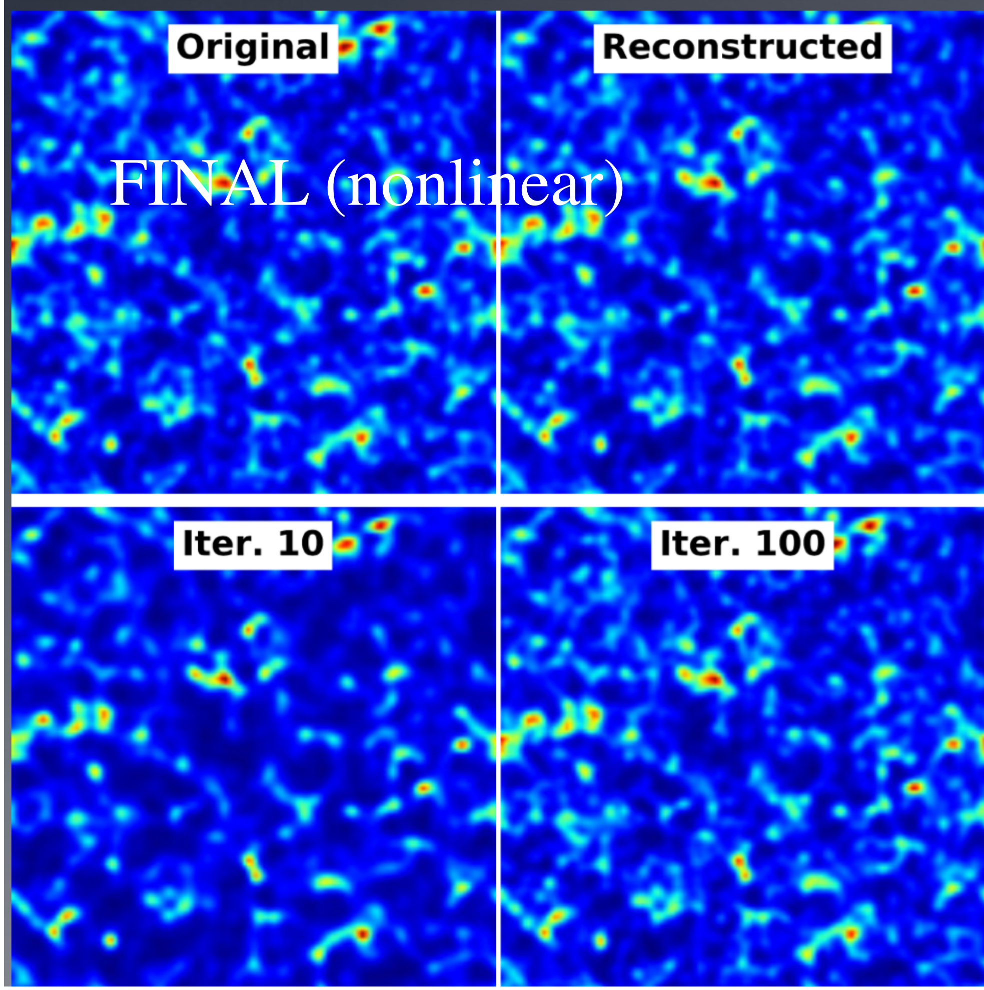


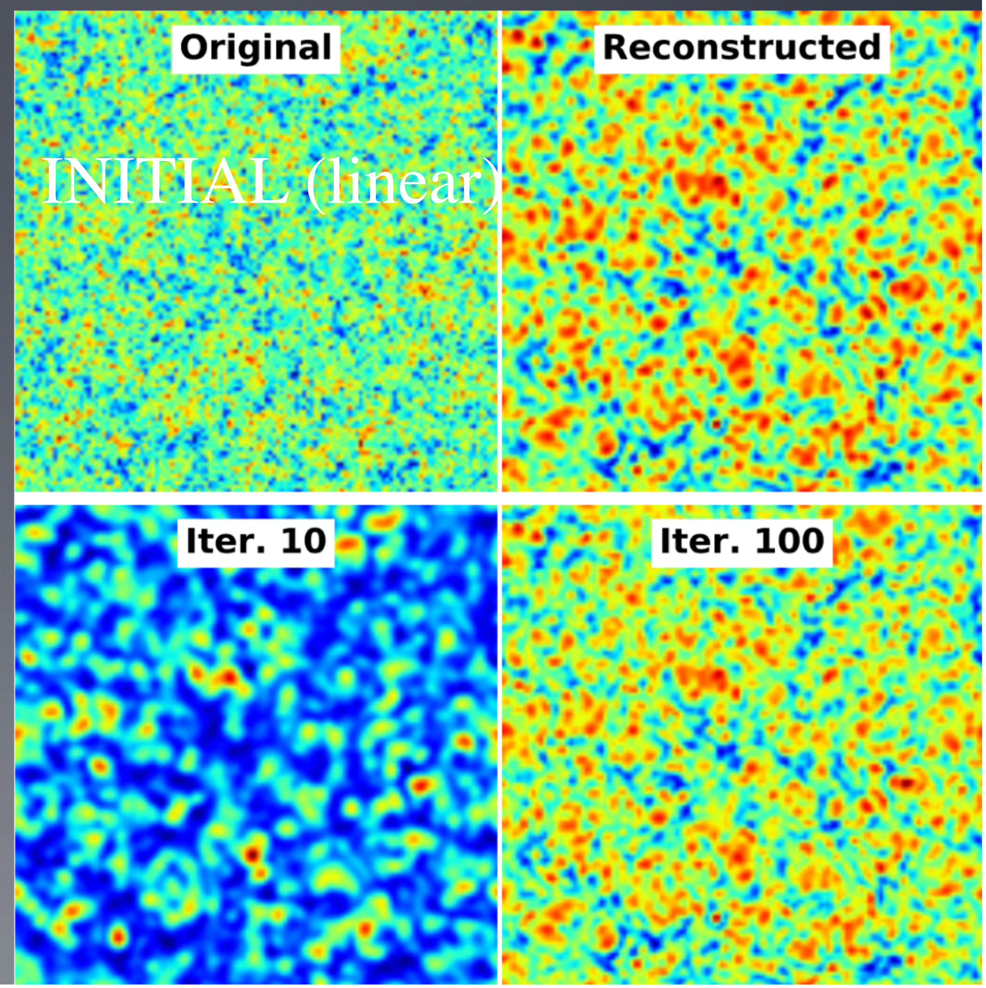
Elena Massara, Yu Feng, US

Comparison against very high resolution simulation: 1-2% accurate for 5 time steps using abundance matching of halos

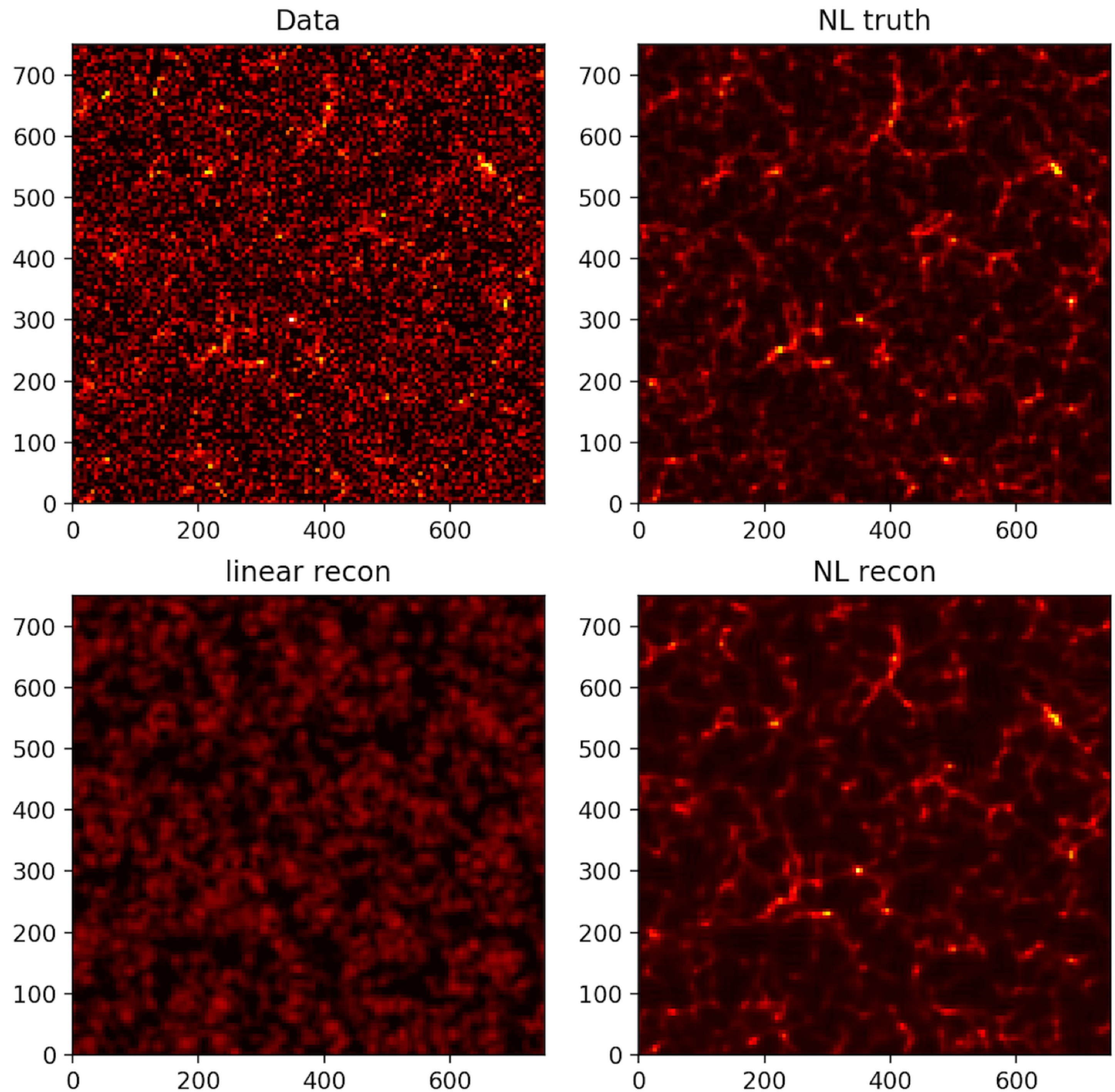


Reconstruction of initial conditions of our universe and final density map for a toy dark matter case with low noise (with Grigor Aslanyan, Yu Feng and Chirag Modi): 230 simulation calls

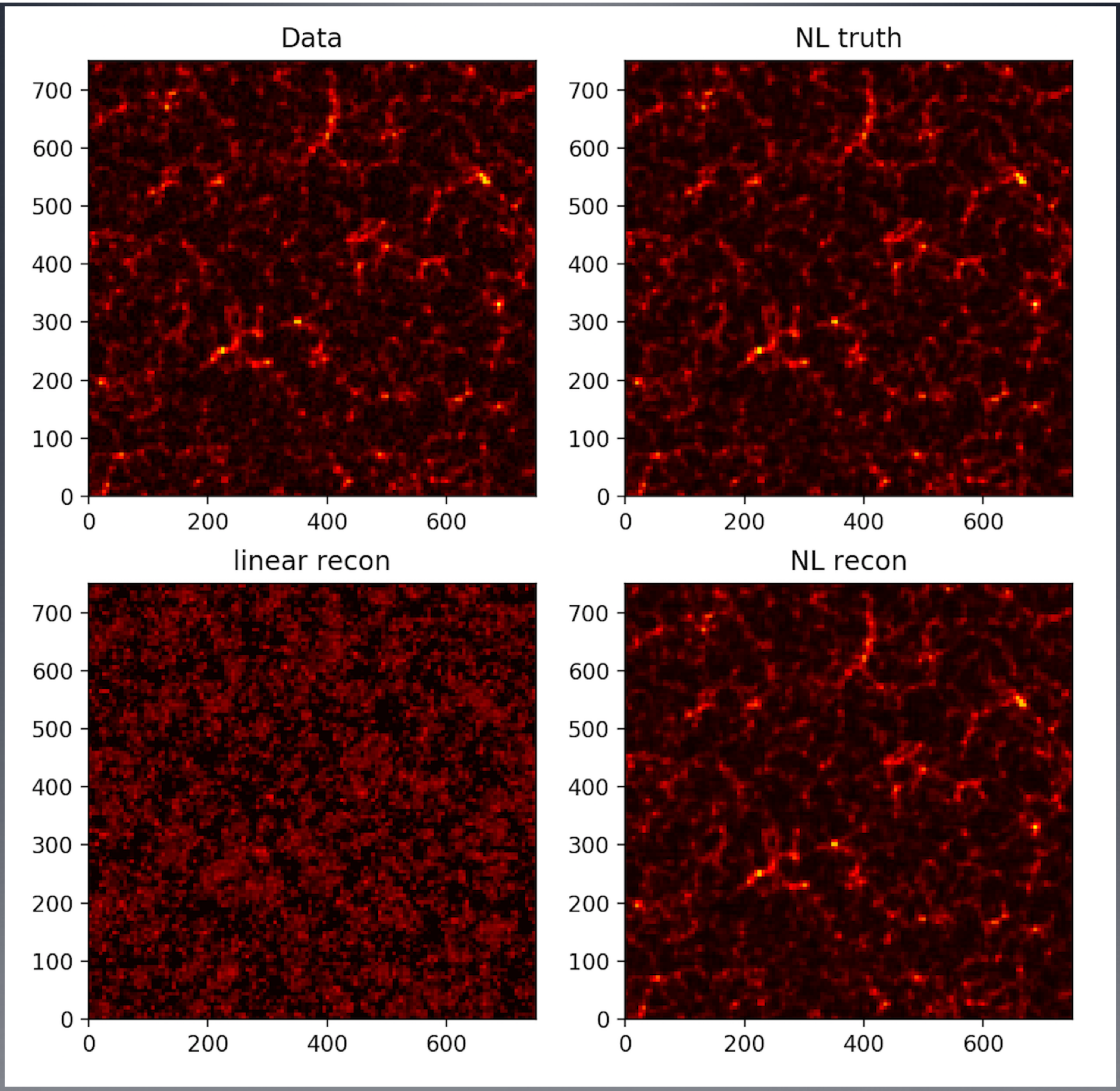




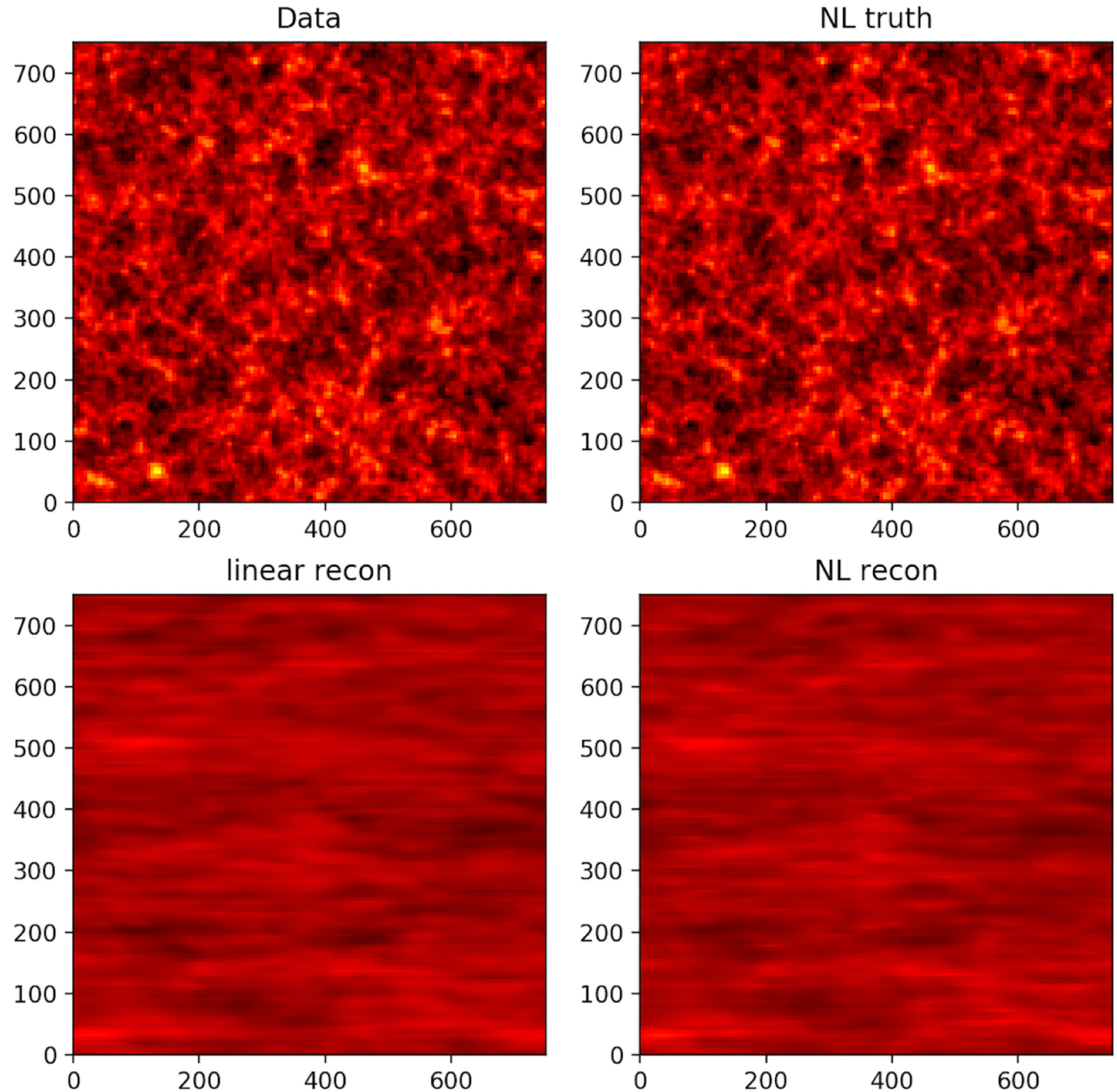
- “high” noise (P=1000Mpc/h^3), low smoothing
- 750Mpc/h box, 128^3
- High k suppressed
- Slices 6Mpc/h



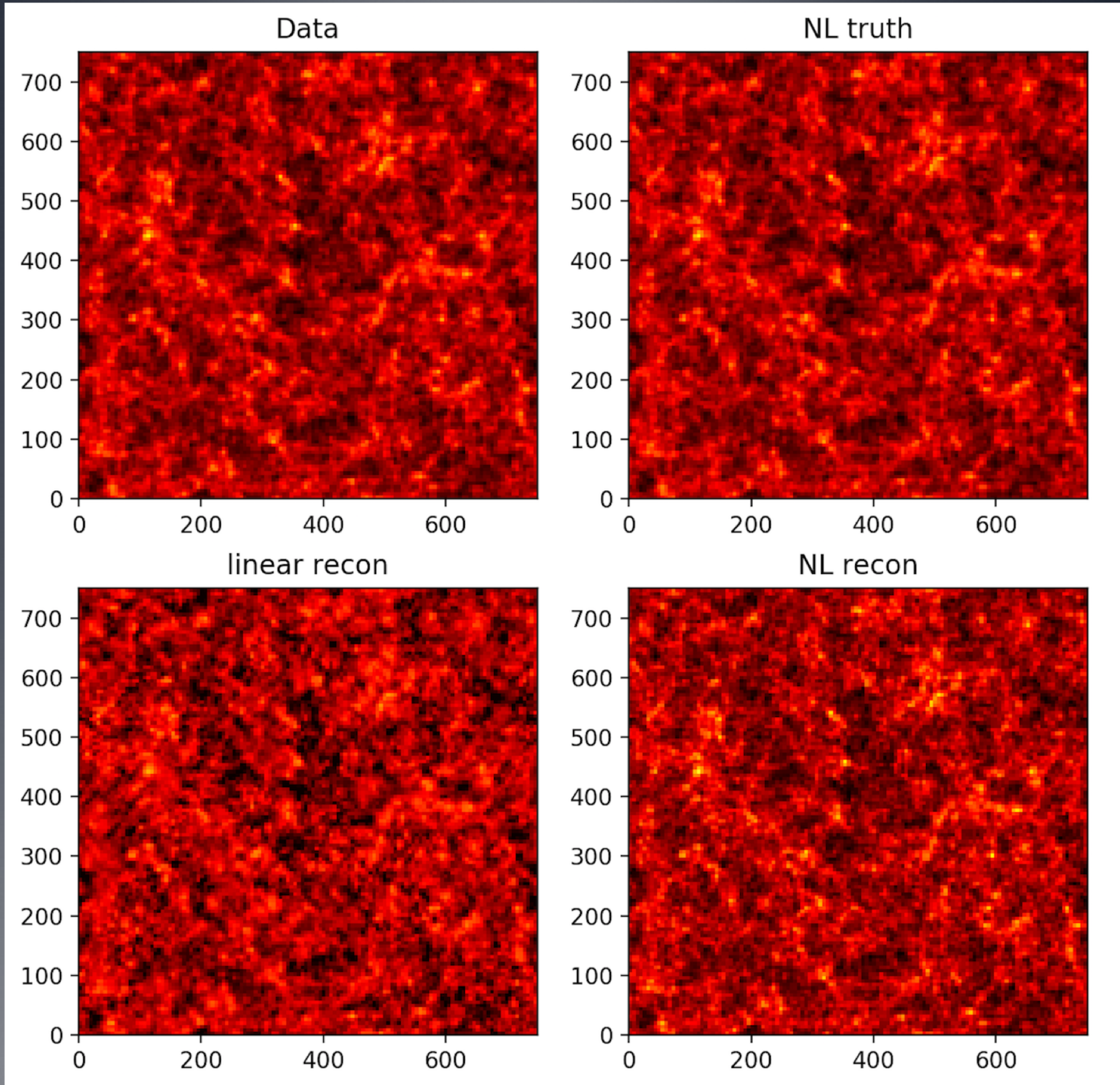
- Low noise, low smoothing
- 750Mpc/h box, 128^3
- reconstructs well all scales



- 2d projections (weak lensing)
- No reconstruction along line of sight, as expected

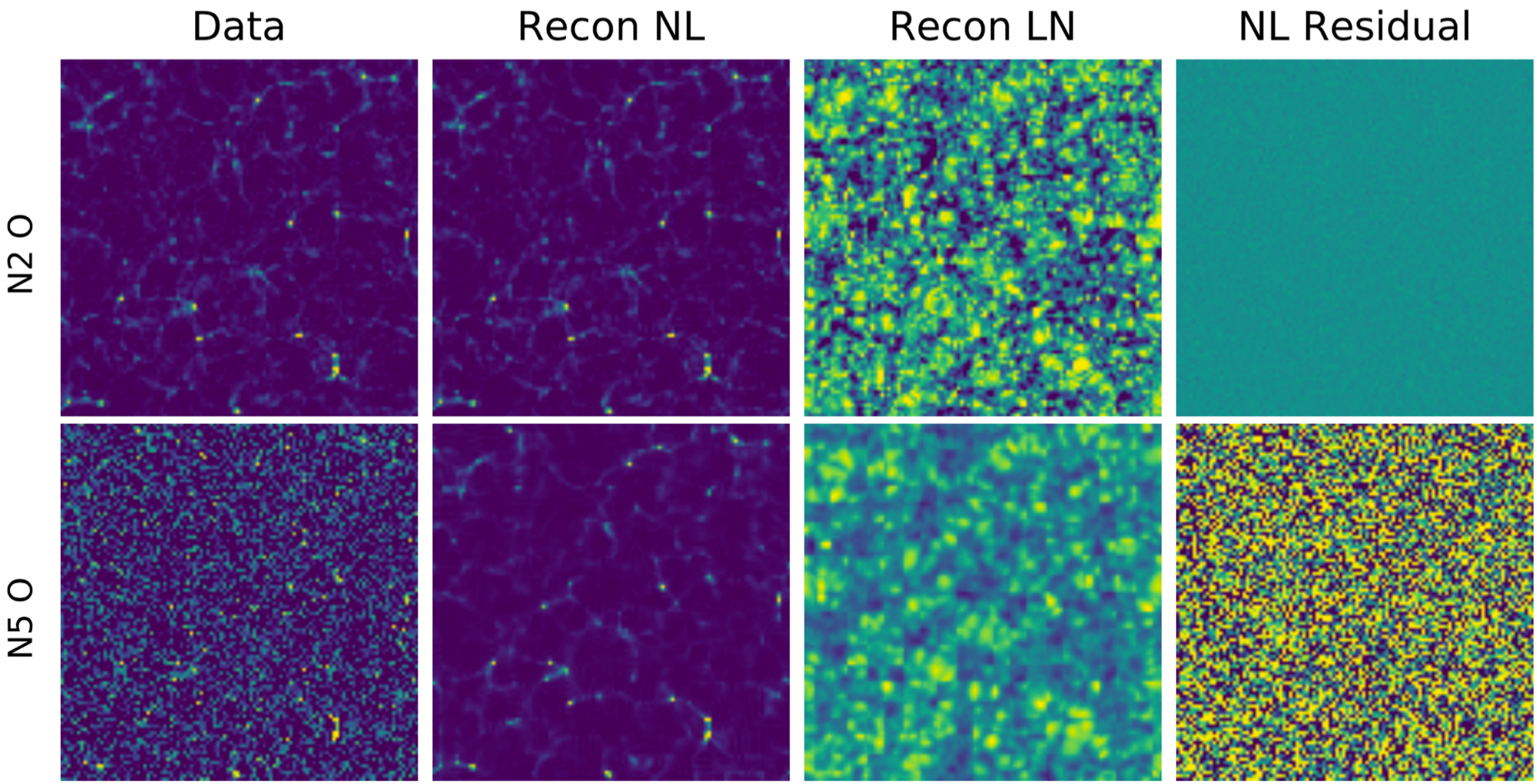


- 2d projections (weak lensing)
- Good reconstruction transverse to line of sight
- More gaussian because of wider projection

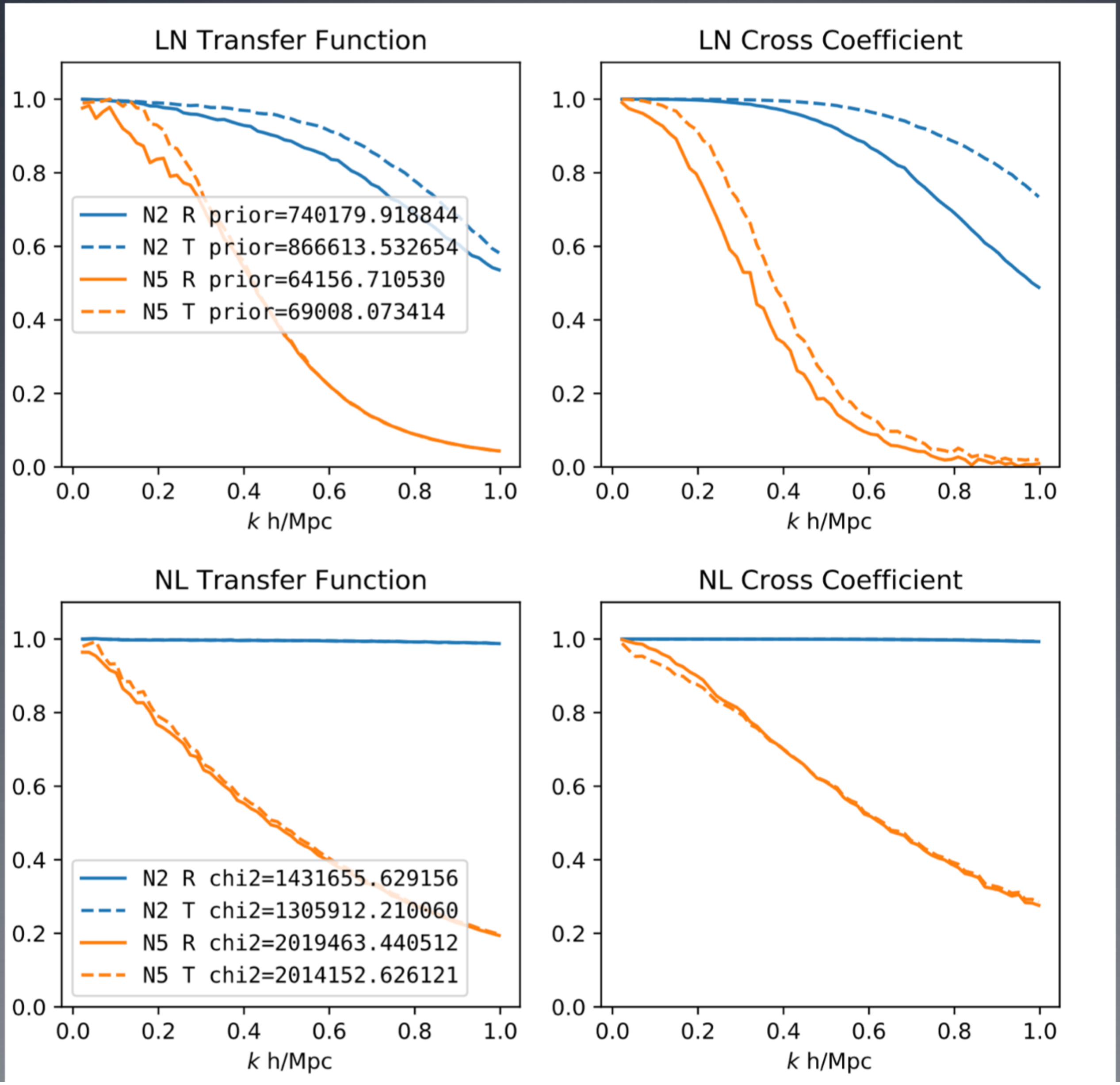


High vs low noise

Dataset name	Noise Power (Mpc/h^3)	Noise seed	Truth seed
N2	1	1234	181170
N3	10	1234	181170
N4	100	1234	181170
N5	1000	1234	181170



Quantifying the results: transfer function and corr. coeff.



From optimal map to optimal power spectrum

- Integrate out the modes around the minimum variance map (multivariate gaussian integrals)

$$
\begin{aligned}
L(\boldsymbol{d}|\boldsymbol{\Theta}) &= \int P(\boldsymbol{s}, \boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})) d^M \boldsymbol{s} \\
&= (2\pi)^{-(M+N)/2} \det(\boldsymbol{S})^{-1/2} \det(\boldsymbol{N})^{-1/2} \exp\left(\frac{1}{2}[\hat{\boldsymbol{s}}^\dagger \boldsymbol{D} \hat{\boldsymbol{s}} - \tilde{\boldsymbol{d}}^\dagger \boldsymbol{N}^{-1} \tilde{\boldsymbol{d}}]\right) \times \\
&\int \exp\{-\frac{1}{2}[\boldsymbol{s} - \hat{\boldsymbol{s}}]^\dagger \boldsymbol{D} [\boldsymbol{s} - \hat{\boldsymbol{s}}]\} d^M \boldsymbol{s} \\
&= (2\pi)^{-N/2} \det(\boldsymbol{S}\boldsymbol{N}\boldsymbol{D})^{-1/2} \exp\left(\frac{1}{2}[\hat{\boldsymbol{s}}^\dagger \boldsymbol{D} \hat{\boldsymbol{s}} - \tilde{\boldsymbol{d}}^\dagger \boldsymbol{N}^{-1} \tilde{\boldsymbol{d}}]\right).
\end{aligned}
$$

- Maximize L(d|Θ) wrt S(Θ) leads to optimal quadratic estimator
- S is inside D and s^

$$
\boldsymbol{D} = \frac{1}{2} \frac{\partial \chi^2}{\partial \boldsymbol{s} \partial \boldsymbol{s}} = \boldsymbol{S}^{-1} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1} \boldsymbol{R}
$$

$$
\ln L(\boldsymbol{\Theta} + \boldsymbol{\delta\Theta}) = \ln L(\boldsymbol{\Theta}) + \sum_l \frac{\partial L(\boldsymbol{\Theta})}{\partial \Theta_l} \delta\Theta_l + \frac{1}{2} \sum_{ll'} \frac{\partial^2 L(\boldsymbol{\Theta})}{\partial \Theta_l \partial \Theta_{l'}} \delta\Theta_l \delta\Theta_{l'}.
$$

Newton's method to maximum likelihood

- S: summary statistics, power spectrum, compressed into bandpowers Θ_l

$$-2\frac{\partial L(\mathbf{\Theta})}{\partial \Theta_l} = \mathbf{\Pi}_l \boldsymbol{S}^{-1}\hat{\boldsymbol{s}}^\dagger \hat{\boldsymbol{s}}\boldsymbol{S}^{-1}\mathbf{\Pi}_l - \boldsymbol{b}_l,$$

$$b_l = \frac{1}{2}\mathrm{tr}\left[\frac{\partial \ln(\boldsymbol{S}\boldsymbol{N}\tilde{\boldsymbol{D}})}{\partial \Theta_l}\right]$$

$$(\boldsymbol{F}\hat{\mathbf{\Theta}})_l = \frac{1}{2}[\mathbf{\Pi}_l \boldsymbol{S}^{-1}\hat{\boldsymbol{s}}^\dagger \hat{\boldsymbol{s}}\boldsymbol{S}^{-1}\mathbf{\Pi}_l - b_l],$$

- Fisher matrix $F_{ll'} = \left\langle \frac{\partial^2 L(\mathbf{\Theta})}{\partial \Theta_l \partial \Theta_{l'}} \right\rangle$
- To solve for QML need the to average squares of minimum variance map s^ within bandpower : D has dropped out
- Also need Fisher matrix (window and covariance matrix) and bias b_l

“Near-optimal” power spectrum estimator

- Can be constructed from minimum variance map s^

$$(\boldsymbol{F}\hat{\boldsymbol{\Theta}})_l = \frac{1}{2}[\boldsymbol{\Pi}_l \boldsymbol{S}^{-1} \hat{\boldsymbol{s}}^\dagger \hat{\boldsymbol{s}} \boldsymbol{S}^{-1} \boldsymbol{\Pi}_l - b_l],$$

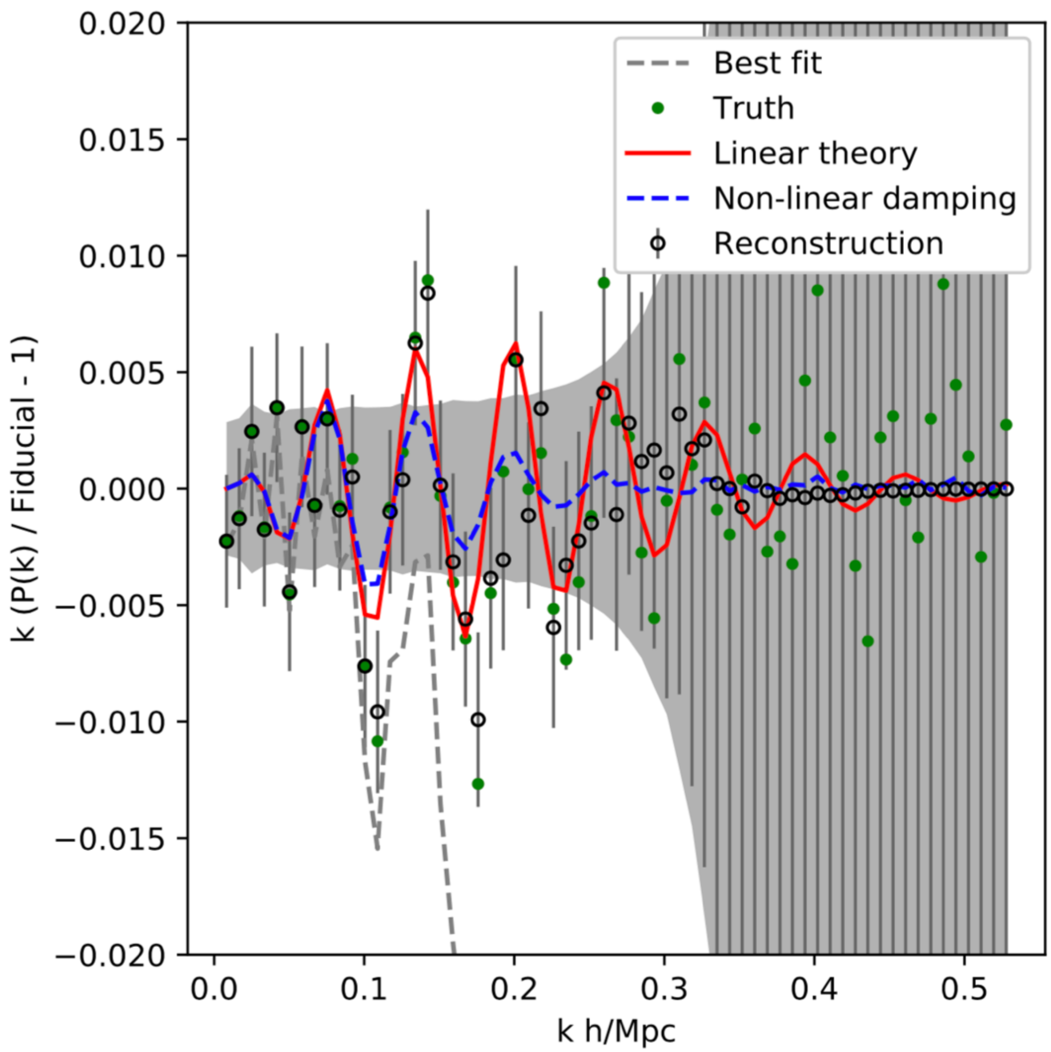
- Fisher matrix gives bandpower mixing and covariance matrix

$$\langle \hat{\boldsymbol{\Theta}} \rangle = \boldsymbol{M}\boldsymbol{F}\boldsymbol{\Theta},$$
$$\langle \hat{\boldsymbol{\Theta}} \hat{\boldsymbol{\Theta}}^\dagger \rangle - \langle \hat{\boldsymbol{\Theta}} \rangle \langle \hat{\boldsymbol{\Theta}}^\dagger \rangle = \boldsymbol{M}\boldsymbol{F}\boldsymbol{M}$$

- Noise bias b_l: squaring noisy modes creates noise bias
- Fisher matrix constructed from bandpower responses

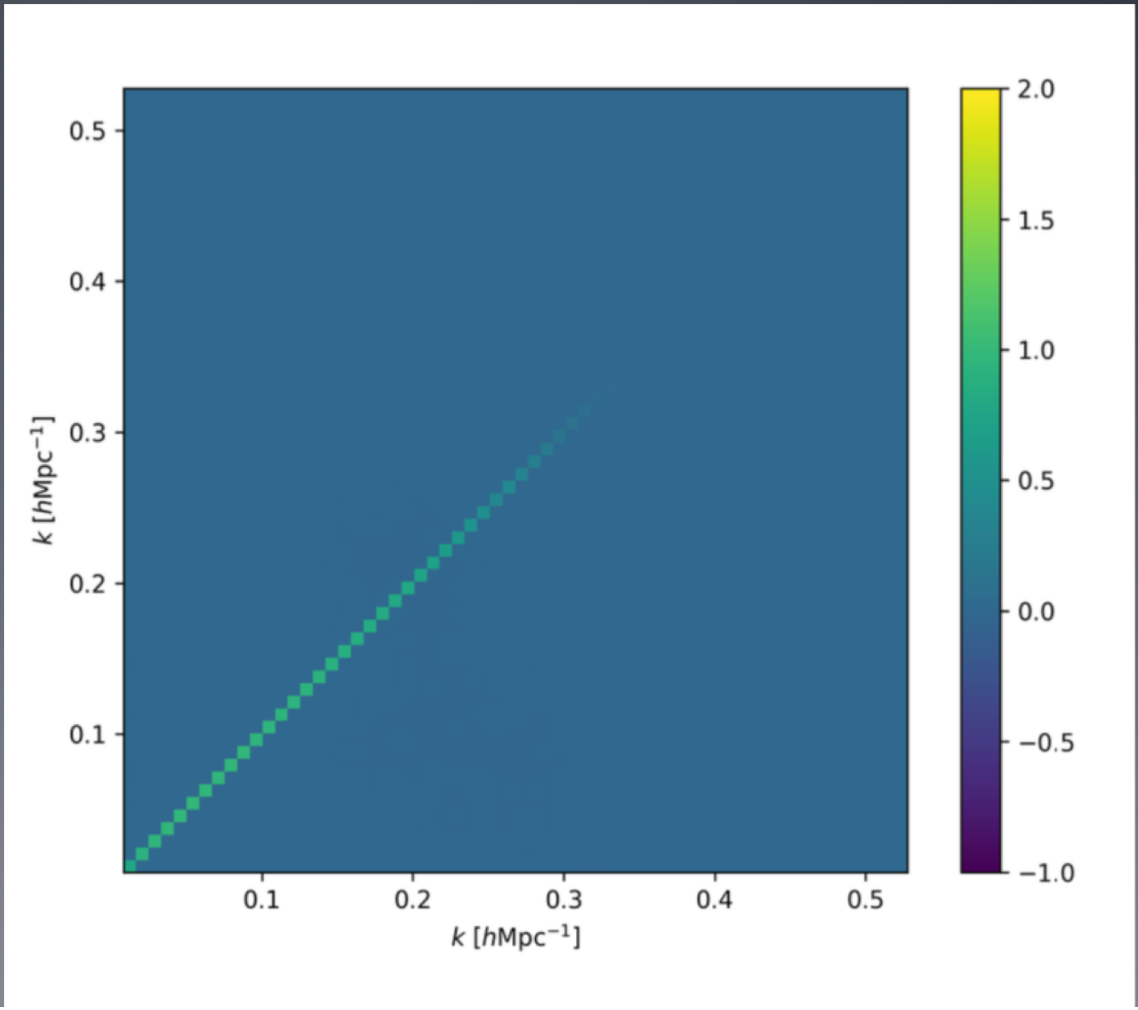
Near optimal power spectrum reconstruction: example

- with Grigor Aslanyan, Yu Feng and Chirag Modi)
- Unbiased P(k) reconstruction, with linear wiggles at high k
- This is the best possible reconstruction of baryonic oscillations or other features
- Noise prevents high k reconstruction



Fisher matrix: inverse covariance and window matrix

- For periodic box it is diagonal even at high k
- At high k noise cause the solution to go to zero
- In the absence of noise we would be able to reconstruct much better



This approach allows for a unification of all LSS methods/observables

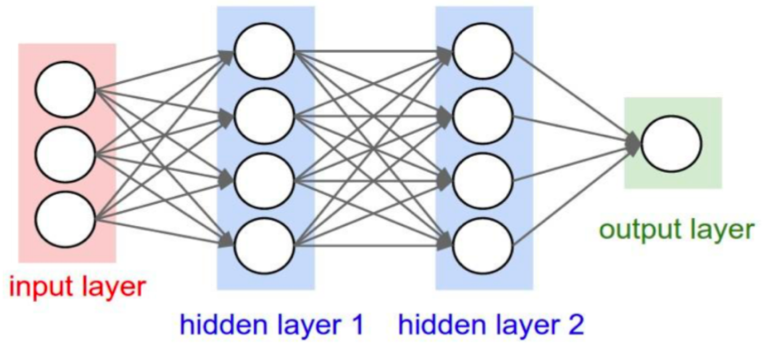
- Easy to add more data d: need a model F(s) and its error (noise)
- Can be directly applied to weak lensing
- Any data can be added: galaxy magnitudes/colors, redshifts, SZ, X-rays, 2d or 1d (spectra)
- Data with less scatter relative to dark matter are better: what is the lowest scatter observable (stellar mass? Luminosity and colors?)
- Automatically includes all LSS methods: e.g. cluster abundance, voids…
- Key is probabilistic description of data and forward modeling
- But how do we handle point objects (e.g. galaxies)?

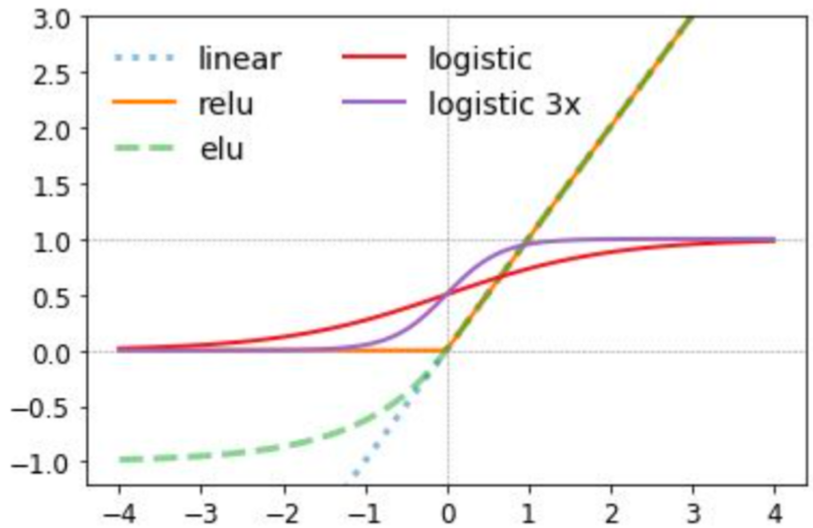
How to make galaxies differentiable?

- Galaxies appear to be discrete points, and this is not (easily) differentiable
- Halo finders are complicated and not (easily) differentiable
- Instead we want to use local properties of dark matter density and velocity to define galaxy observables. These may even be better tracers
- Neural networks are differentiable and can be trained on discrete objects: we feed halo mass or position and the DM information
- First results (Modi etal, in prep)

Neural Networks

- Deep/Convolution NNs
- Simpler ones - Fully Connected Networks
- Classification (NNp) for position
  - Position mask
  - 1 if a neighboring cell has halo
  - relu & logistic activation functions
- Regression (NNm) for mass
  - Halo mass (CIC convolved) at given point
  - elu & linear activation functions
- Model = NNp * NNm
- Y = f (W·X + b)





Identifying Features

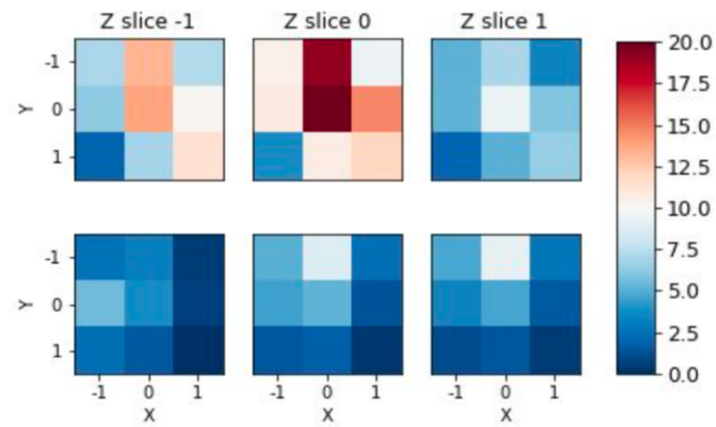
Anything you think affects halo formation!

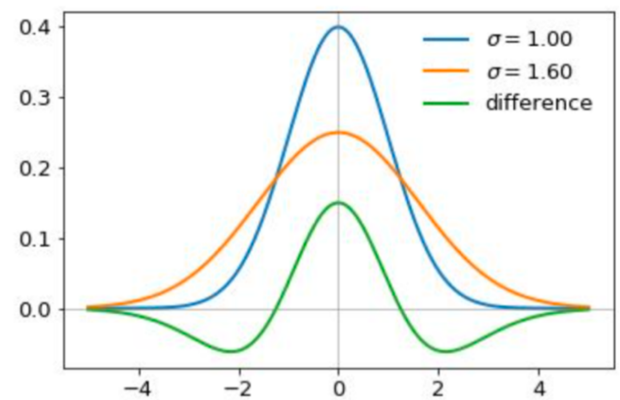
- Density field at different smoothings: $\square_{R0}$, $\square_{R1}$...
- Difference of Gaussian smoothings: $\square_{R1}$ - $\square_{R2}$
- Tidal field, velocity field, gradients

NNp feature array - non local features

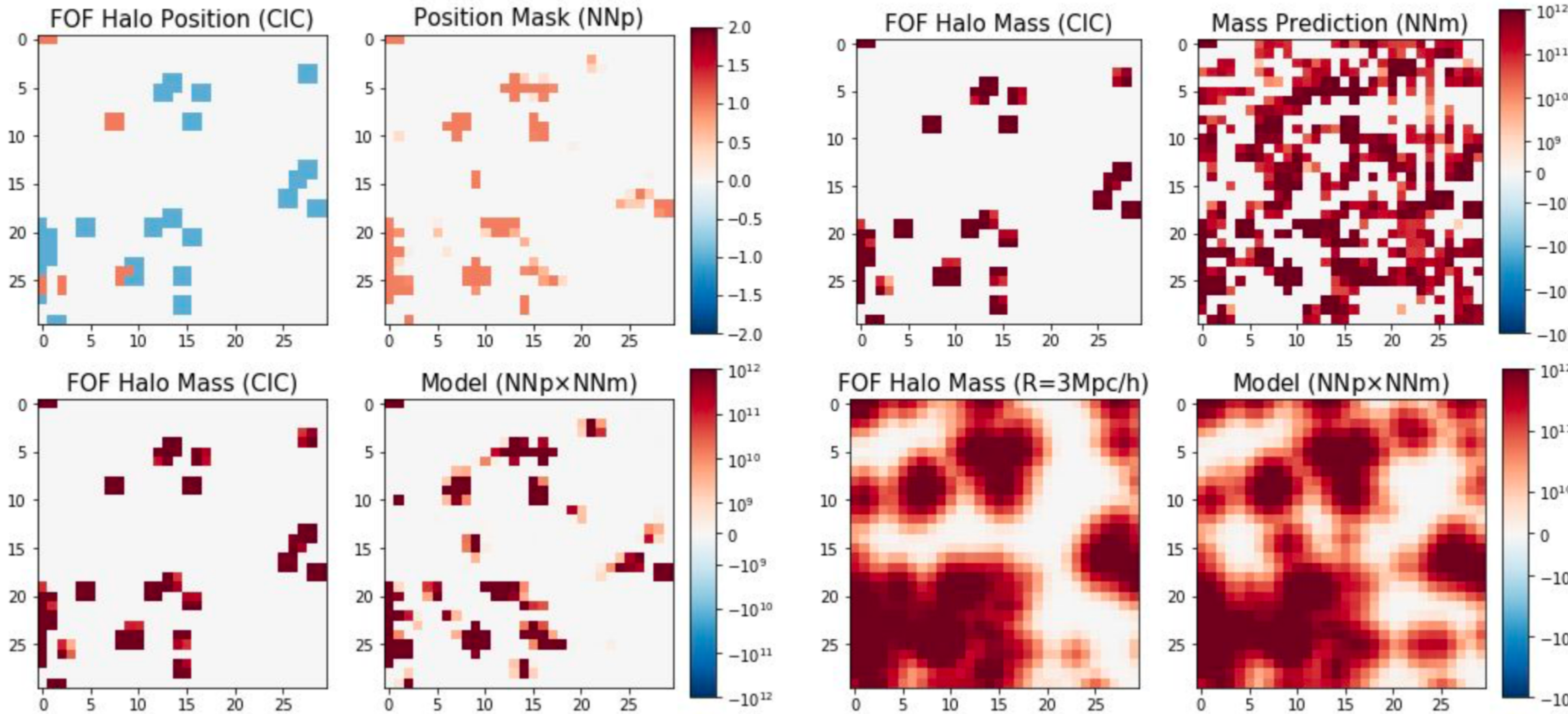
- value at (3*3*3 =) 27 neighbors of every point
- locate maxima with 3 points

These are similar to features identified by CNNs.

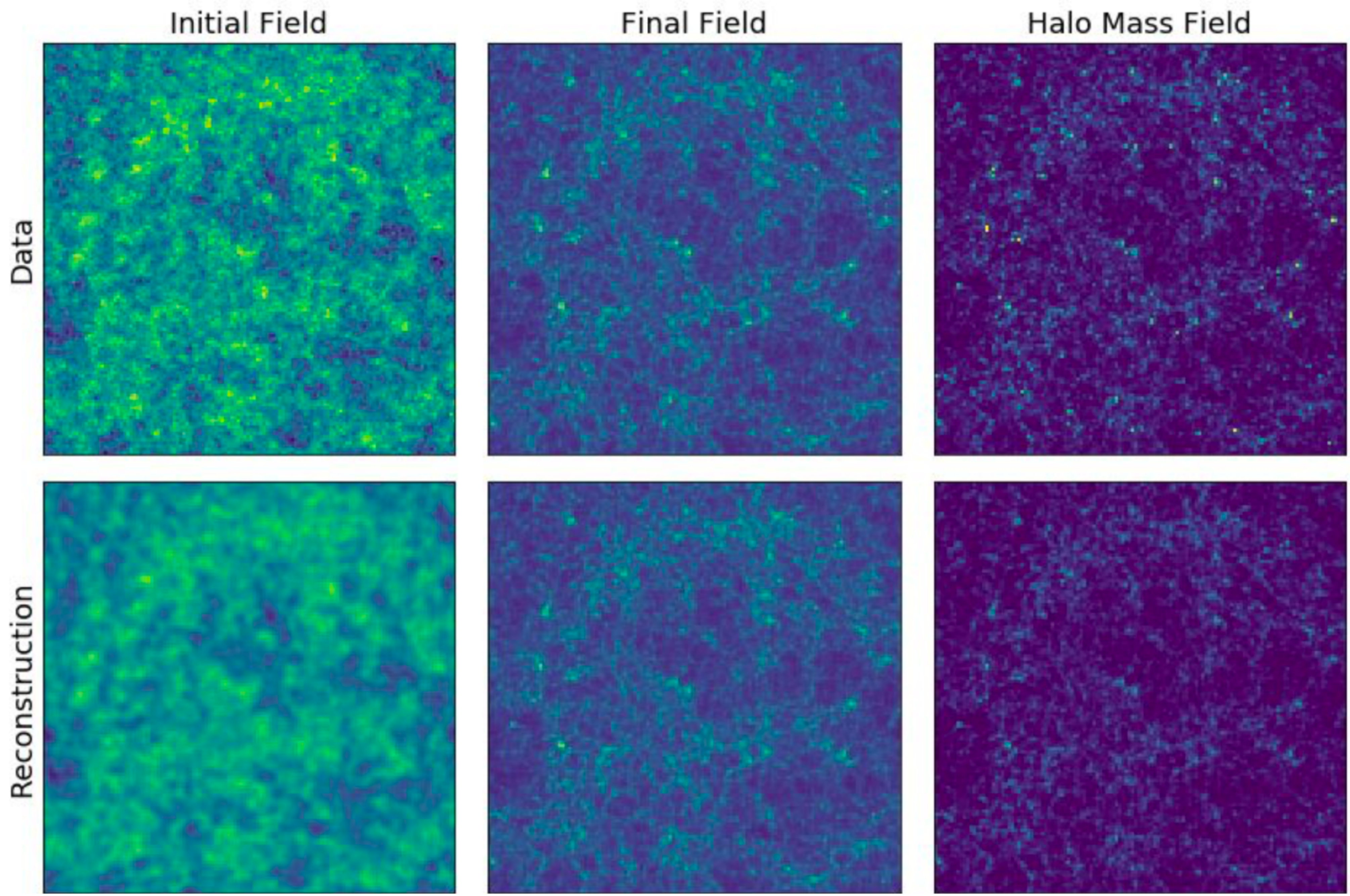




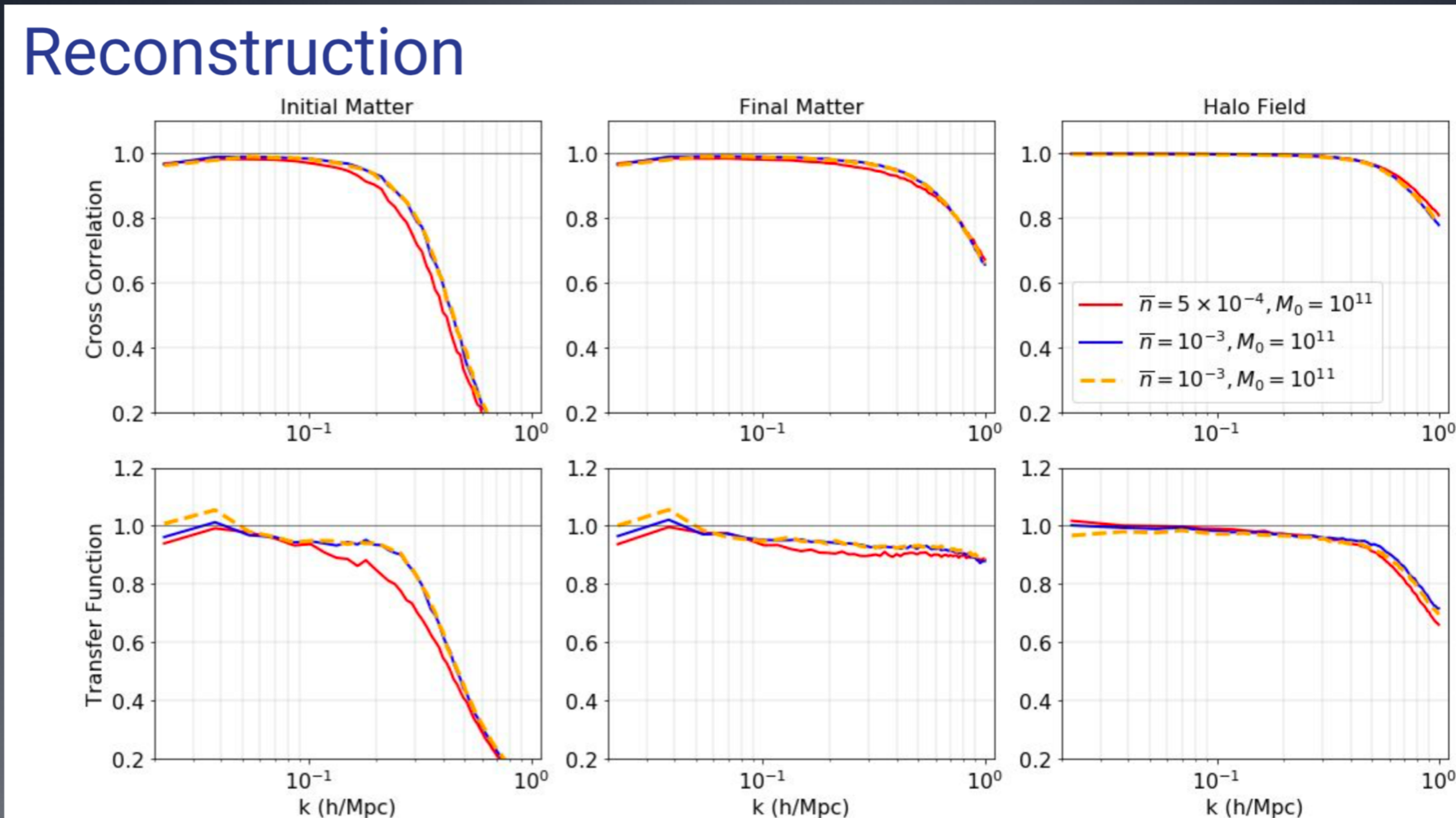
What are the networks doing?



Reconstruction from halos



Power spectrum from halos



These results are encouraging, but more work is needed (RSD…)

Can we push beyond shell crossing?

with Y. Feng and M. Zaldarriaga

- After shell crossing the mapping becomes non-injective: multiple realizations of initial field give identical final density (simple case study: 1d Zeldovich)
- The method selects the solution with the lowest power, which may not be the correct solution: fundamental limit?
- Adding velocity information breaks the degeneracies: phase space dynamics is fully reversible. However, coarse graining due to finite sampling or noise destroys information.
- The pre shell-crossing modes are unaffected
- Numerically due to the non-convex nature of posterior it is very difficult to converge to the global minimum

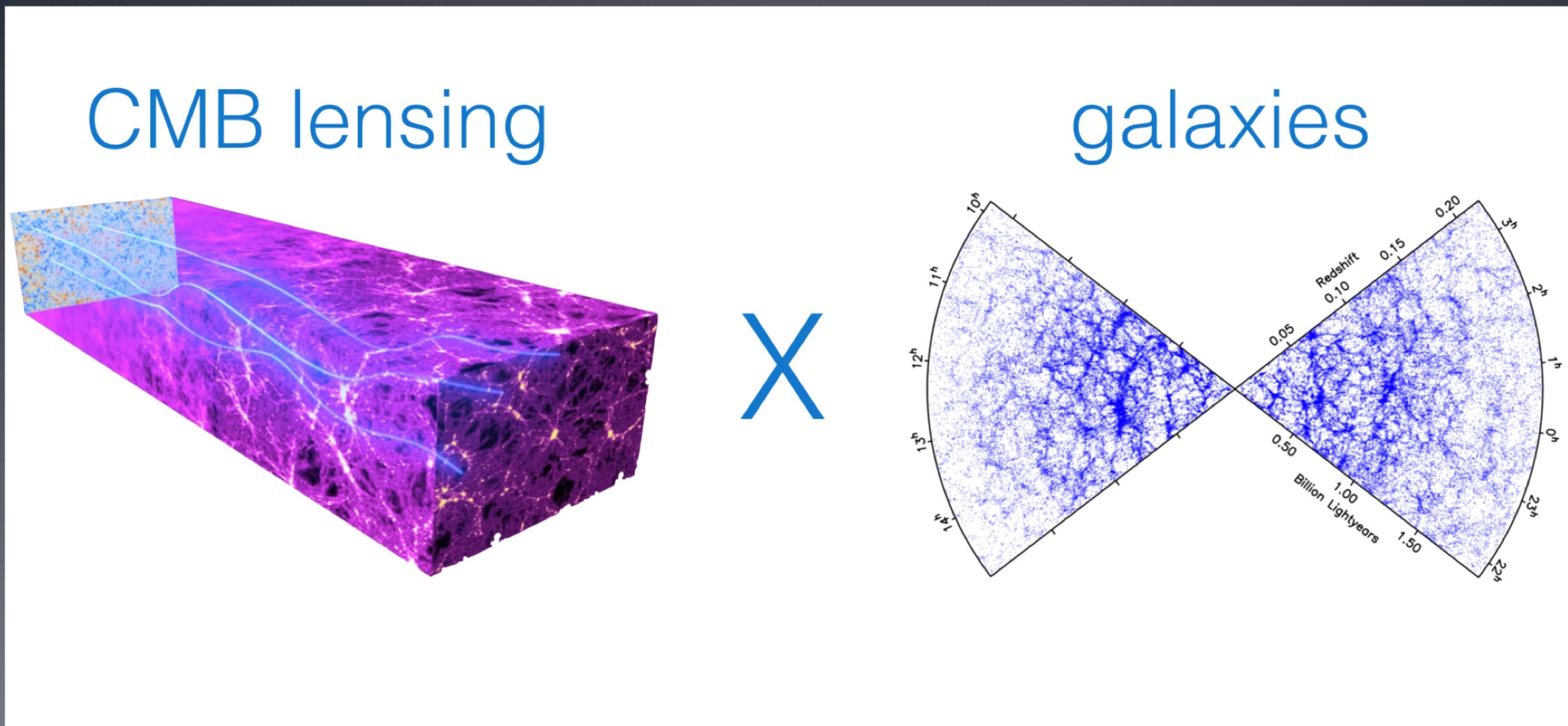
Shell crossings: 1-d Zeldovich

- Reconstructed solution is smoother than true solution even in absence of noise

- What is optimal analysis in this case?

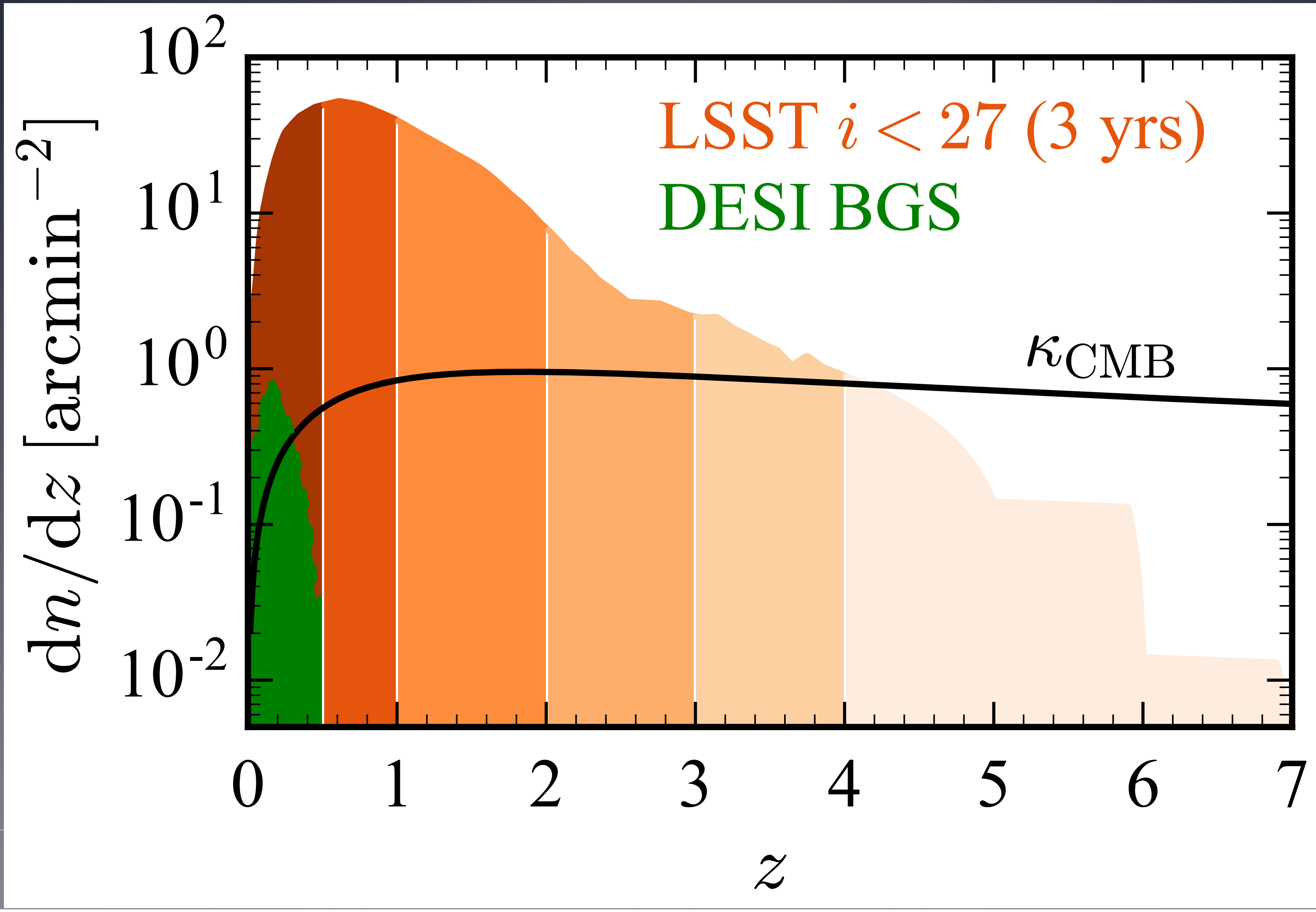
CMB lensing vs galaxies

Schmittfull & US 2017

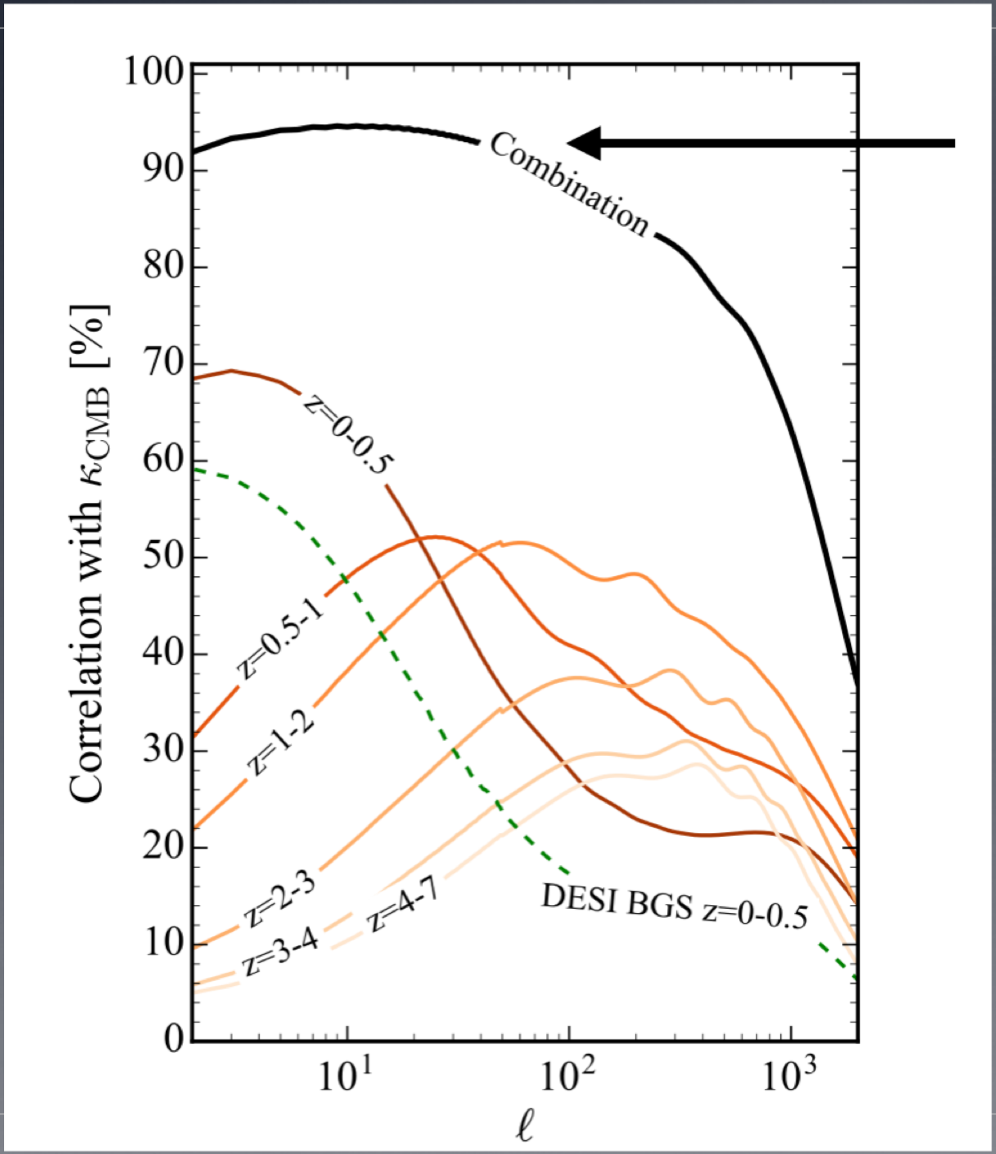


Galaxies are 3d (spectro) or 2.5d (photoz), CMB lensing is 2d
Can we get the best of both? Yes, with cross-correlation

LSST traces CMB kappa to z=4



Cross-correlation coefficient reaches 0.95



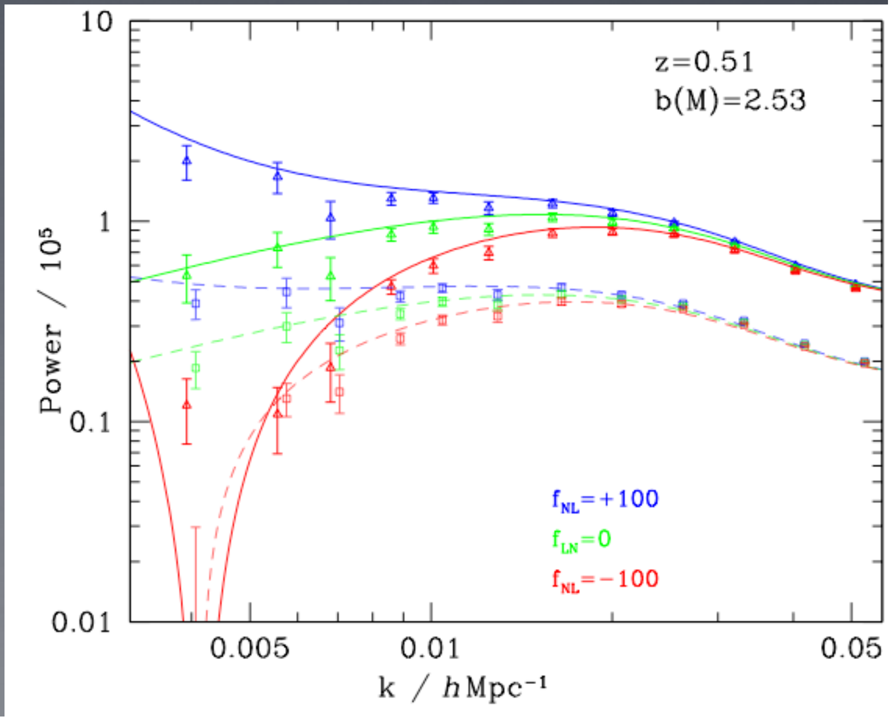
Combined DESI+LSST sample more than 90% correlated with CMB-S4 lensing at L<100

[includes shot noise and lensing noise]

Primordial non-gaussianity

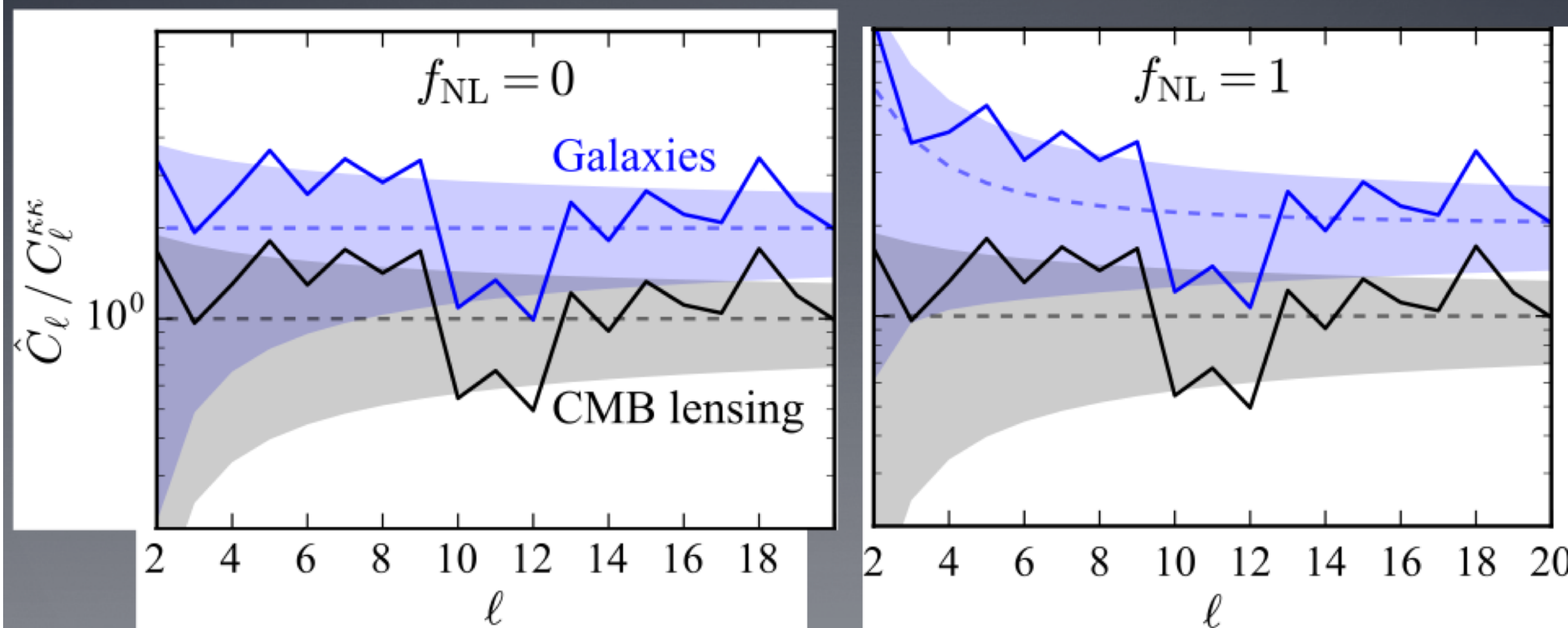
- Local model $$\Phi(x) = \Phi_G(x) + f_{NL}\Phi_G^2(x)$$
- Simple single field slow roll inflation predicts $f_{nl}=0$
- Inflationary models beyond single field slow roll can give $f_{nl}>1$?
- Alternatives to inflation generically give $f_{nl}>>1$?
- Other models give different angular dependence of bispectrum
- Scale dependent bias (Dalal etal 2008)

$$b_{f_{nl}} \propto f_{nl}(b-1)k^{-2}T(k)$$



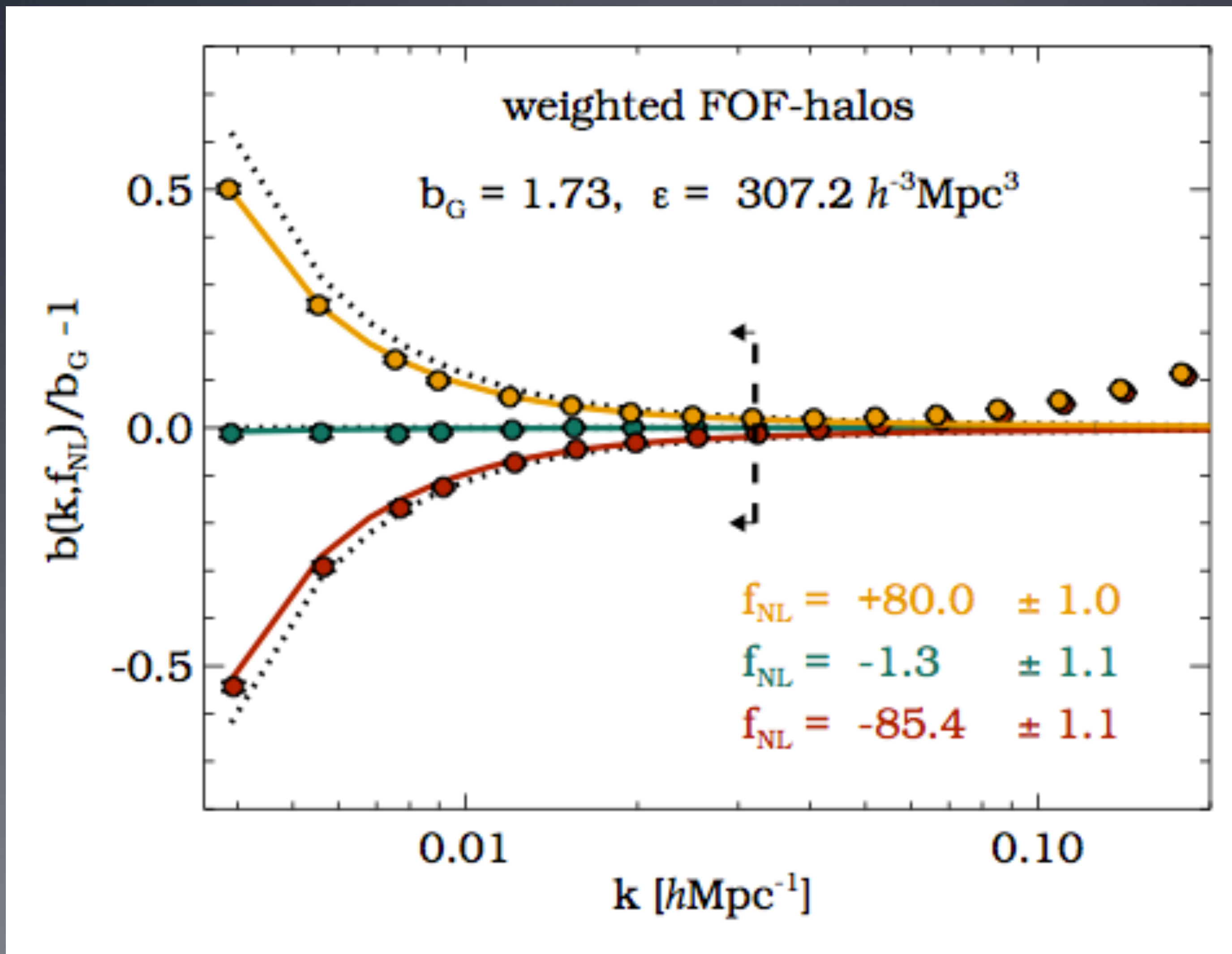
Sampling variance cancellation

- The response scales as b-1: so if we compare biased galaxies with b>1 to unbiased galaxies or dark matter (b=1) we cancel sampling variance (US 2009)



f_{nl} with several tracers: simulations

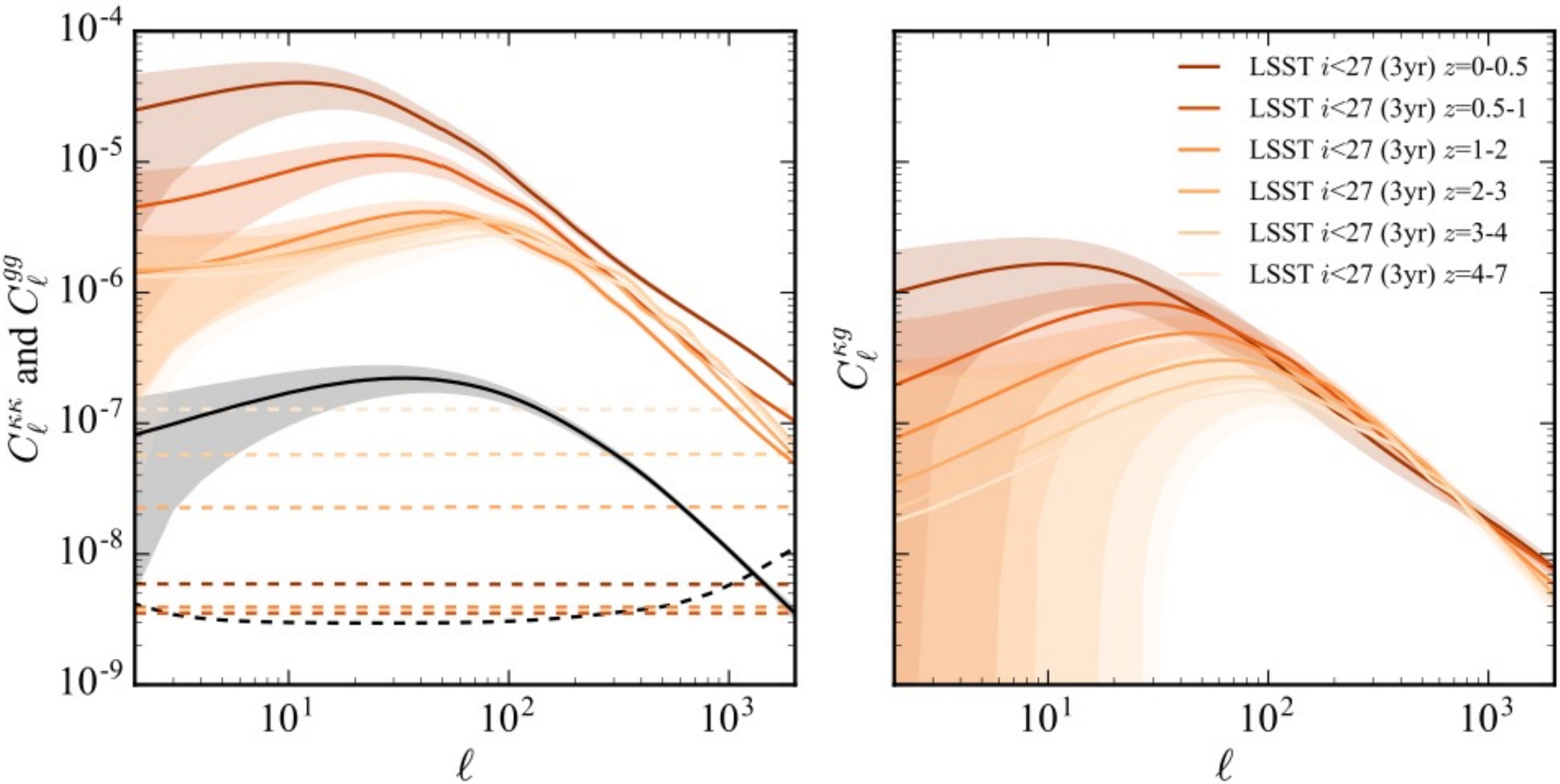
Hamaus, US, Desjacques 2011



Responses need to be calibrated by simulations

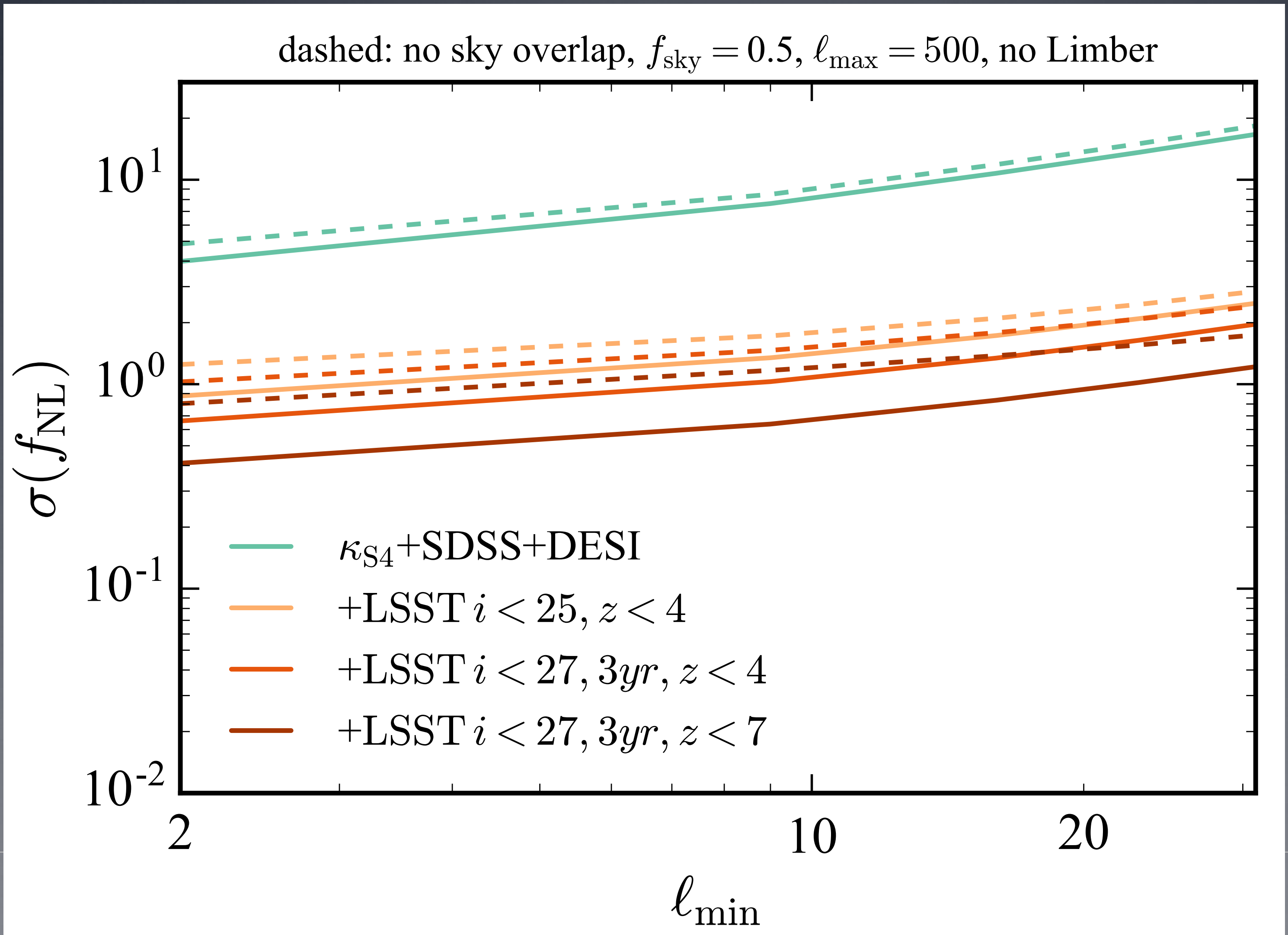
Joint κκ, κg, gg analysis:

sampling variance dominated at low l



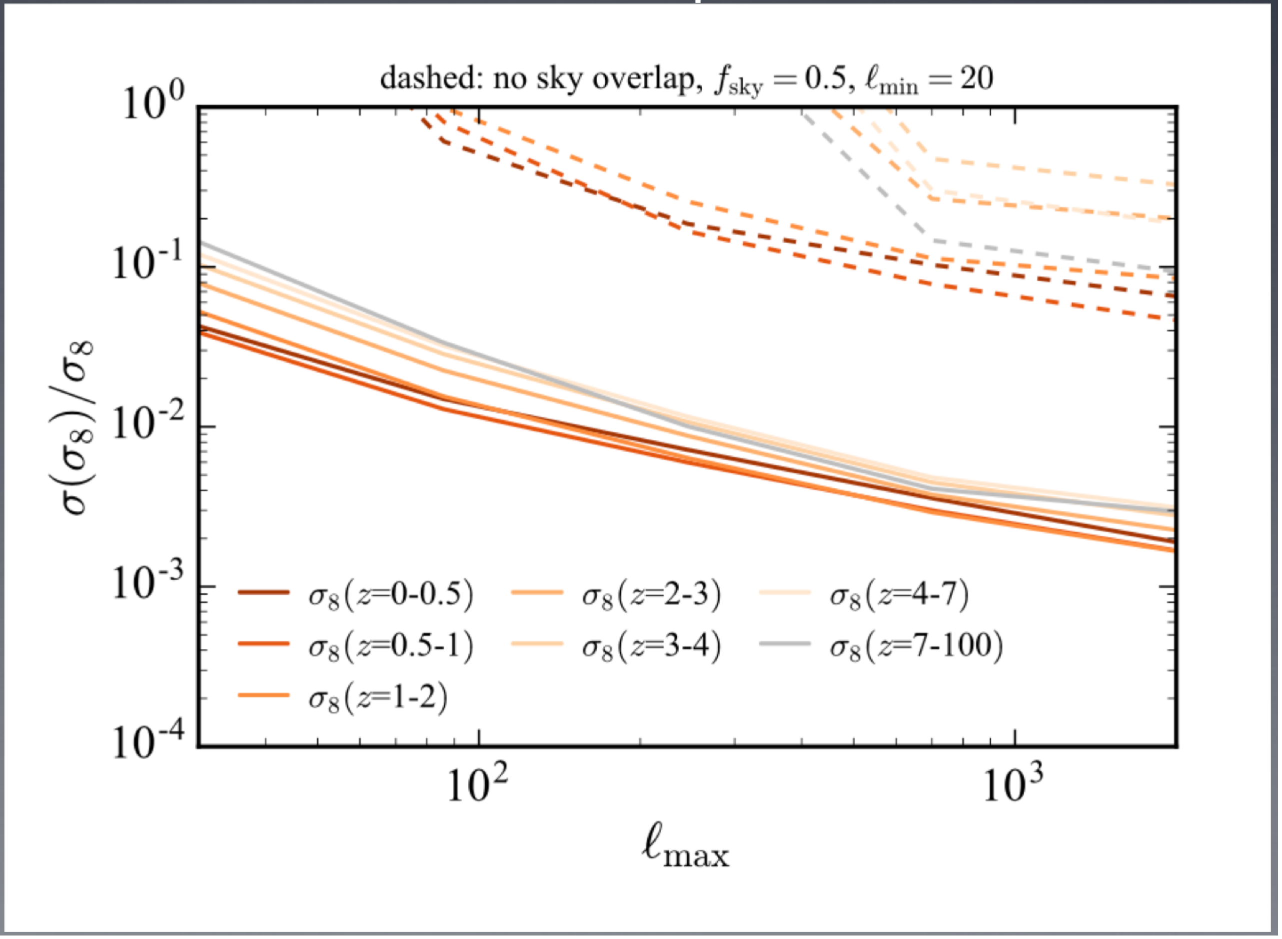
Expected f_{nl} constraints:

factor of 2 improvement from sampling variance
predicted error below 1



What else can we do with it? Tracing amplitude with redshift

- Very tight errors, current analysis assumes linear bias
- Cross-correlations essential to have redshift dependence



Summary

- LSS can probe fundamental physics in many ways: amplitude versus redshift, shape of P(k)
- LSS is not only probing large scales, but scales as small as 100AU (10^{10}km) using SN1A lensing
- LSS is nonlinear, and optimal statistical analysis has many challenges that next generation of analyses should address
- The problem of optimal analysis of LSS remains open, but we think we have a clear, but expensive path forward
- Combining weak lensing with galaxy clustering gives us best of both worlds (DM modes in 3d)
- If we succeed we will have not millions, but billions of linear modes to play with