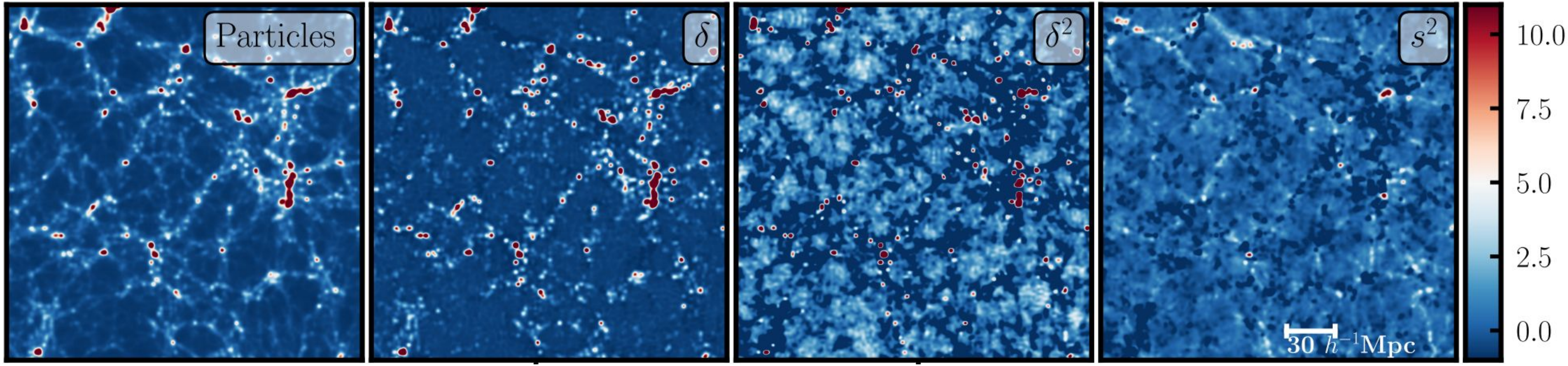
Structure formation at the juncture of simulations and perturbation theory

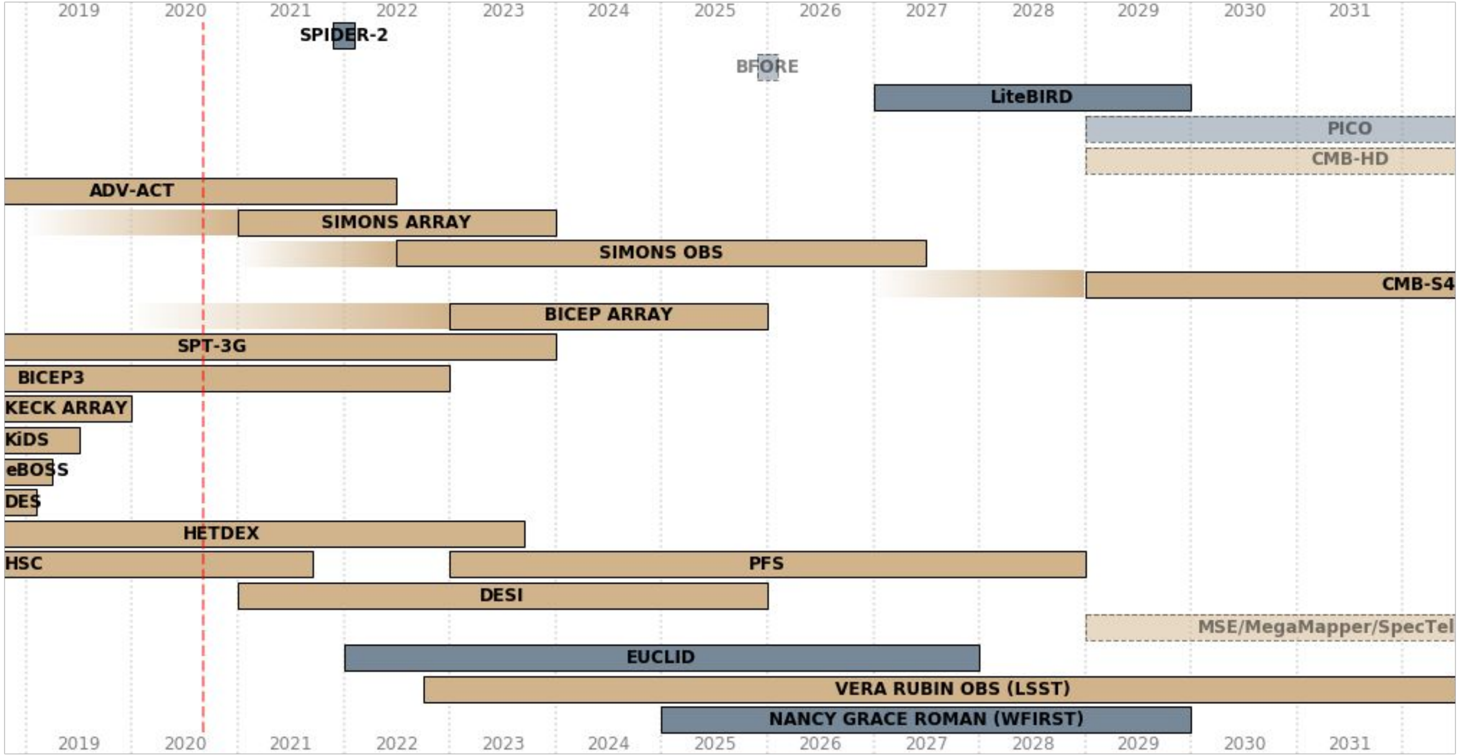


Nickolas Kokron
kokron@stanford.edu

BCCP Seminar 09/27/22

Work done with Tom Abel, Arka Banerjee, Stephen Chen, Joe DeRose, Mark Maus, Risa Wechsler, Martin White

The next decade of surveys will map the sky in an unprecedented way



LSS probes measured:
- -Galaxy positions
- -Galaxy shapes
- -Spec-zs
- -CMB lensing
- -kSZ
- -tSZ
- -More!

Not shown:
- -21cm
- -LIM
- -X-ray
- -More!

Image credit: David Kirkby

Anatomy of cosmological data analysis

(and an asynchronous outline of this talk)

$$P(\theta|\mathbf{d}) \propto^{*} \mathcal{L}\left(\mathbf{d}|\mu(\theta)\right)\pi(\theta)$$

Raw data compressed to summary statistics

-What are the optimal summary statistics beyond two-point functions?

Models must be built for each summary statistic

-How do we efficiently probe smaller scales?

-Are our models accurate enough for next generation surveys?

Priors on model parameters

-If better models require many nuisance parameters, can we gain prior understanding on them?

*Apologies to those thinking about evidences! They're cool too!

$$P(\theta|\mathbf{d}) \propto \mathcal{L}\left(\mathbf{d}|\mu(\theta)\right)\pi(\theta)$$

How can we combine all of the modelling tools at our disposal jointly to tackle these questions?

$$P(\theta|\mathbf{d}) \propto \mathcal{L}\left(\mathbf{d}|\mu(\theta)\right)\pi(\theta)$$

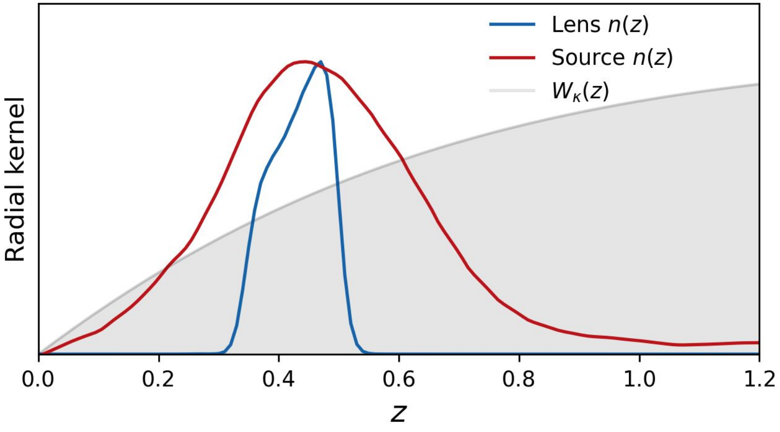
Models must be built for each summary statistic

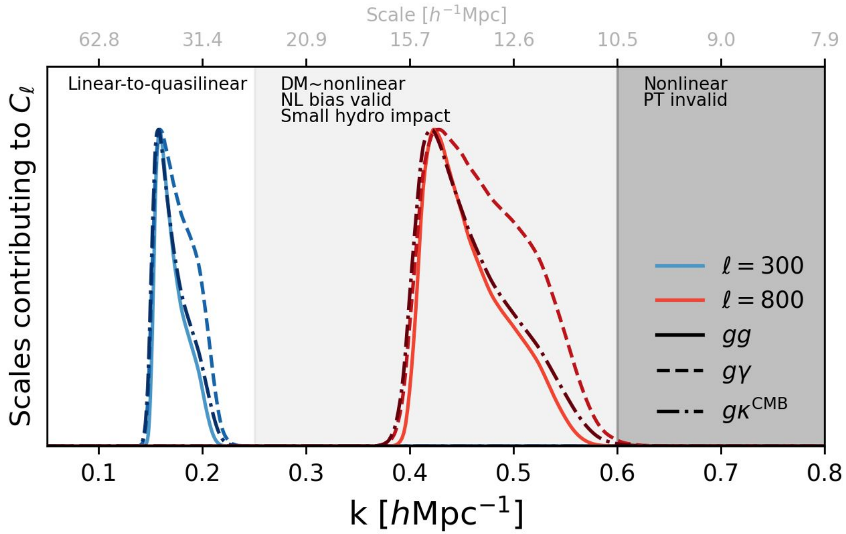
-**How do we efficiently probe smaller scales?**

-Are our models accurate enough for next generation surveys?

What are the scales we'll need to model better?

$$C_\ell^{\alpha\beta} = \int d\chi \frac{W^\alpha(\chi) W^\beta(\chi)}{\chi^2} P_{\alpha\beta}\left(k = \frac{\ell + 1/2}{\chi}, z(\chi)\right)$$





Multiple approaches to modelling structure formation

Analytic/perturbation theory

$$\delta(\mathbf{x},t) \approx \delta^{(1)} + \delta^{(2)} + \delta^{(3)} + \cdots + \delta^{(\mathrm{ct})}$$

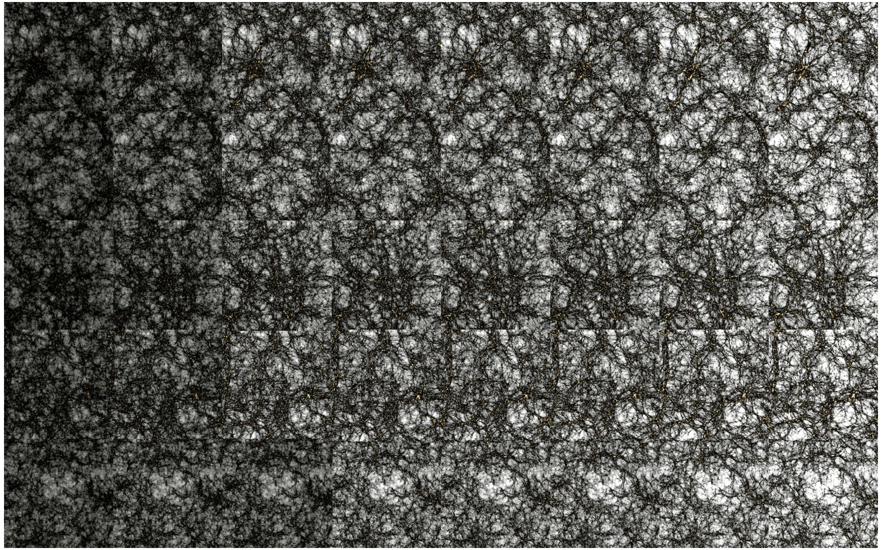
or

$$\mathbf{\Psi}(\mathbf{q},t) \approx \mathbf{\Psi}^{(1)} + \mathbf{\Psi}^{(2)} + \mathbf{\Psi}^{(3)} + \cdots + \mathbf{\Psi}^{(\mathrm{ct})}$$

where

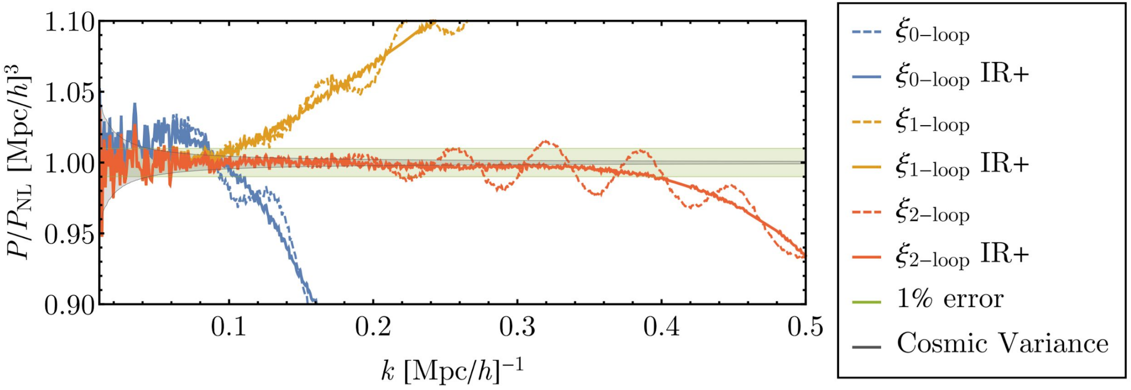
$$\mathcal{O}^{(n)} \sim \int_{\boldsymbol{q}_1\cdots\boldsymbol{q}_n} F^{(n)}(\boldsymbol{q}_1,\cdots,\boldsymbol{q}_n)\mathcal{O}^{(1)}_{\boldsymbol{q}_1}\cdots\mathcal{O}^{(1)}_{\boldsymbol{q}_n}$$

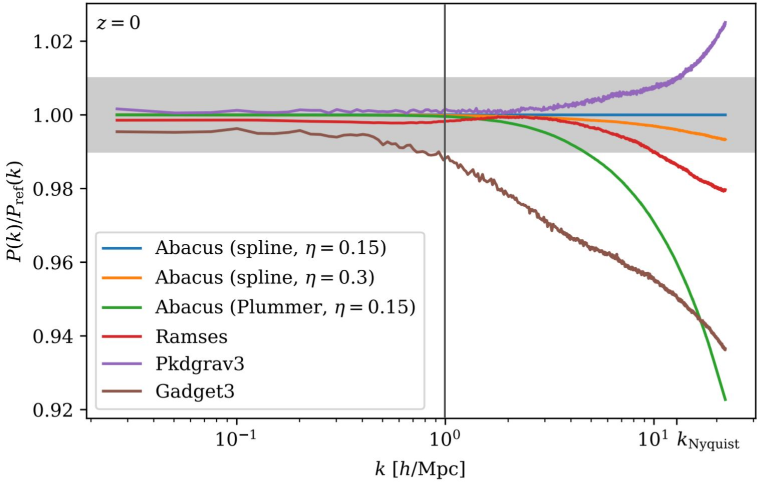
Simulations



Perturbation theory is great, but will inevitably fail!

Simulations, on the other hand, agree well up to scales of k~1 h/Mpc, but face challenges at large scales

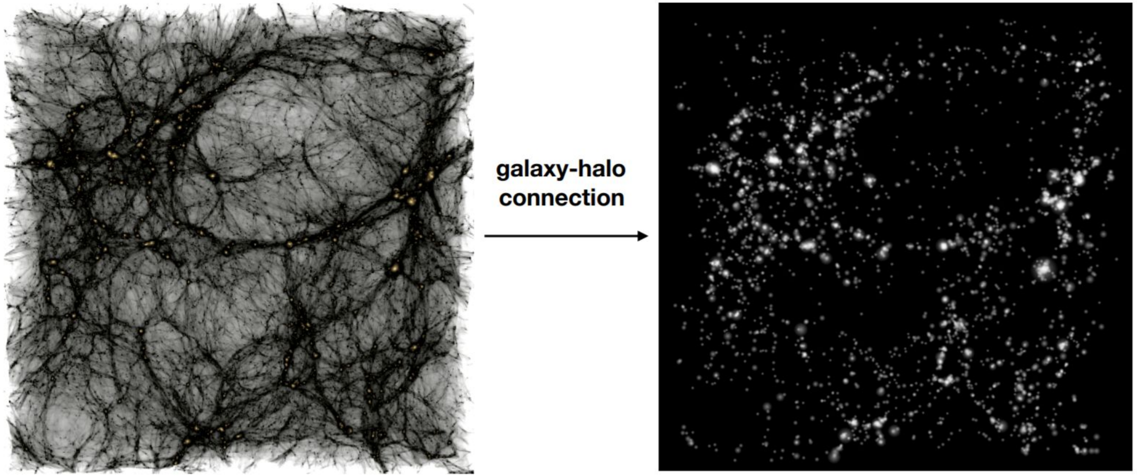




Senatore & Trevisan 2018

Garrison et al 2019

Galaxies live in dark matter, but the relationship is not 1:1



Approaches to modeling the galaxy-halo connection

← physical models　　　　empirical models →

Hydrodynamical Simulations	**Semi-analytic Models**	**Empirical Forward Modeling**	**Subhalo Abundance Modeling**	**Halo Occupation Models**
Simulate halos & gas; Star formation & feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos & subhalos) plus assumptions about galaxy—(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

[Wechsler & Tinker 2018](#)

Mathematically relating the two distributions

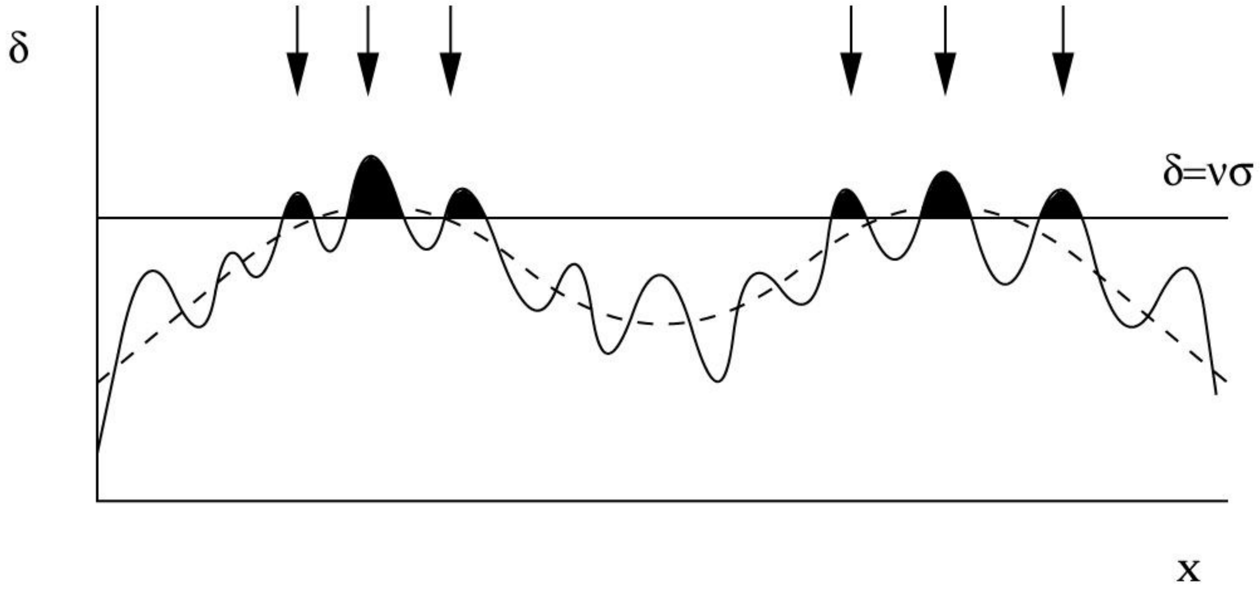
From the previous slide we can tell the density of the dark matter and of galaxies seems to be related. This relation is constrained by symmetries of galaxy formation – Galilean invariance.

$$\delta_g(\boldsymbol{x}, t) = F[\partial_i \partial_j \Phi(\boldsymbol{x}, t)]$$

Expanding F in terms of all contributions allowed by symmetries lead to an "effective theory" of biasing.

Bias, pictorially

High density peaks, above a threshold, will collapse into halos



(Figure courtesy of J. Peacock.)

Biased tracers in Lagrangian Perturbation Theory

The relationship between tracer density and matter density is encoded in the initial conditions of structure formation.

$$1 + \delta_h(\boldsymbol{q}) = F[\partial_i \partial_j \Phi(\boldsymbol{q})]$$

To second order we get

$$\begin{aligned} 1 + \delta_h(\boldsymbol{q}) &= F[\partial_i \partial_j \Phi(\boldsymbol{q})] \\ &\approx 1 + b_1 \delta(\boldsymbol{q}) + b_2 \left(\delta^2(\boldsymbol{q}) - \sigma^2\right) + \\ &\quad b_{s^2}\left(s^2(\boldsymbol{q}) - \frac{2}{3}\sigma^2\right) + \boxed{b_{\nabla^2} \nabla^2 \delta(\boldsymbol{q})} + \mathcal{O}(\delta^3) + \\ &\quad \epsilon(\boldsymbol{q}) + \cdots \end{aligned}$$

Finite-Size (Effective) Correction

“Standard” bias contributions

Stochastic contribution

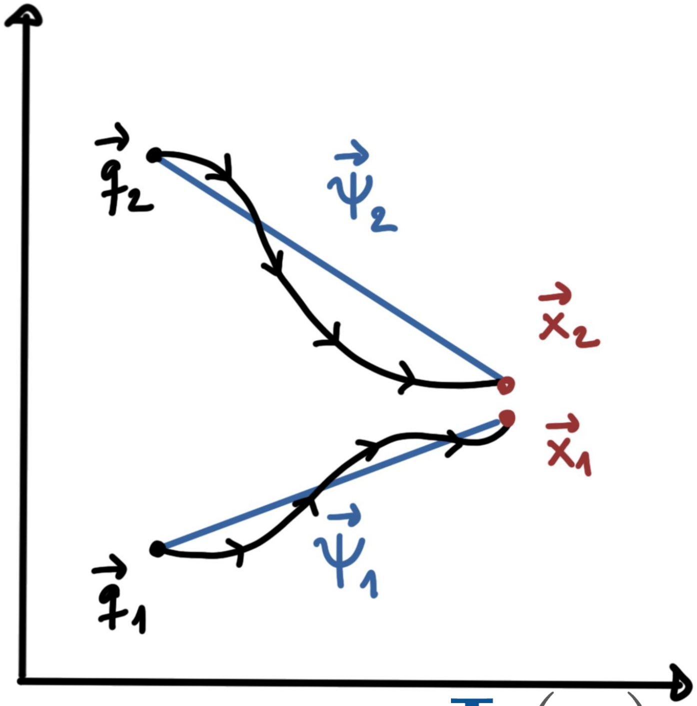
Simulations/analytic approaches describe the *same* physics

Perturbation theory for $\boldsymbol{\Psi}(\mathbf{q}, z)$ and simulations solve for the same quantities.

Proposal (Modi, Chen, White 2020):

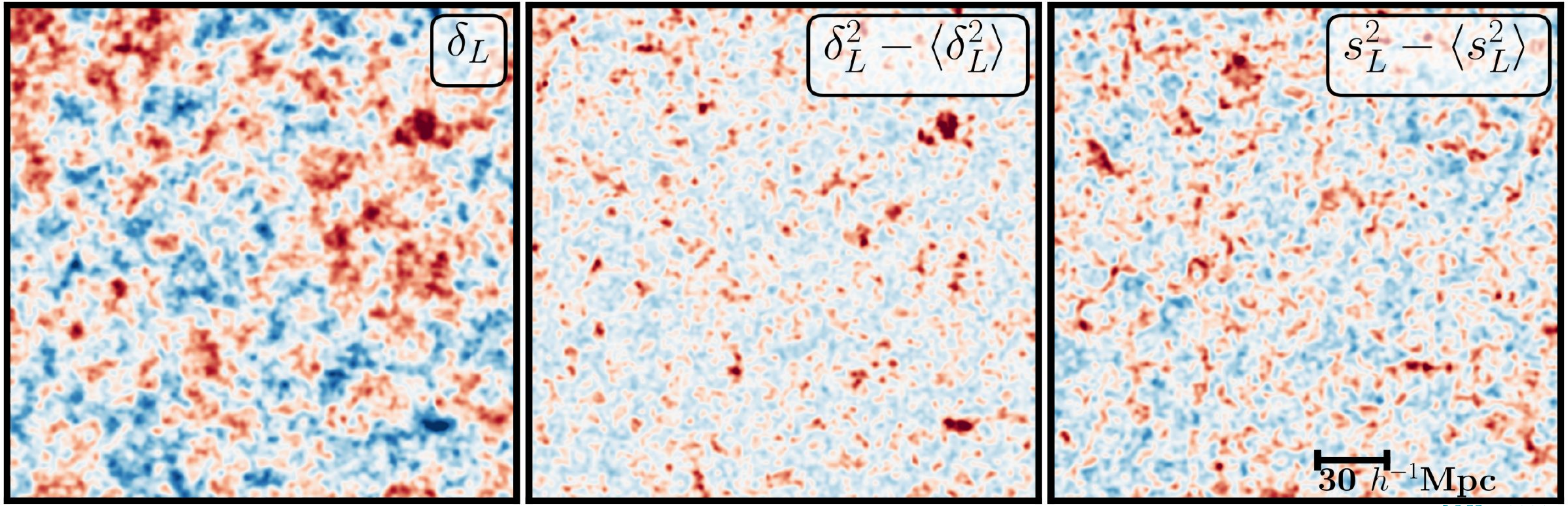
Let perturbation theory inform $F(\mathbf{q})$ as usual, and use $\boldsymbol{\Psi}(\mathbf{q}, z)$ from simulations?*

These models have been termed **hybrid effective field theories** (HEFT).



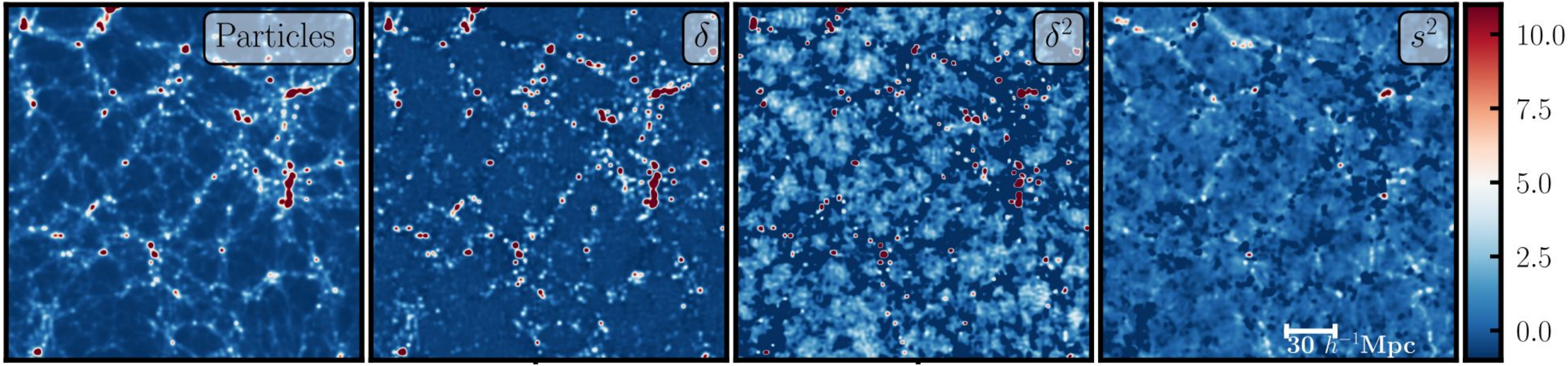
$$\boldsymbol{x} = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q})$$

Hybrid EFT: the Lagrangian fields



N.K.+ 2021a

Hybrid EFT: the late-time fields



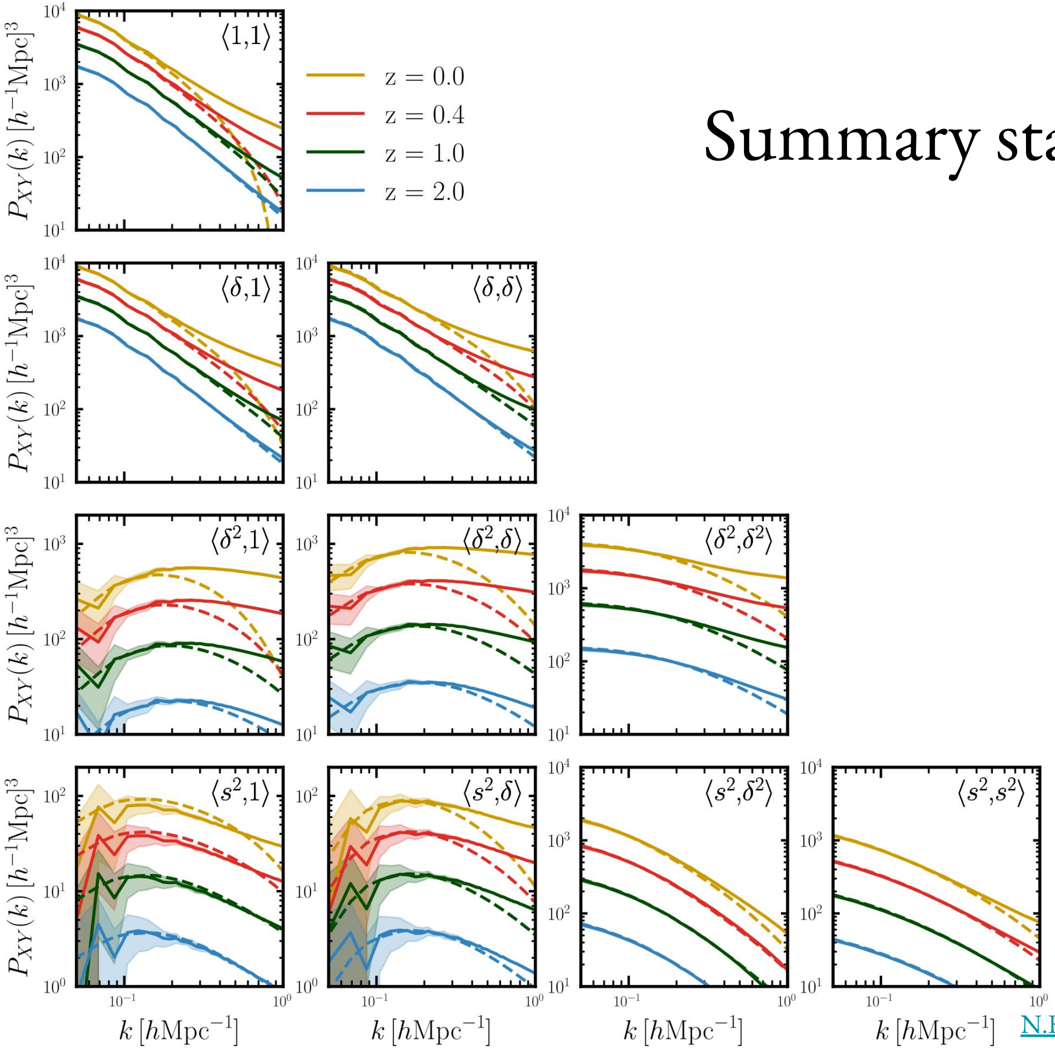
N.K.+ 2021a

$$\delta_g(\boldsymbol{x}) = \delta_m(\boldsymbol{x}) + \boxed{b_1 \mathcal{O}_\delta(\boldsymbol{x}) + b_2 \mathcal{O}_{\delta^2}(\boldsymbol{x}) + b_{s^2}\mathcal{O}_{s^2}(\boldsymbol{x})}$$

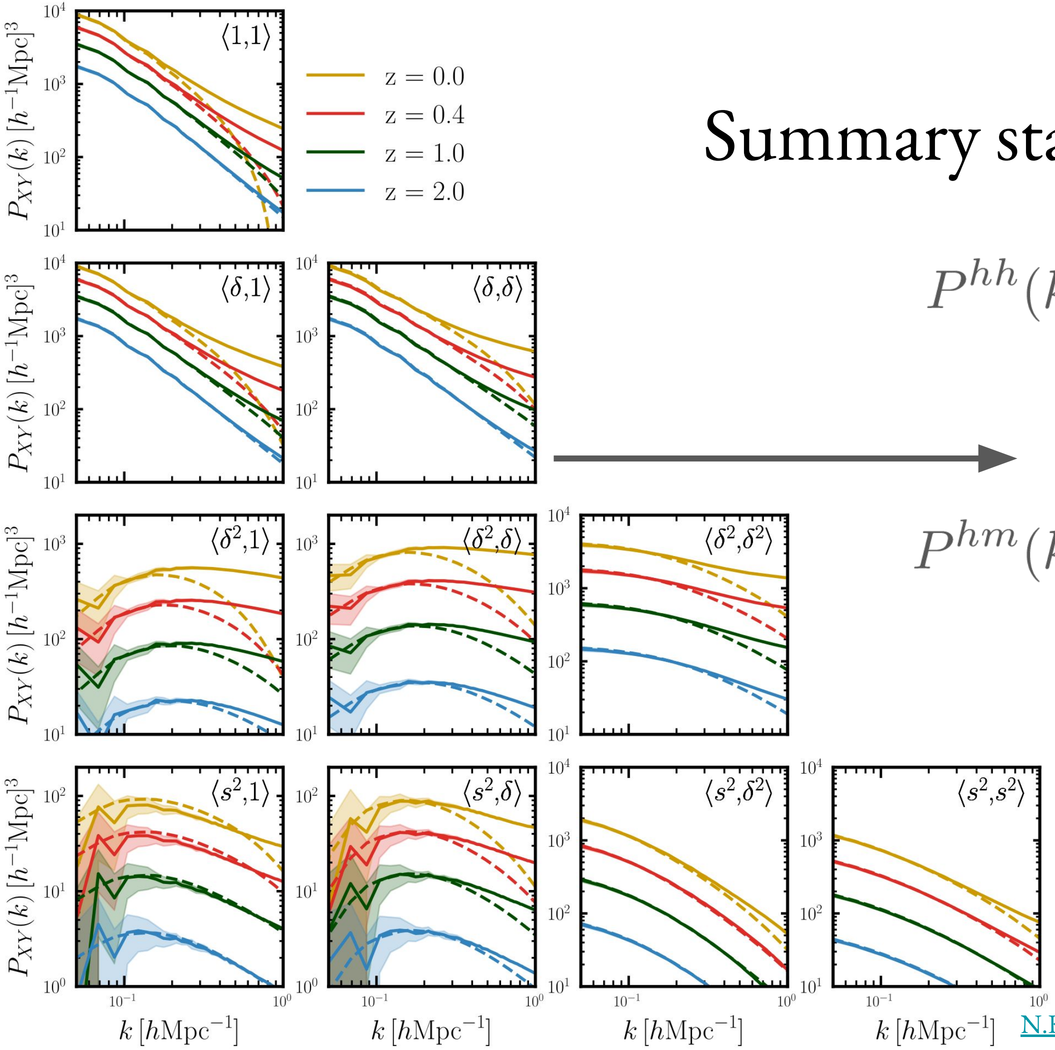
$$+ b_{\nabla^2}\mathcal{O}_{\nabla^2}(\boldsymbol{x}) + \epsilon(\boldsymbol{x}) + \cdots$$

Each panel is showing *the exact same particles*.

Summary statistics consistent with LPT



N.K.+ 2021a



Summary statistics consistent with LPT

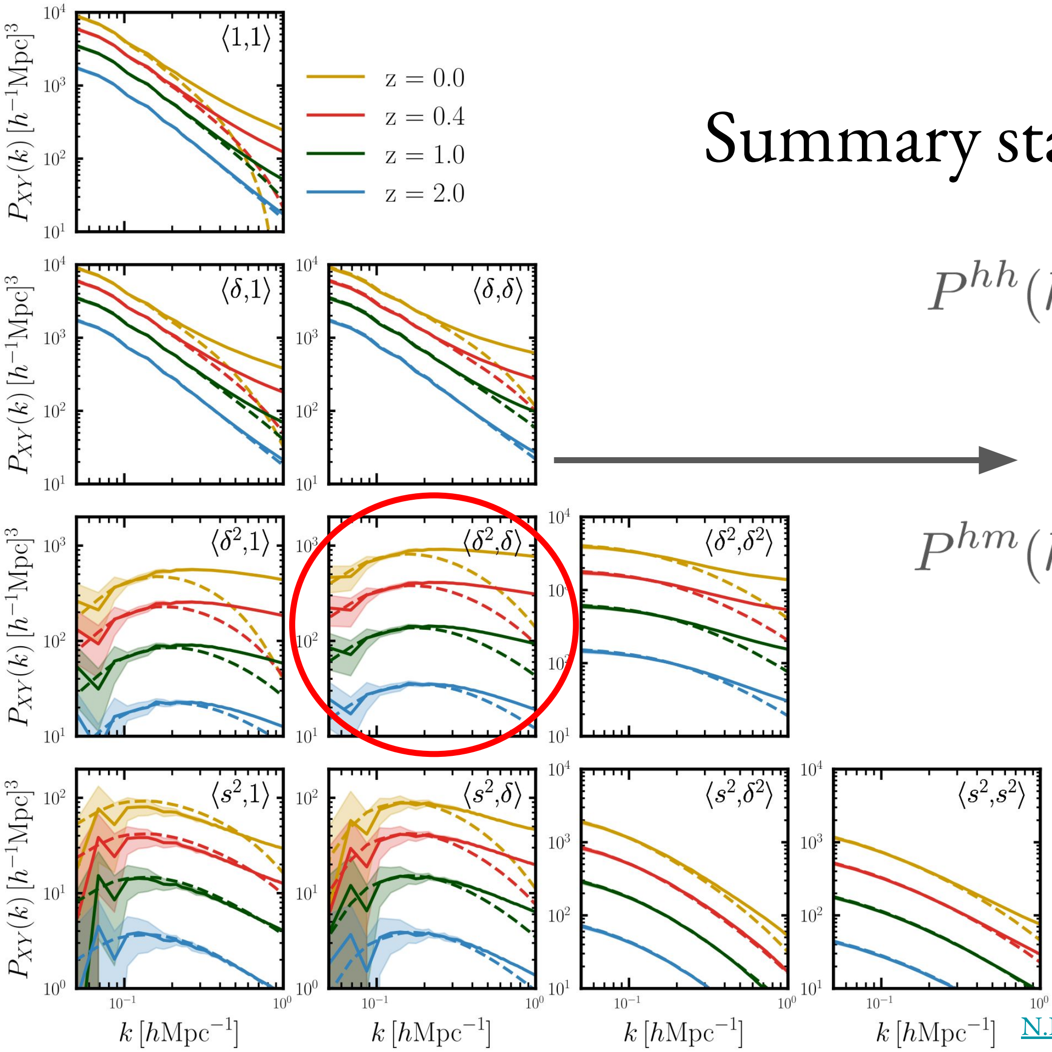
$$P^{hh}(k) = P_{11}(k) + b_1 P_{\delta 1}(k) + b_1^2 P_{\delta\delta}(k) + b_1 b_2 P_{\delta\delta^2}(k) + \cdots$$

$$P^{hm}(k) = P_{11}(k) + b_1 P_{\delta 1}(k) + b_2 P_{\delta^2 1}(k) + b_{s^2} P_{s^2 1}(k) + b_{\nabla^2} P_{\nabla^2 1}(k)$$

Free parameters here are the **same** as in PT-based analyses of RSD surveys.

Jointly analyse small scale lensing and RSD with the same nuisance parameters

N.K.+ 2021a



Summary statistics consistent with LPT

$$P^{hh}(k) = P_{11}(k) + b_1 P_{\delta 1}(k) + b_1^2 P_{\delta\delta}(k) + b_1 b_2 P_{\delta\delta^2}(k) + \cdots$$

$$P^{hm}(k) = P_{11}(k) + b_1 P_{\delta 1}(k) + b_2 P_{\delta^2 1}(k) + b_{s^2} P_{s^2 1}(k) + b_{\nabla^2} P_{\nabla^2 1}(k)$$

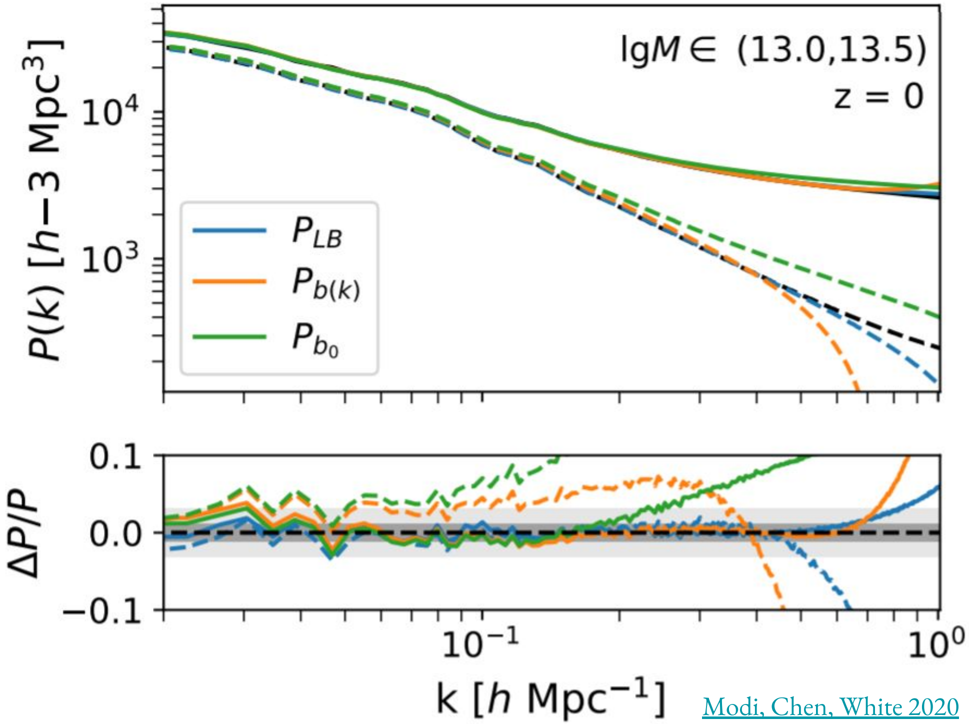
Free parameters here are the **same** as in PT-based analyses of RSD surveys.

Jointly analyse small scale lensing and RSD with the same nuisance parameters

N.K.+ 2021a

The combination is better than the sum of its parts

(and can also describe assembly bias + baryon feedback)



Modi, Chen, White 2020

Hybrid EFT can be used for parameter inference

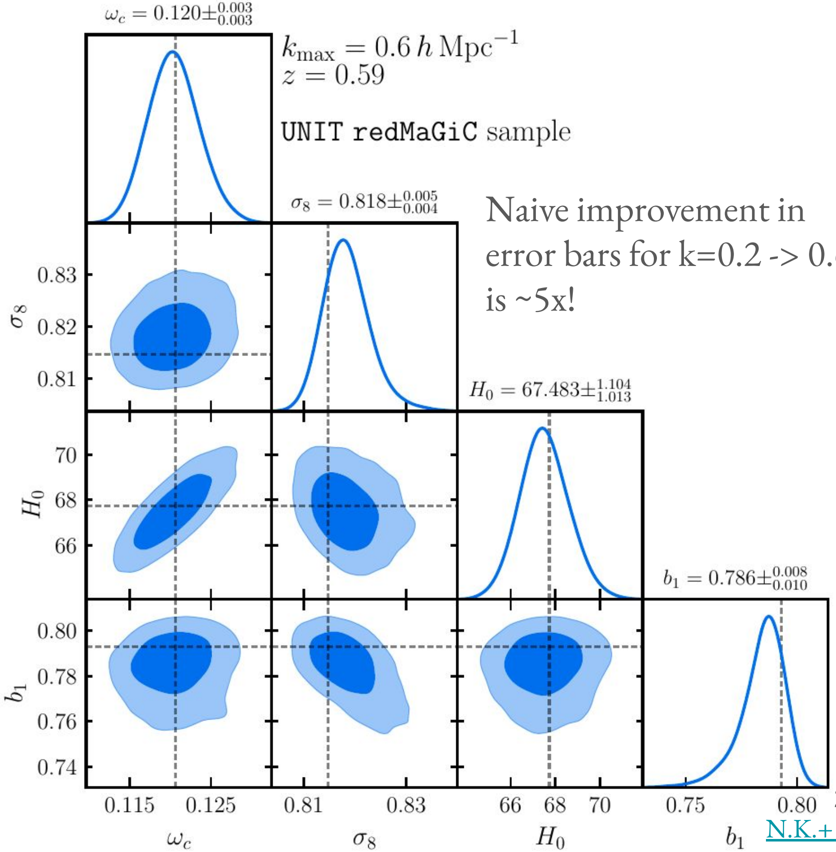
Emulator of 10 Lagrangian basis spectra built using Aemulus, anzu

Mock analysis on red galaxy HOD from independent simulation + redshift slice

Jointly fit P_{gg} and P_{gm} while varying all bias parameters.

Smallest scales are k_{max}~0.6 h Mpc^{-1}

True cosmology recovered to within one σ



Naive improvement in error bars for k=0.2 -> 0.6 is ~5x!

N.K.+ 2021a

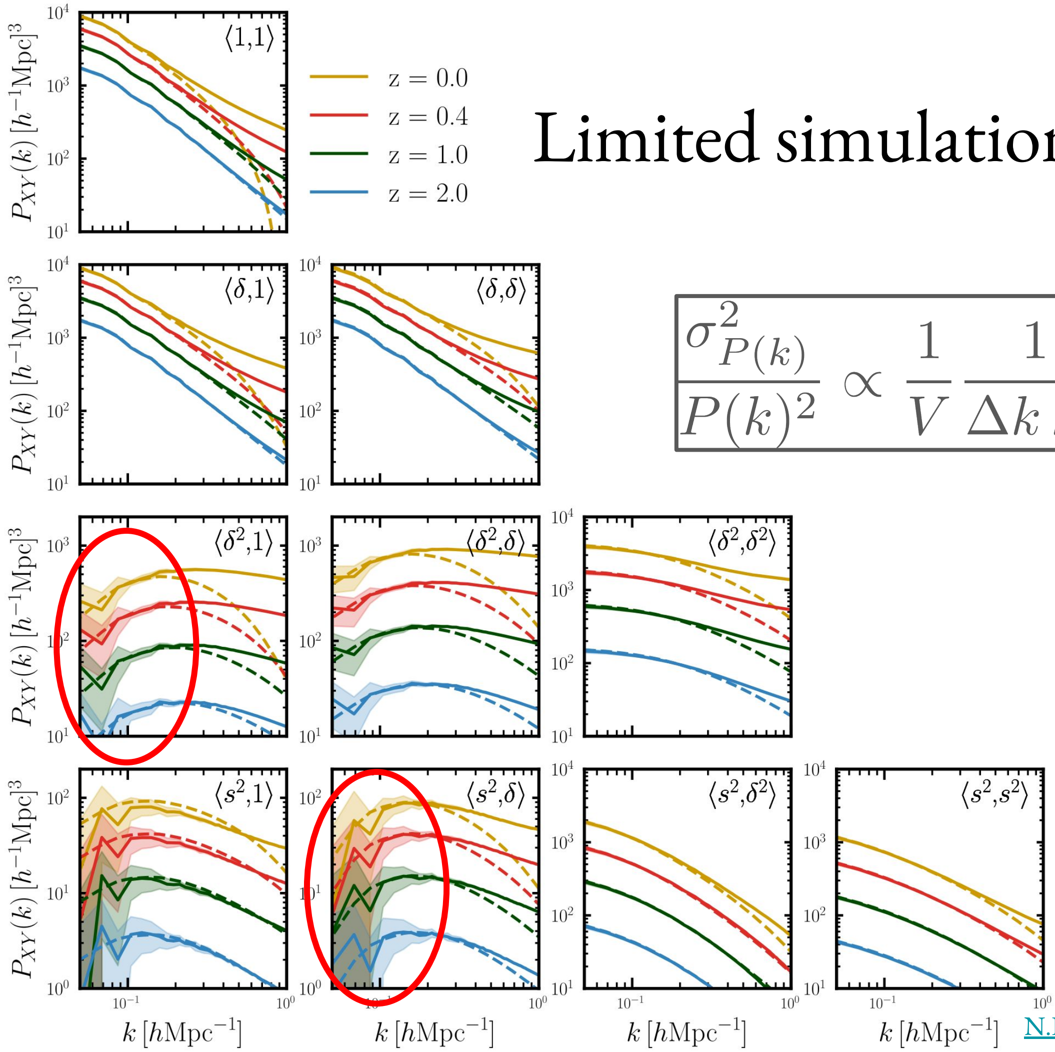
$$P(\theta|\mathbf{d}) \propto \mathcal{L}\left(\mathbf{d}|\mu(\theta)\right)\pi(\theta)$$

Models must be built for each summary statistic

-How do we efficiently probe smaller scales?

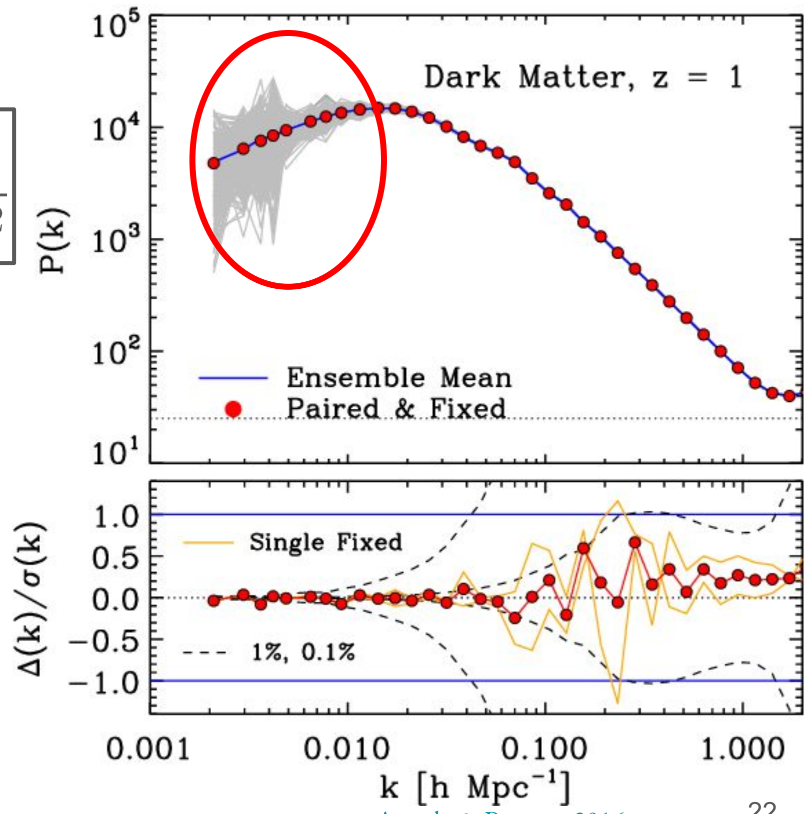
-**Are our models accurate enough for next generation surveys?**

Limited simulation volume hinders modelling



$$\boxed{\frac{\sigma^2_{P(k)}}{P(k)^2} \propto \frac{1}{V}\,\frac{1}{\Delta k\, k^2}}$$

N.K.+ 2021a



Angulo & Pontzen 2016

Control variates and variance reduction

When one wishes to estimate the mean of a noisy quantity (such as a power spectrum) but can produce cheap correlated surrogates, a new estimator can be defined

$$\hat{y} \equiv \hat{x} - \beta(\hat{c} - \mu_c)$$

Minimizing the variance of y gives

$$\beta^\star = \frac{\mathrm{Cov}[\hat{x}, \hat{c}]}{\mathrm{Var}[\hat{c}]}$$

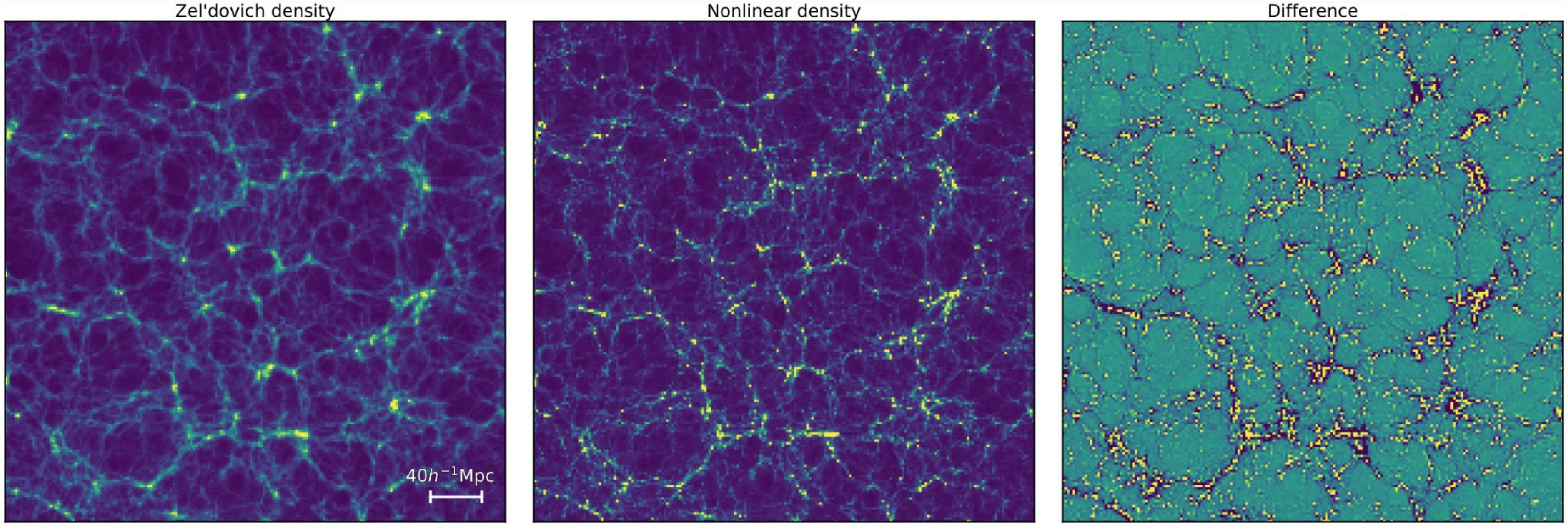
Which leads to a variance reduction that depends on the correlation coefficient

$$\frac{\mathrm{Var}[\hat{y}]}{\mathrm{Var}[\hat{x}]} = 1 - \frac{\mathrm{Cov}^2[\hat{x}, \hat{c}]}{\mathrm{Var}[\hat{x}]\mathrm{Var}[\hat{c}]} = 1 - \rho_{xc}^2$$

High correlation -> large reductions in variance!

The Zel'dovich approximation as a surrogate

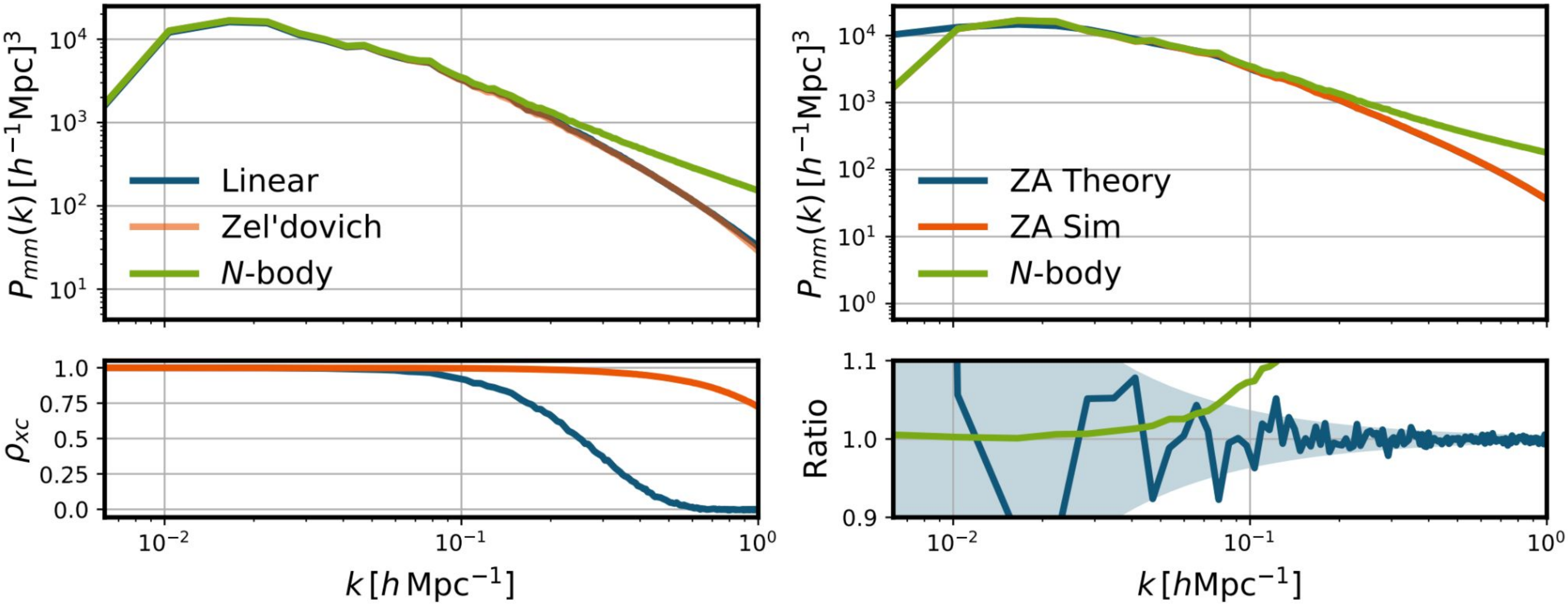
For a full 3D cosmological volume, the agreement on large scales between ZA and N-body is strong
Analytic calculations are well understood in this approximation



IC codes give you the ingredients needed to predict ZA fields at several redshifts **for free***.

N.K.+22

Zel’dovich fields correlate highly with nonlinear simulations

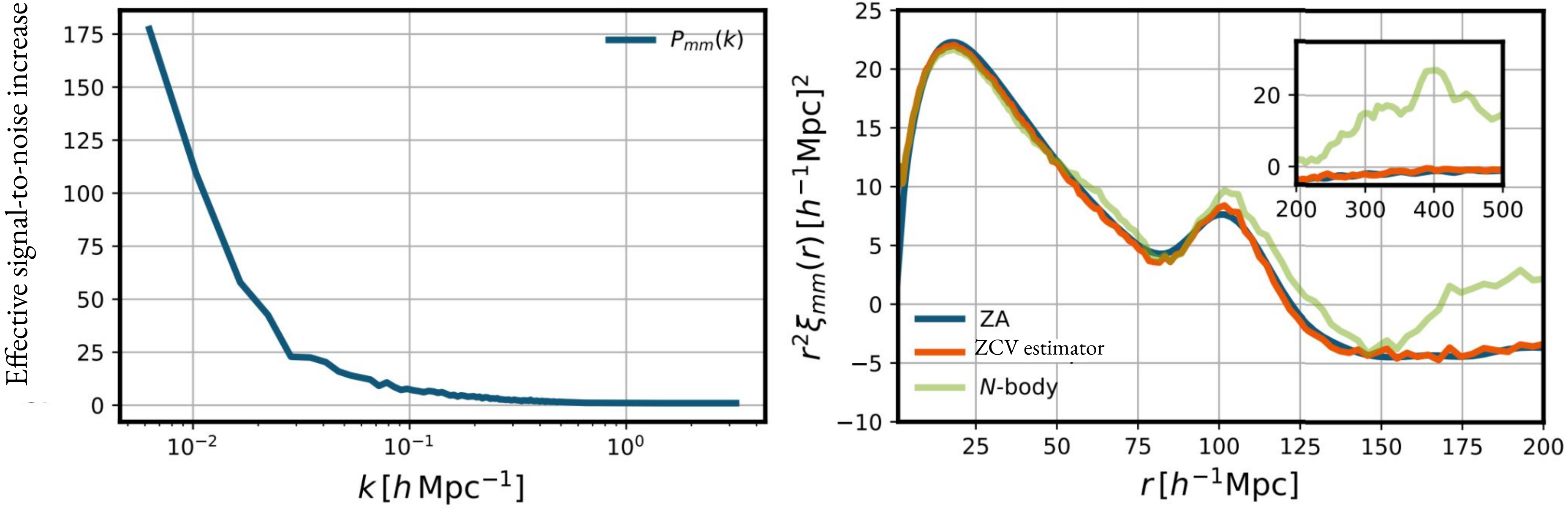


N.K.+22

How powerful of a control variate is the Zel'dovich approximation?

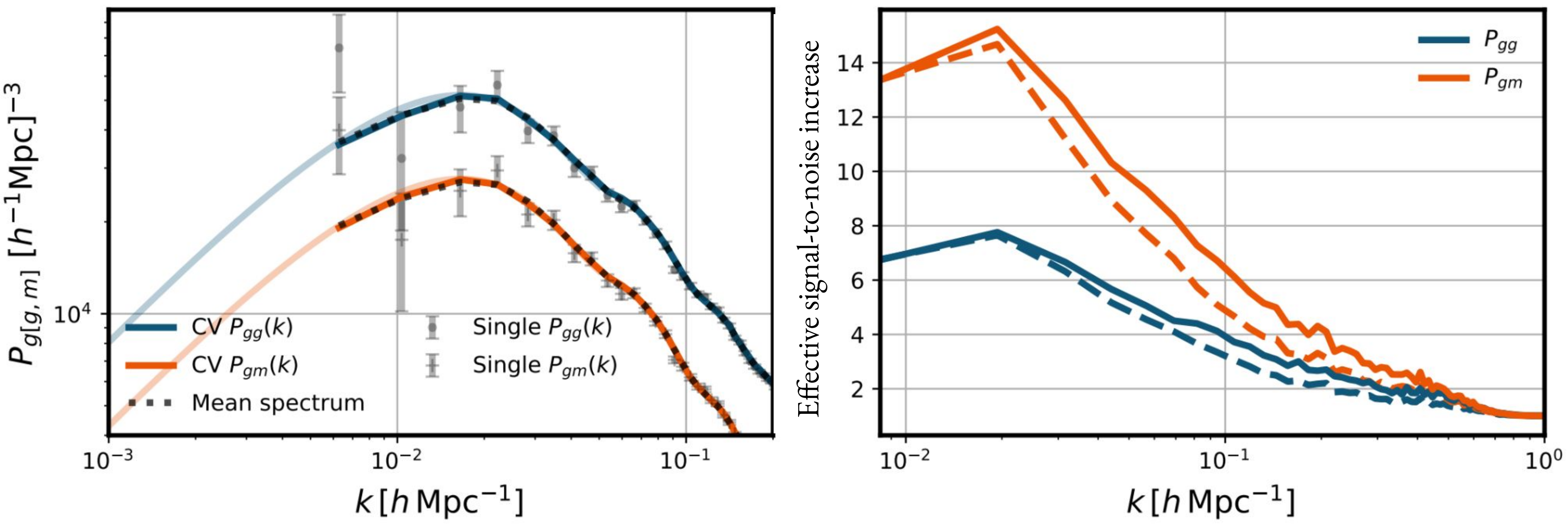
$$\hat{P}^{\mathrm{CV}}(k) = \hat{P}^{\mathrm{Nbody}}(k) - \hat{\beta}(k)\left(\hat{P}^{\mathrm{ZA}}(k) - P^{\mathrm{ZA}}(k)\right)$$

Variance reduction for matter 2-pt statistics



N.K.+22

Variance reduction for galaxy clustering and lensing

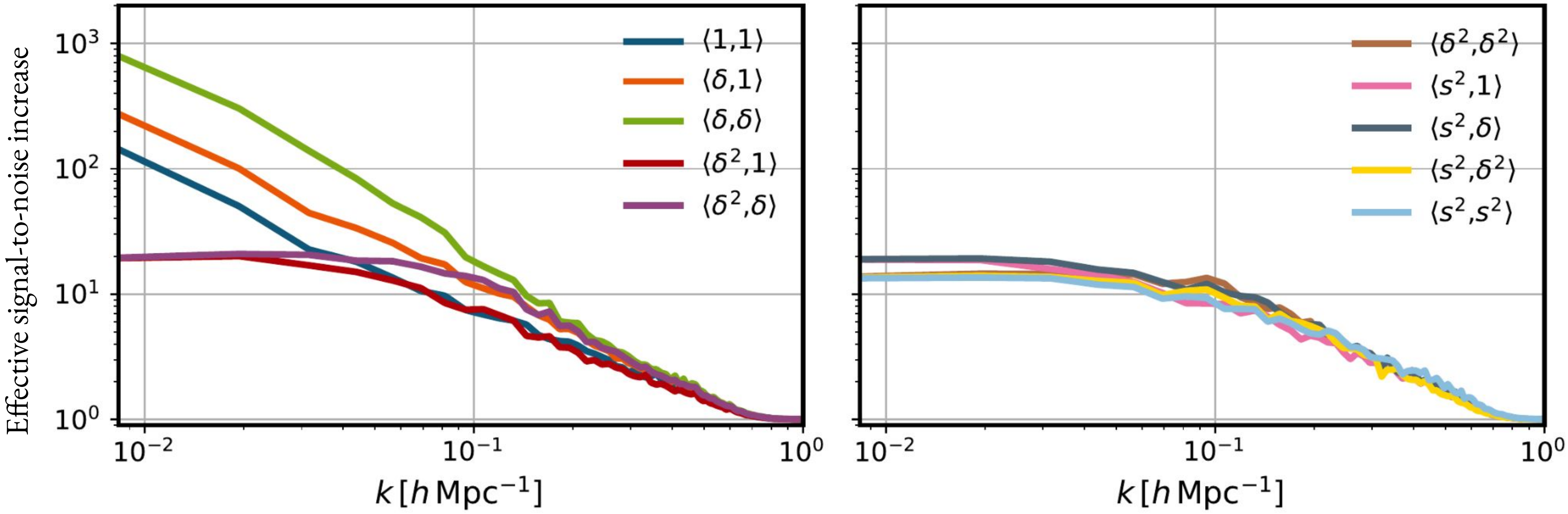


N.K.+22

General variance reduction for biased tracers

Similar reduction in sample variance for all LPT basis spectra

Requires ~50 CPU hours to produce variance-reduced basis functions for $(1024)^3$ sim



N.K.+22

Using simple PT as surrogate N-body sims is *powerful*

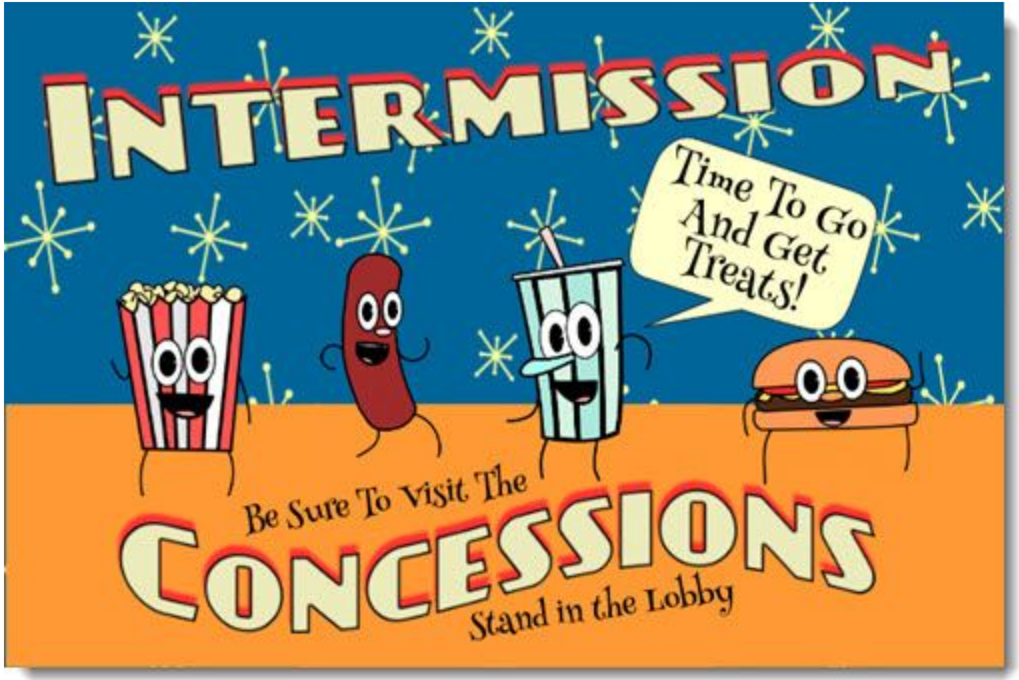
Emulators can be designed with significantly smaller boxes

Accurate mock surveys with smaller volumes

Unlike "paired-fixed" simulations, no additional requirements. Just one run of an IC code.

Being extended to redshift space and other summary statistics!

Questions?



INTERMISSION
Time To Go And Get Treats!
Be Sure To Visit The
CONCESSIONS
Stand in the Lobby

$$P(\theta|\mathbf{d}) \propto \mathcal{L}\left(\mathbf{d}|\mu(\theta)\right)\pi(\theta)$$

Raw data compressed to summary statistics

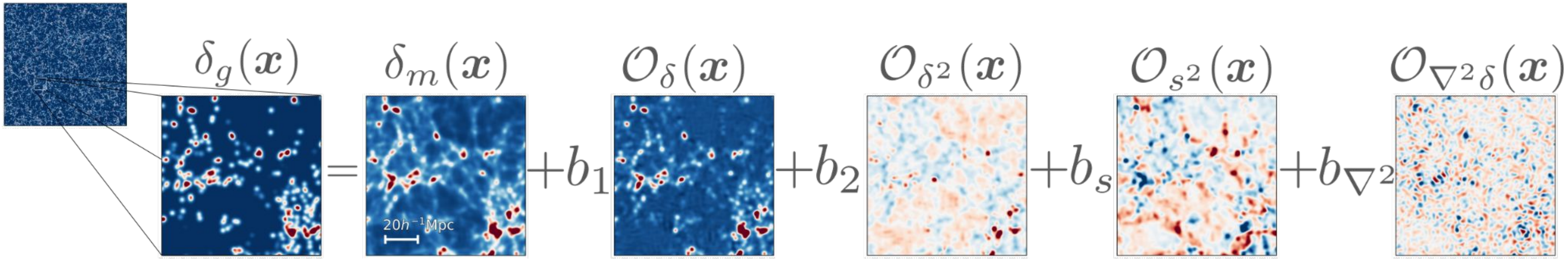
-**What are the optimal summary statistics beyond two-point functions?**

Priors on model parameters

-If better models require many nuisance parameters, can we gain prior understanding on them?

Hybrid EFT is a field-level model for tracers

HEFT should work at more than just the two-point statistic level. How can we test this?



What about beyond two-point?

First example might be to look at higher N-point correlations

$$\xi^{(N)}(\mathbf{r}_1, \cdots, \mathbf{r}_N) \sim \langle \delta_g(\mathbf{r}_1) \cdots \delta_g(\mathbf{r}_N) \rangle$$

But with lots of bias fields, even the bispectrum (3pt function) is combinatorially challenging to model and measure!

Is there something simpler?

Counts in cells are a promising “new” statistic

N-point functions are probing moments of the underlying distribution of galaxies, $\mathcal{P}\left(\delta_g\right)$.

How can we probe the distribution directly?

Look at the histogram of galaxy densities in your survey! These are *counts-in-cells.*

$$\mathcal{P}_{k|V}$$

“Probability of finding *k* counts in a sphere of volume V”

Counts-in-cells date back to at least Hubble (1934), who noted the lognormality of the galaxy density field

The information content of counts-in-cells

Generating function of CiC explicitly probes all connected correlation functions (White, 1979)

$$\mathcal{P}_k(V) = \frac{1}{k!}\left[\left(\frac{\mathrm{d}}{\mathrm{dz}}\right)^k \exp\left[\sum_{N=0}^{\infty} \frac{\bar{n}^N (z-1)^N}{N!} \times \int_V \dots \int_V \mathrm{d}^3\boldsymbol{r}_1 \dots \mathrm{d}^3\boldsymbol{r}_N \xi^{(N)}(\boldsymbol{r}_1 \dots \boldsymbol{r}_N)\right]\right]_{z=0}$$

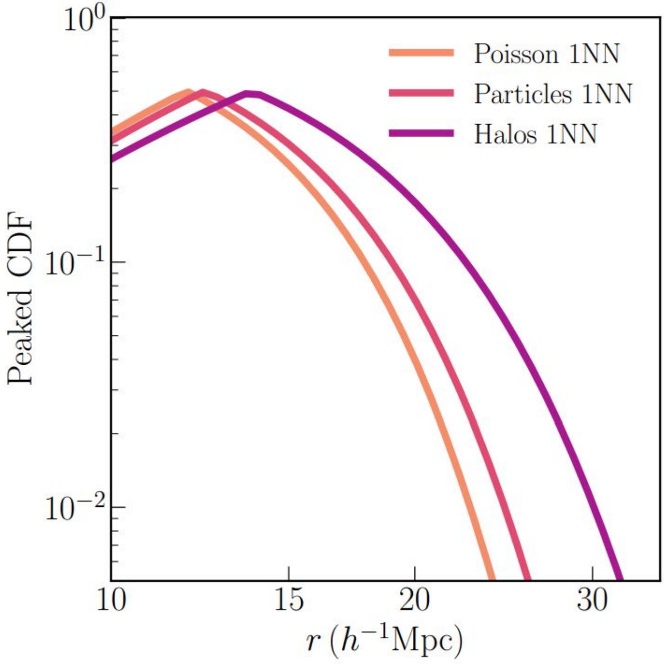
Efficiently probing counts-in-cells statistics

Banerjee & Abel (2020) demonstrated a computationally efficient way to probe CiC statistics using k nearest-neighbour cumulative distribution functions (kNN-CDFs)

$$\mathrm{CDF}_k(r) = \mathcal{P}_{>k-1}(V)$$

$$= 1 - \sum_{i=0}^{k} \mathcal{P}_i(V)$$

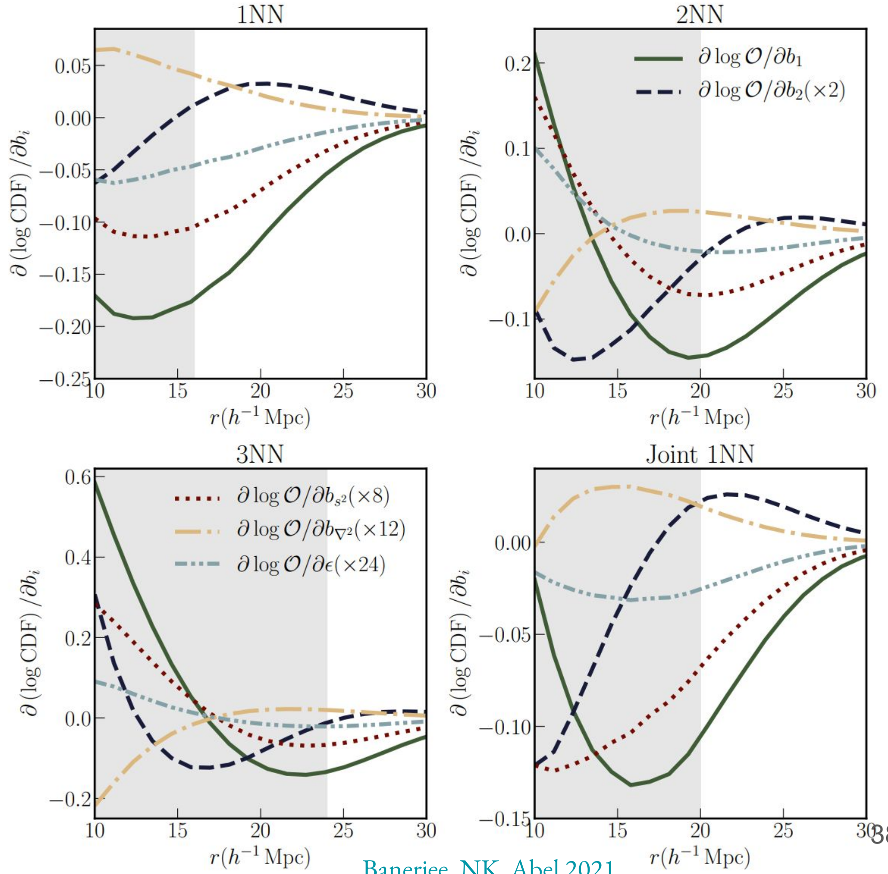
$\mathrm{CDF}_k(r)$ can be measured easily.



Can hybrid EFT describe kNN statistics?

Implicit check of ability to model *all* N-point functions.

Each bias parameter induces a unique small-scale response in kNN-CDFs.

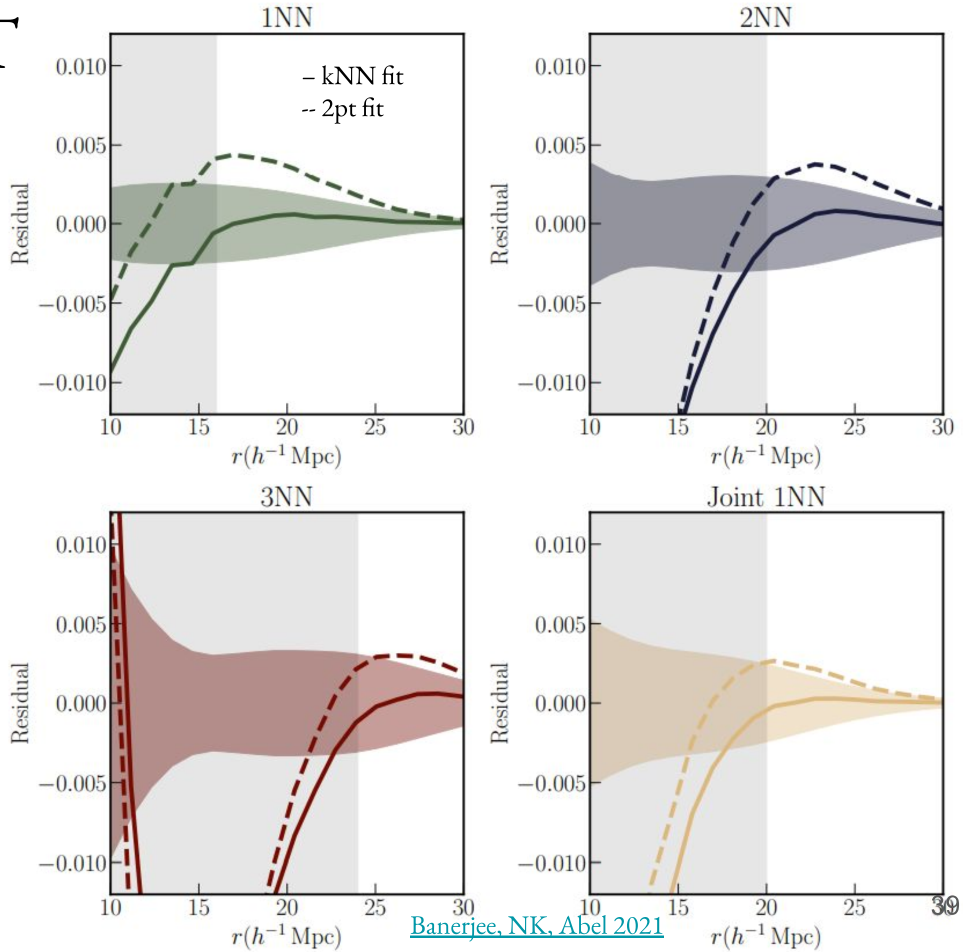


Banerjee, NK, Abel 2021

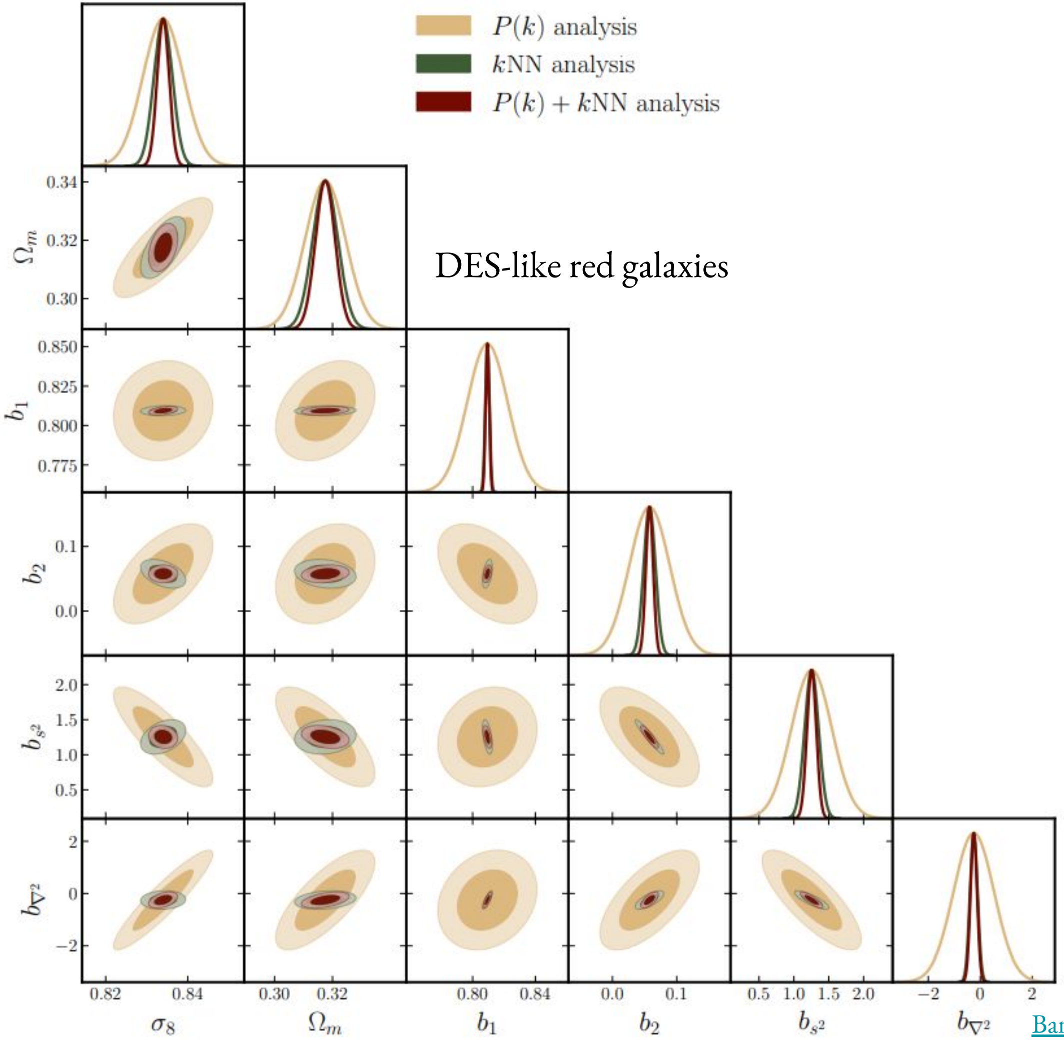
Internal consistency of HEFT

Bias parameters consistent with the clustering and lensing spectra of halos simultaneously describe kNN statistics!

Note: *not* the best-fit biases to two-point statistics



Banerjee, NK, Abel 2021



Forecasted improvements

Combining P(k) with kNN statistics sharply tightens measurements of bias parameters.

Ω_m-σ_8 Figure of Merit is increased by ~3.6.

Follow-up: to what scales are ZA kNNs valid?

[Banerjee, NK, Abel 2021](#)

$$P(\theta|\mathbf{d}) \propto \mathcal{L}\left(\mathbf{d}|\mu(\theta)\right)\pi(\theta)$$

Raw data compressed to summary statistics

-What are the optimal summary statistics beyond two-point functions?

Priors on model parameters

-If better models require many nuisance parameters, can we gain prior understanding on them?

Field-level bias parameters

Previously, we fit to “galaxy” or halo samples by maximizing fits to clustering and lensing correlation functions

$$P_{gg} = \langle \delta_g \delta_g \rangle,$$
$$P_{gm} = \langle \delta_g \delta_m \rangle$$

These parameters don’t necessarily describe all other statistics. Analysis affected by sample variance. Consider the residuals between the galaxy field and the bias fields

$$\epsilon(\boldsymbol{x}) = \delta_g(\boldsymbol{x}) - \delta_m(\boldsymbol{x}) - \sum_i b_i \mathcal{O}_i(\boldsymbol{x})$$

Minimizing $\langle \epsilon^2 \rangle$ leads to an analytic estimator for bias parameters that is easy to implement

Probing stochasticity with field-level bias models

With an estimate of best-fit parameters we can then directly probe the residuals of these bias models for *any* model of the galaxy-halo connection

$$\hat{\epsilon}(\mathbf{k}) = \delta_g(\mathbf{k}) - \delta_m(\mathbf{k}) - \sum_i \hat{b}_i \mathcal{O}_i(\mathbf{k})$$

$$P_{\mathrm{err}} = \langle \hat{\epsilon}\hat{\epsilon} \rangle$$

From symmetries, this function should broadly behave as (Desjacques et al 2018)

$$P_{\mathrm{err}} = \frac{1}{\bar{n}} \left[a_1 + a_2 k^2 + \cdots \right],$$

$a_1 = 1$ is the Poisson shot-noise prediction. Is the Poisson approximation good enough?
For samples of red galaxies, how well does the above form hold?

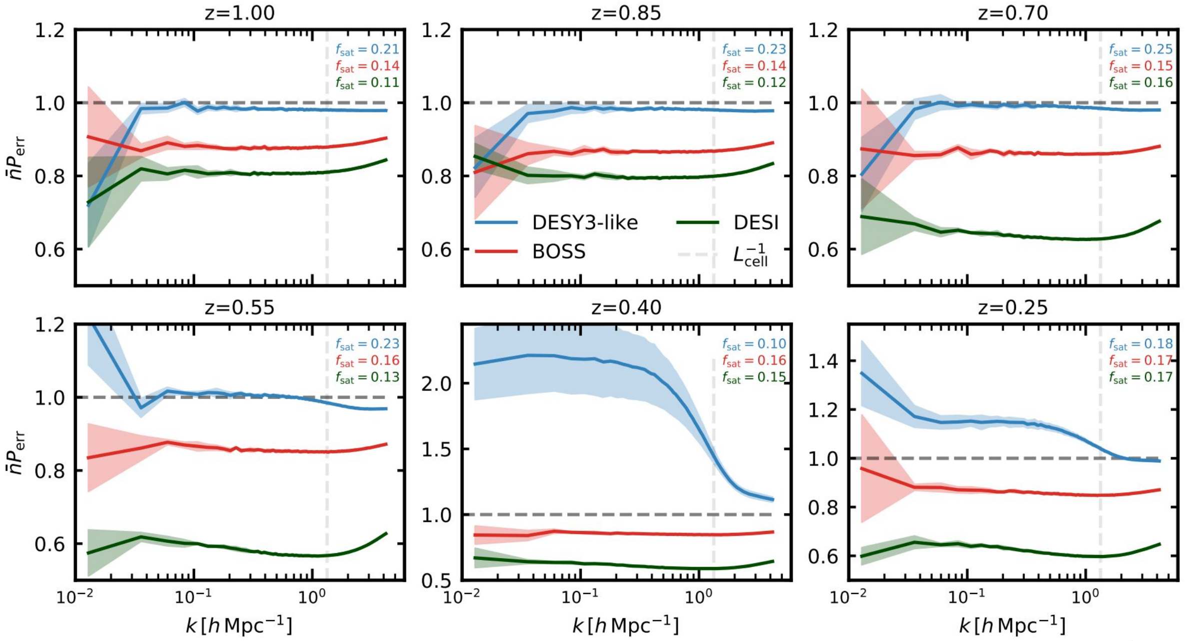
Red galaxy stochasticity

Large scale P_{err} spans from sub to super-Poisson!

Deviations within ~30%->Informative priors on a_0

No necessary trend with derived parameters like f_{sat}

See paper (or ask me!) about a potential explanation of this phenomenology



N.K.+21b

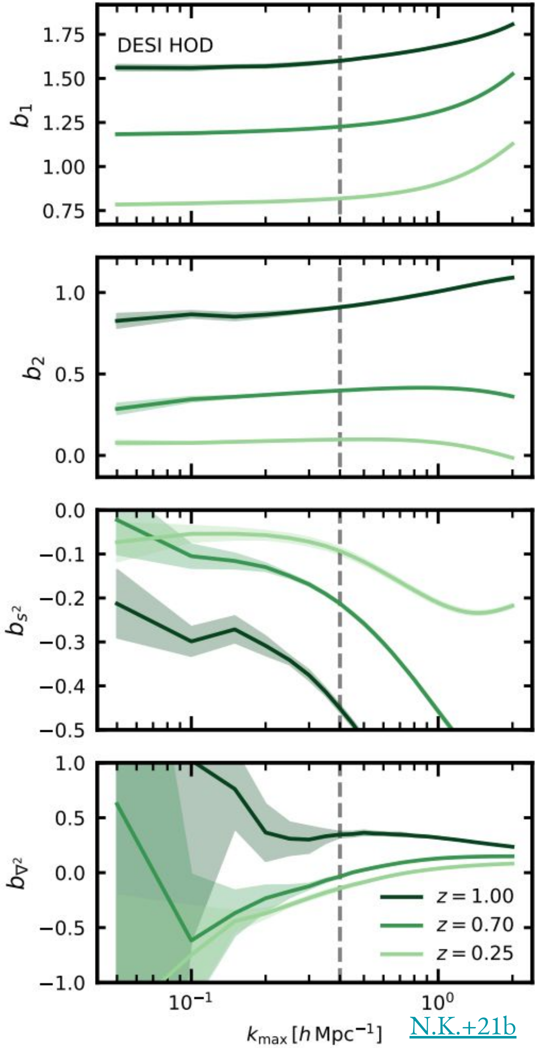
Biases themselves

Beyond studying stochasticity, we can also use field-level inference to study bias parameters themselves.

Priors on higher-order biases are crucial to not dilute cosmological information using these analyses (Zennaro+21, Cabass+22, Philcox+22)

Lagrangian bias is also a "Rosetta stone" of galaxy-halo connection models!

-Ongoing work with Mahlet Shiferaw, a 2nd year Stanford PhD student

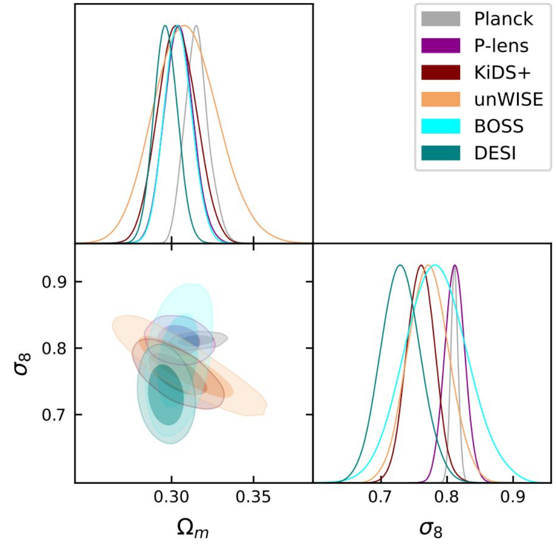


N.K.+21b

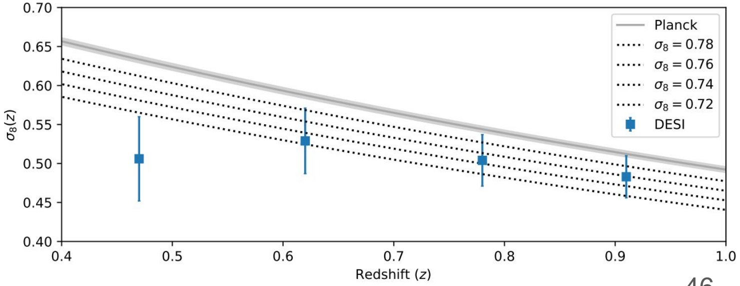
What do priors enable?

As models become more and more complicated, better priors on parameters are needed.

Informed priors will maximize cosmology results.



Improved models
+
Priors on stochasticity
+
Priors on bias parameters



White et al (incl N.K.), arXiv:2111.09898

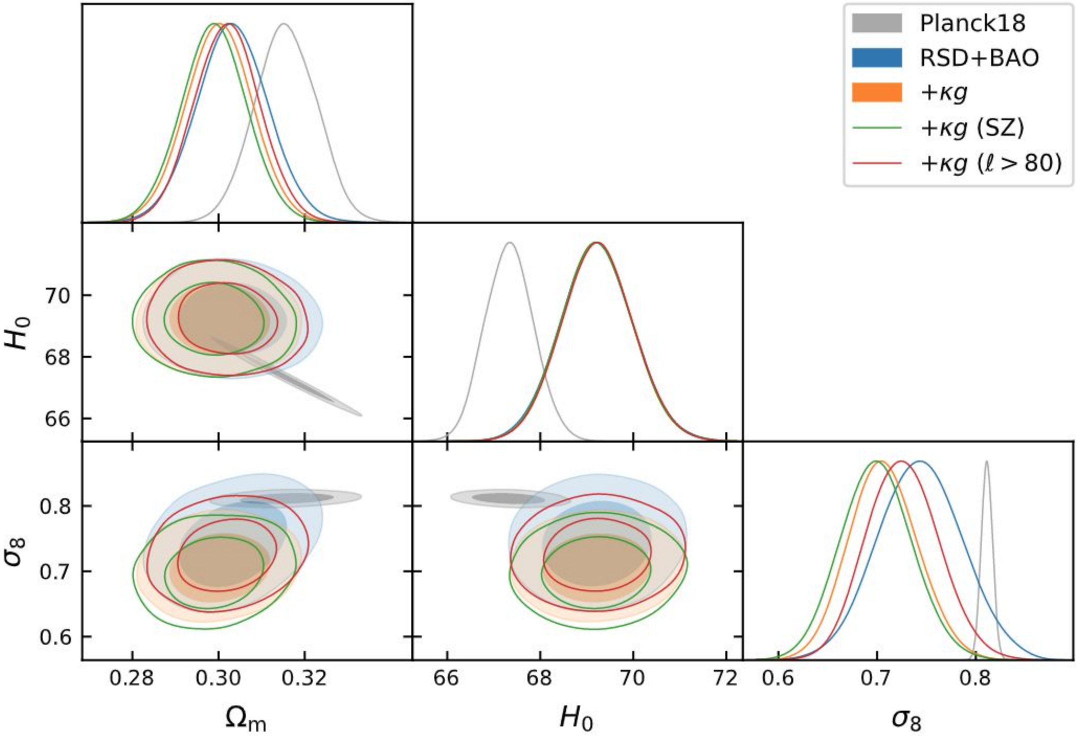
All together: BAO + RSD + CMB lensing

All observables self-consistently modelled in LPT with the same set of bias parameters.

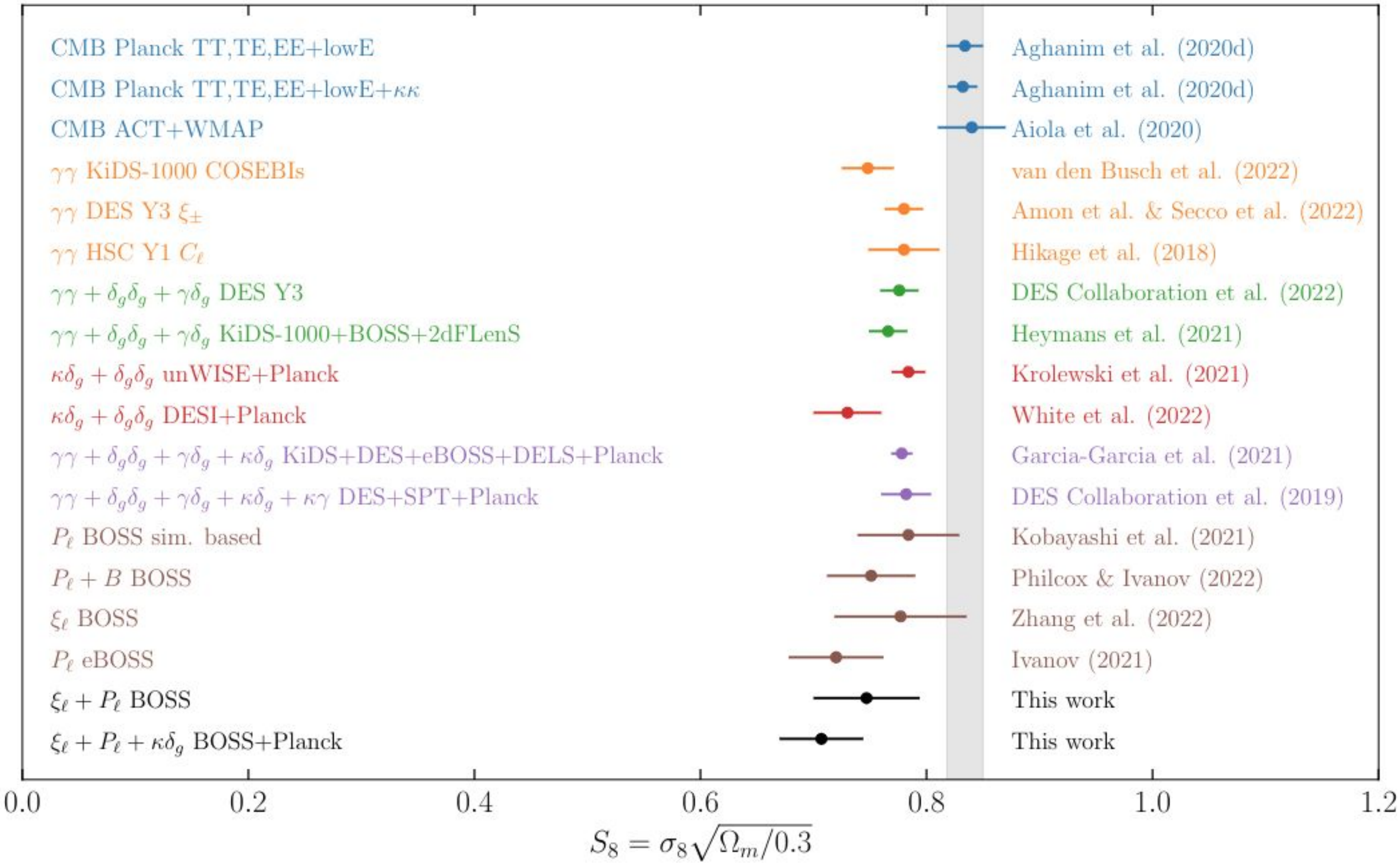
Limited by degeneracies in biases.

Beter data allows us to probe smaller scales, priors help in degeneracy-breaking

Picture will sharpen significantly in coming years!



Chen, White, DeRose, N.K 2022.



CMB Planck TT,TE,EE+lowE
Aghanim et al. (2020d)
CMB Planck TT,TE,EE+lowE+$\kappa\kappa$
Aghanim et al. (2020d)
CMB ACT+WMAP
Aiola et al. (2020)
$\gamma\gamma$ KiDS-1000 COSEBIs
van den Busch et al. (2022)
$\gamma\gamma$ DES Y3 $\xi_\pm$
Amon et al. & Secco et al. (2022)
$\gamma\gamma$ HSC Y1 C_ℓ
Hikage et al. (2018)
$\gamma\gamma + \delta_g\delta_g + \gamma\delta_g$ DES Y3
DES Collaboration et al. (2022)
$\gamma\gamma + \delta_g\delta_g + \gamma\delta_g$ KiDS-1000+BOSS+2dFLenS
Heymans et al. (2021)
$\kappa\delta_g + \delta_g\delta_g$ unWISE+Planck
Krolewski et al. (2021)
$\kappa\delta_g + \delta_g\delta_g$ DESI+Planck
White et al. (2022)
$\gamma\gamma + \delta_g\delta_g + \gamma\delta_g + \kappa\delta_g$ KiDS+DES+eBOSS+DELS+Planck
Garcia-Garcia et al. (2021)
$\gamma\gamma + \delta_g\delta_g + \gamma\delta_g + \kappa\delta_g + \kappa\gamma$ DES+SPT+Planck
DES Collaboration et al. (2019)
P_ℓ BOSS sim. based
Kobayashi et al. (2021)
$P_\ell + B$ BOSS
Philcox & Ivanov (2022)
ξ_ℓ BOSS
Zhang et al. (2022)
P_ℓ eBOSS
Ivanov (2021)
$\xi_\ell + P_\ell$ BOSS
This work
$\xi_\ell + P_\ell + \kappa\delta_g$ BOSS+Planck
This work
0.0
0.2
0.4
0.6
0.8
1.0
1.2
$S_8 = \sigma_8\sqrt{\Omega_m/0.3}$

Wrapping up

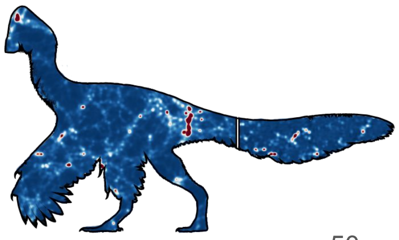
Modern cosmological inference is challenging on many fronts.

Thinking about problems at each step of an analysis from a multitude of angles can lead to solutions that are more powerful than any individual tool!

Models of galaxy survey observables can be made more efficient, more accurate, and more constraining by thinking about them from a perspective that combines simulations and analytic techniques.

This juncture of PT and simulations is a powerful lens through which to think about the formation of large-scale structures.

Thanks for having me!



Code available here!