Title: Analytic calculation of the power spectrum covariance: speedup by four orders of magnitude

Speakers: Digvijay Wadekar

Series: Cosmology & Gravitation

Date: November 10, 2020 - 11:00 AM

URL: http://pirsa.org/20110007

Abstract: In order to infer cosmological parameters from galaxy survey data, we typically use summary statistics such as the power spectrum and we need an accurate estimate of their covariance matrix. The traditional process of obtaining the covariance involves simulating thousands of mocks. I will present an analytic approach for the covariance matrix which is more than four orders of magnitude faster than mocks and show its validation with an analysis of the BOSS DR12 data. Furthermore, our analytic approach is free of sampling noise which makes it useful for upcoming surveys like DESI and Euclid. Towards the end, I will change gears and talk about some recent work on the assembly bias of neutral hydrogen.

Analytic calculation of power spectrum covariance

(Jay) Digvijay Wadekar

New York University

DW, Roman Scoccimarro (arXiv 1910.02914)
DW, Misha Ivanov, Roman Scoccimarro (aXiv:2009.00622)

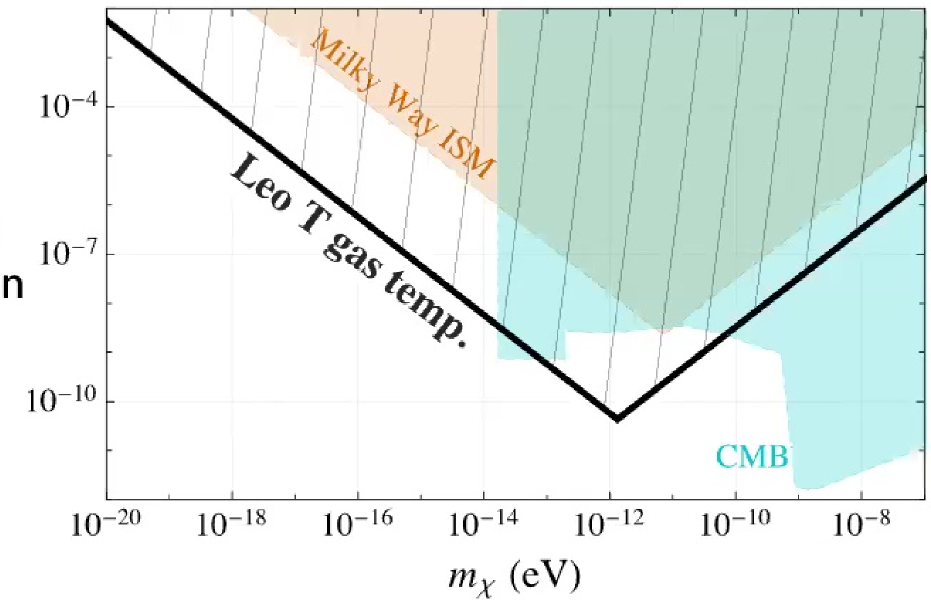
1

Research overview

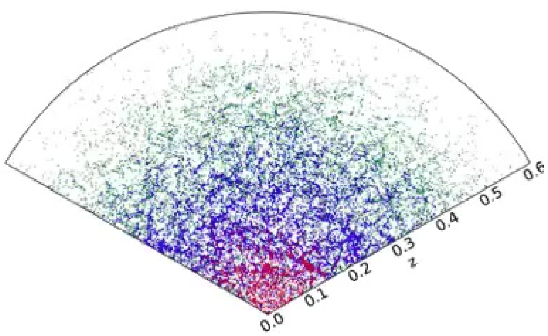
- Dwarf galaxy observations to constrain alternatives to cold dark matter (CDM)

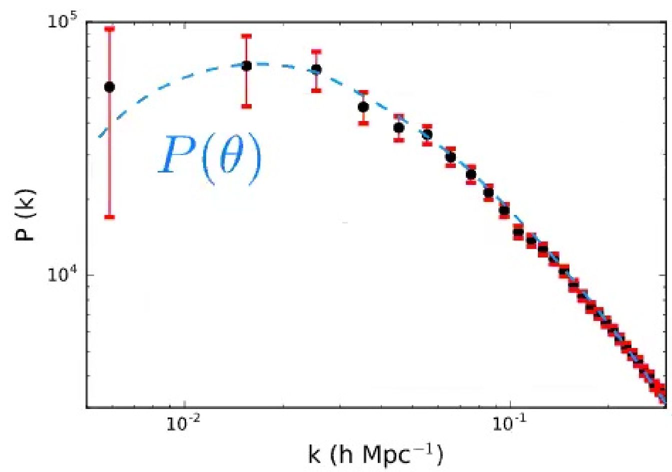
DW & Farrar 19
Farrar et al. 19

Ultra-light dark photon coupling



Galaxy power spectrum covariance





$$H_0 \quad \left\{ \begin{matrix} \Omega_m \\ \sigma_8 \end{matrix} \right\} \quad b_1 \quad \ldots.$$

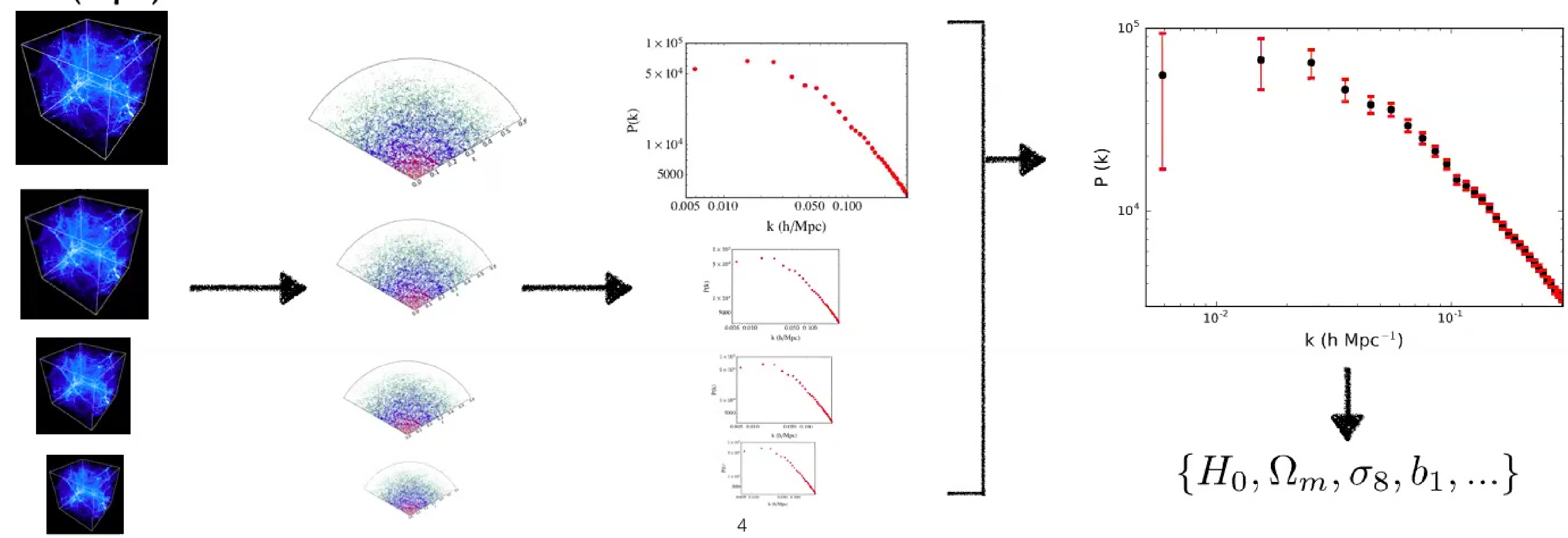
$$p(P_d|\theta) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C(\theta)}} \exp\left[-\frac{1}{2}(P_d - P(\theta))^T C(\theta)^{-1}(P_d - P(\theta)) \right]$$

Covariance from mock catalogs

- Need to simulate mock surveys (~ thousands)

$$C_{1,2} = \frac{1}{N_{\rm mock} - 1} \sum_{i=1}^{N_{\rm mock}} (\hat{P}_i(k_1) - \bar{P}(k_1))(\hat{P}_i(k_2) - \bar{P}(k_2))$$

~$\mathcal{O}$ (Gpc)^3



Covariance from mock catalogs

- As survey volume increases, mock catalogs become tougher to simulate (DESI, LSST, Euclid and others)

- Dependence of covariance on cosmology and bias parameters is computationally prohibitive

$$p(P_d|\theta) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C(\theta)}} \exp\left[-\frac{1}{2}(P_d - P(\theta))^T C(\theta)^{-1}(P_d - P(\theta))\right]$$

Covariance from mock catalogs

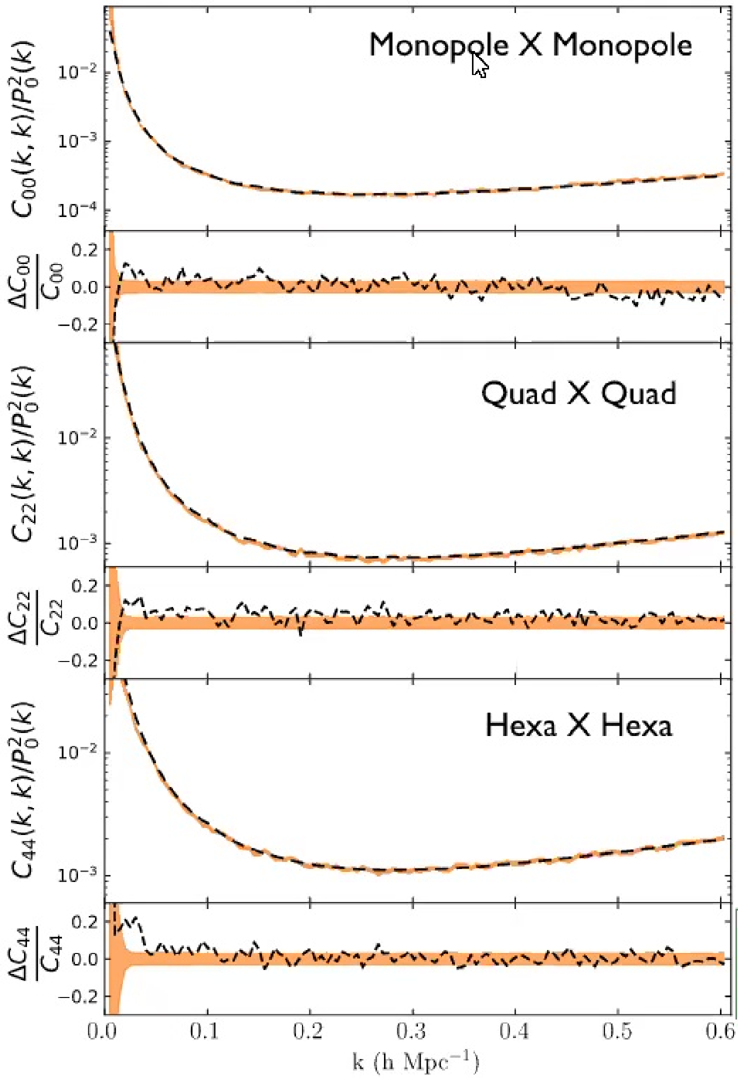
- As survey volume increases, mock catalogs become tougher to simulate (LSST, DESI, Euclid and others)

- Dependence of covariance on cosmology and bias parameters is computationally prohibitive

$$p(P_d|\theta) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C(\theta)}} \exp\left[-\frac{1}{2}(P_d - P(\theta))^T C(\theta)^{-1}(P_d - P(\theta)) \right]$$

- Mocks suffer from sampling noise
  - Need to artificially inflate constraints

Results

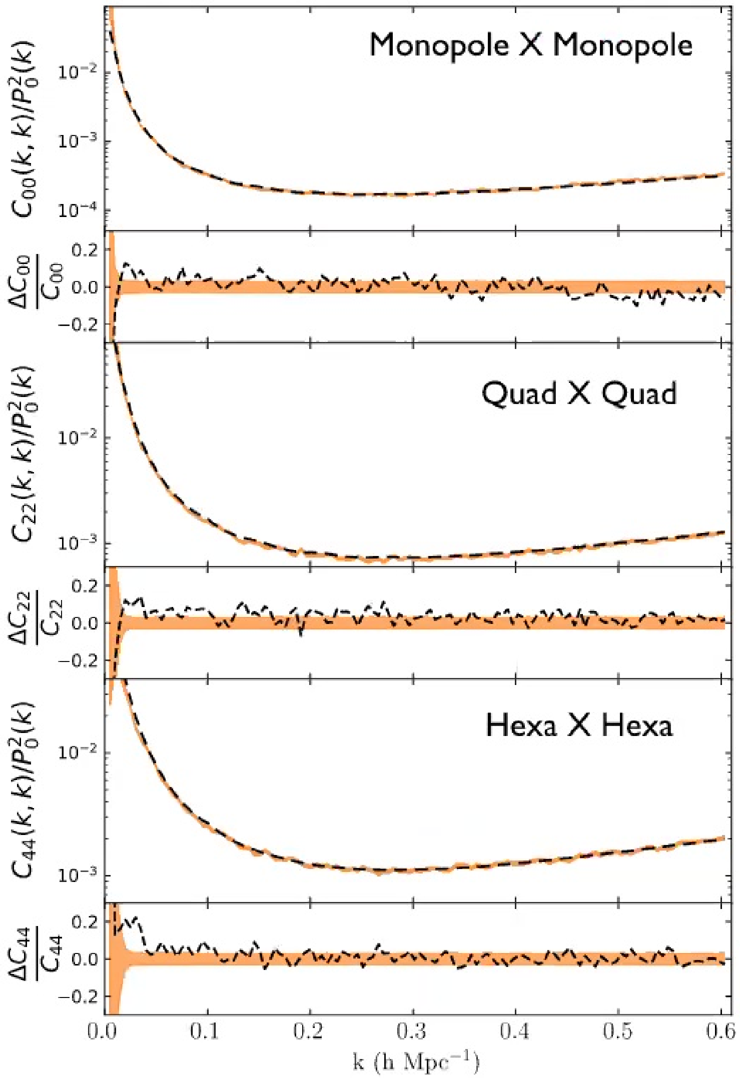


- - - - - ***Our analytic method***

Patchy Mocks
(state-of-the-art mocks used for SDSS BOSS parameter estimation)

DW & Scoccimarro 19
(Cova-PT public code)

Results

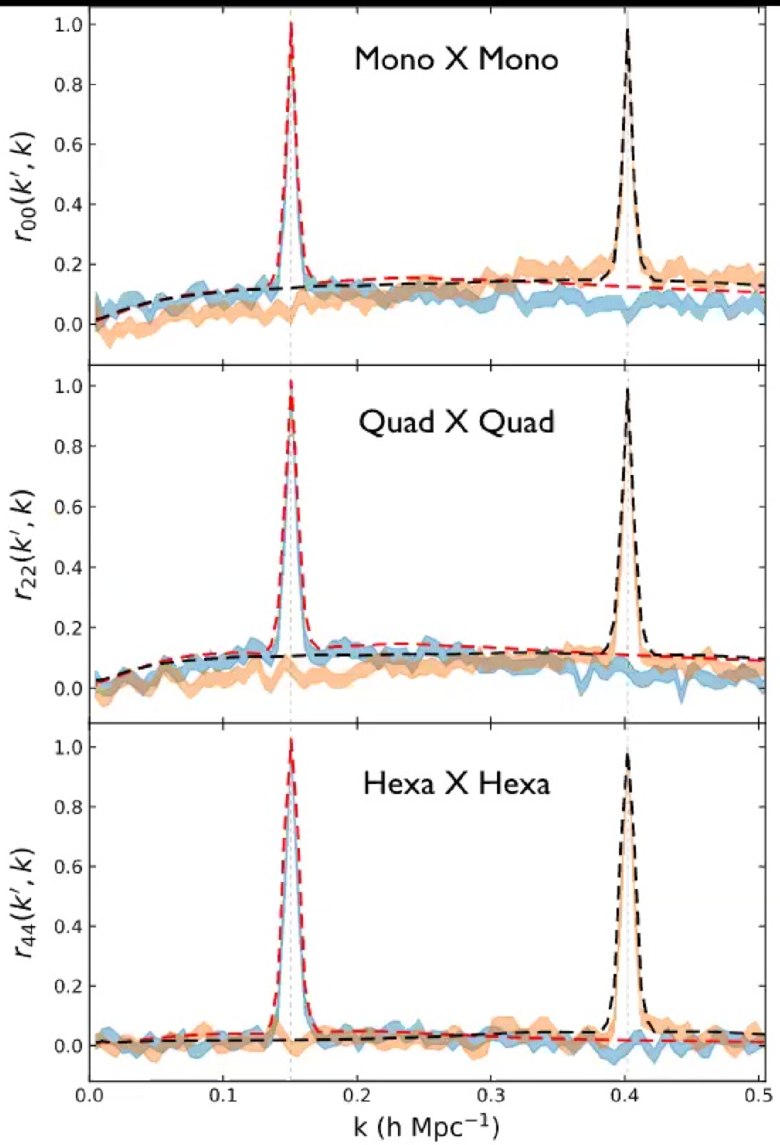


----- ***Our analytic method***
($\leq$ MINUTE)

Patchy Mocks ***(MONTHS)***
(state-of-the-art mocks used for SDSS BOSS parameter estimation)

DW & Scoccimarro 19

8



Results

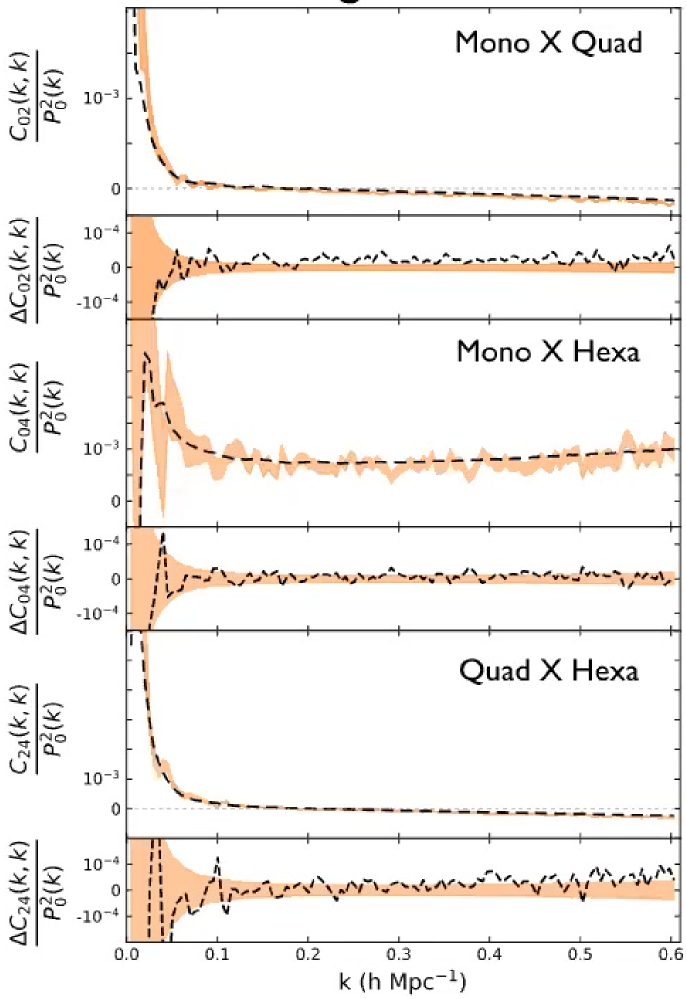
(off-diagonals)

(2 rows compared in fig)

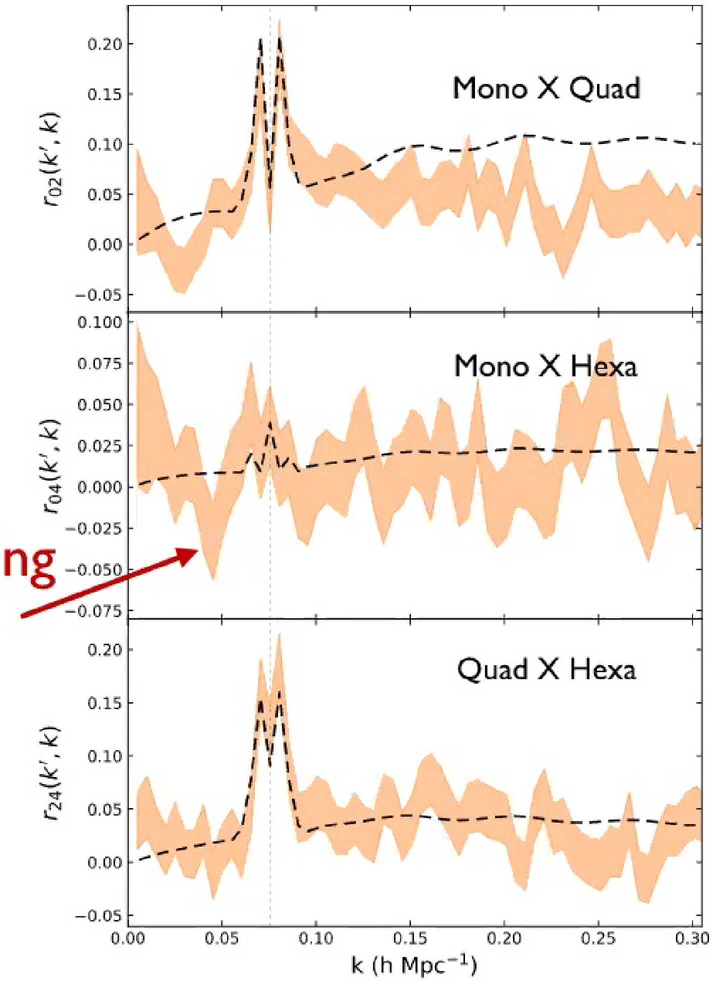
9

Cross-covariance

Diagonals



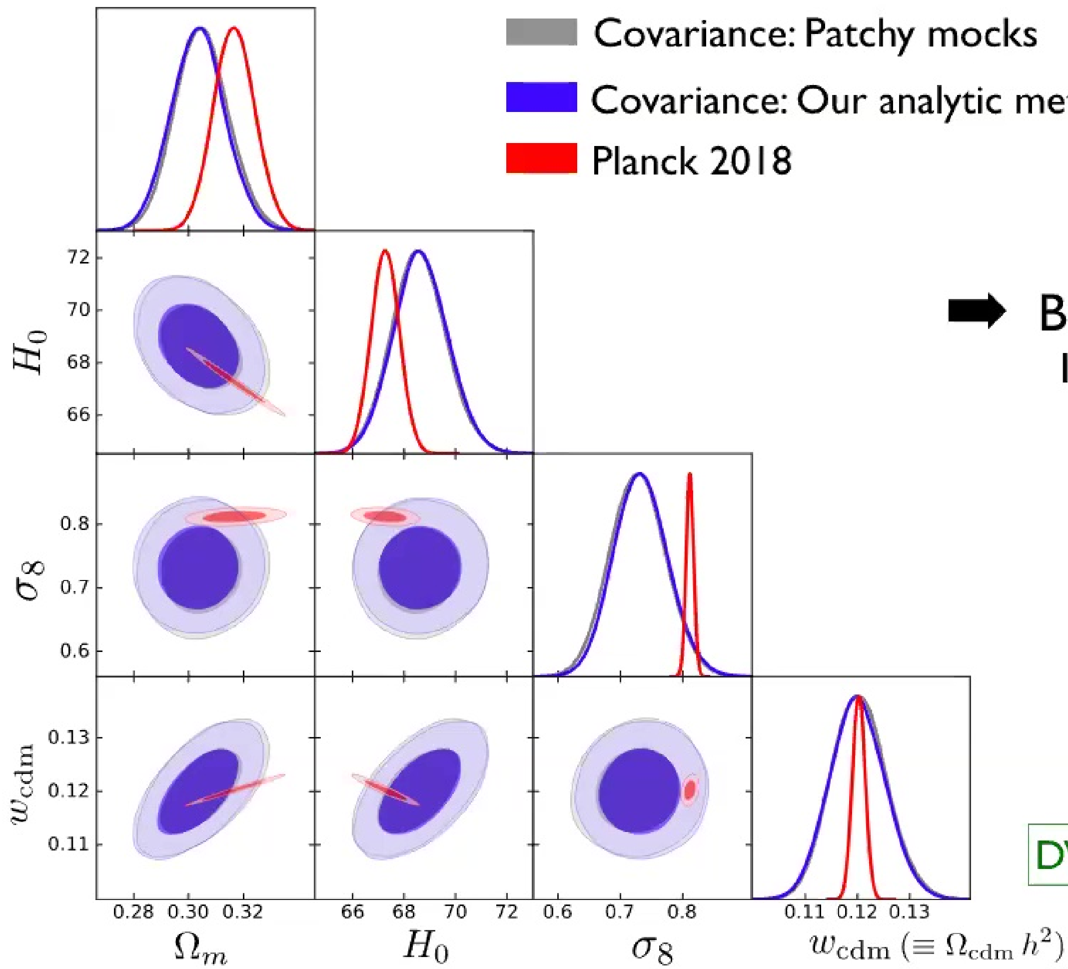
Row of matrix



Large sampling noise in mocks

10

Results: BOSS DR12 full-shape analysis



➡ Based on BOSS analysis pipeline of

Ivanov, Simonovic, Zaldarriaga, JCAP 20

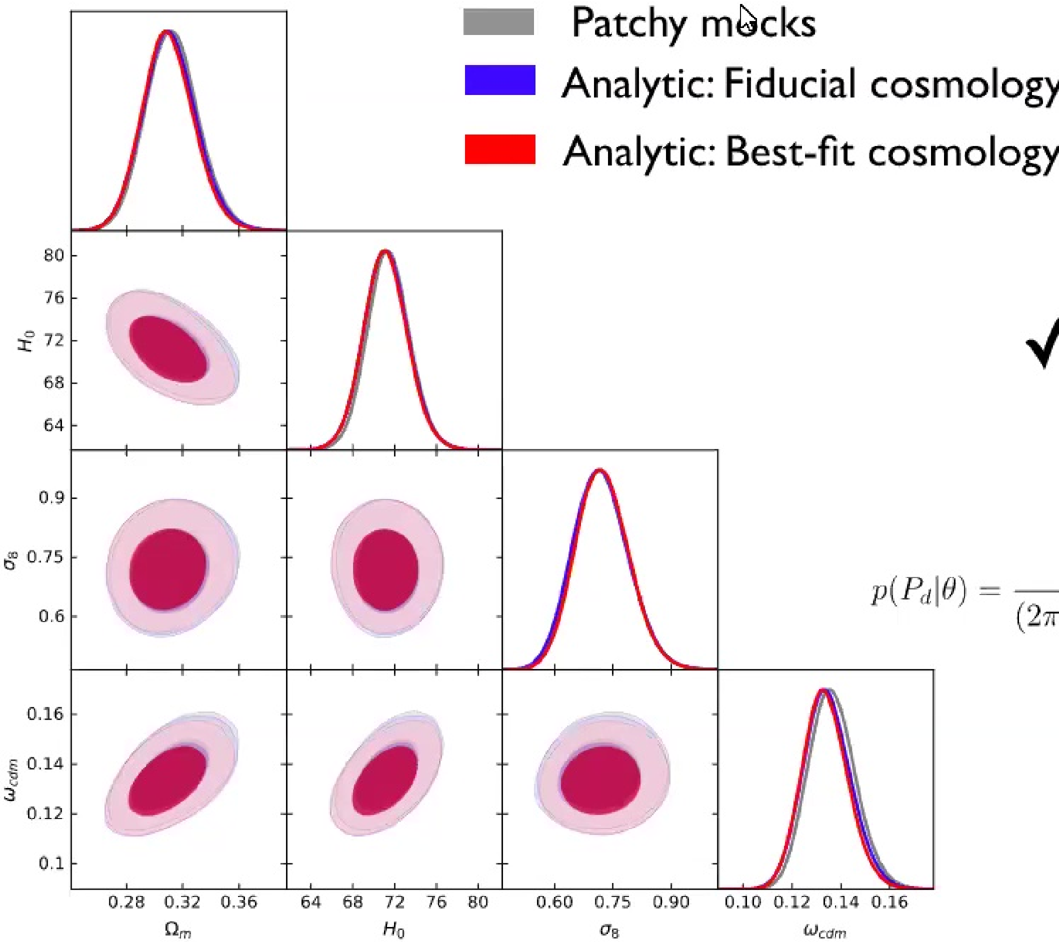
Philcox et al. 2020

Ivanov et al 2020 (CLASS-PT)

DW, Ivanov & Scoccimarro, 2020

11

Case study: BOSS sample of NGC high-z

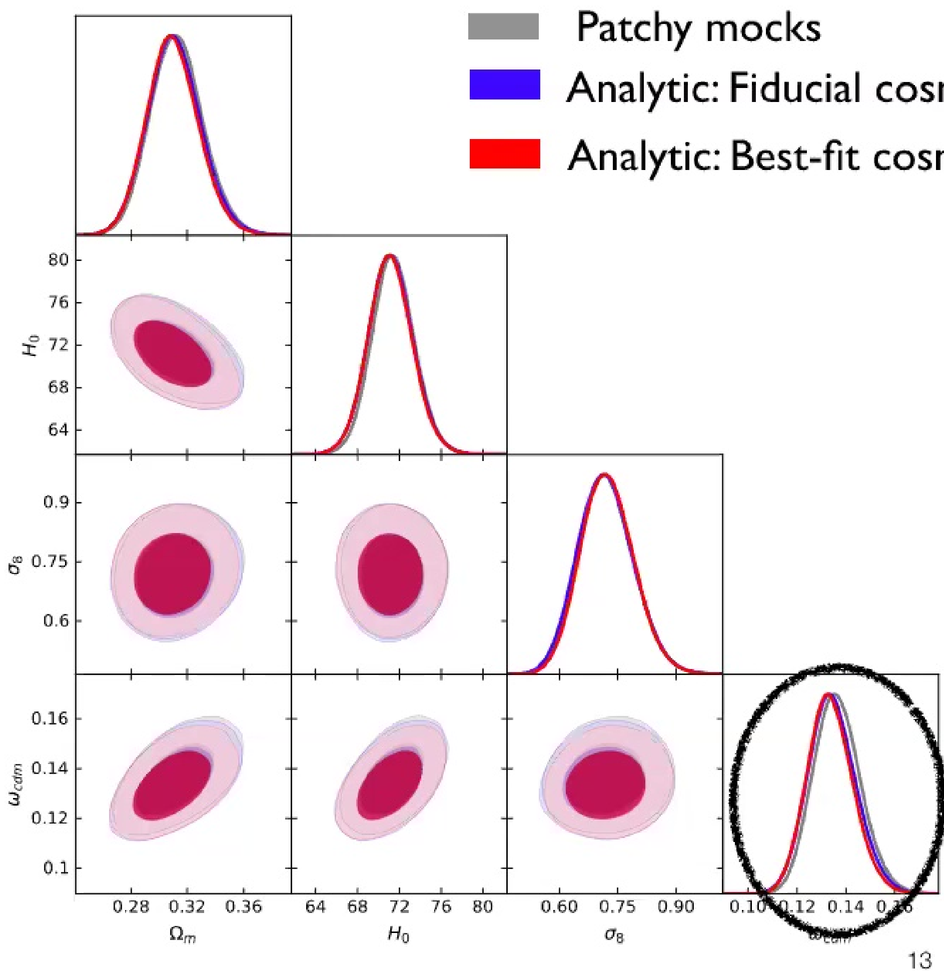


✓ BOSS results are robust w.r.t change in cosmology of covariance matrix

$$p(P_d|\theta) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C(\theta)}} \exp\left[-\frac{1}{2}(P_d - P(\theta))^T C(\theta)^{-1}(P_d - P(\theta))\right]$$

12

Case study: BOSS sample of NGC high-z

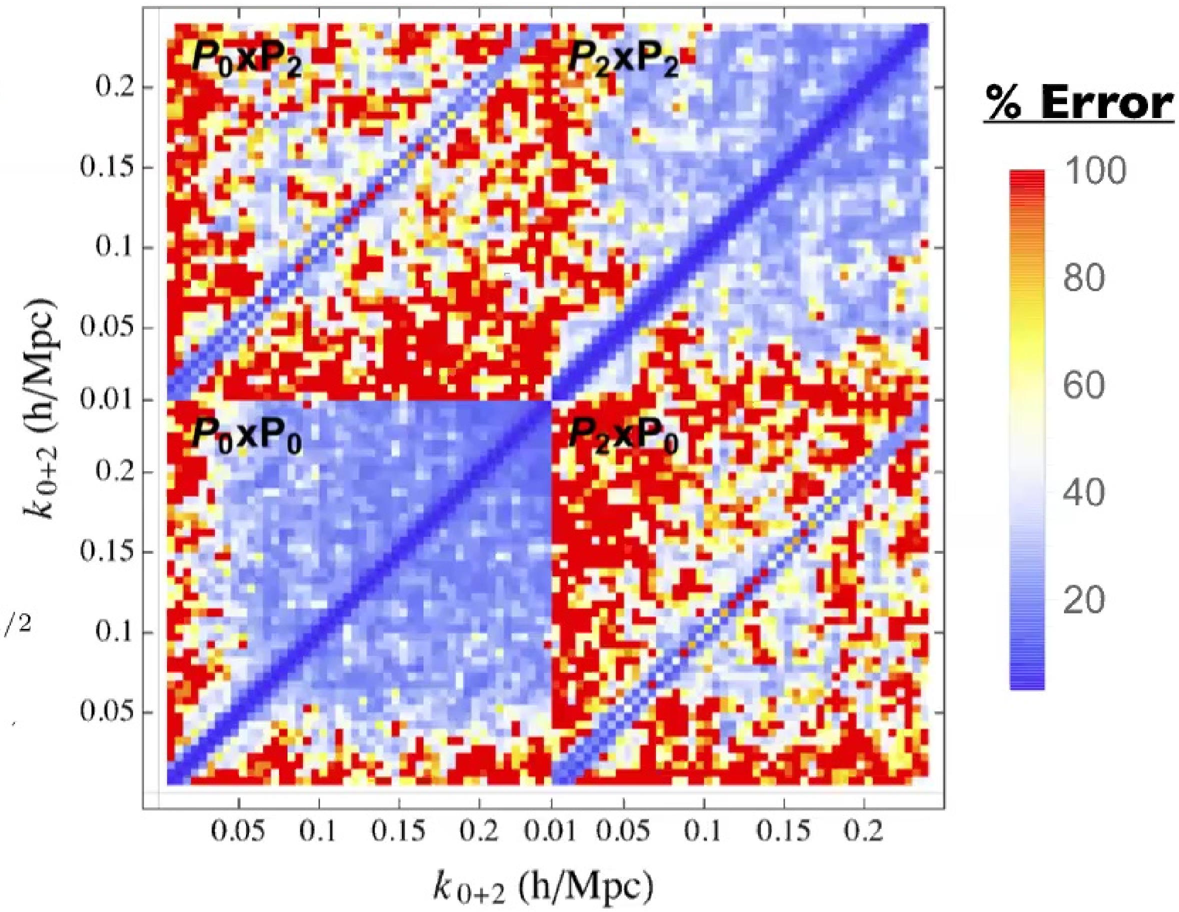


✓ BOSS results are robust w.r.t change in cosmology of covariance matrix

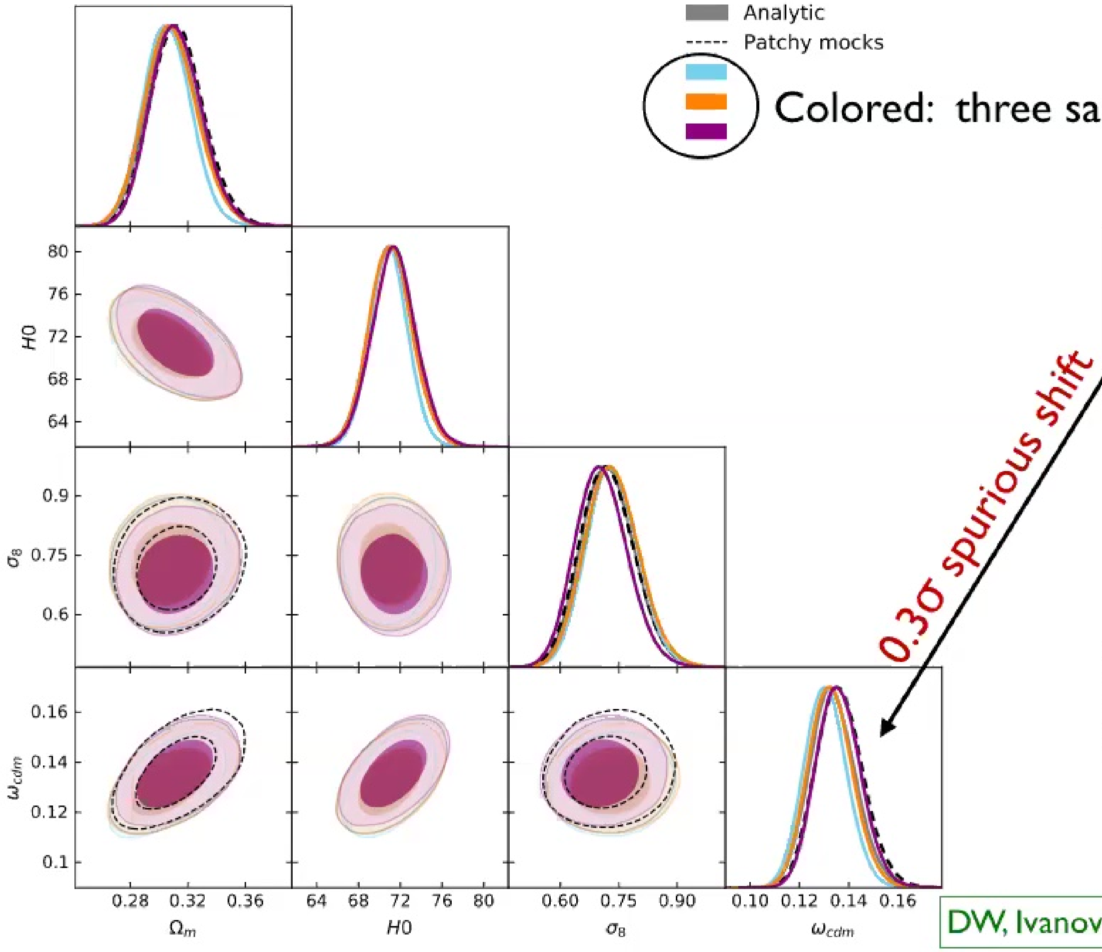
➡ Small shifts (0.2σ) because of sampling noise

Large sampling noise in off-diagonals & cross-covariance from 2048 Patchy mocks

$$\frac{\Delta \mathbf{C}_{\ell_1\ell_2}(k_i,k_j)}{\mathbf{C}_{\ell_1\ell_2}(k_i,k_j)} \sim \left[\frac{1}{\mathrm{N_m}}\frac{\mathbf{C}_{\ell_1\ell_1}(k_i,k_i)\mathbf{C}_{\ell_2\ell_2}(k_j,k_j)}{\mathbf{C}^2_{\ell_1\ell_2}(k_i,k_j)}\right]^{1/2}$$



Sampling noise in covariance from 2048 mocks

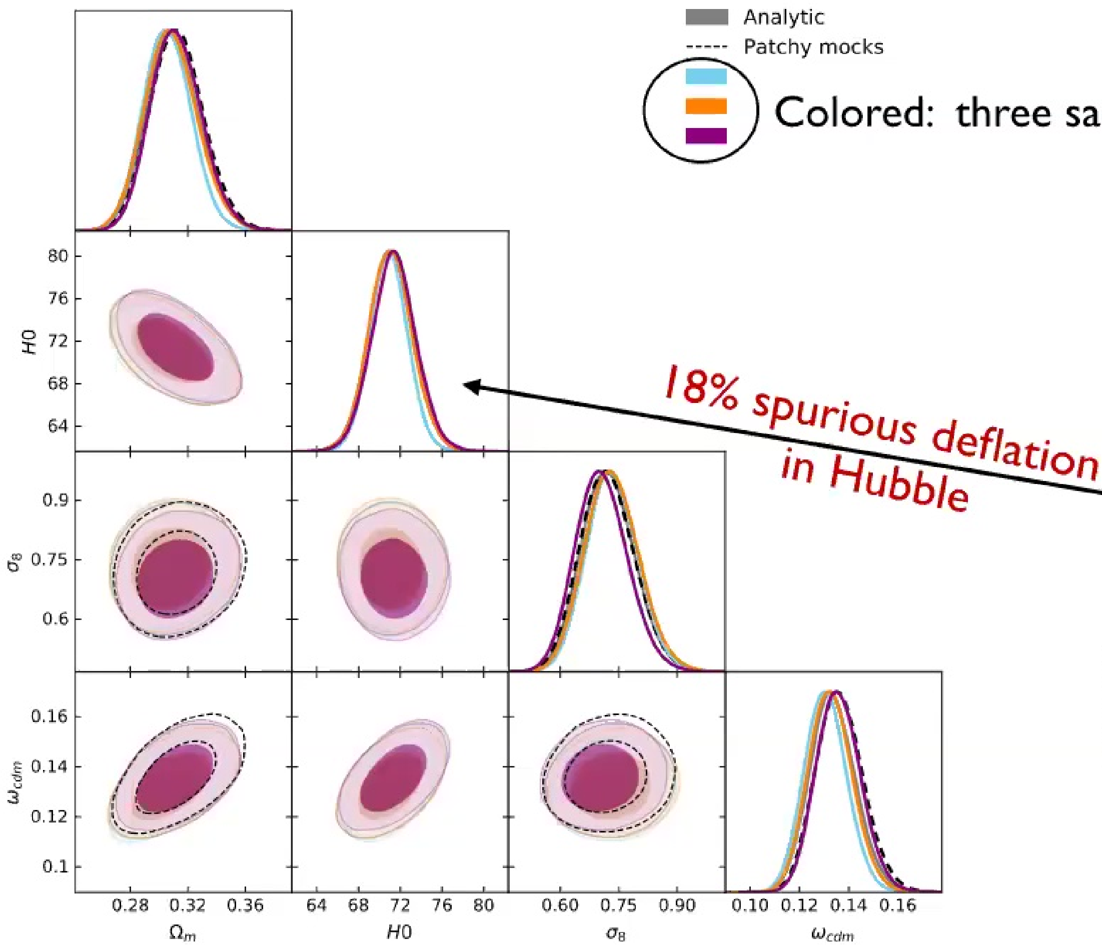


1. Shifts in parameter constraints

Dodelson & Schneider 2013
Percival et al. 2014

DW, Ivanov & Scoccimarro, 2020

Sampling noise in covariance from 2048 mocks



1. Shifts in parameter constraints

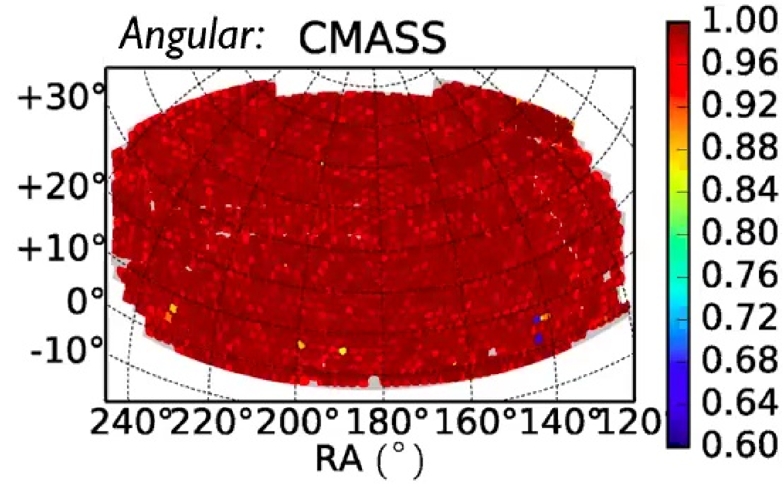
2. Inflation/deflation of parameter error bars

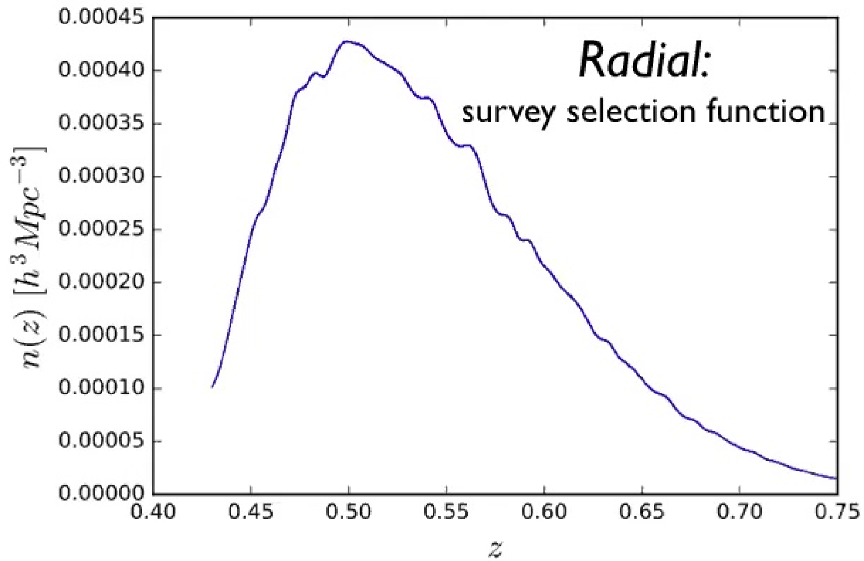
Dodelson & Schneider 2013
Percival et al. 2014

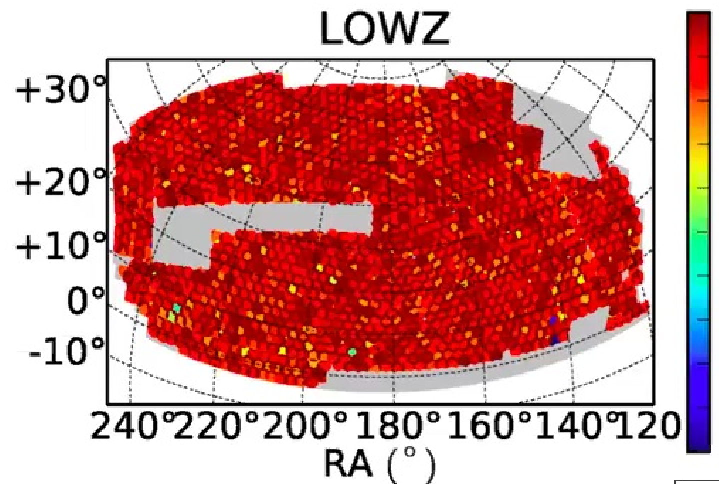
What are the challenges to analytically calculate the covariance?

Challenge I:
Highly non-trivial survey window

$$\delta_W(\mathbf{x}) \equiv W(\mathbf{x})\delta(\mathbf{x})$$

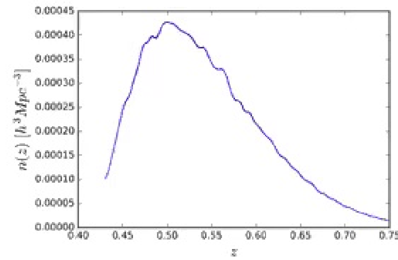


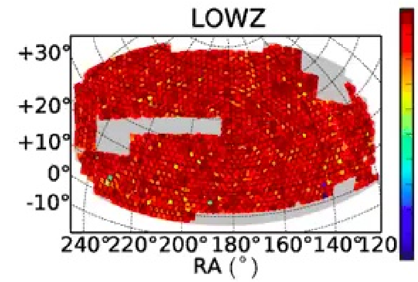




Reid et al. 2015

Survey window enters covariance





$$C(k_1, k_2) = \langle \hat{P}(k_1)\hat{P}(k_2)\rangle - \langle \hat{P}(k_1)\rangle\langle \hat{P}(k_2)\rangle$$

$$\begin{aligned}\langle \hat{P}(k_1)\hat{P}(k_2)\rangle =& \frac{1}{V_2^2}\int_{\hat{\mathbf{k}}_1,\hat{\mathbf{k}}_2} \langle \delta_W(\mathbf{k}_1)\delta_W(-\mathbf{k}_1)\delta_W(\mathbf{k}_2)\delta_W(-\mathbf{k}_2)\rangle \\ =& \frac{1}{V^2{}_2}\int_{\hat{\mathbf{k}}_1,\hat{\mathbf{k}}_2,\mathbf{p}_1,\mathbf{p}'_1,\mathbf{p}_2,\mathbf{p}'_2} W(\mathbf{k}_1-\mathbf{p}_1)W(-\mathbf{k}_1-\mathbf{p}'_1)W(\mathbf{k}_2-\mathbf{p}_2)W(-\mathbf{k}_2-\mathbf{p}'_2) \\ & \times \langle \delta(\mathbf{p}_1)\delta(\mathbf{p}'_1)\delta(\mathbf{p}_2)\delta(\mathbf{p}'_2)\rangle\end{aligned}$$

18 dimensional integral

Solution: separate clustering and window terms

Contains all dependence on cosmology and bias parameters

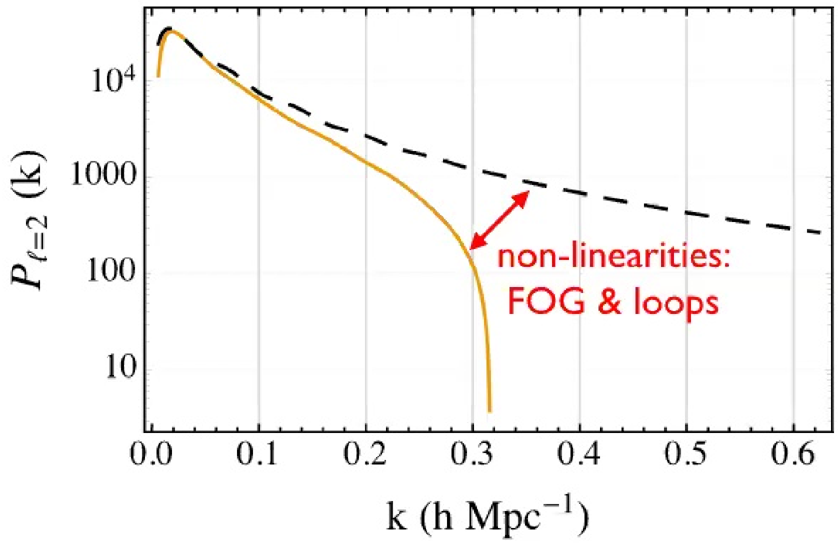
$$\mathbf{C}^{\mathrm{G}}_{\ell_1\ell_2}(k_1,k_2) \simeq \sum_{\ell_1',\ell_2'} P_{\ell_1'}(k_2)P_{\ell_2'}(k_1)\left\{\frac{(2\ell_1+1)(2\ell_2+1)}{\mathrm{I}_{22}^2}\int_{\hat{\mathbf{k}}_1,\hat{\mathbf{k}}_2,\mathbf{x}_1,\mathbf{x}_2} W_{22}(\mathbf{x}_1)W_{22}(\mathbf{x}_2)\, e^{-i(\mathbf{x}_1-\mathbf{x}_2)\cdot(\mathbf{k}_1-\mathbf{k}_2)} \right.$$

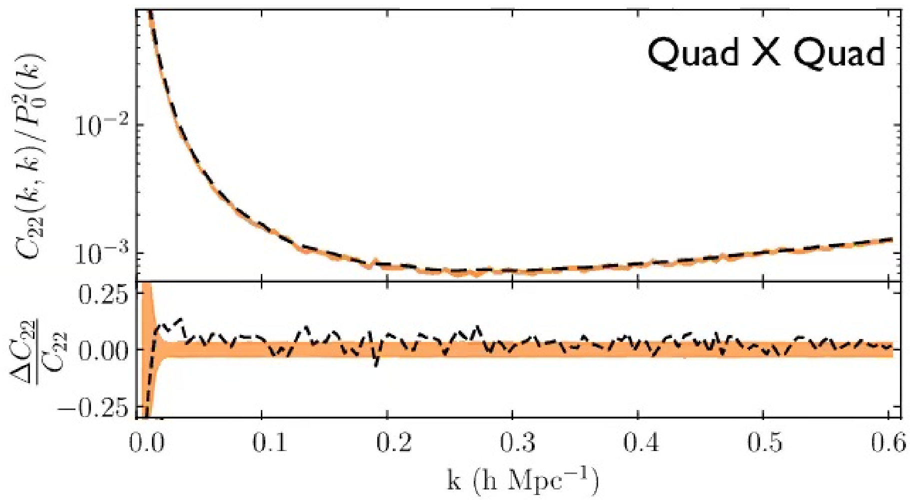
$$\left. \times \mathcal{L}_{\ell_1}(\hat{\mathbf{x}}_1\cdot\hat{\mathbf{k}}_1)\mathcal{L}_{\ell_1'}(\hat{\mathbf{x}}_1\cdot\hat{\mathbf{k}}_1)\mathcal{L}_{\ell_2'}(\hat{\mathbf{x}}_2\cdot\hat{\mathbf{k}}_2)\left[\mathcal{L}_{\ell_2}(\hat{\mathbf{x}}_2\cdot\hat{\mathbf{k}}_2)+\mathcal{L}_{\ell_2}(\hat{\mathbf{x}}_1\cdot\hat{\mathbf{k}}_1)\right]\right\}$$

{ Computed from survey random catalog }

See also:
Li et al. 19

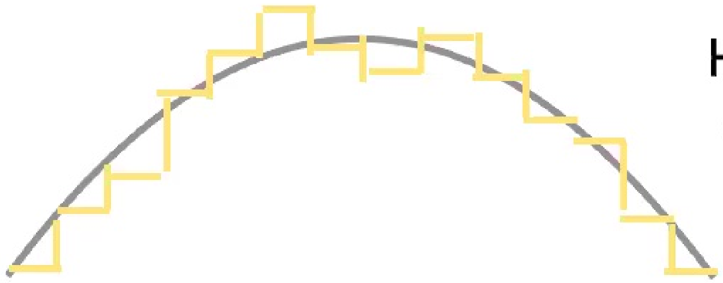
Challenge II:
Analytic modeling in the non-linear regime



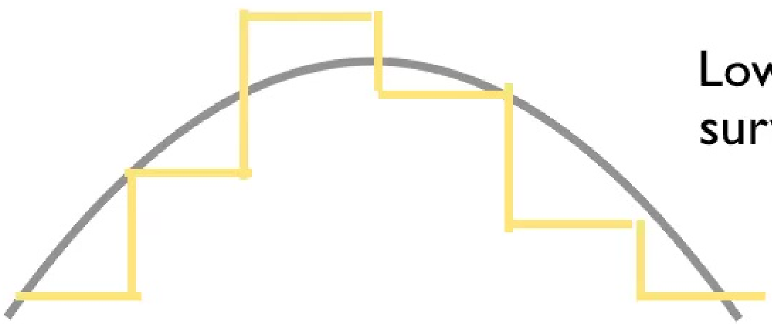


Analytic covariance works
very well at high-k.
WHY!?

Why does analytic work in the non-linear regime?



High $\bar{n}$ survey



Low $\bar{n}$ survey

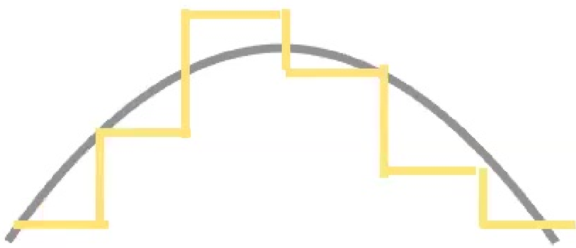
- Poisson fluctuations dominate the error bars at small scales:

↓

✓ Can be well modeled analytically

Why does analytic work in the non-linear regime?

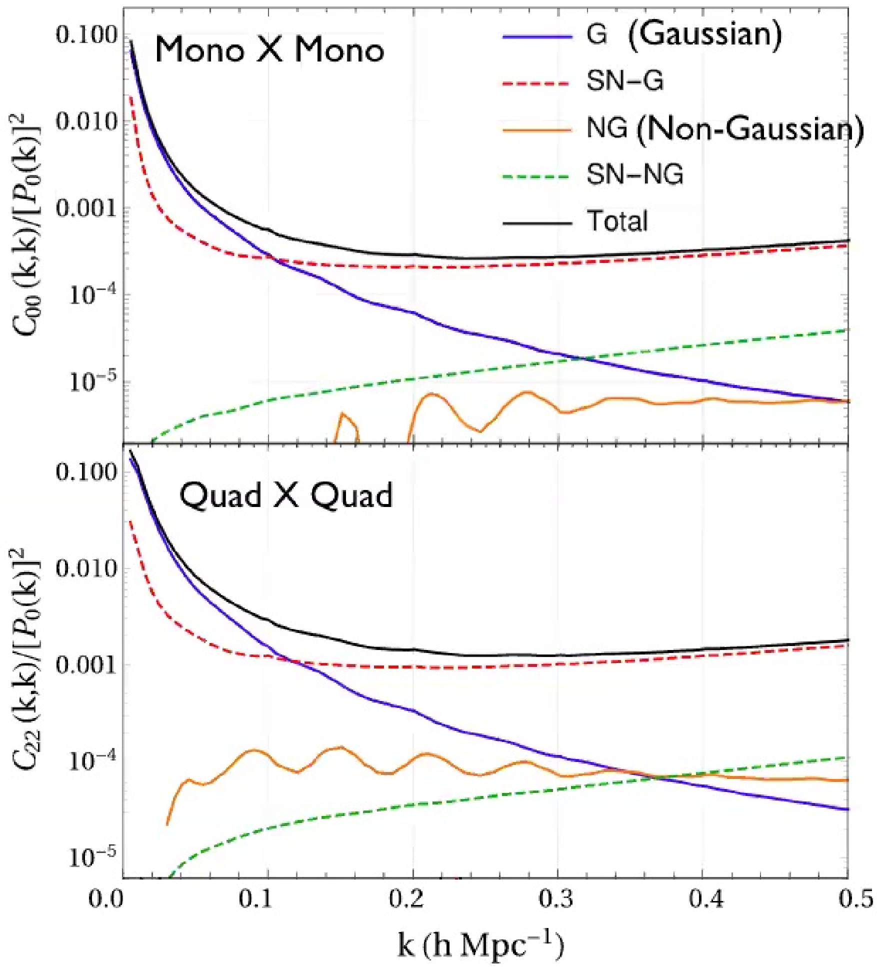
Shot noise (dashed) dominates at high-k



$$\mathbf{C}(k_1, k_2) = \langle \delta(k_1)\delta(-k_1)\delta(k_2)\delta(-k_2)\rangle - \langle \delta(k_1)\delta(-k_1)\rangle\langle\delta(k_2)\delta(-k_2)\rangle$$

$$\mathbf{C}^{\mathrm{G}}(k_1, k_2) \simeq 2\,\langle \delta(k_1)\delta(-k_2)\rangle\,\langle\delta(k_2)\delta(-k_1)\rangle$$

$$\mathbf{C}^{\mathrm{NG}}(k_1, k_2) = \langle \delta(k_1)\delta(-k_2)\delta(k_2)\delta(-k_1)\rangle_c$$



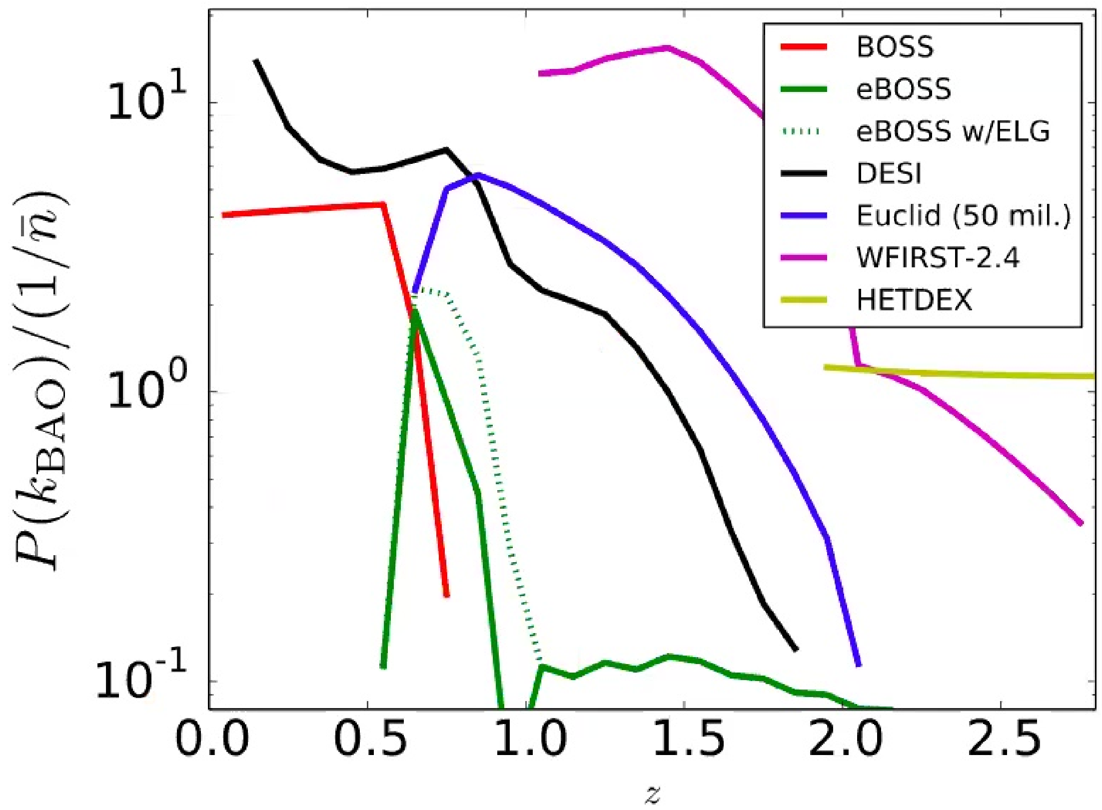
23

Analytic cov. should work at small scales for upcoming surveys

- Shot noise level of upcoming surveys is comparable to BOSS

↓

✓ Shot noise will dominate covariance for upcoming surveys



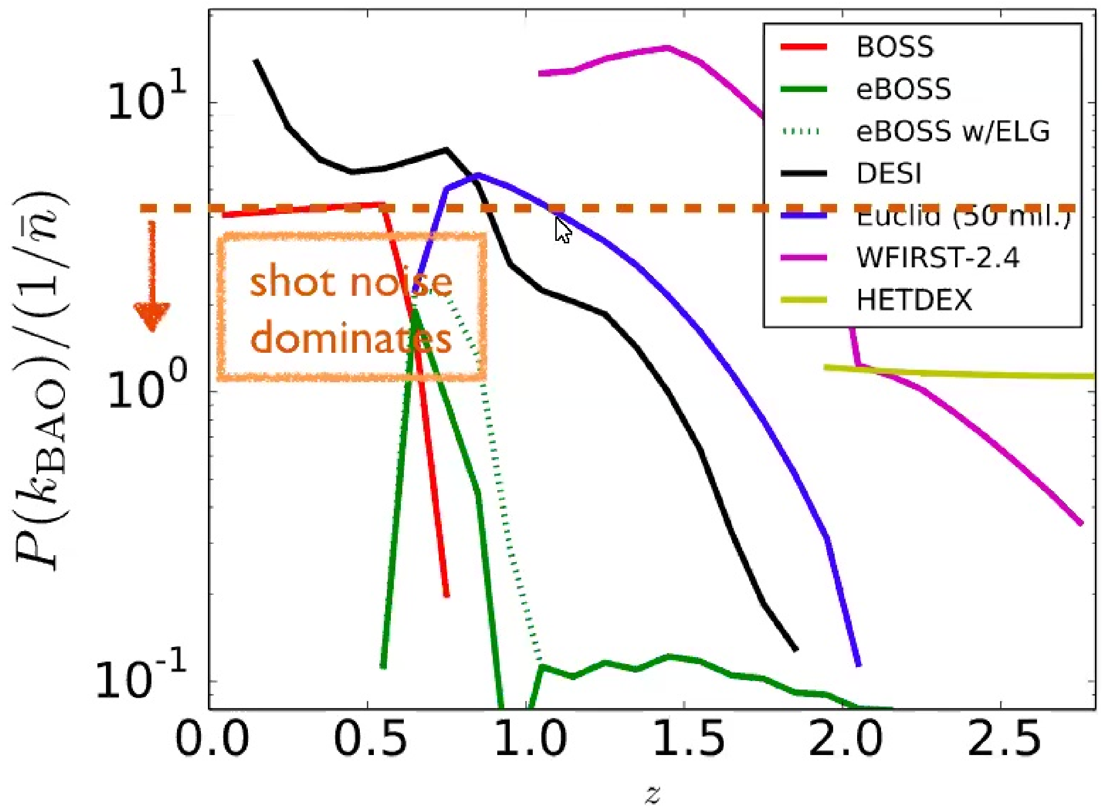
Font-Ribera et al. 2014

Analytic cov. should work at small scales for upcoming surveys

- Shot noise level of upcoming surveys is comparable to BOSS

↓

✓ Shot noise will dominate covariance for upcoming surveys



Font-Ribera et al. 2014

Analytic covariance is crucial for going beyond a 2-point analysis

- Number of mock simulations: $\mathcal{O}(1000)$
- For low sampling noise:

  size of data vector $\ll$ no. of mocks

Analytic covariance is crucial for going beyond a 2-point analysis

✓ Number of k-bins in power spectrum ~100

- Number of mock simulations: $\mathcal{O}(1000)$
- For low sampling noise:

size of data vector $\ll$ no. of mocks

27

Analytic covariance is crucial for going beyond a 2-point analysis

✓ Number of k-bins in power spectrum ~100

◉ Number of triangles in bispectrum (3-pt) ~6000
- Bottleneck for BOSS
  (Gil-Marin et al 17 could only use ~800 triangles)

- Number of mock simulations: $\mathcal{O}(1000)$
- For low sampling noise:

size of data vector $\ll$ no. of mocks

Analytic covariance is crucial for going beyond a 2-point analysis

✓ Number of k-bins in power spectrum ~100

◉ Number of triangles in bispectrum (3-pt) ~6000
- Bottleneck for BOSS
  (Gil-Marin et al 17 could only use ~800 triangles)

➡ Also important for other areas with high-dimensionality of the covariance matrix:
for e.g., 3x2pt analysis in photometric surveys
or combining cluster counts with correlation fns

- Number of mock simulations: $\mathcal{O}(1000)$
- For low sampling noise:

size of data vector $\ll$ no. of mocks

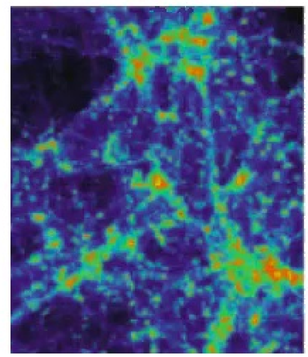
Neutral hydrogen (HI) from dark matter with machine learning and symbolic regression

DW, Paco Villaescusa-Navarro,
Shirley Ho & Laurence Perrault-Levasseur
(aXiv:2007.10340 & aXiv:2011.xxxx)

Emulation of hydro sims for future surveys

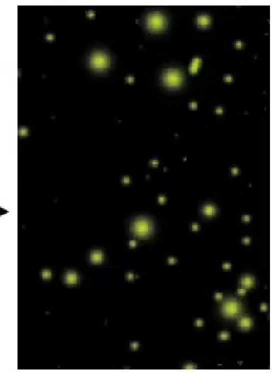
- Volume of upcoming surveys:
  ~$\mathcal{O}$ (100 Gpc³)
- Hydro sims are expensive:
  ~10 million CPU hours
  for (0.001 Gpc³)

N-body (DM)



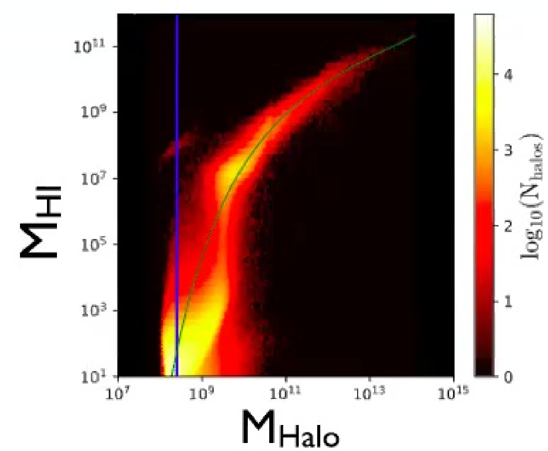
Quickly →

HI (neutral hydrogen)



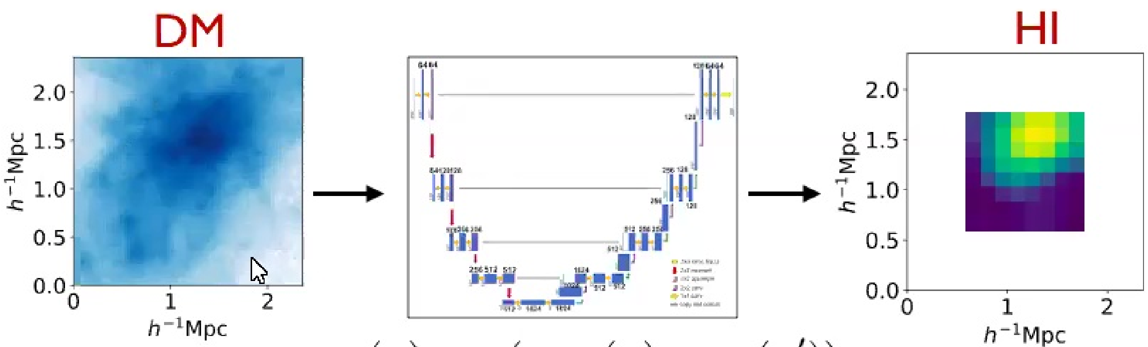
HOD:
(Halo
Occupation
Distribution)

- Identify DM halos
- Fill HI using:
  $M_{HI} = f\,(\,M_{halo}\,)$
- assembly history, environmental info. neglected

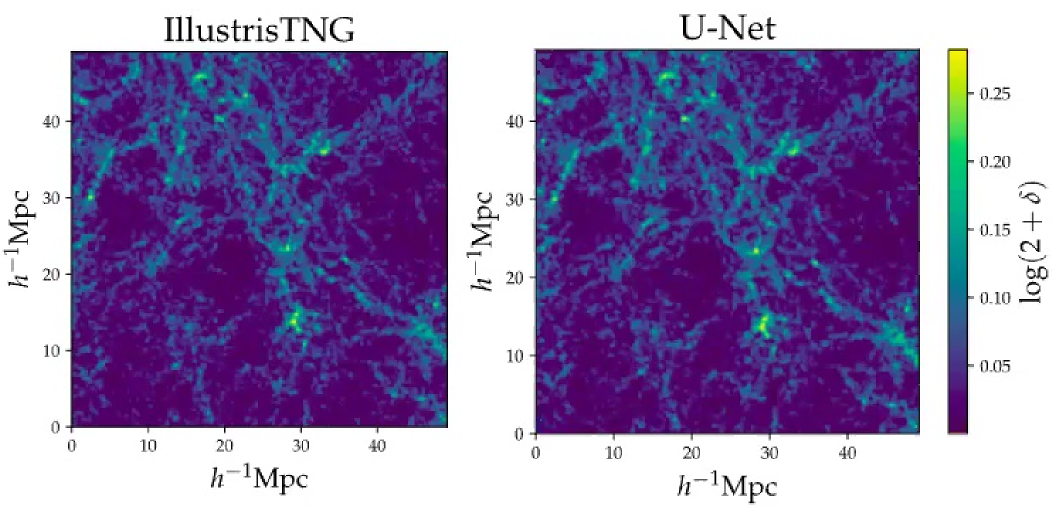


Villaescusa-Navarro et al. 2019

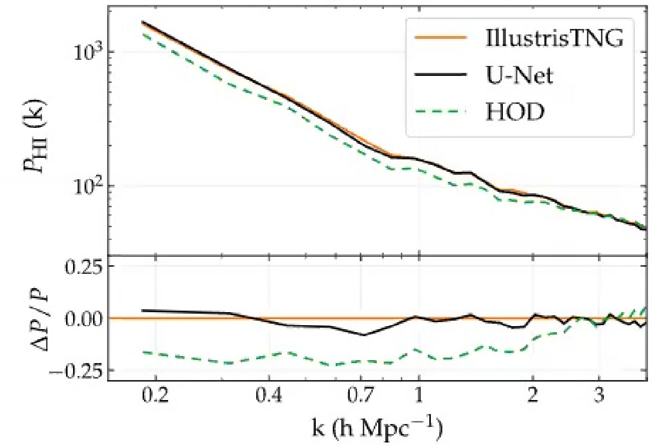
Neural networks as emulators



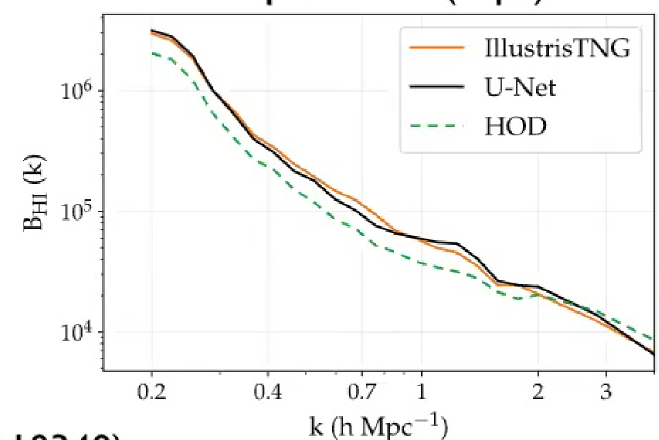
$$\rho_{\rm HI}(\mathbf{x}) = g\left(\rho_{\rm DM}(\mathbf{x}), \rho_{\rm DM}(\mathbf{x}')\right)$$



Power spec. (2 pt)



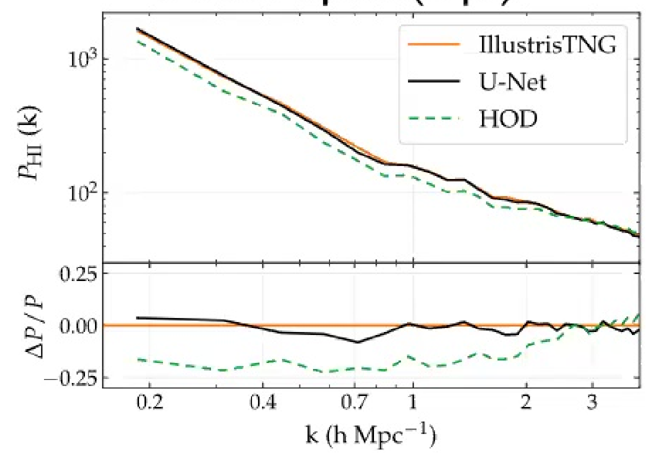
Bispectrum (3 pt)



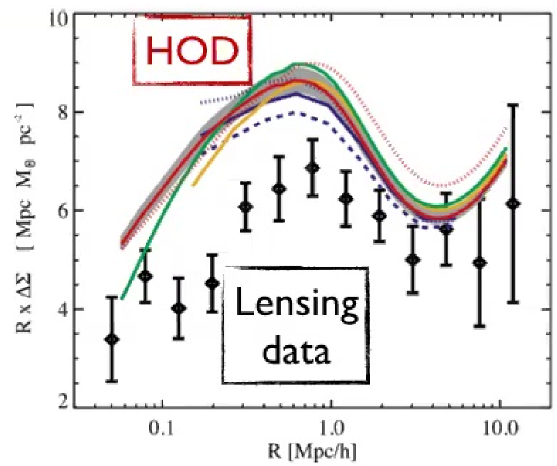
(arXiv:2007.10340)

Can we interpret what the network has learnt?

Power spec. (2 pt)



HI

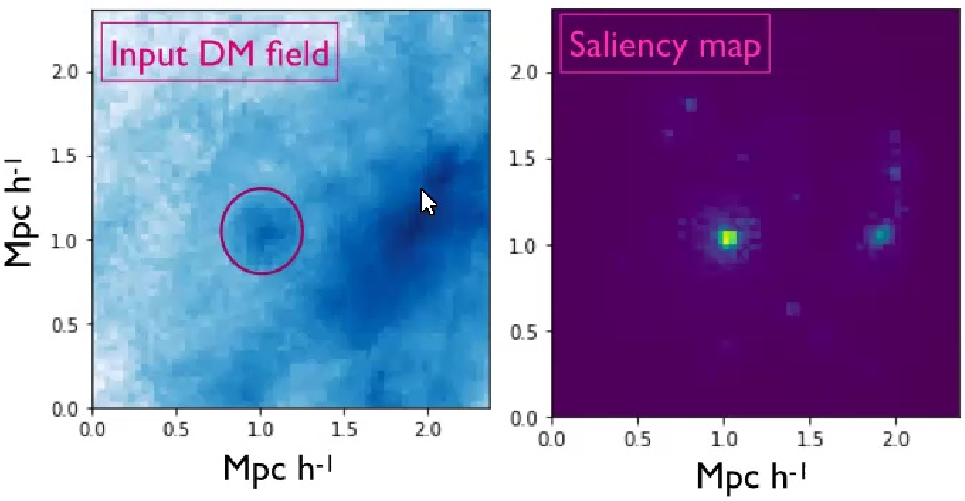


Galaxies from SDSS-CMASS

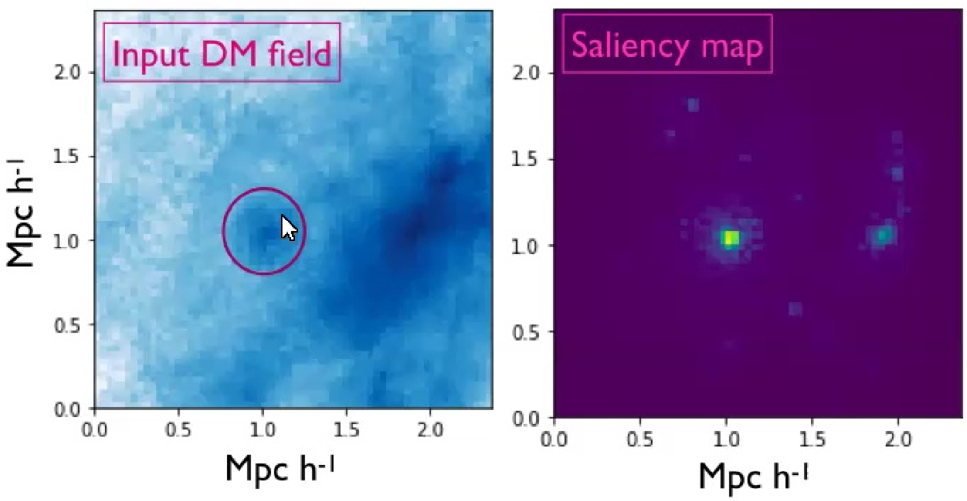
"Lensing is low" problem (Leauthaud et al. 17)

33

Network has learnt to include env. info



Network has learnt to include env. info



- Network lowers M_{HI} in a cluster-like environment

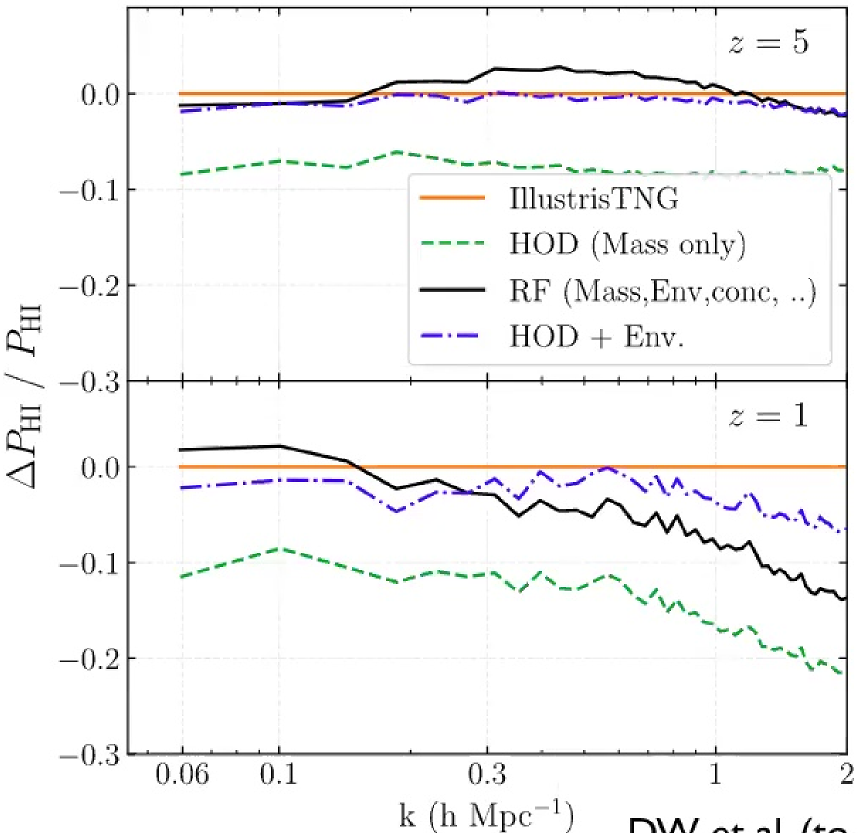
  (ram pressure stripping)

Modeling the halo HI mass with symbolic regression

HI mass of halo $= f$ (halo mass, ?)

- Halo env. overdensity (R)
- Env. anisotropy (R)
- Halo concentration
- Halo spin
- Halo assembly history
- Halo shape
- Velocity dispersion anisotropy
- …., …., ….

Results: Modeling the halo HI mass with symbolic regression



DW et al. (to appear)

Env. overdensity at 0.5 Mpc

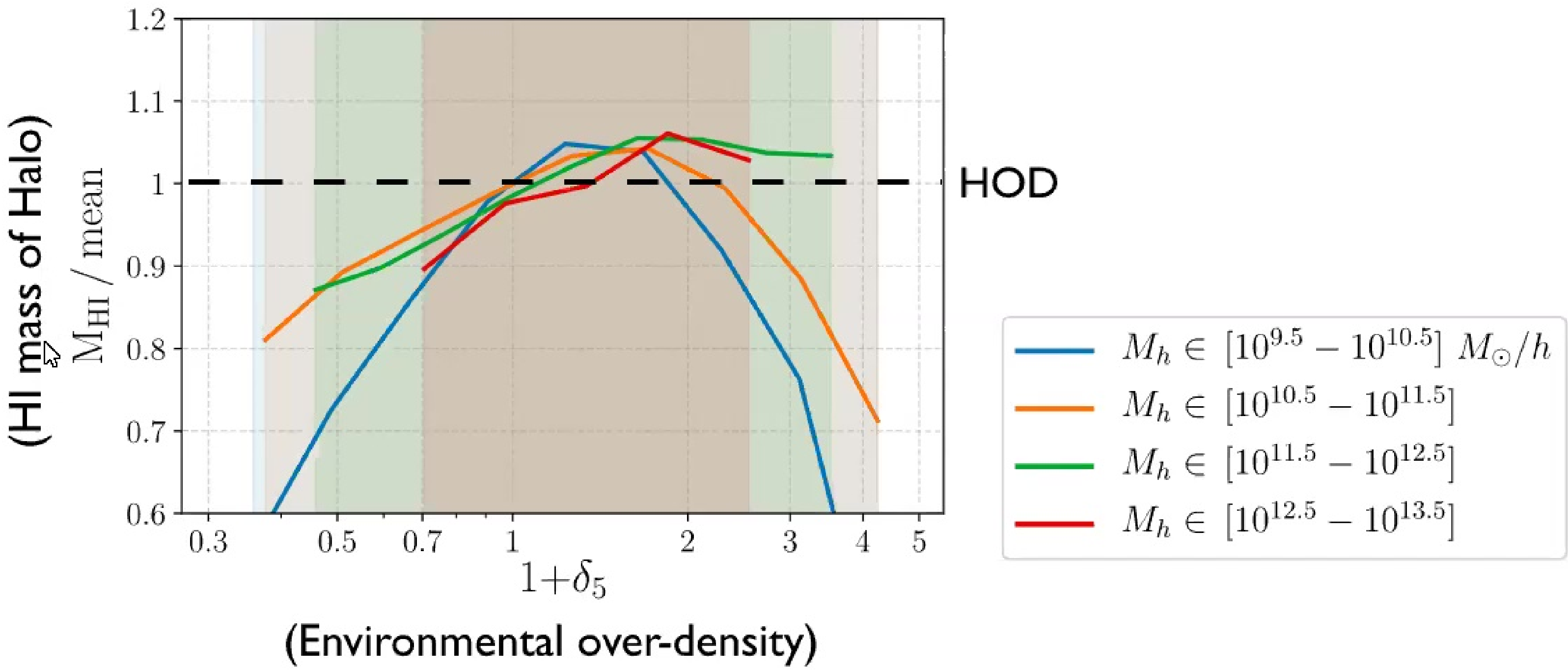
Env. anisotropy

$$z = 5: \quad \frac{M_{HI}}{M_{HOD}} = 0.95 + \alpha'_{0.5}\,\delta'_{0.5}\,(\alpha'_{0.5} + \delta'_{0.5})$$

$$z = 1: \quad \frac{M_{HI}}{M_{HOD}} = 0.8 + 1.4\,\alpha'_{0.5} m_{10} - 0.6\,(\alpha'^2_{0.5} m^2_{10} + \alpha'_{0.5}\delta'_5)$$

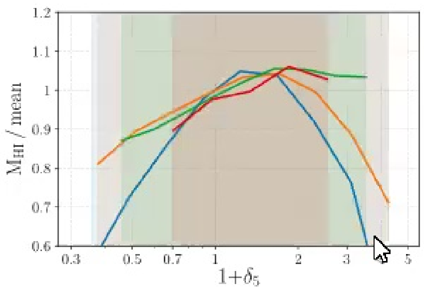
https://github.com/MilesCranmer/PySR

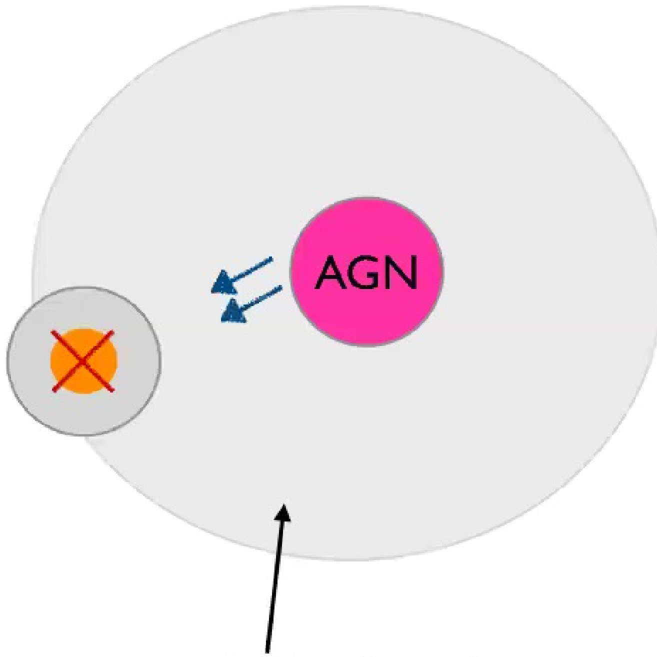
Understanding the effect of halo environment on HI: Don't baryons just follow dark matter?



DW et al. (to appear)

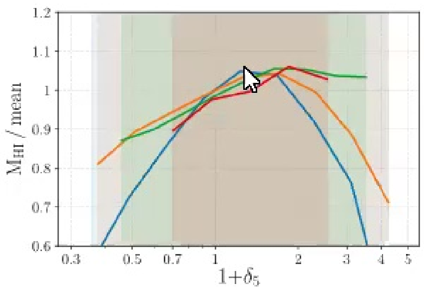
Understanding the effect of halo environment on HI: Don't baryons just follow dark matter?

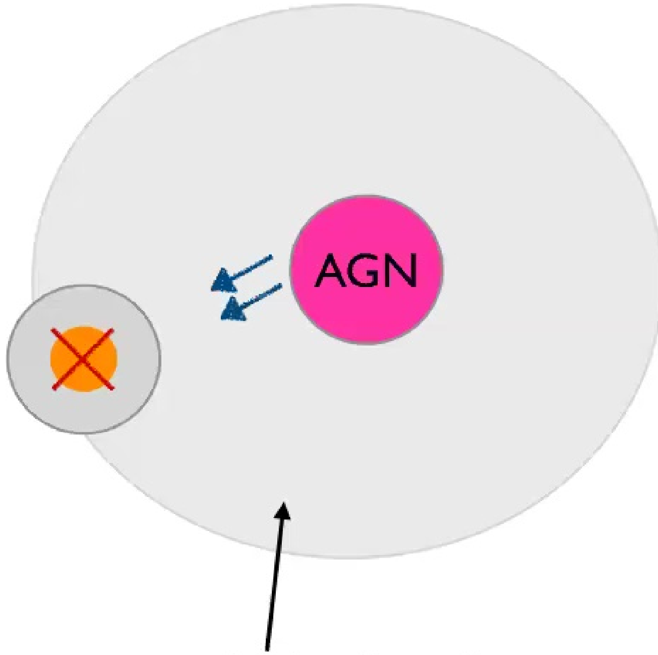




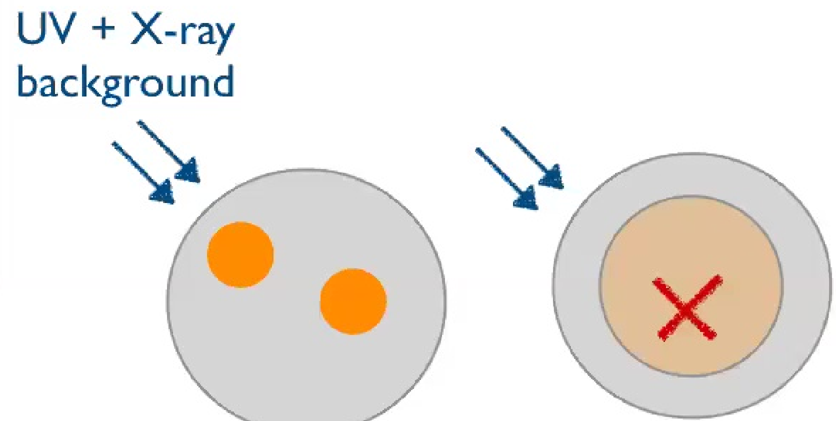
Denser env.: Ionized medium
(ram pressure stripping)

Understanding the effect of halo environment on HI:
Don't baryons just follow dark matter?





Denser env.: Ionized medium
(ram pressure stripping)



Denser env.
(more mergers)

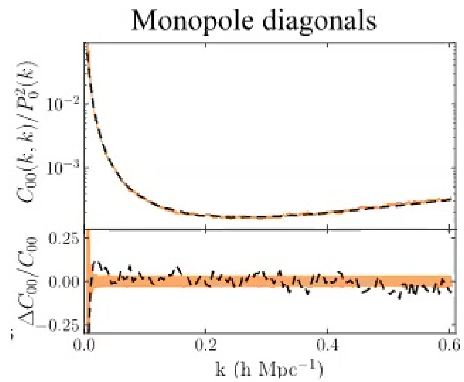
Summary

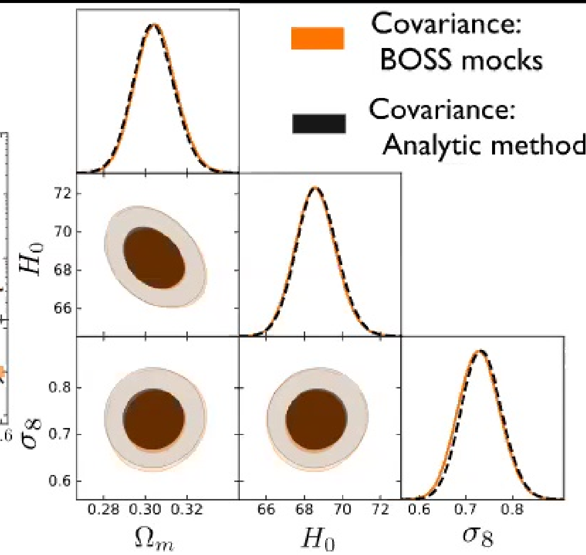
★ Analytic covariance is an excellent alternative to mock simulations for upcoming spectroscopic surveys

1. Very good agreement with the state-of-the-art mocks up to non-linear scales

2. Immense computational speedup (~ 10^4)

3. No sampling noise effects





Summary

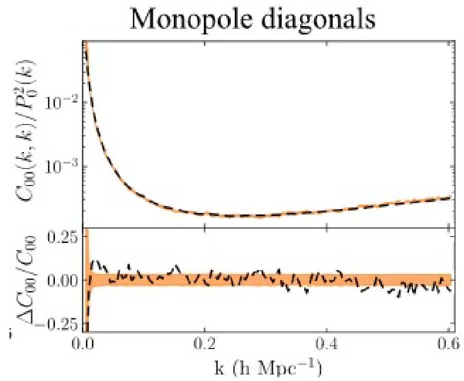
★ Analytic covariance is an excellent alternative to mock simulations for upcoming spectroscopic surveys

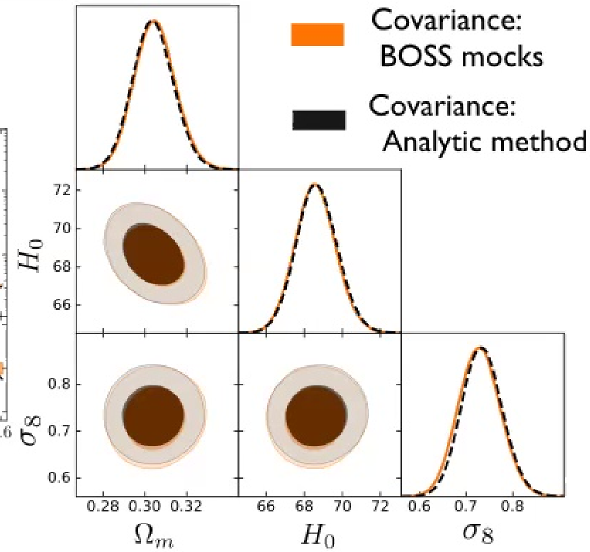
1. Very good agreement with the state-of-the-art mocks up to non-linear scales

2. Immense computational speedup (~ 10^4)

3. No sampling noise effects

★ Symbolic regression can be used to model assembly bias from hydro sims. and guide its detection in survey data





$$\frac{M_{\rm HI}}{M_{\rm HOD}} = 1.1 + 0.8\,\alpha_{0.5}\,(\log_{10}(M_{\rm halo}) - 10.6)$$