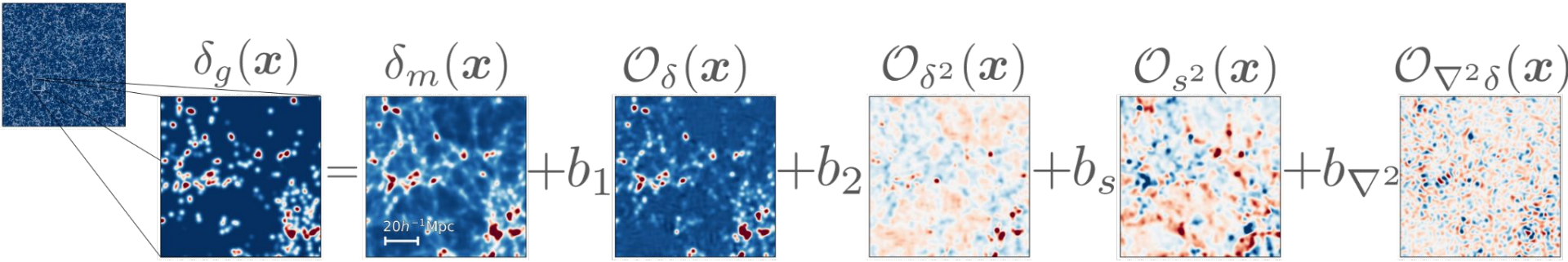
Pen-and-paper meets supercomputing: building accurate models for cosmological surveys



Nickolas Kokron
kokron@stanford.edu
INPA Seminar, LBNL
11/16/2022

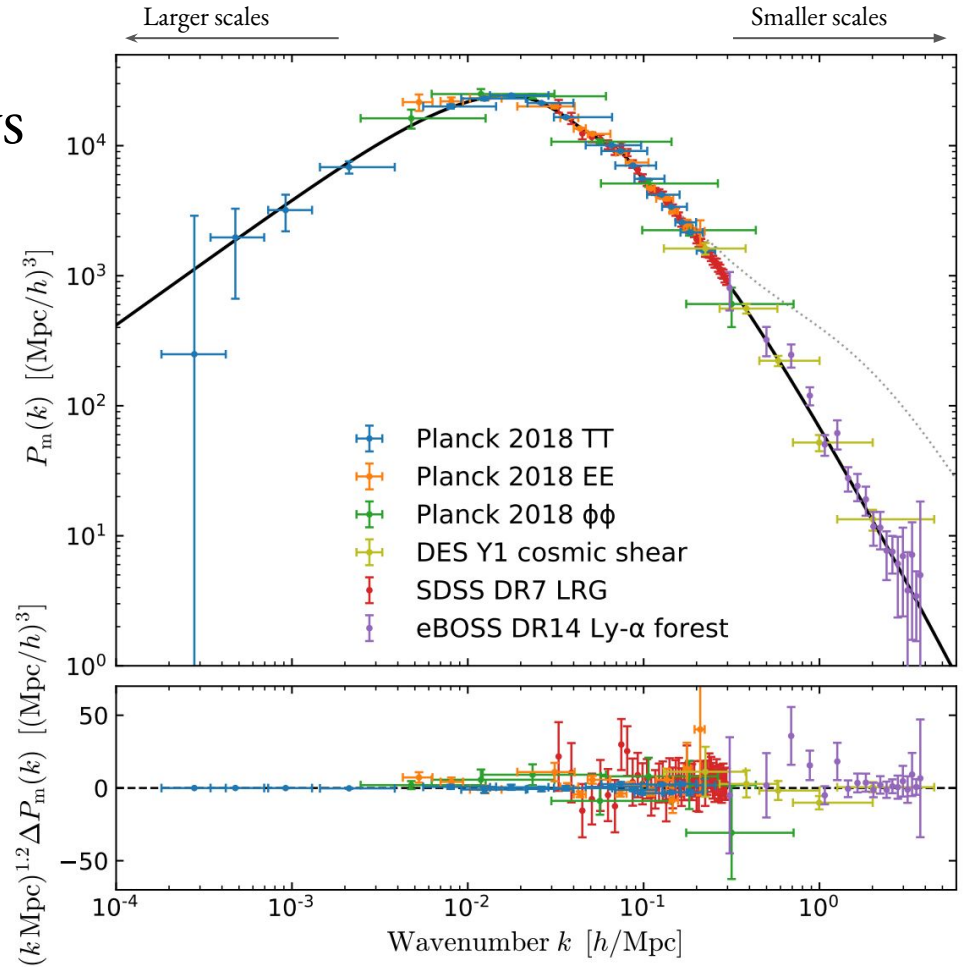
Work done with Tom **A**bel, Arka **B**anerjee, Stephen **C**hen, Joe **D**eRose, Mark Maus, Risa Wechsler, Martin White

The last decade of cosmic surveys

The last decade has seen a resounding confirmation of the concordance cosmological model, LCDM.

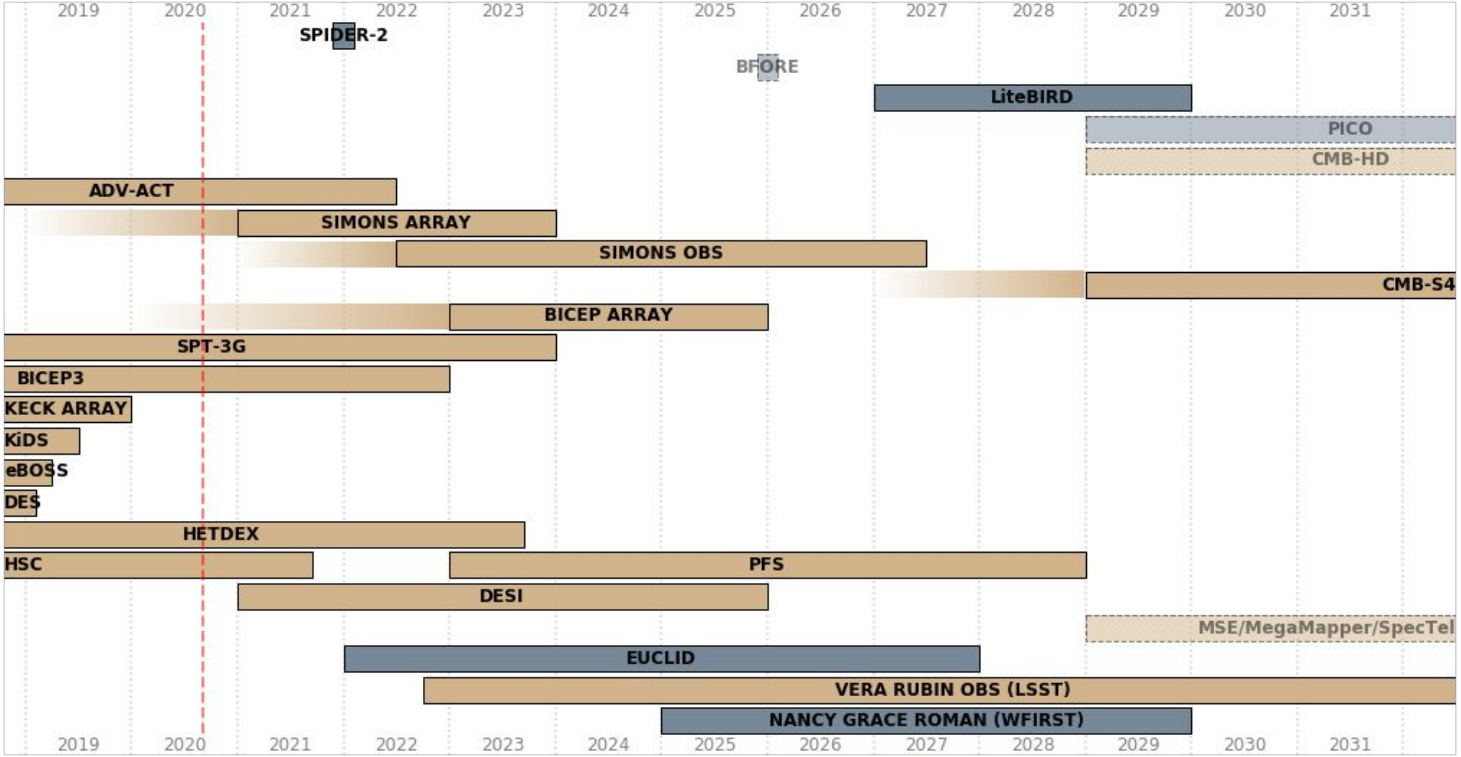
We are now approaching a new era, asking precision questions:

1. What is the sum of neutrino masses?
2. What is the nature of dark energy?
3. What is the nature of cosmic inflation?
4. What is the nature of **tensions** that have arisen between early and late Universe experiments?



Chabanier+19

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

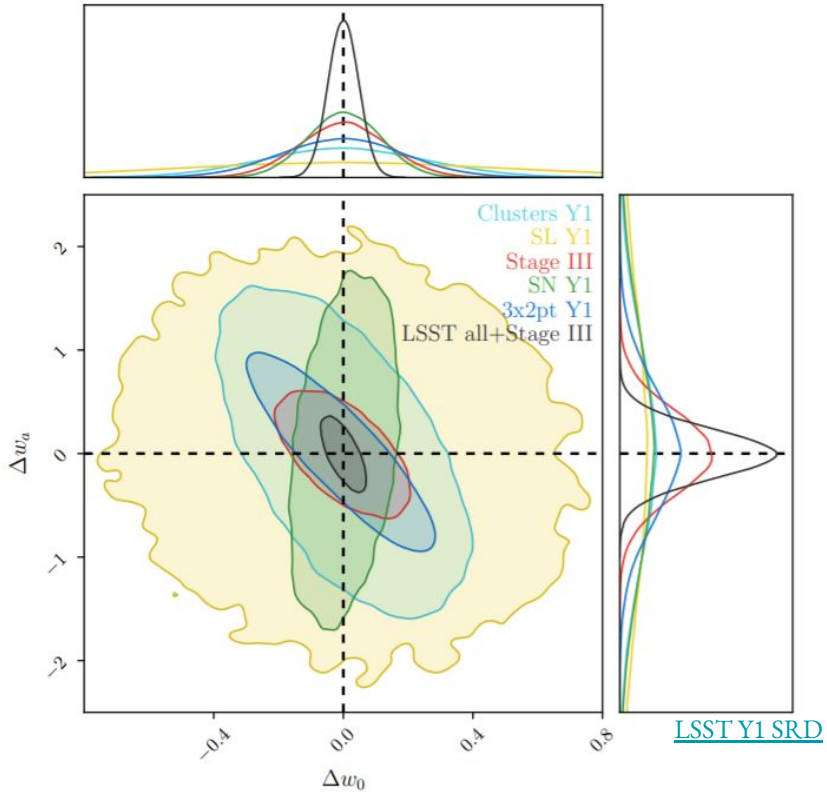


LSS probes measured:
-Galaxy positions
-Galaxy shapes
-Spec-zs
-CMB lensing
-kSZ
-tSZ
-More!

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

Image credit: David Kirkby

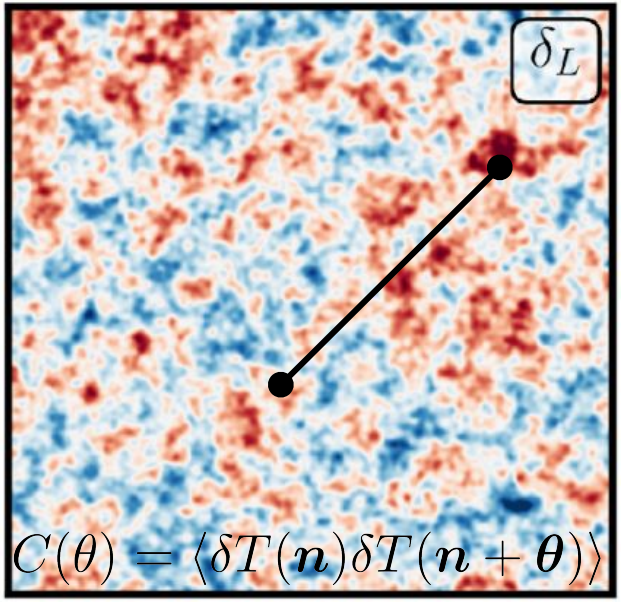
Next-generation clustering and lensing surveys promise unparalleled statistical power



LSST Y1 SRD

M.O.: Large data sets are compressed to summary statistics

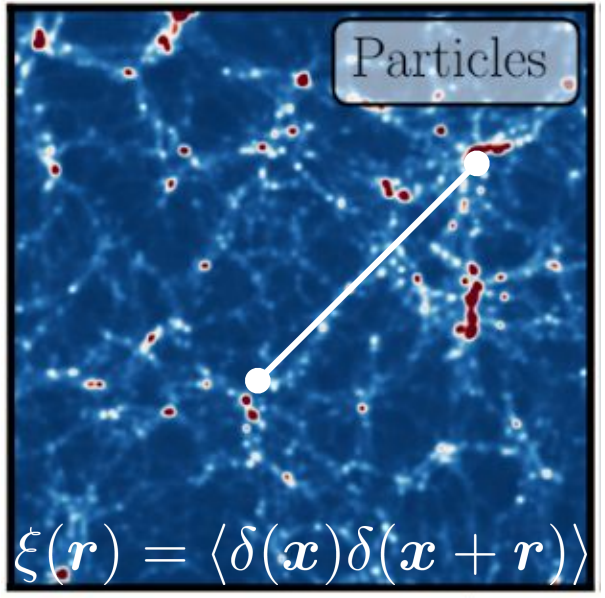
Density (temperature) fluctuations in the early Universe



$$\delta T(\boldsymbol{n}) = \frac{T(\boldsymbol{n}) - \bar{T}}{\bar{T}}$$

Gravitational collapse

Density fluctuations in the late Universe

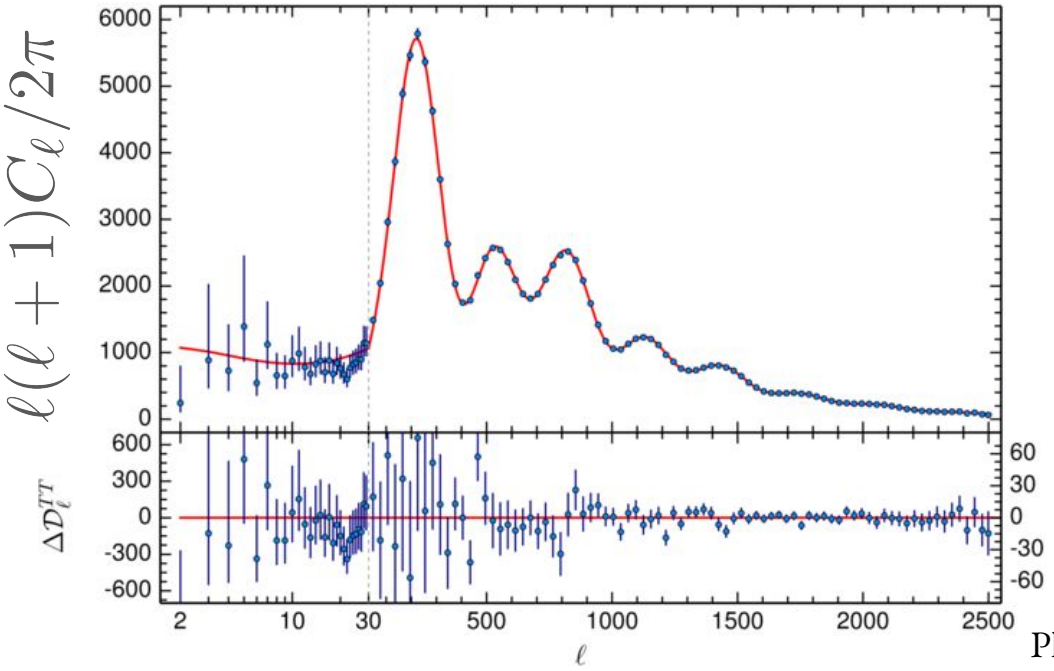


$$\delta(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x}) - \bar{\rho}}{\bar{\rho}}$$

M.O.: Large data sets are compressed to summary statistics

Alternatively look at the *spectrum* of these fluctuations.

For 2D cross-correlations, $C(\theta) \rightarrow C_\ell$



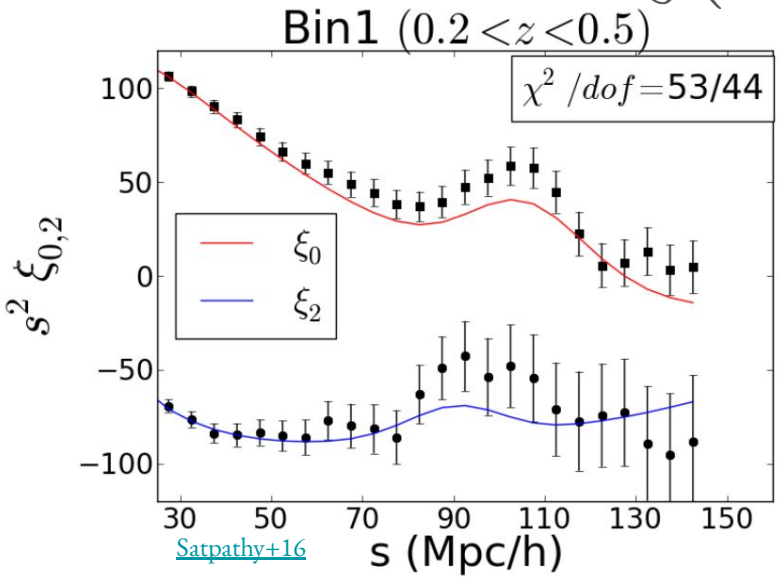
Planck 2015

M.O.: Large data sets are compressed to summary statistics

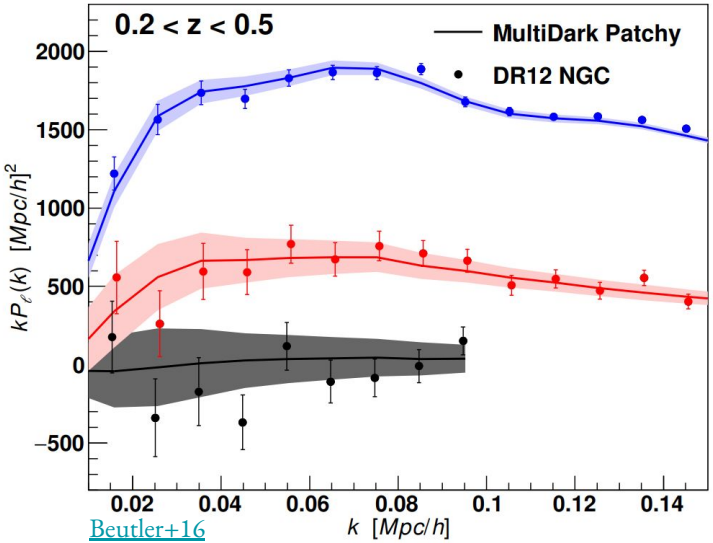
Alternatively look at the *spectrum* of these fluctuations.

For 3D data sets, correlation functions Fourier transform to *Power spectra*

$$\xi(s) \rightarrow P(k)$$

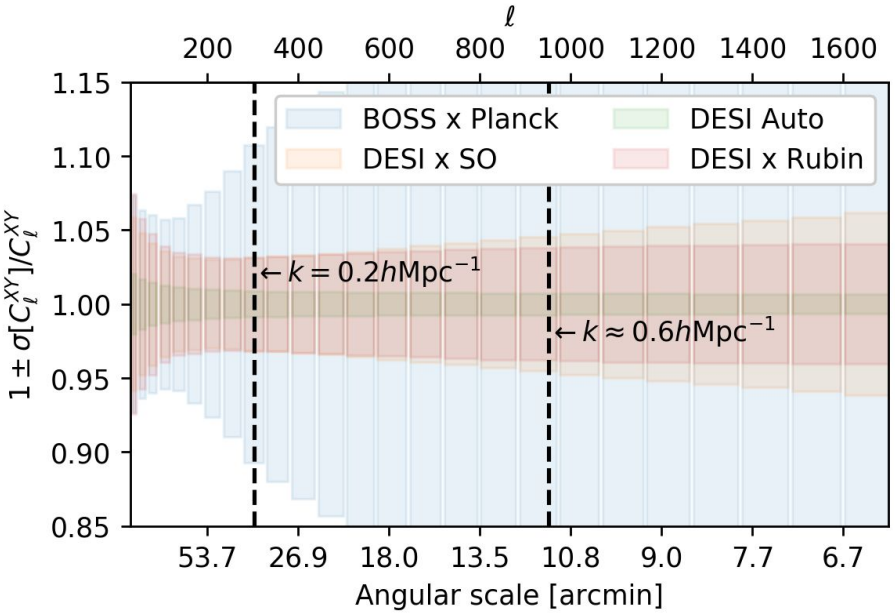


Satpathy+16



Beutler+16

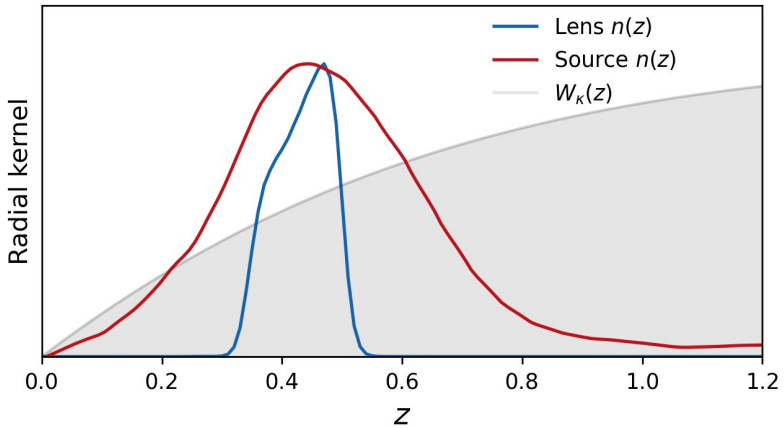
Future power spectrum measurements are **precise**

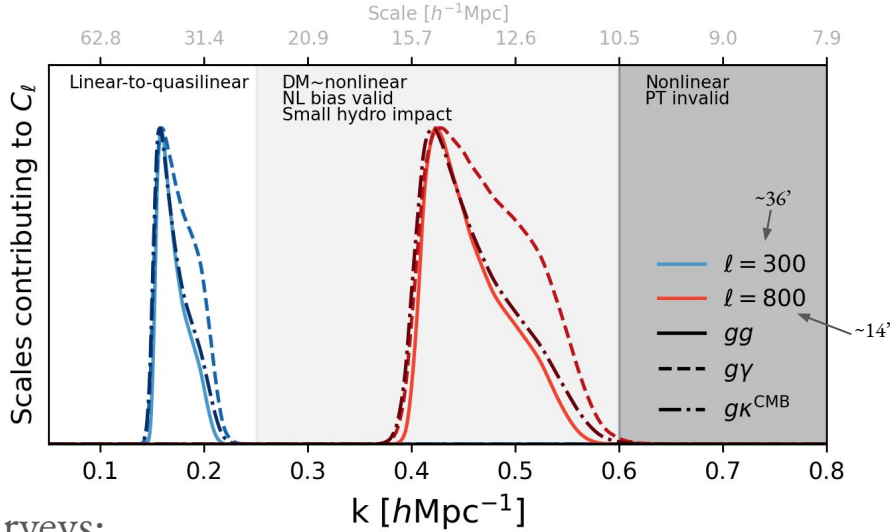


"With great statistical power comes great systematic responsibility"
- Daniel Gruen

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

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





Two challenges must be solved for next generation surveys:

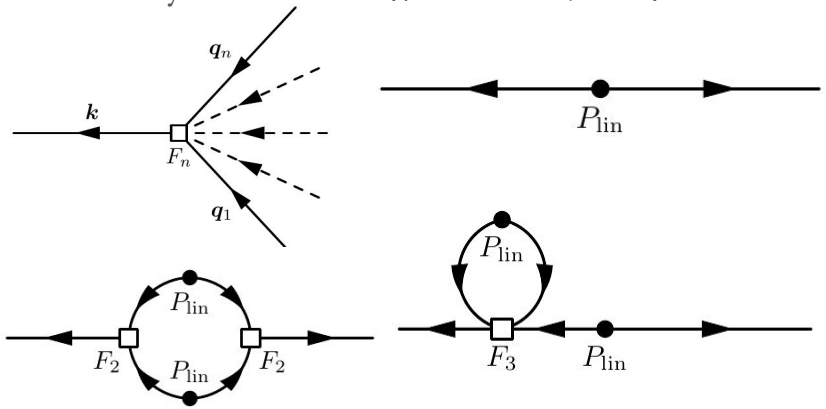
1. Higher volumes/depths/area require more accurate models for the same physical scales.
2. Going to smaller scales promises great statistical reward, if we can conquer the more challenging modelling requirements.

Multiple approaches to modelling structure formation

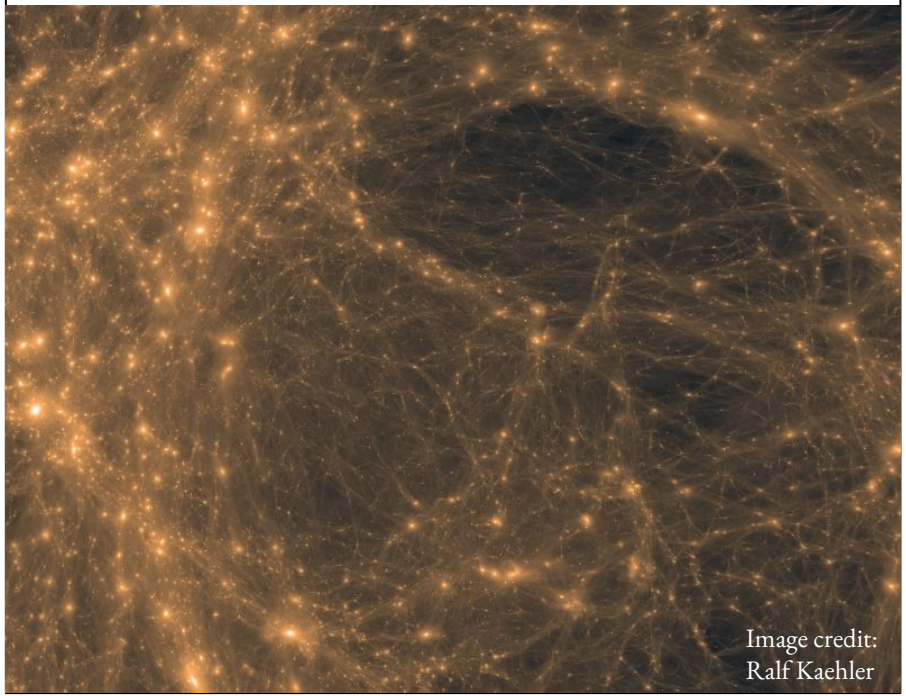
Analytic/perturbation theory

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

And summary statistics are algorithmically computed:

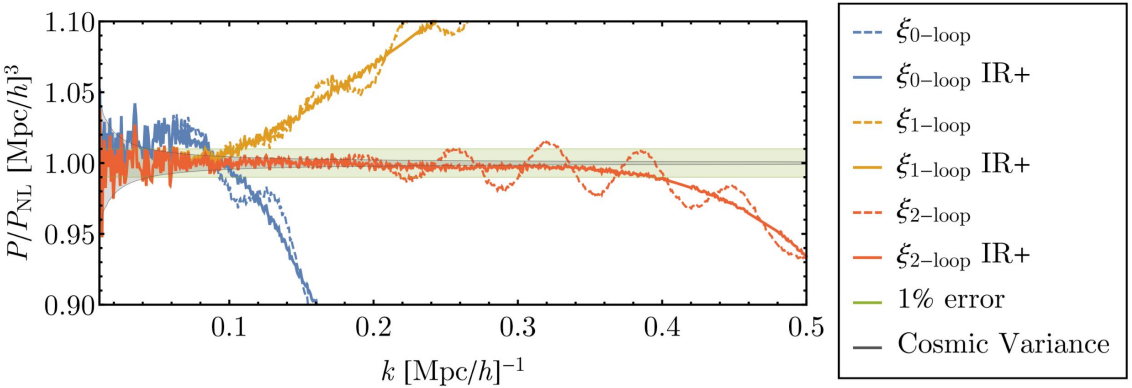


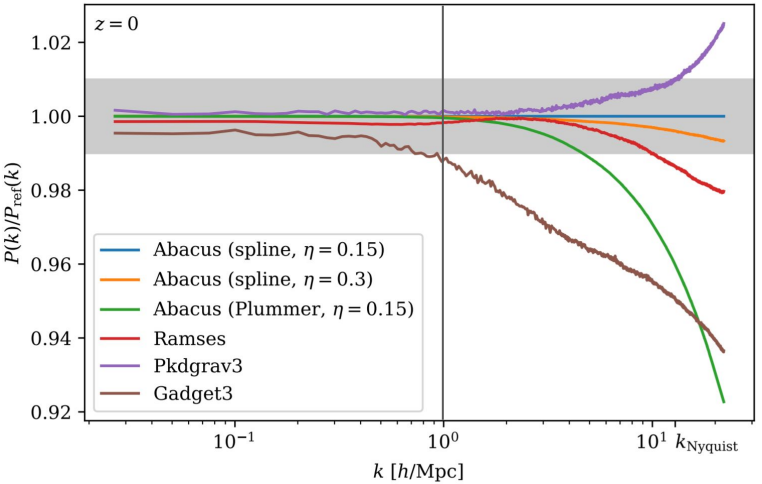
Simulations



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

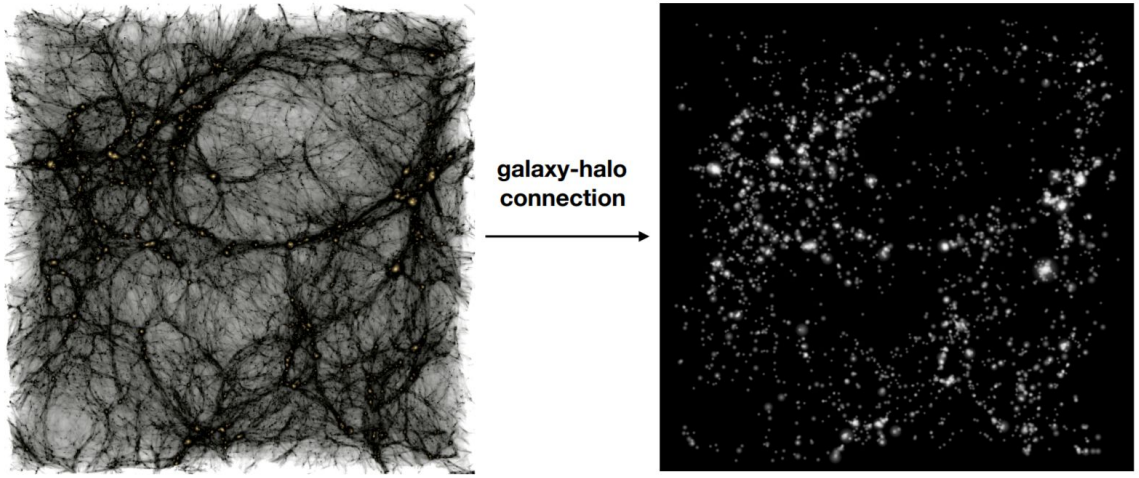




[Senatore & Trevisan 2018]

[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

Bias expansion – mathematically relating the two distributions

Ansatz: the relationship between the galaxy and dark matter distributions depends only on symmetries of galaxy formation

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

Expanding F in a series of all contributions allowed by symmetries lead to an “effective field theory” of biasing.

- Pros: General and rigorous. Fully specified by fundamental symmetries of problem. Easy to include additional physics such as PNG, relative velocity effect, neutrinos.
- Cons**:** Limited regime of applicability, like other perturbation theories

The second-order bias expansion

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

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

$$\approx 1 + b_1 \delta(\boldsymbol{q}) + b_2 \left(\delta^2(\boldsymbol{q}) - \sigma^2\right) + b_{s^2}\left(s^2(\boldsymbol{q}) - \frac{2}{3}\sigma^2\right) + b_{\nabla^2}\nabla^2\delta(\boldsymbol{q}) + \mathcal{O}(\delta^3) + \epsilon(\boldsymbol{q}) + \cdots$$

"Standard" bias contributions

Finite-Size (Effective) Correction

Stochastic contribution

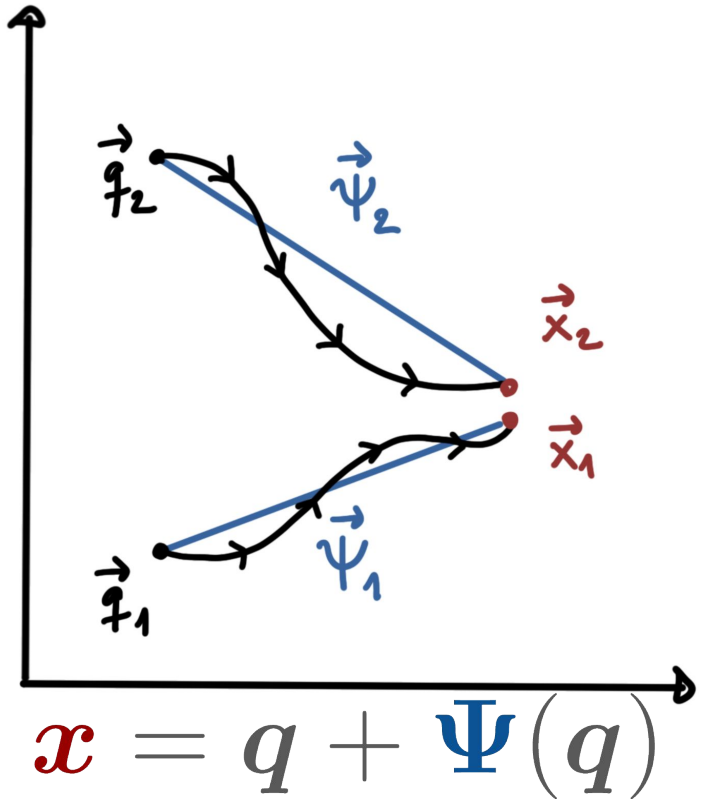
Simulations/analytic approaches describe the *same* physics

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

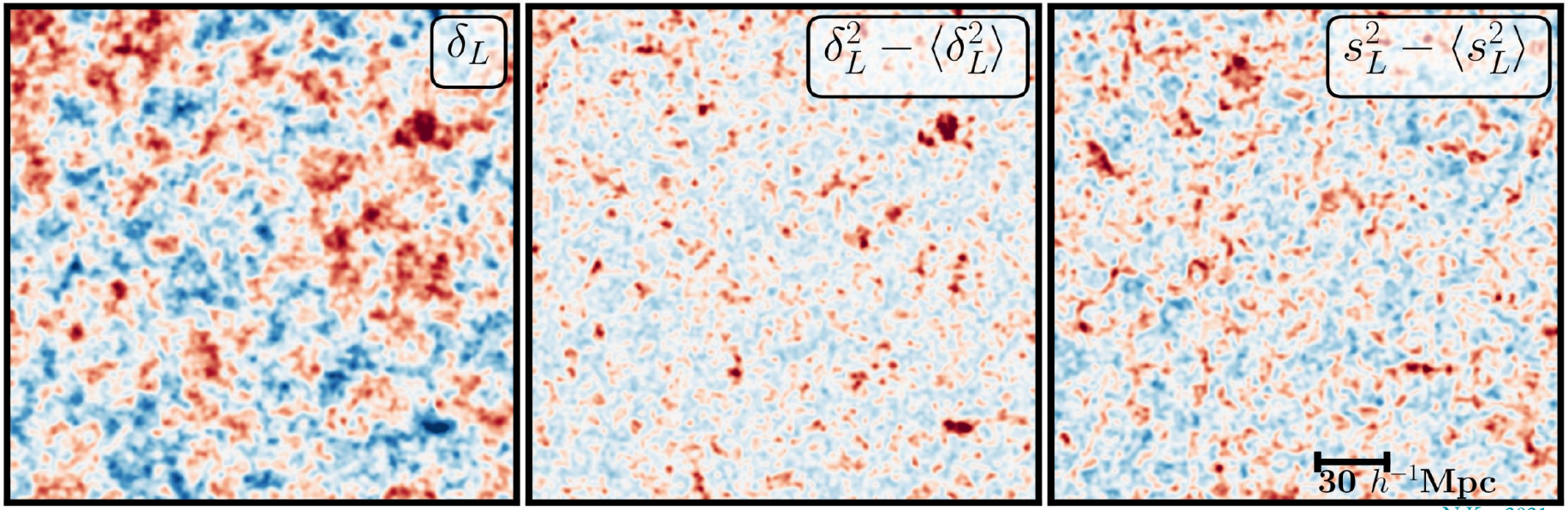
Proposal (Modi, Chen, White 2020):

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).

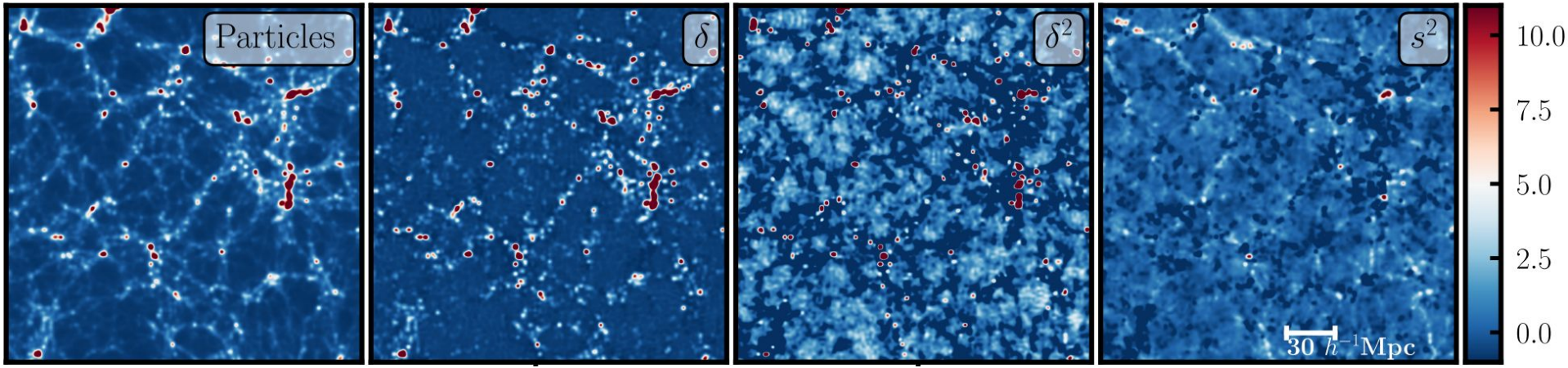


Hybrid EFT: the Lagrangian fields



N.K.+ 2021a

Hybrid EFT: the late-time fields

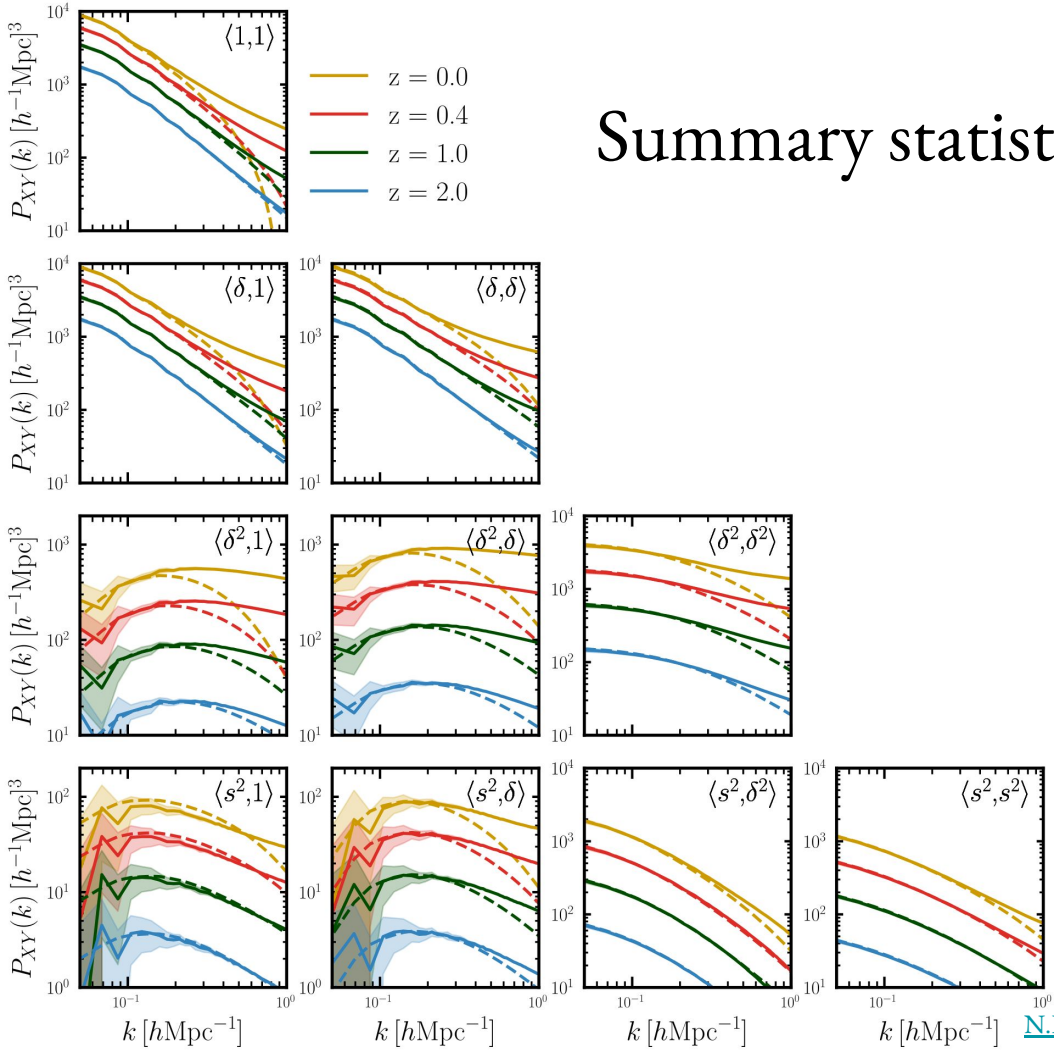


N.K.+ 2021a

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

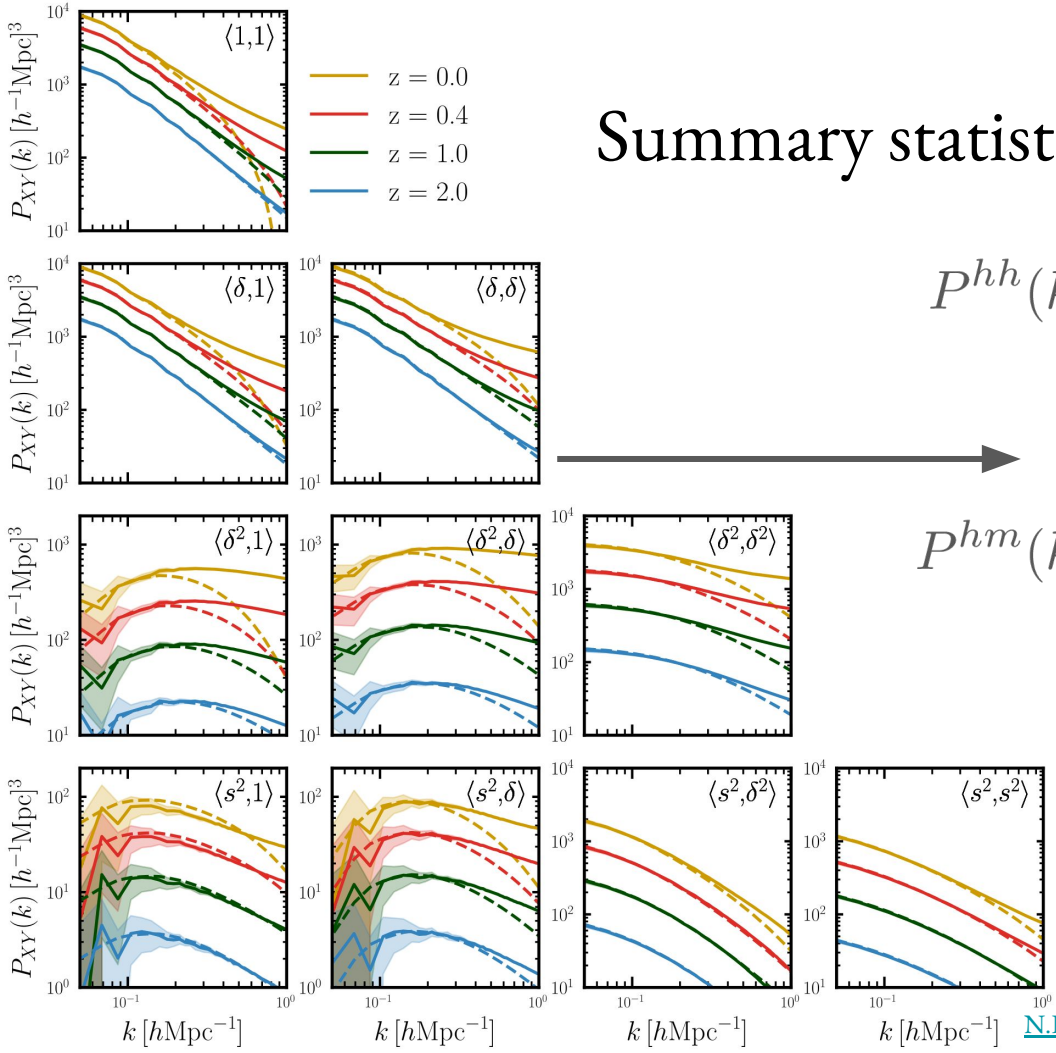
Each panel is showing *the exact same particles*.

Summary statistics consistent with analytics



N.K.+ 2021a

Summary statistics consistent with analytics



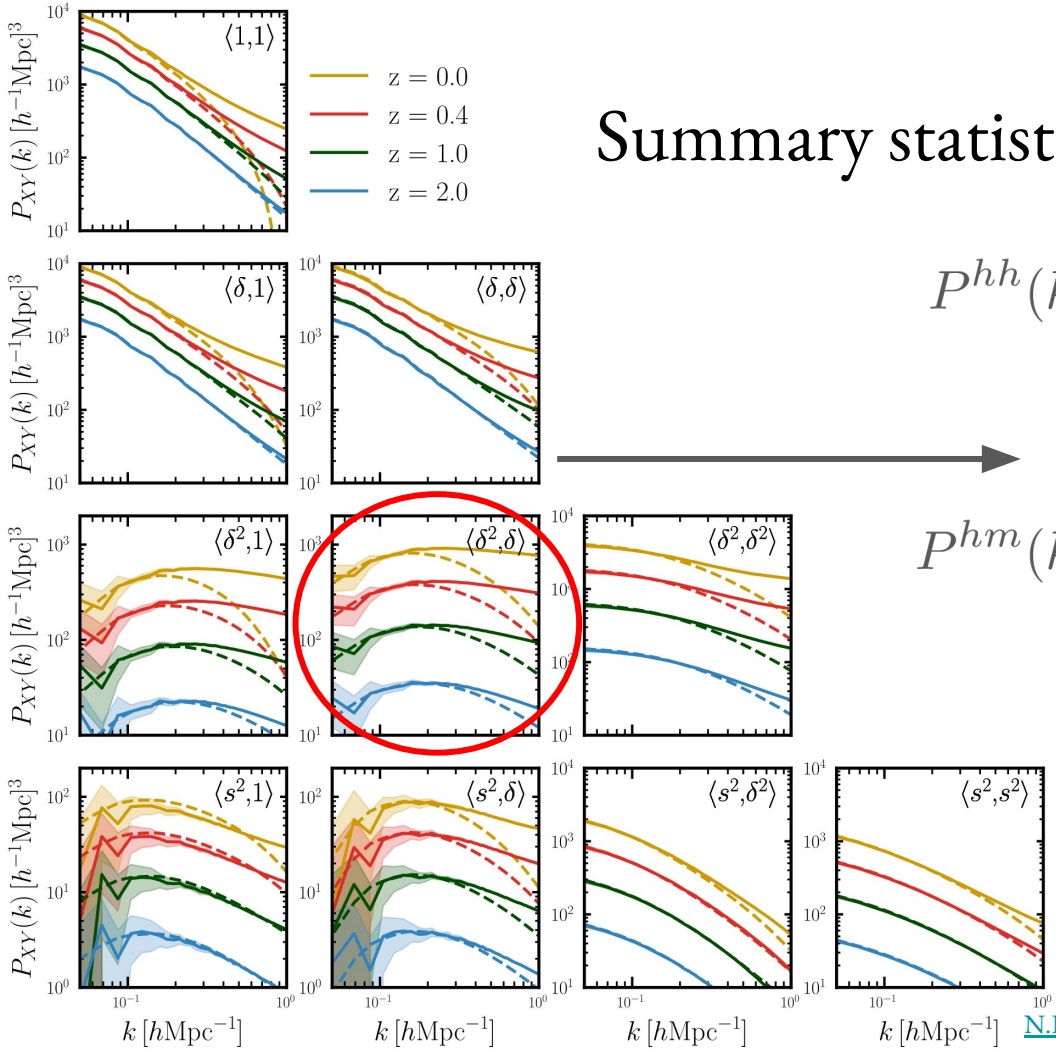
$$P^{hh}(k) = P_{11}(k) + b_1 P_{\delta 1}(k) + b_1^2 P_{\delta\delta}(k) + b_1 b_2 P_{\delta\delta^2}(k) + \cdots$$

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

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

N.K.+ 2021a

Summary statistics consistent with analytics



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

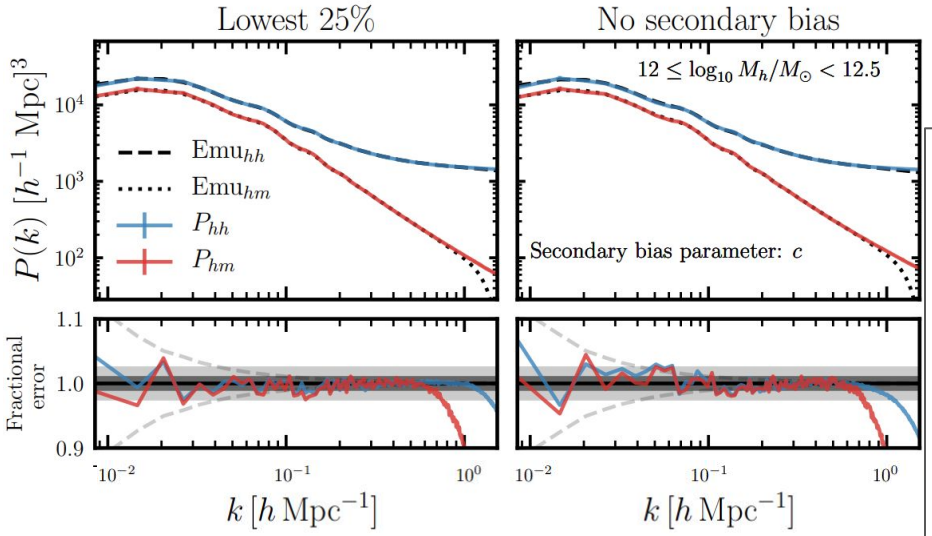
$$P^{hm}(k) = P_{11}(k) + b_1 P_{\delta 1}(k) + b_2 P_{\delta^2 1}(k) + b_{s^2} P_{s^2 1}(k) + b_{\nabla^2} P_{\nabla^2 1}(k)$$

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

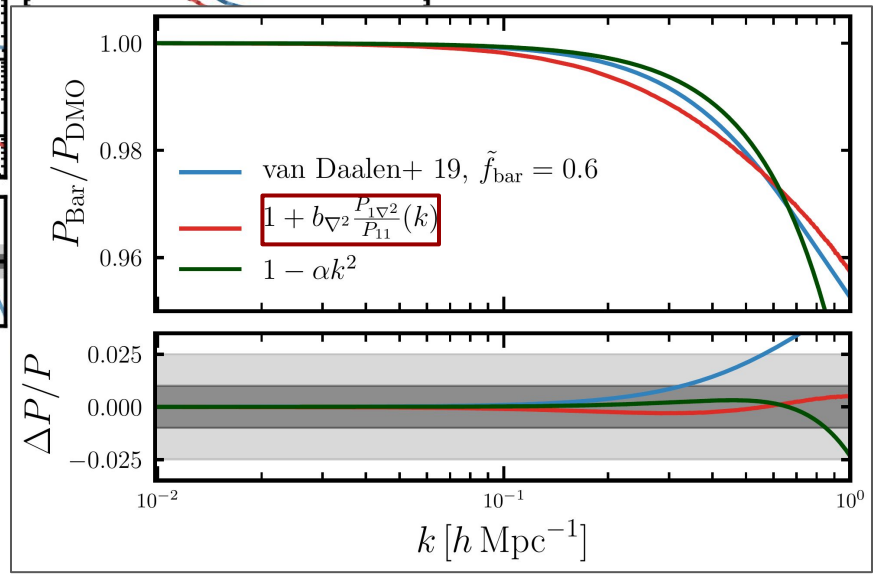
N.K.+ 2021a

The combination is better than the sum of its parts

(describes the statistics of clustering and lensing cross-correlations at 1% accuracy to significantly smaller scales)



Even with (halo) assembly bias!

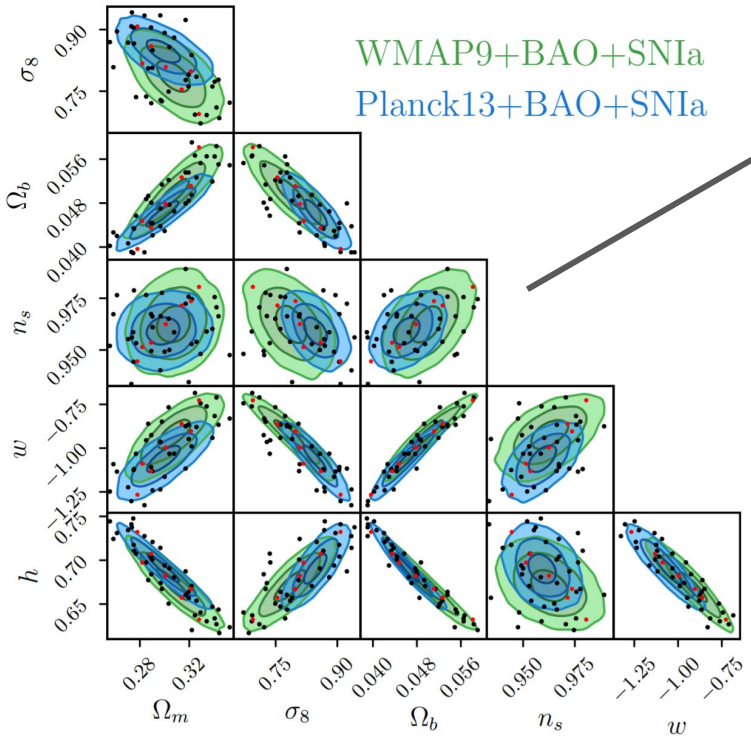


And mild baryons too!

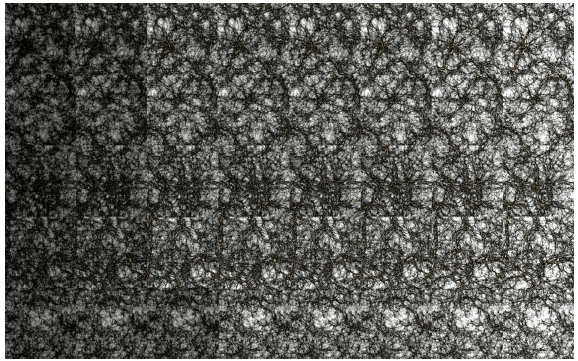
$$\delta_g(\boldsymbol{x}) = \delta_m(\boldsymbol{x}) + b_1\mathcal{O}_\delta(\boldsymbol{x}) + b_2\mathcal{O}_{\delta^2}(\boldsymbol{x}) + b_{s^2}\mathcal{O}_{s^2}(\boldsymbol{x}) + b_{\nabla^2}\mathcal{O}_{\nabla^2}(\boldsymbol{x}) + \epsilon(\boldsymbol{x}) + \cdots$$

N.K.+ 2021a

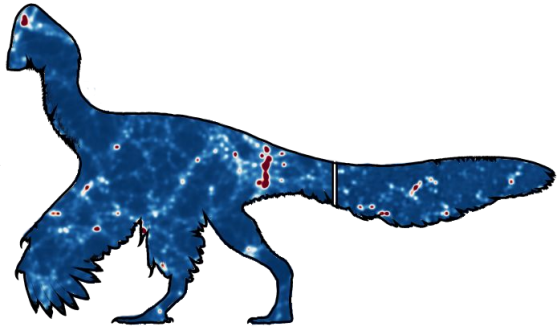
Simulations ‘emulated’ by leveraging modern statistical learning



DeRose et al 2018



Aemulus-α: Supercomputer simulations at 40 cosmologies and 10 redshifts



anzu , a code for the cosmology dependence of hybrid EFT spectra for clustering and lensing modeling

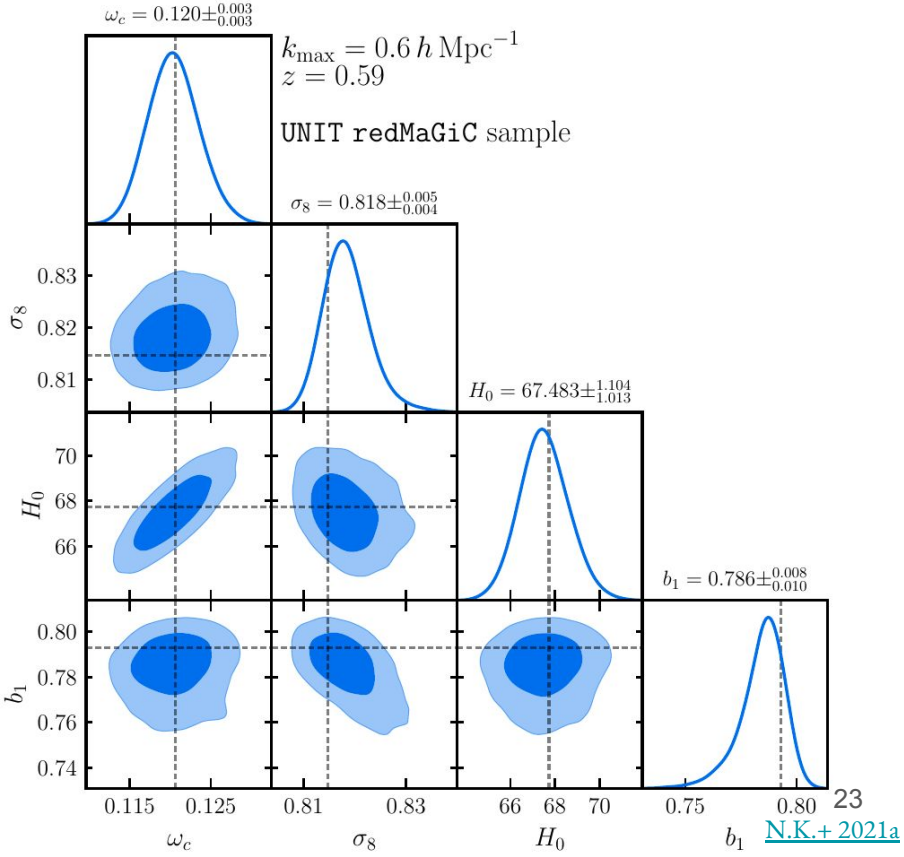
Hybrid EFT can be used for parameter inference

Emulator of 10 Lagrangian basis spectra built using Aemulus, anzu

Recover unbiased cosmology in independent simulation and redshift

Smallest scales are $\mathbf{k_{max}}$**~0.6 h Mpc^{-1}**

Naively, 5x reduction in error bars from $k_{max}=0.2 \rightarrow 0.6$!



N.K.+ 2021a

Research questions spurred by this hybrid approach

1) How do cosmic surveys benefit from the use of these models?

2) How can we understand the galaxy-halo connection with these models?

3) What are novel directions where simulations and perturbation theory can be combined?

Research questions spurred by this hybrid approach

1) **How do cosmic surveys benefit from the use of these models?**

2) How can we understand the galaxy-halo connection with these models?

3) What are novel directions where simulations and perturbation theory can be combined?

Novel survey cross-correlations

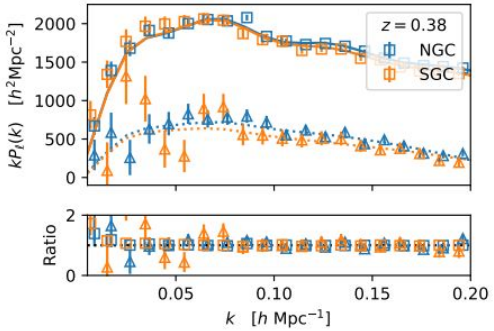
Large-scale BAO and RSD described by a PT model

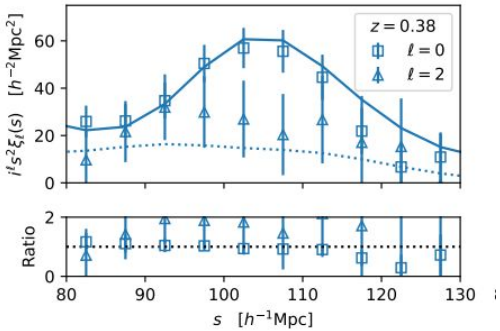
Analyses with 2D surveys (CMB lensing, cosmic shear) -> **Small-scale lever arm** from hybrid EFT

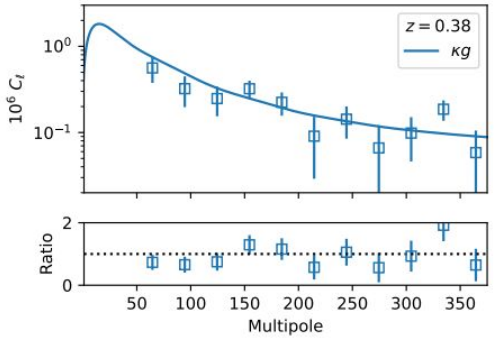
Planck CMB lensing is too noisy to gain from small scales. **ACT and SO will probe these scales**.

Similar gains to be had from “3D + 3x2” analysis of DESI and lensing surveys, where smaller scales **are** probed!

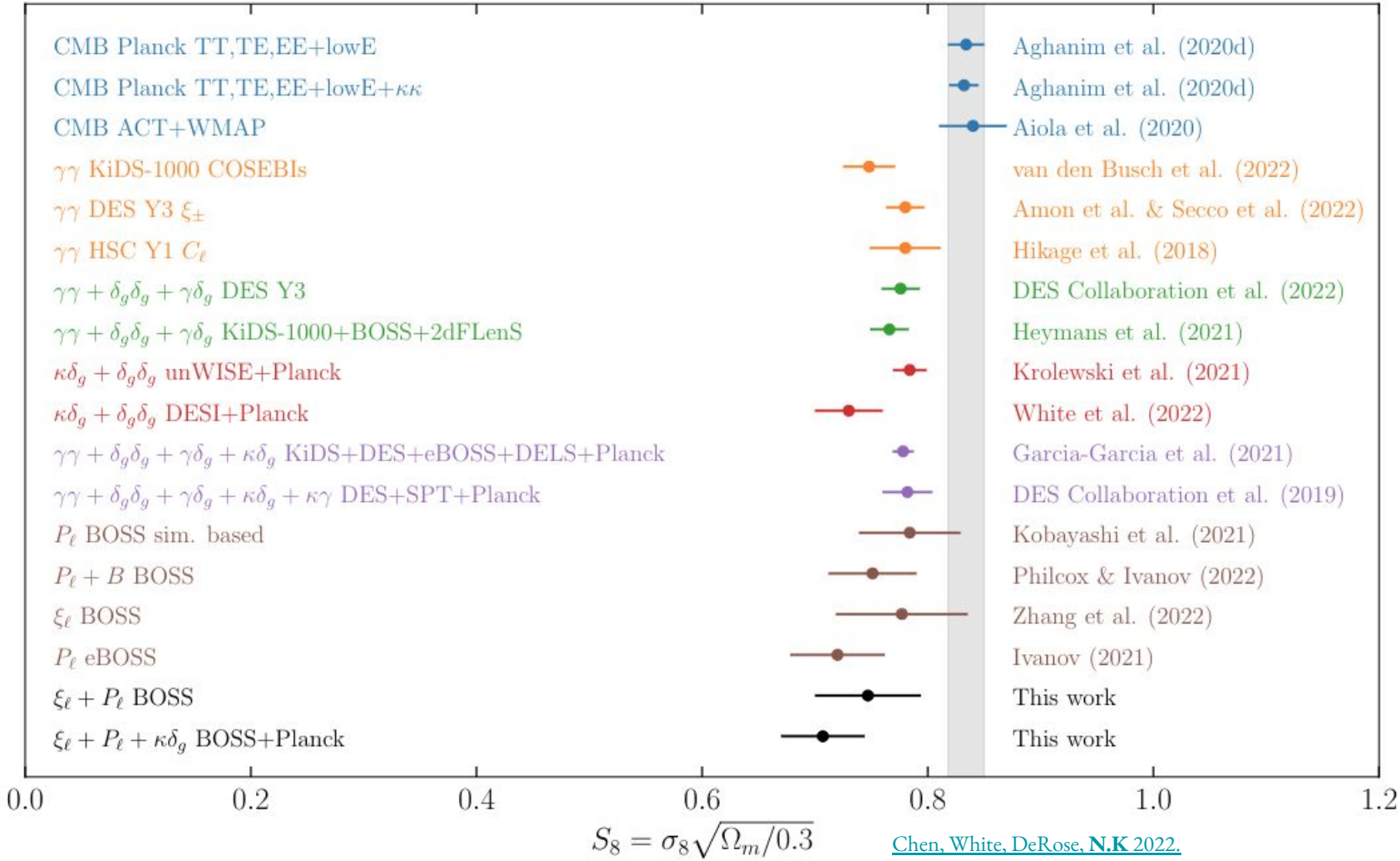
Analyses will weigh in on the “S8 tension”.







Chen, White, DeRose, **N.K** 2022.



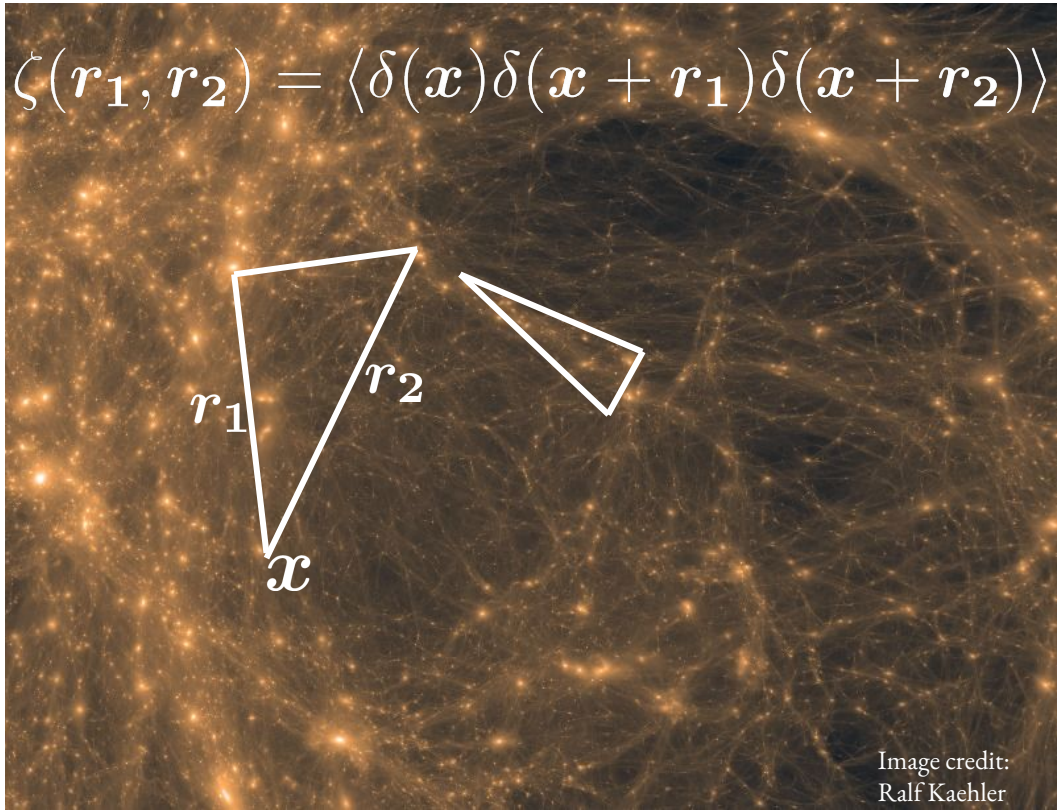
Chen, White, DeRose, **N.K** 2022.

Going beyond the “two-point” paradigm

For non-Gaussian fields (like the distribution of galaxies in the Universe!) the two-point paradigm does not capture all information.

Skewness is captured by the *three*-point function.

Many “beyond two-point” statistics exist.



Counts in cells as a beyond two-point 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*.

"Probability of finding *k* counts in a cell 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)

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

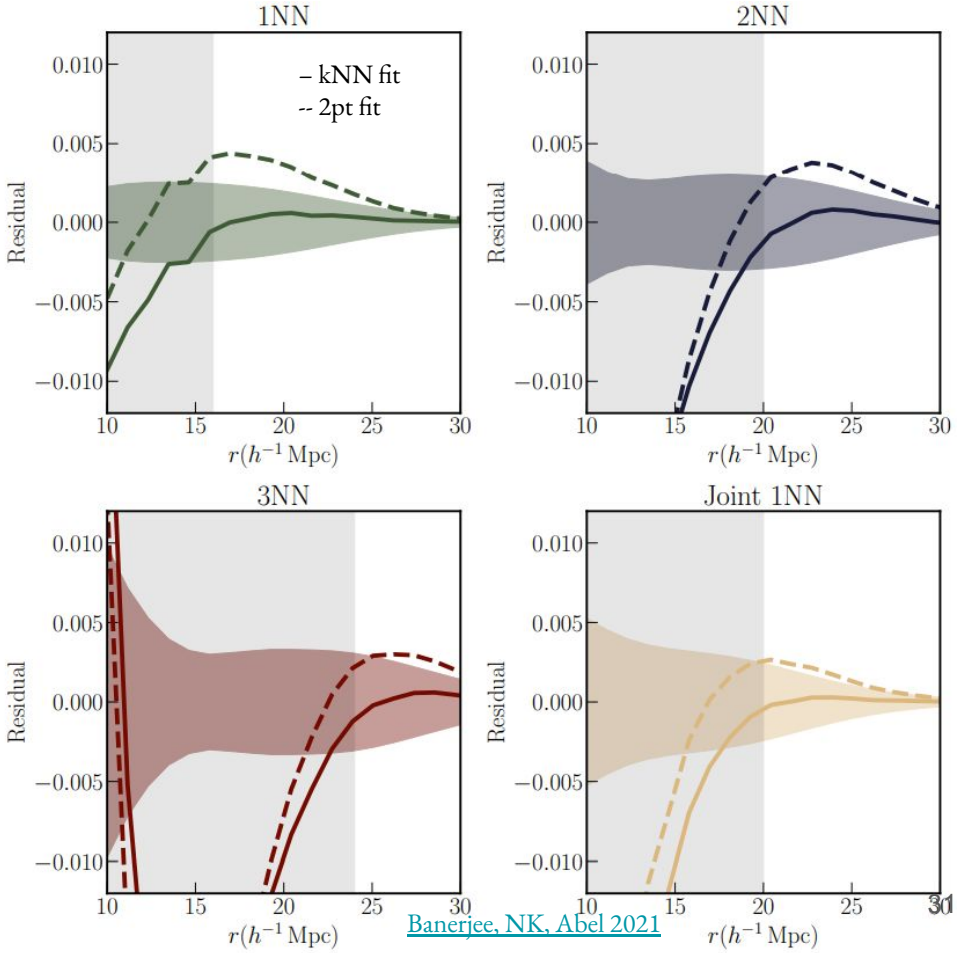
N-point correlation function

Modeling beyond 2-point

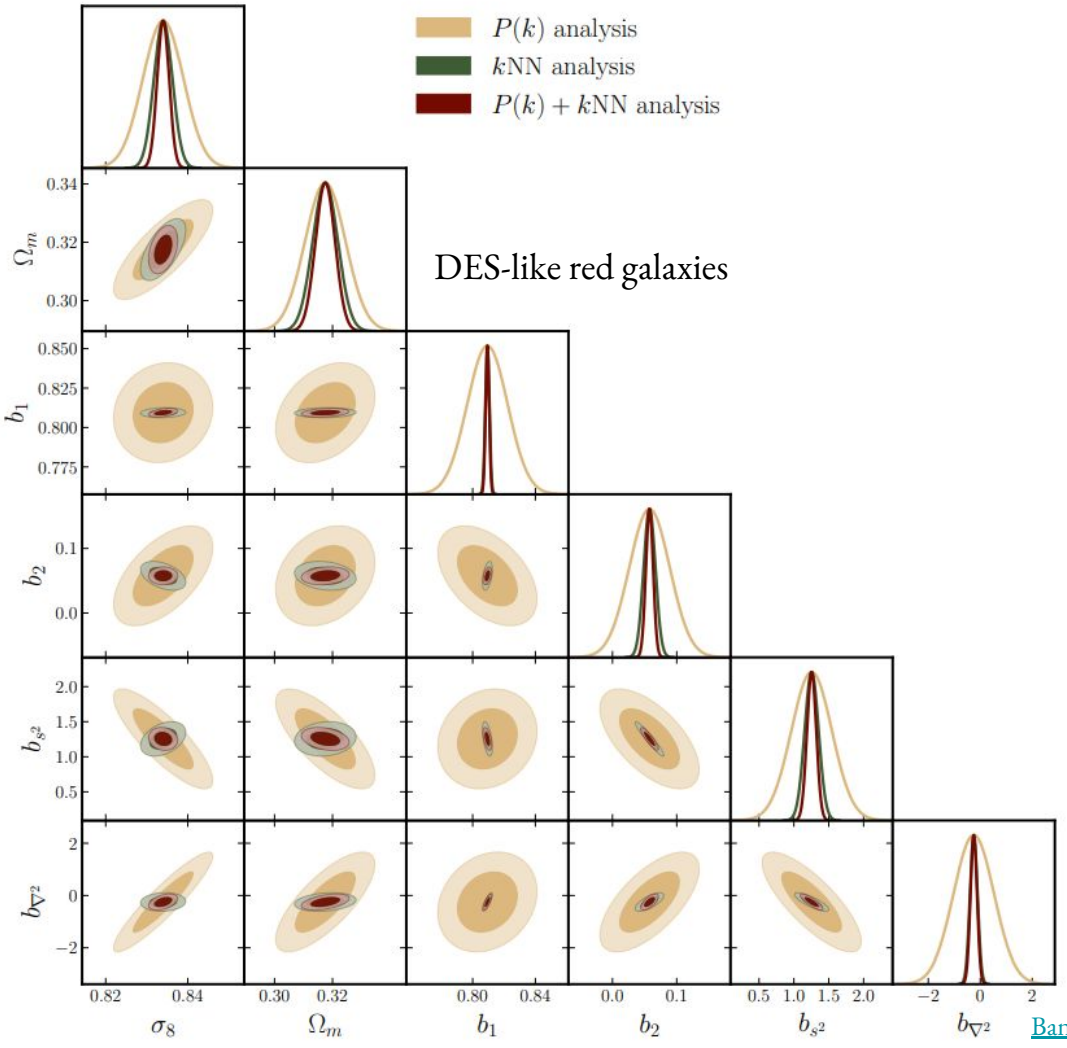
Investigated modeling the “k Nearest Neighbor Cumulative Distribution Functions”, a cousin of Counts in Cells.

Hybrid models self-consistently describe 2-point and kNN statistics!

(Implicitly) describes all N-point functions.



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 can hybrid approaches be used to make models for the **bispectrum** and **trispectrum**?

Banerjee, NK, Abel 2021

Research questions spurred by this hybrid approach

1) How do cosmic surveys benefit from the use of these models?

2) **How can we understand the galaxy-halo connection with these models?**

3) What are novel directions where simulations and perturbation theory can be combined?

Potential challenges with bias models in the future

Extended versions of bias models begin to face challenges (especially in redshift space)

1. At fourth order in bias there are **~~44~~ free parameters** which heavily dilute constraining power (Philcox+22)

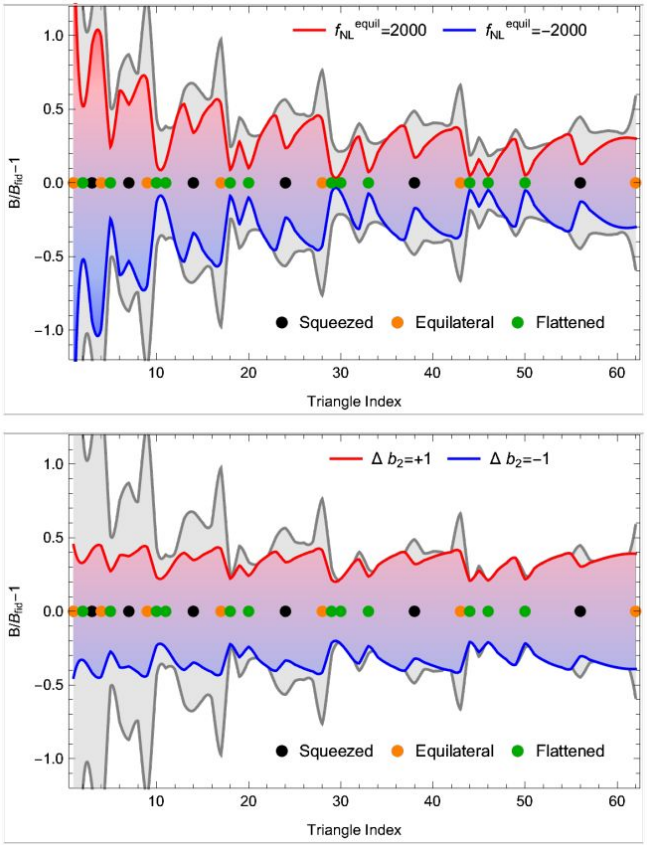
$$\{b_1, b_2, b_{\mathcal{G}_2}, b_3, \gamma_2^\times, \gamma_3, b_{\Gamma_3}, \gamma_{21}^\times, \gamma_{211}, \gamma_{22}, \gamma_{31}\}$$
$$\times\{c_0, c_2, c_4, \tilde{c}, \beta_{B,a}, \beta_{B,b}, \beta_{B,c}, \beta_{B,d}, \beta_{B,e}, C_i[i = 1...9]\}$$
$$\times\{P_{\rm shot}, a_0, a_2, B_{\rm shot}, A_{\rm shot,0}, A_{\rm shot,1}, S_i[i = 0...7]\},$$

(Unless well-motivated priors are imposed)

Potential challenges with bias models in the future

Extended versions of bias models begin to face challenges (especially in redshift space)

1. At fourth order in bias there are **44 free parameters** which which heavily dilute constraining power (Philcox+22)
2. The signature of the quadratic bias b_2 is **degenerate** with non-local primordial non-Gaussianity (Cabass22+)



Potential challenges with bias models in the future

Extended versions of bias models begin to face challenges (especially in redshift space)

1. At fourth order in bias there are **44 free parameters** which which heavily dilute constraining power (Philcox+22)
2. The signature of the quadratic bias b_2 is **degenerate** with non-local primordial non-Gaussianity (Cabass22+)
3. Key assumption about galaxy formation and local PNG has been **shown to be not hold at the accuracy needed for DESI++**

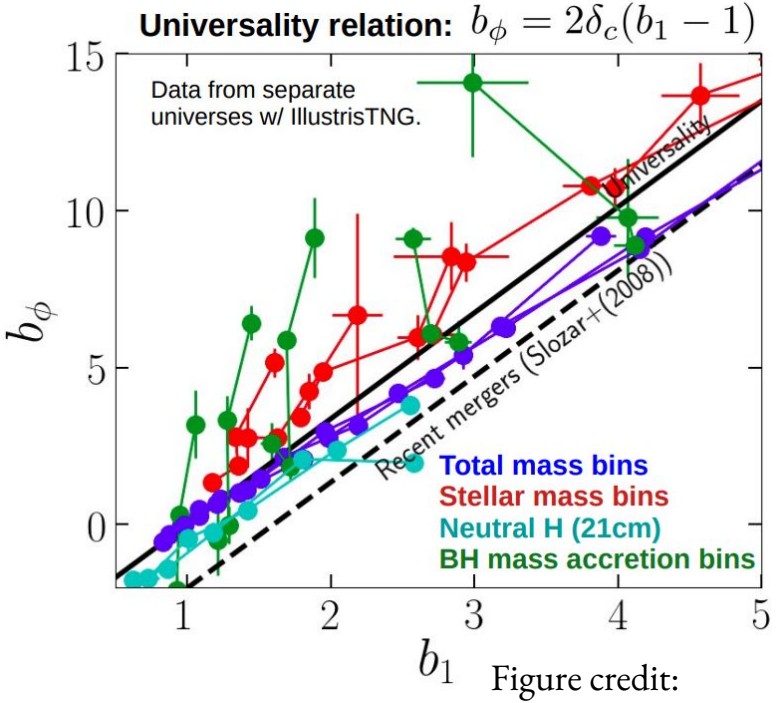


Figure credit: Alexandre Barreira

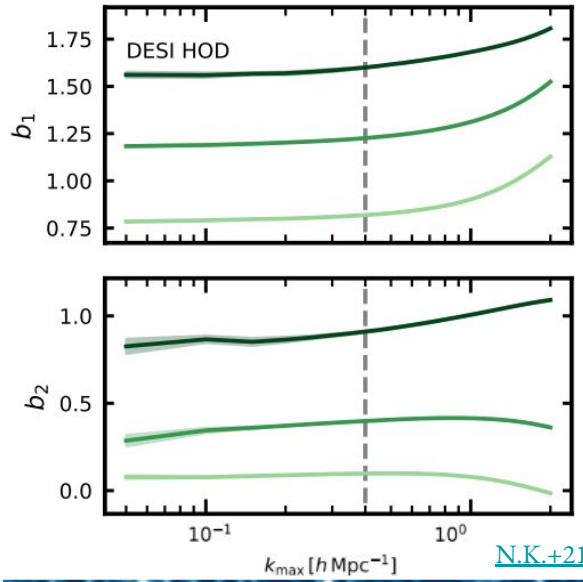
Hybrid tools can sharpen this picture!

We recently derived techniques that can infer these bias parameters in simulated galaxies at high precision.

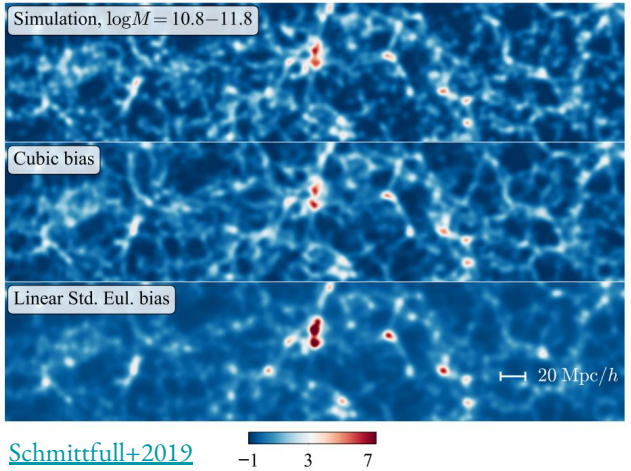
Used these to **precisely measure the biases** of simulated DESI luminous red galaxies.

Realistic simulated populations of galaxies (LRGs, ELGs, LBGs, QSOs, etc) can lead to informed priors for analyses of galaxy surveys.

Directly applicable to probe the response of galaxy formation to primordial non-Gaussianity.



N.K.+21b



Schmittfull+2019

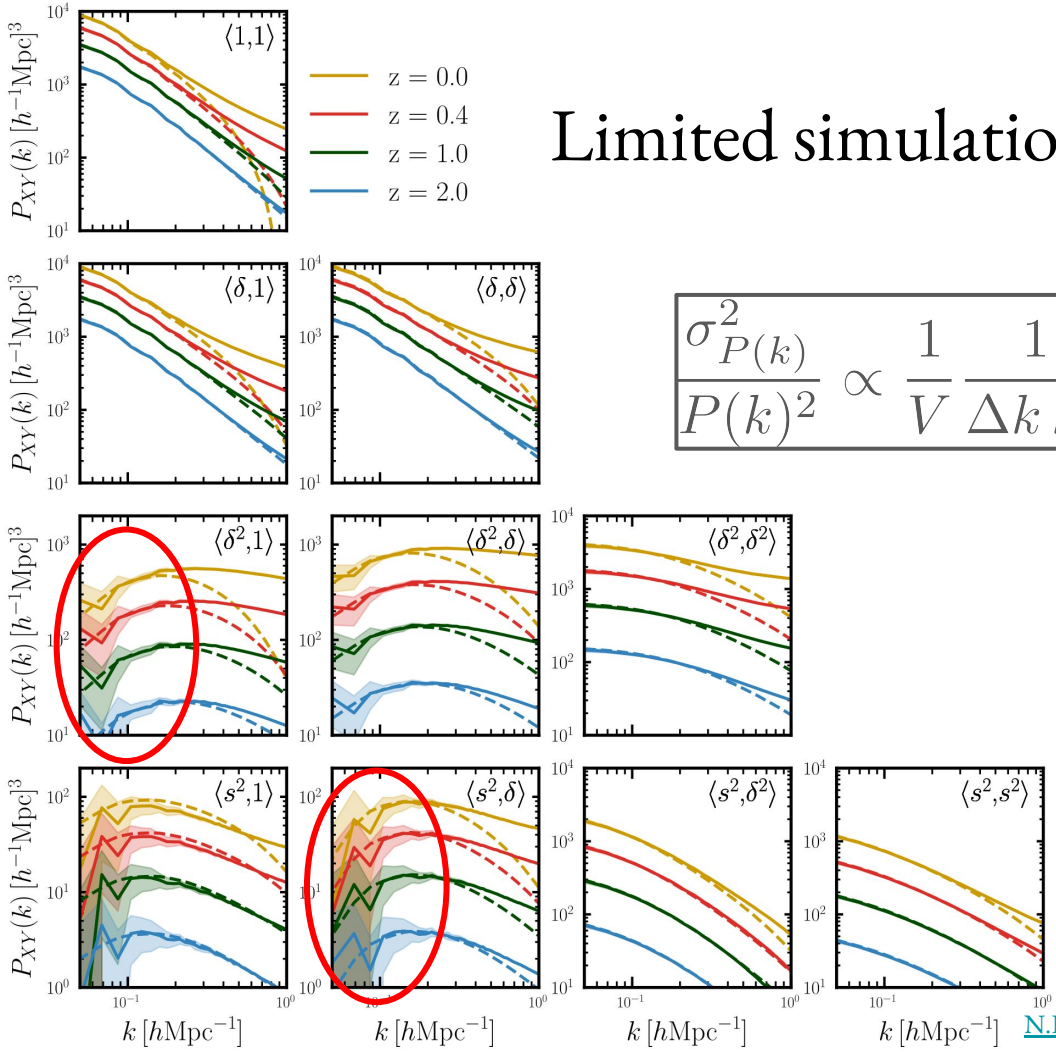
Research questions spurred by this hybrid approach

1) How do cosmic surveys benefit from the use of these models?

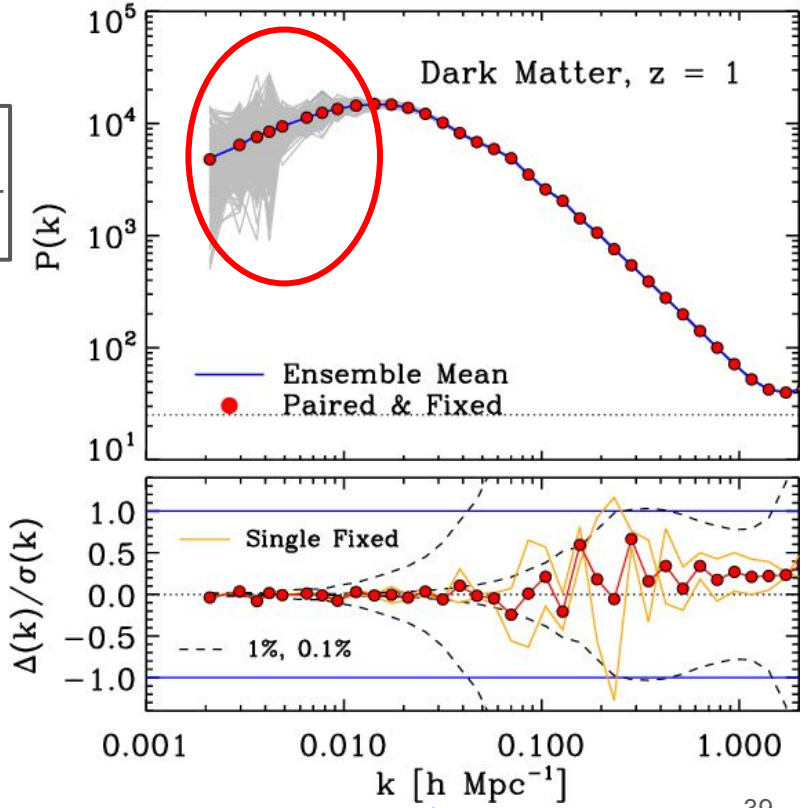
2) How can we understand the galaxy-halo connection with these models?

3) **What are novel directions where simulations and perturbation theory can be combined?**

Limited simulation volume hinders modeling



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



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

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

High correlation -> large reductions in variance!

Control variates in cosmology

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

In cosmology, known as the *CARPool* technique (Chartier et al 20, Chartier & Wandelt 21, Chartier & Wandelt 22).

Surrogates used are statistics from approximate N-body solvers like COLA or FastPM.

Substantial computational requirements:

1. 1500 COLA sims for μ_c
2. 500 pairs of simulations and surrogates to estimate β

The DESI-FastPM (Ding et al 22) project simulated ~500 FastPM mocks with 810GB of data products per simulation. **24 million total CPU-hours!**

Control variates in cosmology

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

In cosmology, known as the *CAP Peel* technique (Chartier et al 20, Chartier & Wandelt 21, Chartier & Wand

Surro

Subst

1.
2.

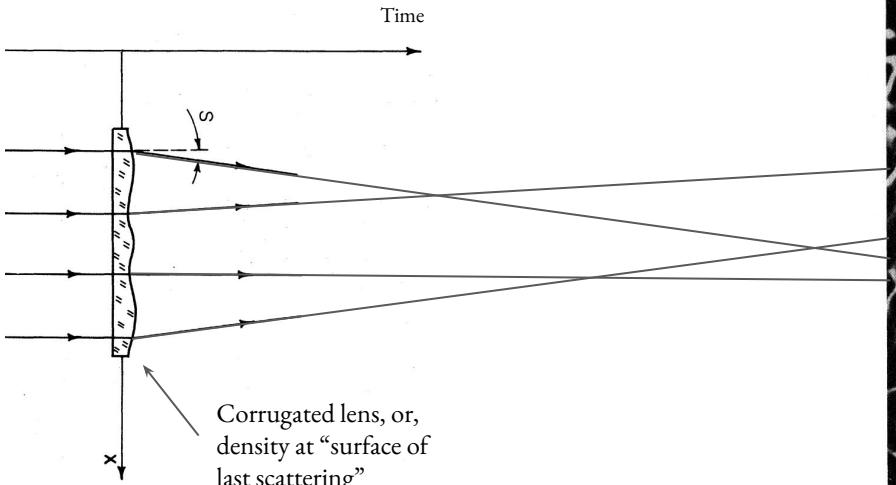
Wanted: a control variate that is analytically tractable, but also correlates strongly with the nonlinear distribution of dark matter.

The DESI-FastPM (Ding et al 22) project simulated ~500 FastPM mocks with 810GB of data products per simulation. **24 million total CPU-hours!**

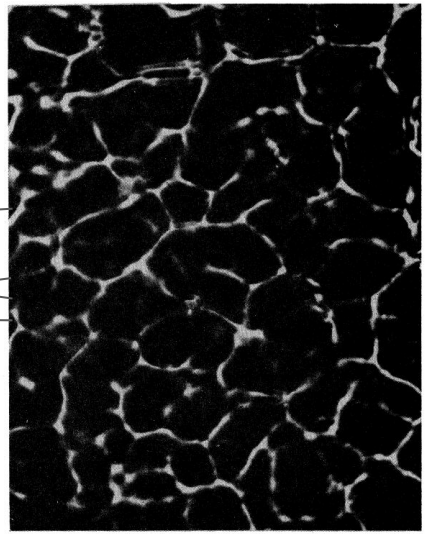
The Zel'dovich approximation for structure formation

Conceptually: a homogeneous distribution of particles is scattered by the initial distribution of matter, and then travels in straight lines. Strong analogies with geometric optics (Zel'dovich & Shandarin 1989)

While extremely simple, this approximation predicts the formation of cosmic structures such as halos, pancakes and filaments of the cosmic web!

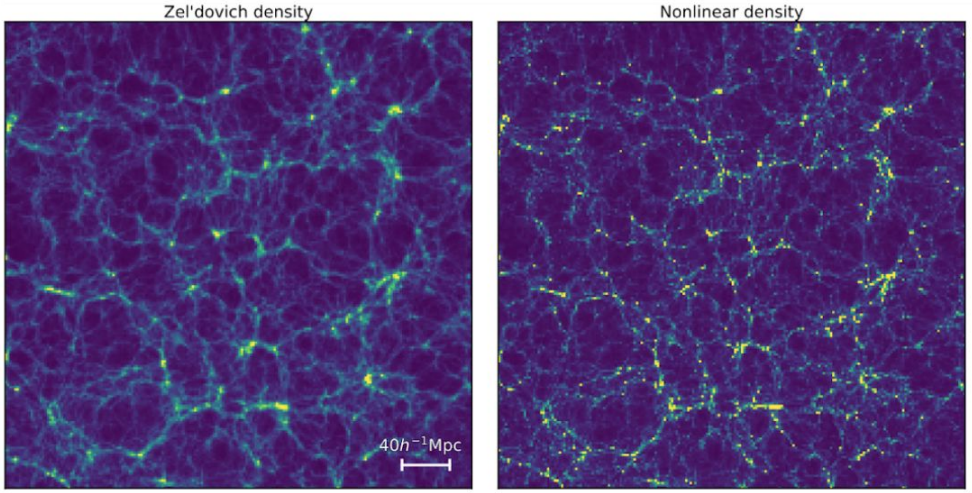


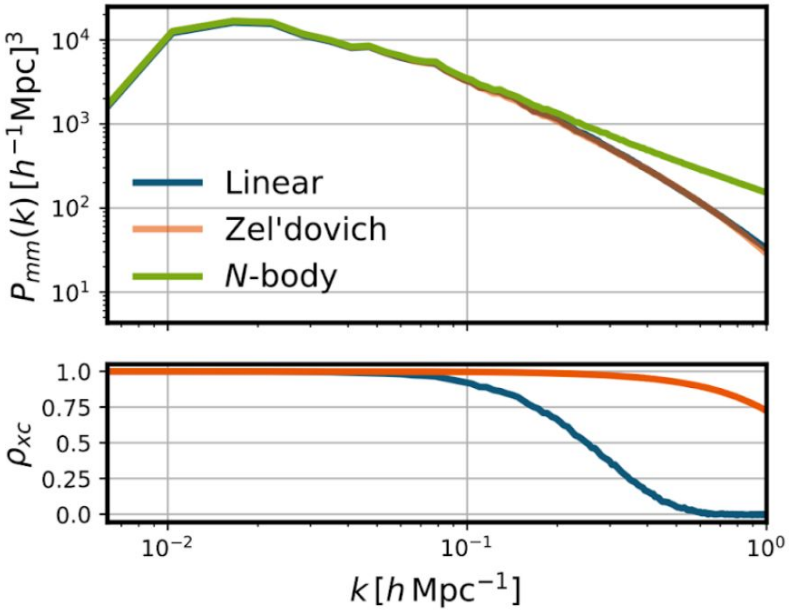
Corrugated lens, or, density at "surface of last scattering"



Projection of light rays on a screen, or, the cosmic web

Zel'dovich fields correlate highly with nonlinear simulations





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***

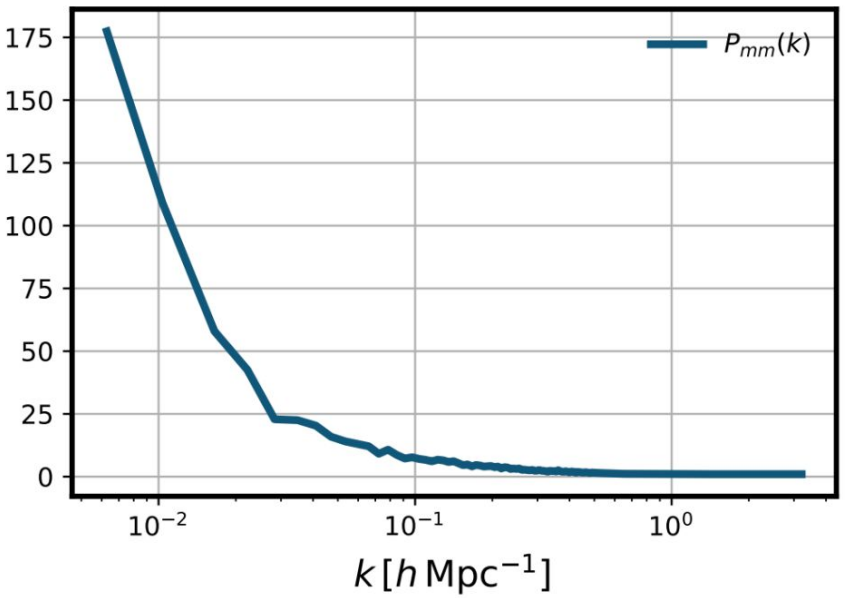
N.K.+22

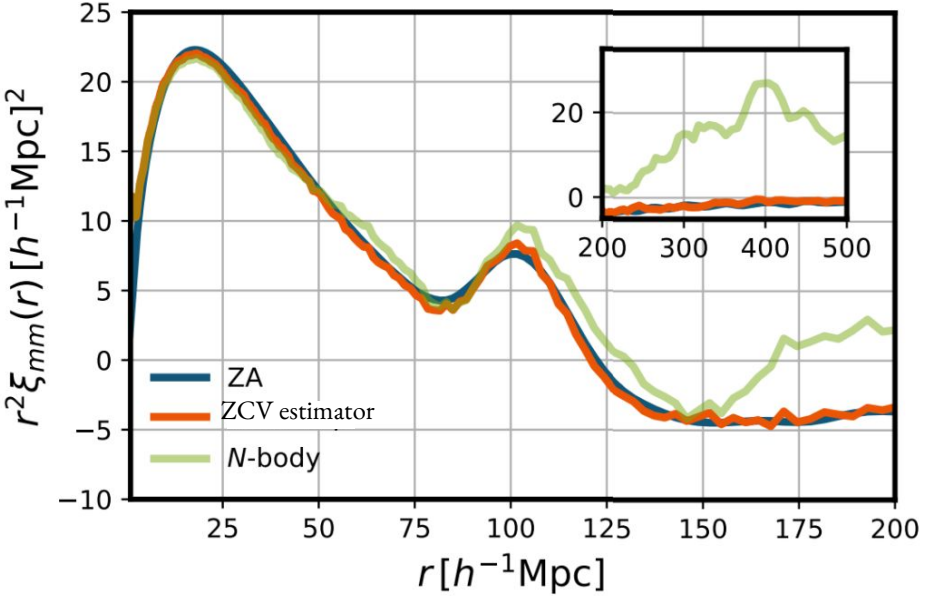
How powerful of a control variate is the Zel'dovich approximation?

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

Variance reduction for matter 2-pt statistics

Effective signal-to-noise increase



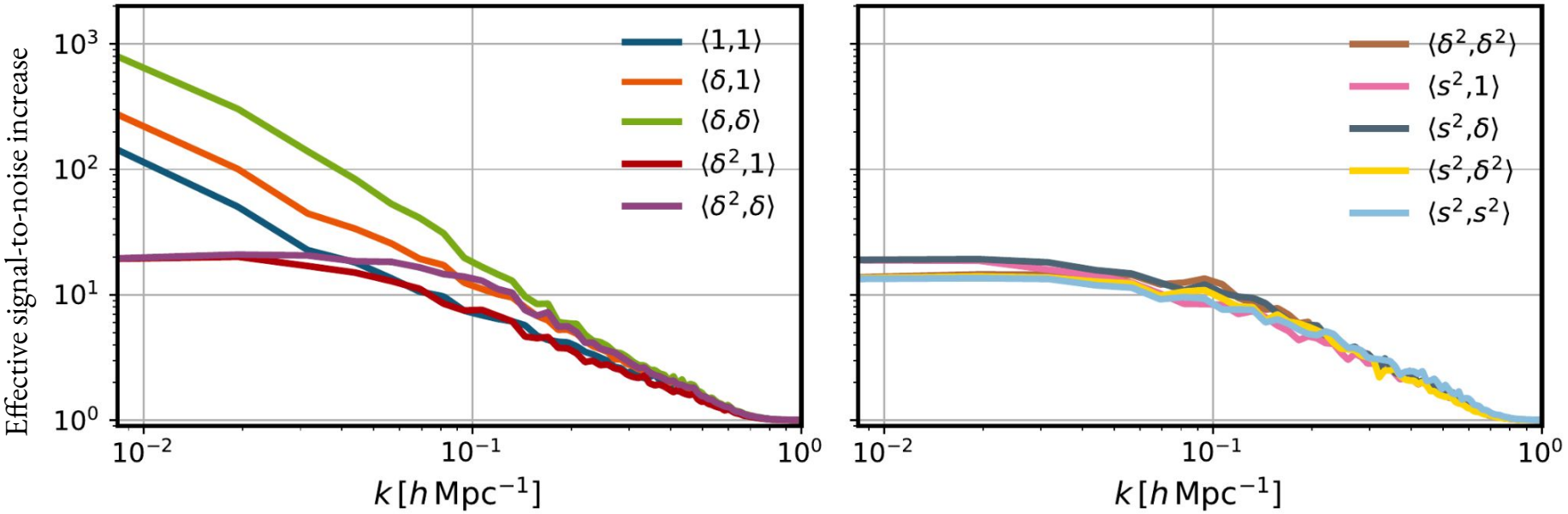


N.K.+22

General variance reduction for biased tracers

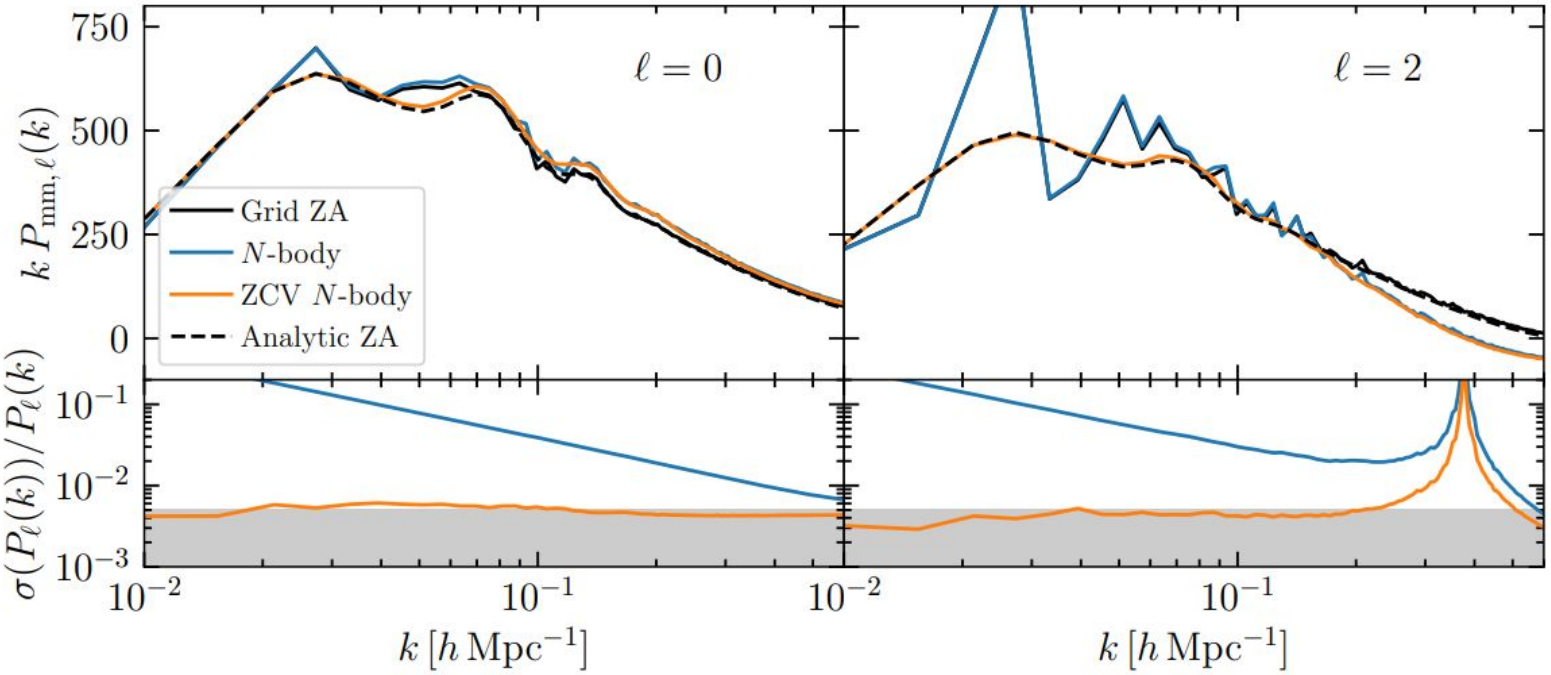
Similar reduction in sample variance for all hybrid bias spectra

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



N.K.+22

Also works in redshift space!



DeRose, Chen, **N.K.**, White 2022

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

Emulators can be designed with significantly smaller boxes, **more efficient exploration of non-linear beyond LCDM phenomena**

Accurate mock surveys with smaller volumes, **mock catalogs for survey validation**

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

Can be extended to other summary statistics such as the bispectrum and the covariance matrix of the power spectrum.

Conclusions

Modern cosmological inference is challenging on many fronts.

The meeting of pencil-and-paper and supercomputing is a powerful lens through which to think about the formation of large-scale structures.

In the future, these techniques will enable:

1. State of the art models for galaxy clustering and cross-correlations with lensing surveys
2. Accurate models for analyses beyond the two-point paradigm
3. A precision characterization of the connection between DESI galaxies and their dark matter halos, including under the presence of primordial non-gaussianity
4. Significant reduction in the variance of simulation-based models

Thank you for having me!