Modeling large-scale structure in the golden age of cosmological surveys

Martin White

(UC Berkeley, LBNL)

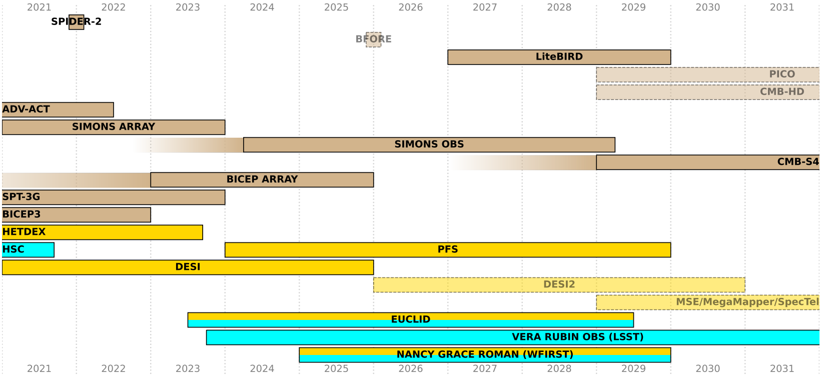
Higgs Colloquium, 17 Jun 2022

Outline

- The golden age of cosmological surveys
  - We are gathering vast amounts of data that can inform us about many interesting questions in physics and astrophysics ... given suitable models!
- Modeling the evolution of large-scale structure
- Future directions for the field
- Concluding thoughts

Golden age of surveys

We are living in the "golden age" of cosmological surveys, with survey capabilities increasing exponentially ... (Moore's law)

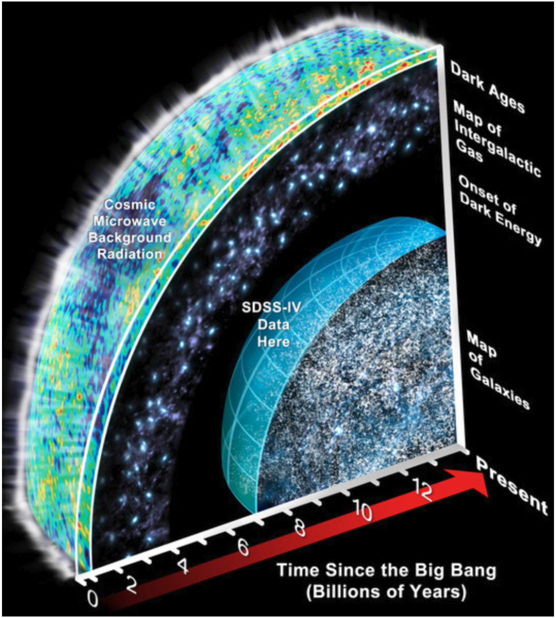


D. Kirkby

Large-scale structure (LSS)

When we view the Universe today, we find structure on scales from the cosmological horizon to planetary systems.

Courtesy: SDSS/eBOSS



The story we tell ...

- Something like inflation turned quantum fluctuations (in the "inflaton" field) into classical perturbations in the density of all species at early times.
- Fluctuations grow over time through a process of gravitational instability to form all of the structure we see today.
- 14 Gyr of evolution shapes the spectrum.
  - We see these fluctuations at $z \simeq 10^3$ ($t \sim 400,000\,\mathrm{yr}$) as fluctuations in the temperature of the cosmic microwave background (CMB) radiation with $\Delta T/T \sim 10^{-4}$.
  - and "today" with galaxy surveys with $\delta \gg 1$.
  - Probing a wide range of energy densities, temperatures, ...
- Growth is a competition between gravity and expansion
  - Depends on the laws of gravity (general relativity).
  - Depends upon the expansion of the Universe (metric).
  - Depends on constituents and their properties.

Probe metric, particle content and **both** epochs of accelerated expansion ... with high precision

The standard model

- We have a “standard model”, based on General Relativity, inflation, dark matter (DM) and dark energy (DE).
- The model is stunningly successful, but completely phenomenological.
- We don’t have a 1^{st} principles understanding of much of the model.
    - ... or even a 2^{nd} or a 3^{rd} ...
- Need to test each piece to see what are only approximations, or perhaps what’s “wrong” (test GR, inflation, DM and DE).

In the absence of a clear signal of new physics currently ... I will consider high-precision tests of the SM with a focus on large-scale structure (LSS; where some “tensions” have arisen)

Models of large-scale structure (LSS)

How do people model measurements of large-scale structure?

- There are two broad classes of approaches to modeling LSS: numerical and analytical.
- Numerical approaches (simulations)
  - Techniques for solving systems of pure dark matter are well developed; though the combination of volume and resolution required by next-gen surveys is very demanding.
  - The best way to deal with the complexities of galaxy formation, hydrodynamics and multiple species is still an open research problem.
- I will discuss analytic approaches based on perturbation theory (PT) – which have seen a renaissance in recent years.
- Most practitioners use some combination ...
  - All N-body codes use PT for initial conditions.
  - N-body can be used to test PT for fiducial models.
  - New ideas for combining the two: "best of both worlds".

Perturbation theory (PT) history

- Cosmology deals with relativistic gauge field theories, like many other sub-fields of physics.
- The equations of motion are both non-linear and non-local.
- PT developed starting in the 1960's, reached its "classical" form in the early 1990's.
  - At this point surveys could only probe well k-modes for which perturbation theory was barely applicable at all.
  - And in 1992 anisotropies in the CMB were discovered (queue Nobel Prize) for which linear theory is all that is needed.
  - The 1990's also saw the introduction of 'commodity' supercomputing and a huge advance in our ability to simulate large-scale structure.
- While it never fully died out, use of PT only really blossomed again within the last 5-10 years ...

A funny thing happened ...

- Cosmology is riding the Moore's law/big data revolution like many other fields.
- Even though computing/simulation is becoming a bigger component of the analysis toolkit, modern surveys are empowering theorists as never before ...
- We have the technology to survey very large volumes at larger distance (i.e. earlier times).
  - Large scales have undergone less processing, more correlated with ICs.
  - Fluctuations are linear, or quasi-linear ($\delta \lesssim 1$).
  - Such modes are under good "theoretical control" using PT.
  - We're now computing small corrections to "almost linear" quantities.
  - Bigger surveys demand higher precision.

Challenge I: dynamics

- Standard perturbative techniques familiar from QM, condensed matter or particle theory.
  - Relativistic gauge field theory: Green's functions, diagrams, tree level, loops, ...
  - Some technical subtleties due to peculiarities of our specific problem.
  - Long-range order (like stat.mech. – can't shield gravity!)
- For purely Newtonian gravity the "UV completion" of our theory is known – but not calculable (non-linear, chaotic, ...).
- Once hydrodynamics, star formation, etc. are included the problem is even more intractable analytically.
- We work with effective field theories
  - Language will be familiar from condensed matter or particle theory.
  - Regularization, renormalization, running, counterterms, ...
- Long wavelength modes are numerically very important.
  - Need for IR resummation.

Challenge II: bias

- Different "tracers" of large-scale structure are related to the underlying perturbations in density and potential differently.
- The connection between how the tracer clusters and how the matter clusters is known as <u>bias</u>, and dealing with bias is one of the big challenges in modeling large-scale structure.
  - For example, more luminous galaxies tend to be more clustered than less luminous galaxies, even though both trace the same underlying density field.

Bias, peaks and EFT

- Write $\delta_{\rm gal}$ as a functional of the initial (long wavelength) density, velocity and potential fields: $\delta_{\rm gal}[\delta, \partial \mathbf{v}, \partial\partial\Phi, \cdots]$
- Coefficients of an expansion in e.g. δ are bias coefficients.

$$\delta_{\rm gal}(\mathbf{x}) = b_1\delta(\mathbf{x}) + b_2\delta^2(\mathbf{x}) + \cdots + \mathrm{stochastic} + \cdots$$

- Bias coefficients incorporate our uncertainty about complicated galaxy formation physics in addition to UV effects.
    - Dark matter halo formation, merger history, ...
    - Chemistry and gas cooling.
    - Star formation, SNe, AGN
    - Thermal and kinetic feedback
    - Background radiation

Bias, peaks and EFT

- While the process that form and shape galaxies and other objects are complex, all such objects arise from simple initial conditions acted upon by physical laws which obey well-known symmetries.
- For non-relativistic tracers these are
  - the equivalence principle
  - translational, rotational and
  - Galilean invariance.
- This highly restricts the kinds of terms that can arise in a bias expansion, no matter how complex the history.

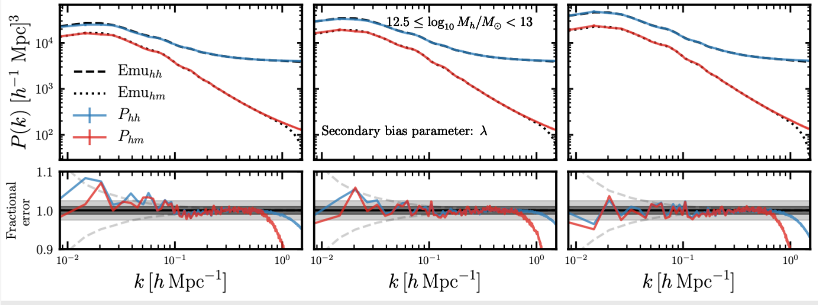
Symmetry arguments are extremely powerful for bias since we really don't understand the small-scale physics of bias.

Aside: Simulations and Symmetries

- We can simulate structure formation in a DM-only Universe pretty well.
- It's the baryonic component that is "hard"!
  - Don't understand cooling, star-formation, feedback, ...
  - Resort to parameterized models (when to stop adding parameters, how to test for numerical convergence?)
- Symmetries-based thinking is ubiquitous in PT studies and very powerful.
- PT folks and simulators are trying to solve the same problems ...
- Can we have the best of both worlds?
  - Use dynamics from N-body simulations, but the "galaxies" (symmetries-based bias technique) from perturbation theory [Modi+20, Kokron+21, Hadzhiyska+21, Zennaro+21,...].

Can push into the non-linear regime

Can fit mock catalog data for "3 × 2pt analyses" to 1-2% even for samples with assembly bias and other complex selections and even including hydrodynamics.



Now we can simply "emulate" the basis spectra using standard techniques (no need to emulate the bias parameters – analytic)!

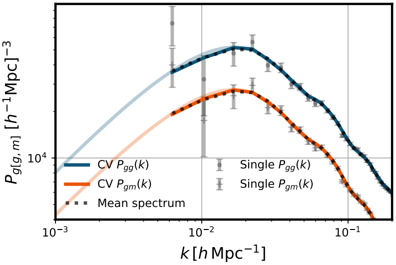
Kokron+21

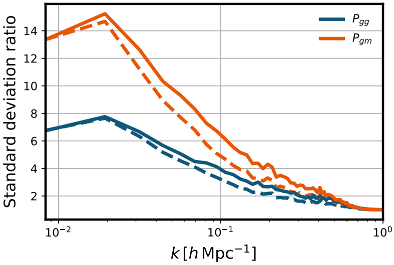
Aside: Zeldovich Control Variates

- Simulations always have limited dynamic range; in particular large scales are often "noisy" due to sample variance.
  - Especially true for simulations of high resolution, or including hydrodynamics, or RT, where boxes tend to be 'small'.
- PT works very well on large scales!
- If you start your simulation using Lagrangian PT (e.g. the Zeldovich approximation, or higher order) then you already have a surrogate field that is well correlated with your final density field (of matter, galaxies, gas, electrons, ...).
- Use control variates to reduce sample variance!
  - For the matter power spectrum this can give gains equivalent to averaging hundreds of simulations, for other spectra it can be tens.

Accurate predictions from small boxes!

An example of measuring P_{gg} and P_{gm} from a single $1\,h^{-1}$Gpc box, after applying CV and compared to the average of 100 such boxes (light lines extend to low k using PT).
[Even better performance for P_{mm}; not shown.]

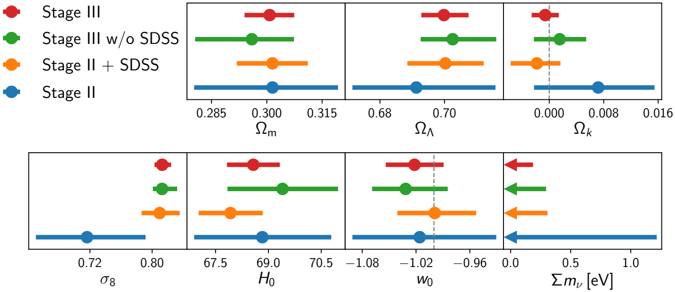




Kokron+22

A decade of progress

The SDSS surveys have used these kinds of perturbative models for their cosmology results.
This figure shows a decade of progress in using cosmology to constrain parameters:

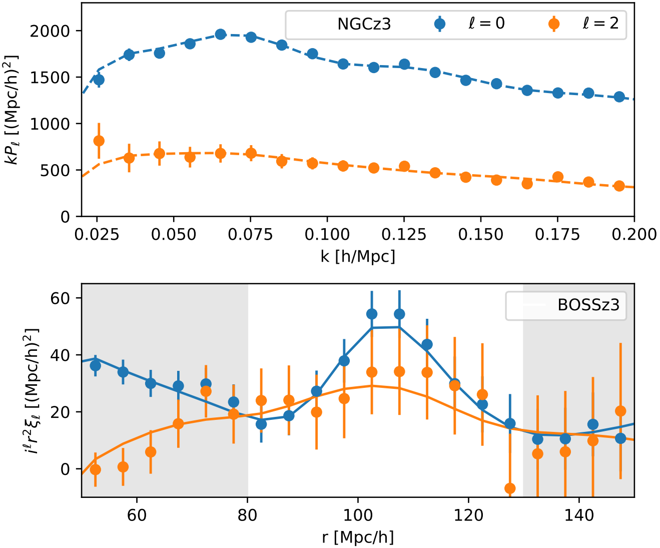


eBOSS collaboration (2020)

Why PT?

- Perturbation theory provides clean predictions for
  - matter (lensing) and biased tracers (galaxies, QSOs, ...)
  - in real and redshift space.
  - pre- and post-reconstruction
- Robust to uncertainties in small-scale physics ("integrated out").
  - No additional assumptions about halos, galaxies, etc. needed beyond the (minimal) set of bias parameters dictated by fundamental symmetries.
- Consistent predictions of $P_\ell(k)$, $\xi_\ell(s)$, $C_\ell^{\kappa g}$, $C_\ell^{\kappa\kappa}$, ... using the same parameters.

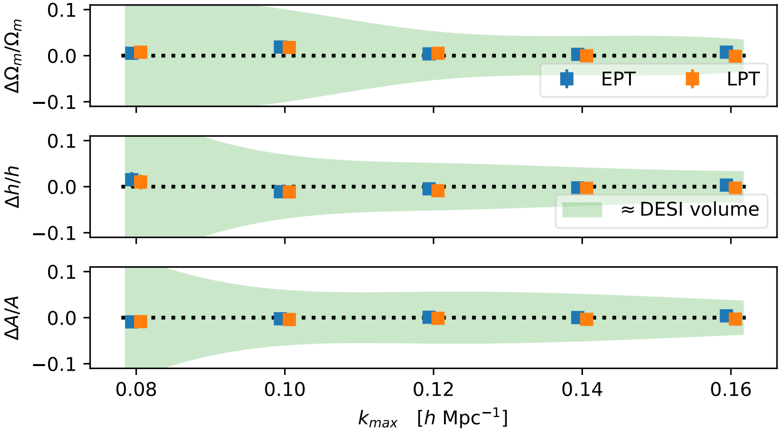
Models fit current data well



Chen et al. (2022)

PT blind challenge

Inferring parameters from fits to mock survey data 100× larger than physically achievable volumes:



"PT blind challenge", Nishimichi+20

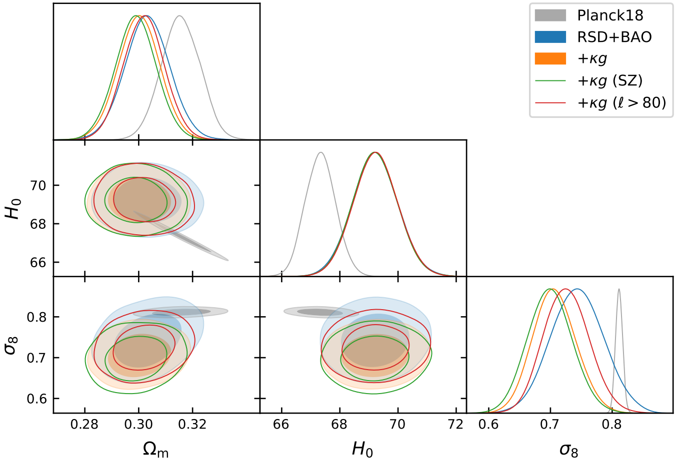
Velocileptors

- We have a public, Python package for these models.
- Being used in a number of surveys and data analyses now.
- Many ways to combine velocities and densities in power spectra: direct PT expansion, moment expansion, Gaussian streaming model, Fourier streaming model.
- Available in both LPT and EPT variants (allowing cross-checks!)
- Works in Fourier and configuration space.
- Fast and “easy to use”; works with Cobaya.

http://github.com/sfschen/velocileptors

BOSS×Planck lensing

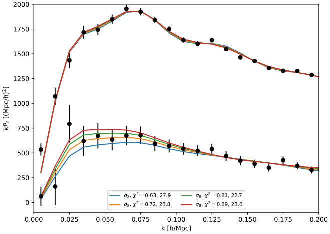
A preview of what we'd like to do with DESI : BAO (with reconstruction) plus RSD plus lensing, including massive neutrinos (carefully) all in a single PT framework with NN acceleration.

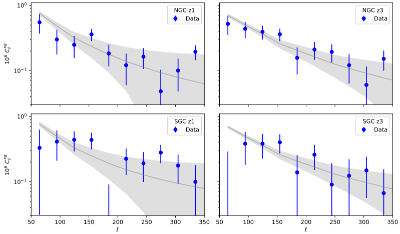


Chen et al. (2022)

The PT view of data

Constraining power comes from large scales



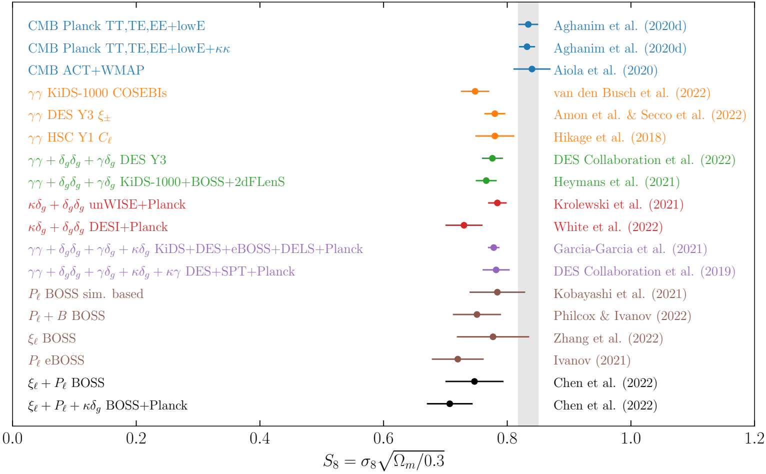


... but small scales help constrain (well-defined) transition to non-linearity.

Note σ_8 constraints from large scales, small scales are uncertain due to bias.

RSD strongly predicts $C_\ell^{\kappa g}$ at large scales (low ℓ) – in tension with the data (strengthens conclusions from previous analyses of the same surveys based on linear theory: Pullen+16, Doux+18)!

Comparison with other (recent) data



Chen et al. (2022)

Tensions in the current model

- These tensions are the focus of a lot of effort in the field!
- They resist 'easy' solution.
  - I (for one) am pretty mystified as to what is going on!
- The evidence is not as robust as we'd like, but more data like this are coming very soon!
- They have only arisen as we've shrunk the error bars: "precision" cosmology.
  - 'Hubble tension' and 'growth tension' represent $\mathcal{O}(10\%)$ shifts in parameters.
  - Seeing such things at $> 5\,\sigma$ requires $\sigma \simeq 1 - 2\%$

Since the model is working "pretty well" any signatures of BSM physics or deviations from ΛCDM are likely to be subtle ...

The “LSS program”: planning for what comes next

Probe metric, particle content and **both** epochs of accelerated expansion ... with high precision

- Expansion history and curvature (metric)
- Primordial non-Gaussianity ($f_{NL}^{\rm loc}$, $f_{NL}^{\rm eq}$, $f_{NL}^{\rm orth}$)
- Primordial or induced features, running of n_s
- Dark energy during MD
- DM interactions, light relics ($N_{\rm eff}$) and neutrinos
- etc.

Maximizing S/N

I want to maximize the S/N for new, BSM, physics

- There are many possible extensions to our SM (ΛCDM+GR).
- To my mind none are more compelling than others.
- If theory can't give us guidance, maybe phenomenology can?
  1. Work where inference is clean.
  2. Look where we haven't looked before (frontier!).
  3. If you don't know how to maximize S, then minimize N!

Push to higher redshift, in the epochs before cosmic noon ($z \simeq 2$)!

Advantages of high z

Moving to higher z gives us four simultaneous advantages:

1. Wide z range leads to rotated degeneracy directions.
2. Larger volume.
   - More than $3\times$ as many "linear" modes in the $2 < z < 6$ Universe than $z < 2$.
   - Large volume $\Rightarrow$ small errors at "low" k, increased dynamic range to break degeneracies.
3. More linearity and correlation with ICs.
   - Get "unprocessed" information from the early Universe.
4. High precision theory.
   - Low k modes are under good "theoretical control" using PT, little need for "nuisance parameter marginalization".
   - Everyone loves PT when you can use it – QED, Fermi liquids, CMB, ... LSS!
   - Theory becoming very advanced: lots of cross-fertilization with GR, CM and theory colleagues.

LSS at high-z offers many of the advantages of CMB anisotropy!

What probes of the $2 < z < 6$ Universe will we have?

Continuous advances in detector technology and experimental techniques are pushing us into a new regime, enabling mapping of large-scale structure in the redshift window $2 < z < 6$ using both relativistic and non-relativistic tracers ...

CMB = lensing at high z

We are witnessing a rapid scaling up of CMB experimental sensitivity as we move into the era of million-detector instruments!

- A natural "by-product" of next generation CMB surveys to constrain primordial gravitational waves is high fidelity CMB lensing maps – probing the matter back to $z \simeq 1100$.
- Lensing is sensitive to mass, not light, and by using a relativistic tracer it gives access to the Weyl potential.
- But lensing is projected ...
- ... want to do cross-correlation with samples of known redshift.
- Lensing + galaxy surveys offer redshift specificity, higher S/N and lower systematics. Natural synergies: total greater than sum of the parts!

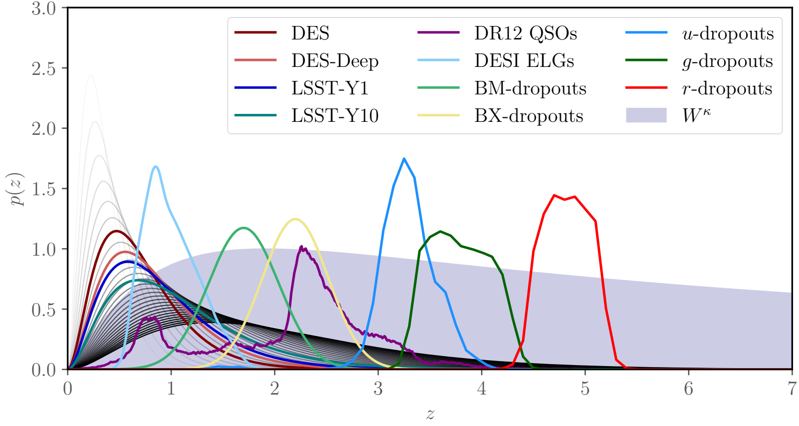
The promise of cross-correlations is that they enable new science as well as increased robustness of the core science of each project!

Tracers of LSS at $2 < z < 6$

There are lots of galaxies at high z, and we have pretty efficient ways of selecting them.

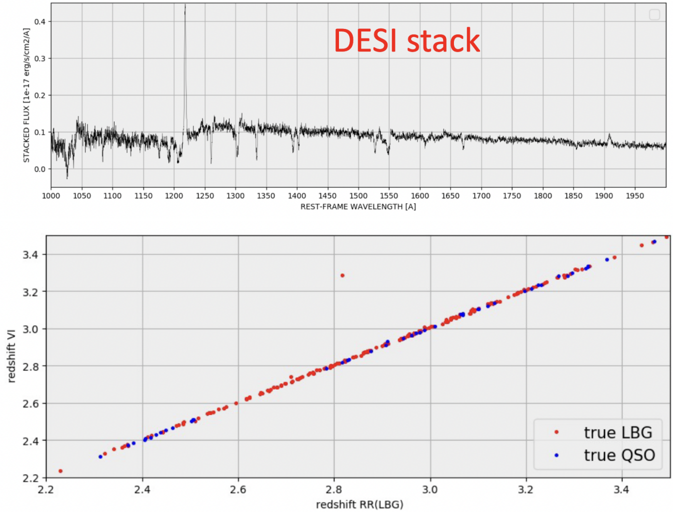
- Dropout, or Lyman Break Galaxy (LBG) selection targets the steep break in an otherwise shallow F_ν spectrum bluewards of 912Å.
- These objects have been extensively studied (for decades!).
- Selects massive, actively star-forming galaxies – and a similar population over a wide redshift range.
- LBGs lie on the main sequence of star formation and UV luminosity is approximately proportional to stellar mass.
- A fraction of these objects have bright emission lines (LAEs). The LAE population tends to be lower bias, younger and with low dust. If you can select them, they're easy to redshift!

Galaxies over the whole range



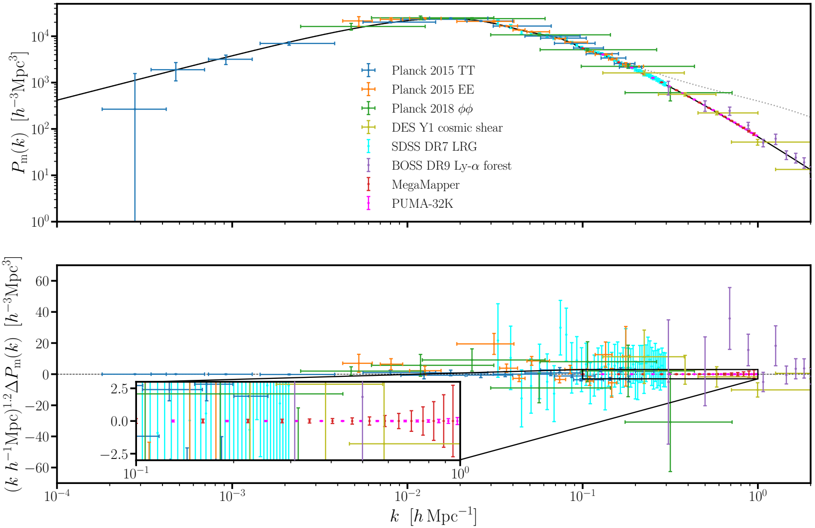
Wilson & White (2019)

Lyman Break Galaxies with DESI



Courtesy A.Raichoor & Ch.Yeche

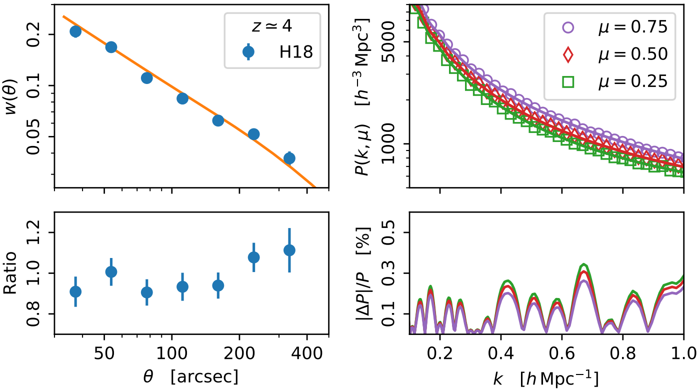
Very small error bars!



Sailer et al (2021)

Can we achieve high precision?

Out-of-the-box comparison of two, public, theory modeling codes



Over half the sky, within $3.5 < z < 4.5$ there are over a billion modes out to $k = 1\,h\,\mathrm{Mpc}^{-1}$!

Conclusions I

We are in the midst of the "golden age of cosmological surveys".

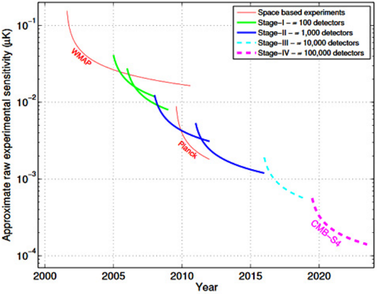
- Increasing survey power is driving a renaissance in analytic models of large-scale structure.
  - More perturbative modes at higher precision!
  - Form and techniques familiar from other areas of physics.
  - A few "cosmology" wrinkles.
- The models are well motivated and work well on current data.
  - Well motivated inference problem.
  - Allow us to forecast performance of future surveys reliably.
  - Survey optimization.
- Adding "beyond standard model" physics or new probes is an active area of research.
- The benefits and disadvantages of the different approaches are still not fully understood ...

Conclusions II

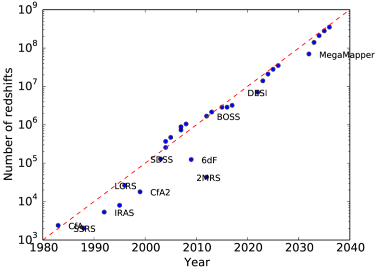
- There are **many** (quasi-)linear modes left to map!
- These will allow precisions tests of SM and GR, and improve constraints on parameters by substantial factors (or find something new!).
  - Already (several) percent-ish level constraints at lower z are turning up much-discussed "tensions".
  - These tensions resist easy solution, and are seen even on large scales – which should be easy to model.
- When looking for BSM physics, if theory can't give us guidance maybe phenomenology can?
  - Work where inference is clean.
  - Look where we haven't looked before.
  - If you don't know how to maximize S, then minimize N!
- The best observational approaches are still TBD.
  - Pilot programs and R&D

The End!

Moore's law for instruments

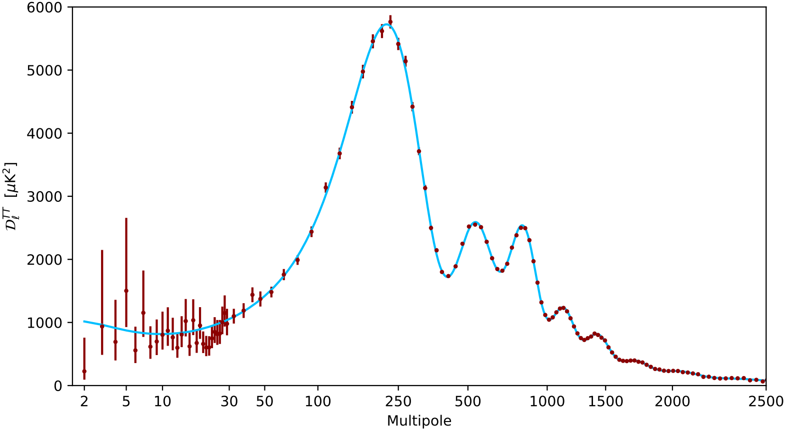


CMB-S4 collab.



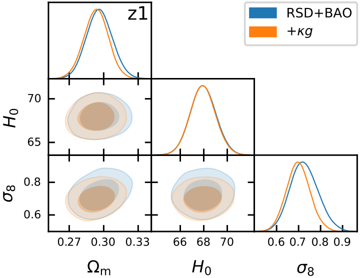
MegaMapper collab.

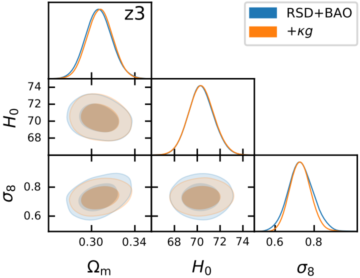
The power spectrum: photons, early time



Planck+(2020)

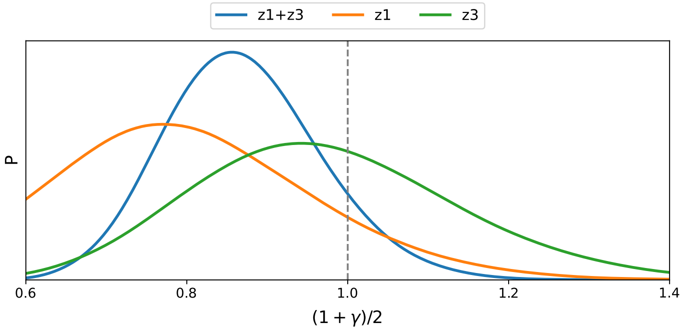
Where is the low σ_8 coming from?





Low redshift, large scales and lensing!

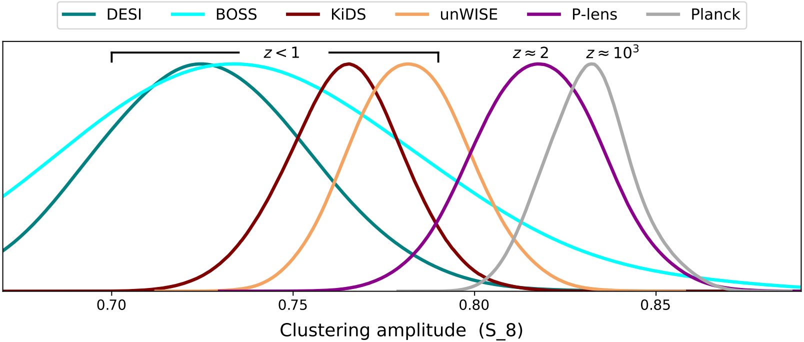
Lensing is low???



Chen et al. (2022)

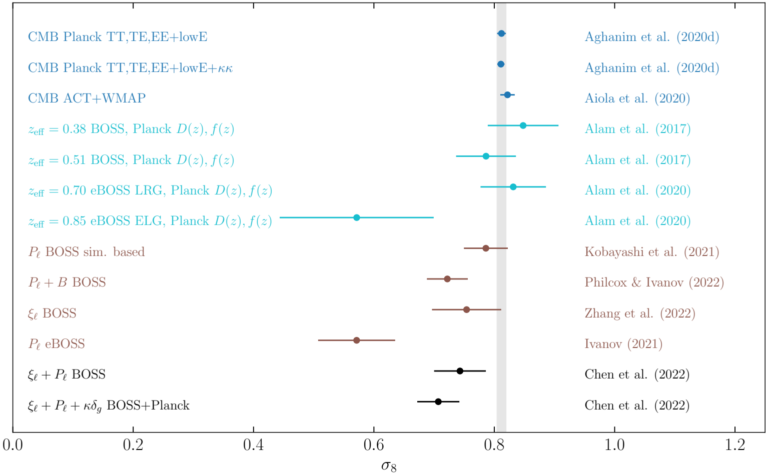
Tensions in the current model

Structure isn't growing the way we think it should be ...



White et al. (2022)

Comparison with other (recent) data



Chen et al. (2022)

Foreground and systematics

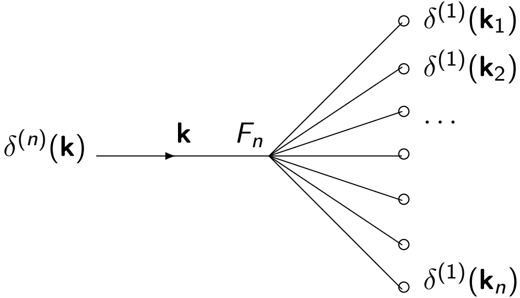
- The BOSS, DESI and Planck maps have all been investigated for numerous systematics.
- We did a number of additional tests, cutting on hemisphere, extinction, redshift, etc.
- Cross-correlated the Planck κ map with the BOSS systematics weights.
  - Find κ correlates with w at the 10% level – would need 10% residuals after weights to produce a 1% effect.
- Cross-correlated BOSS galaxies and the Planck κ map with $E(B-V)$
  - Sub-percent bias in galaxy auto-spectra (as BOSS found).
  - Could be up to a few percent in cross-spectra!

Systematics mitigation for cross-correlation science require some 'additional' care! Need to be thinking about this as we move into our golden age!

Eulerian PT

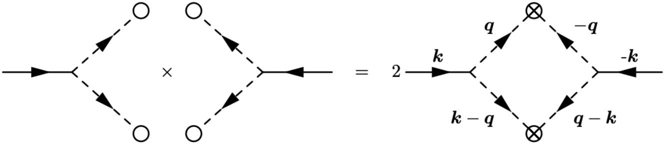
Expand $\delta(\mathbf{k})$ as a power series in the linear solution:

$$\delta^{(n)}(\mathbf{k}) = \int \prod_{i=1}^{n} \frac{d^3 k_i}{(2\pi)^3} (2\pi)^3 \delta^{(D)} \left(\sum \mathbf{k}_i - \mathbf{k} \right) F_n \left(\{\mathbf{k}_i\} \right) \delta(\mathbf{k}_1) \cdots \delta(\mathbf{k}_n)$$



Eulerian PT

Taking the expectation value “joins” pairs of $\delta^{(1)}$ together to form a power spectrum. So the propagator at 1-loop contains a contribution like e.g.:

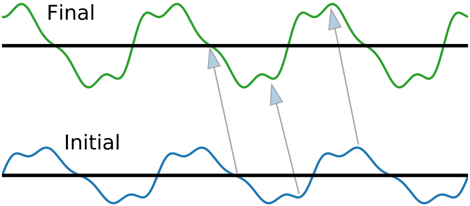


$$= 2 \int \frac{d^3 q}{(2\pi)^3} F_2(\boldsymbol{q}, \boldsymbol{k}-\boldsymbol{q}) F_2(-\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{k}) P_0(q) P_0(|\boldsymbol{k}-\boldsymbol{q}|)$$

Many other rules look similar to particle or condensed matter physics – use path integrals, generating functions, cutoffs, EFT, RG flows, etc.

IR resummation

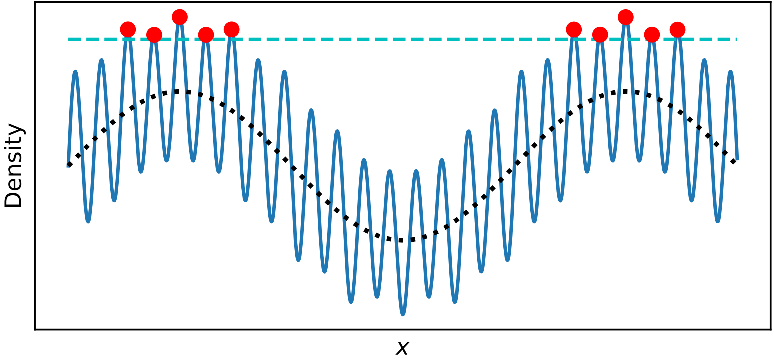
- In addition to problems in the UV, there are issues in the IR.
- A lot of the difference between $\delta^{(\mathrm{non-lin})}(\mathbf{x})$ and $\delta^{(1)}(\mathbf{x})$ comes from displacement (advection).



- The displacement is driven by large-scale tidal fields.
- In "standard" perturbation theory this effect converges slowly.
- Need to "resum" the long-wavelength displacements
  - The Lagrangian formulation of PT is ideally suited to understanding "IR resummation".
  - Impacts topics like "reconstruction", primordial features, relative velocity effect, ...

Bias, peaks and EFT

- To make contact with galaxies, QSOs, 21 cm, Lyα, etc. we need to include bias.
- Simplest (toy) model: galaxies form at peaks in the initial density field:



Neural network acceleration

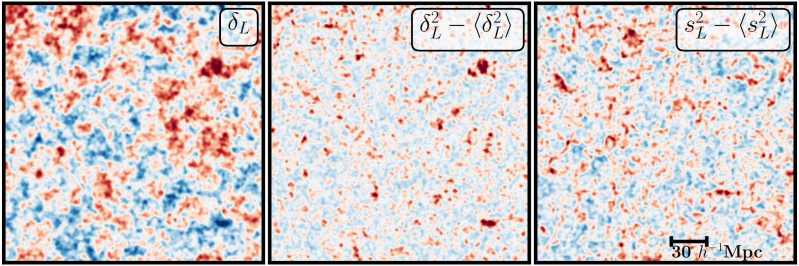
- We use ‘standard’ Markov-Chain Monte Carlo techniques to explore parameter space with the `Cobaya` sampler.
- Due to the multiple samples and nuisance parameters we have a high (20) dimensional parameter space.
- To speed up the chains we first emulate all of the theory computations (including CAMB/CLASS) using neural networks.
  - Within PT, bias parameters are ‘analytic’ so need only emulate the cosmology dependence.
  - Very easy to generate training data – can be arbitrarily accurate.
  - Pay price “up front” once, then theory is “free”.

Massive neutrinos

- Galaxies probe the $c + b$ field while lensing probes the matter.
- At linear level use $P_{cb}(k)$ for galaxies and $P_{cb,m}$ for galaxy-lensing cross-correlation.
  - Good to sub-percent level (e.g. Bayer+21)
- If care is taken with normal ordered bias operators, can use $P_{cb,m}$ in loops with corrections of order $f_\nu P_{\rm lin}^2 \ll 1$ and be correct even in the "transition regime" from clustered to free-streaming neutrinos.

The hybrid EFT procedure in pictures

Generate initial conditions as per usual ... from δ_L you can also compute δ_L^2 and the shear field, s_{ij}:

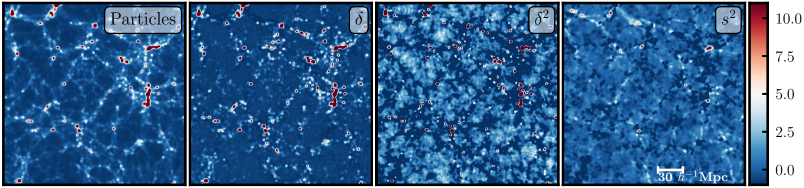


Each particle is assigned the δ_L, ... at its initial position.

Kokron+21

The hybrid EFT procedure in pictures

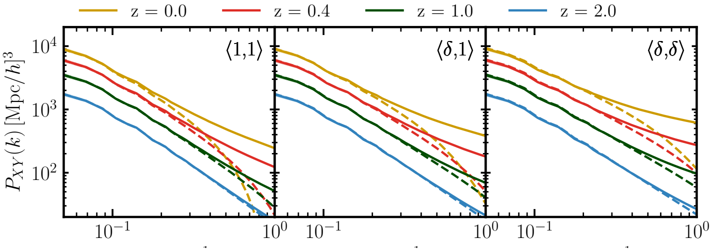
Advect the particles to their final positions using the full N-body dynamics (i.e. run the simulation), and bin using weights 1, δ_L, δ_L^2, etc.



(No need for halo or subhalo finding, merger trees, etc.)

The hybrid EFT procedure in pictures

Take all of the cross-spectra, $P_{XY}(k)$ using standard FFT methods, e.g.



The power spectrum for any biased tracer, or the cross-spectrum between any two tracers, is a linear combination of these "basis spectra" (10 in all) with analytic "bias dependence": $\sum_{ij} b_i b_j P_{ij}$.

Statistical inference

Since inhomogeneity arose from stochastic (QM) fluctuations, all inferences are statistical.

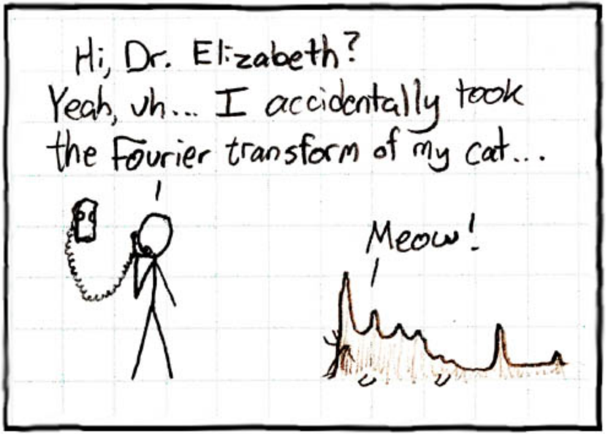
- The particular location of any given galaxy or star is not relevant ... we look at correlations.
- Compare correlators of temperature, density, velocity, etc.
- For this talk my focus will be on 2-point functions, e.g. $\langle\rho(\mathbf{x}_1)\rho(\mathbf{x}_2)\rangle$.

An irreducible contribution to our uncertainty on $\langle\cdots\rangle$ comes from the number of "independent samples" (modes) in our realization (survey) ... want to push to larger volumes (i.e. earlier times)!

Statistical inference

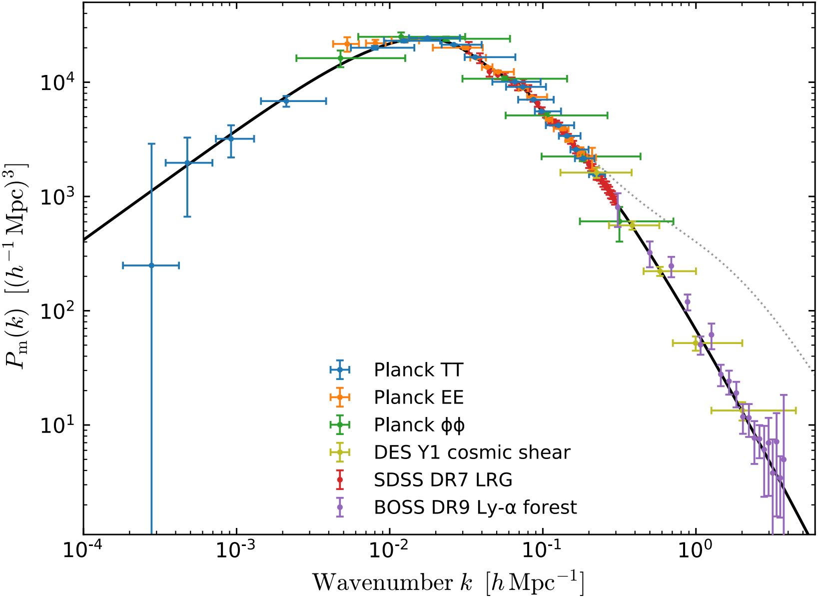
The underlying processes are translationally invariant, so we tend to work in Fourier space:

$$\rho(\mathbf{x}) = \bar{\rho}\,(1+\delta(\mathbf{x})) \quad , \quad \langle\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\rangle = (2\pi)^3\delta^{(D)}(\mathbf{k}_1+\mathbf{k}_2)\,P(k)$$



xkcd.com/26

The power spectrum: multiple tracers, late time



Planck+(2020)