Analysing LSS datasets with COBRA

Thomas Bakx (PhD Candidate at Utrecht University)

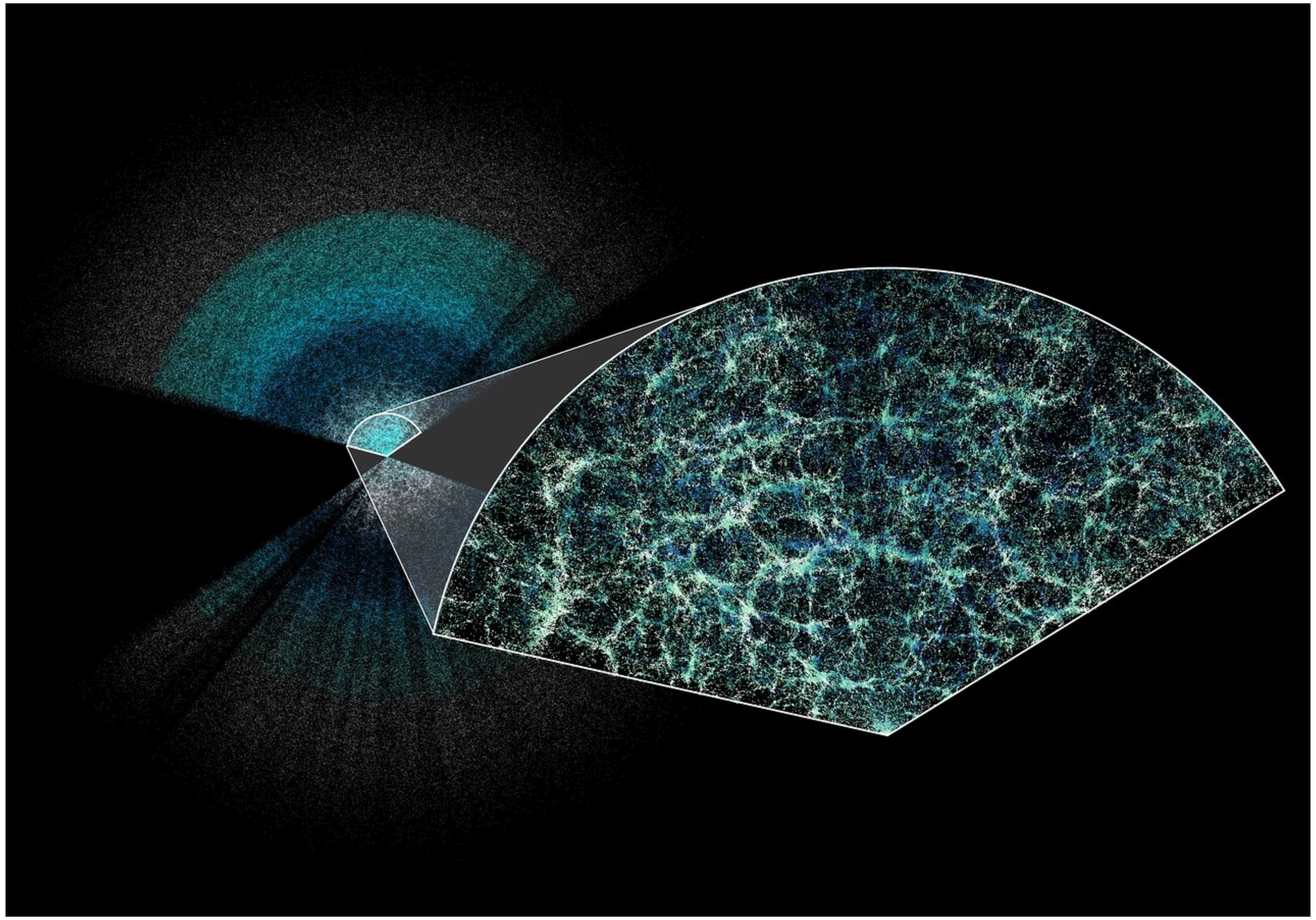
TB, Vlah, Chisari (2024)

GGI - 09/2025

The ‘Large Scale Structure Era’

- DESI (Euclid, SphereX, …)
- 40 million spectra
- Powerful statistics of galaxy clustering
- Test of concordance ΛCDM model
    - Expansion history: $\Omega_m, H_0, w(z)$
    - Growth rate: $f\sigma_8$
- Also use info from *scale-dependence*

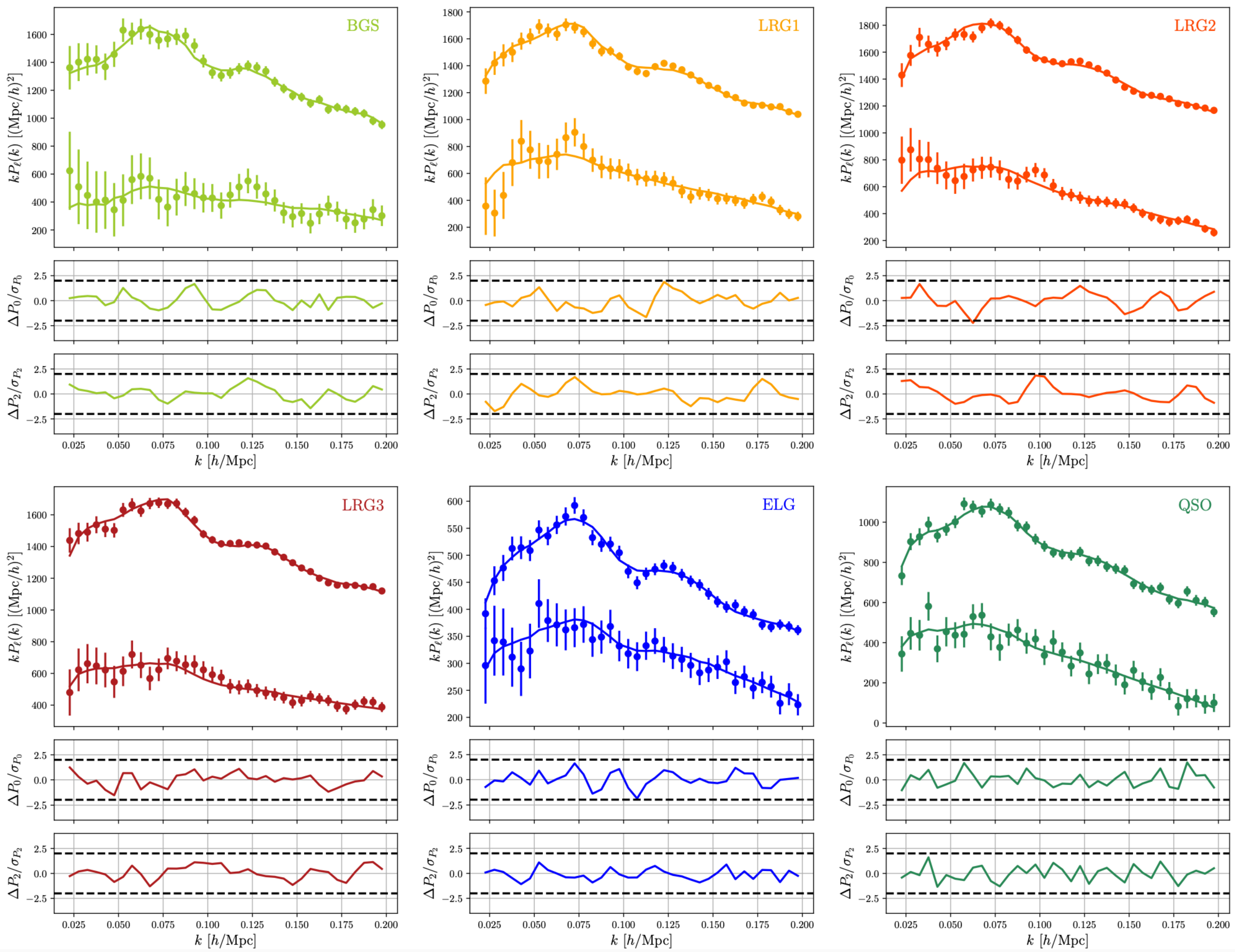
d’Amico++ (2019)
Ivanov++ (2019)
Brieden++ (2021)



What do we measure?

- Correlations between galaxies
- Use *summary statistics* to compress

DESI Collaboration (2411.12021)



$$\langle \delta_g(\mathbf{k}_1)\delta_g(\mathbf{k}_2)\rangle = (2\pi)^3 P_{gg}(\mathbf{k})\delta^D(\mathbf{k}_1+\mathbf{k}_2)$$

$$\langle \delta_g(\mathbf{k}_1)\delta_g(\mathbf{k}_2)\delta_g(\mathbf{k}_3)\rangle = (2\pi)^3 B_{ggg}(\mathbf{k})\delta^D(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3)$$

And so on…

Perturbation Theory for LSS (in brief)

$$\delta_g(\mathbf{x},\tau) = \sum_{\mathcal{O}} b_{\mathcal{O}}(\tau)\mathcal{O}(\mathbf{x},\tau) = b_1\delta + b_{\delta^2}\delta^2 \ldots$$

Baumann, Carrasco, Hertzberg, Senatore, Zaldarriaga, Pajer, Schmidt, Vlah, ...

- Galaxy overdensity δ_g is a function of the dark matter overdensity δ: 'Taylor expand'
- Cannot push to small separations, but at least on large scales it is robust
- Most general expansion for the galaxy overdensity that obeys required symmetries
- Free parameters $b_{\mathcal{O}}$ describe our lack of knowledge of galaxy formation

Current state-of-the-art (1)

Class-PT / OneLoop

velocileptors

PyBird

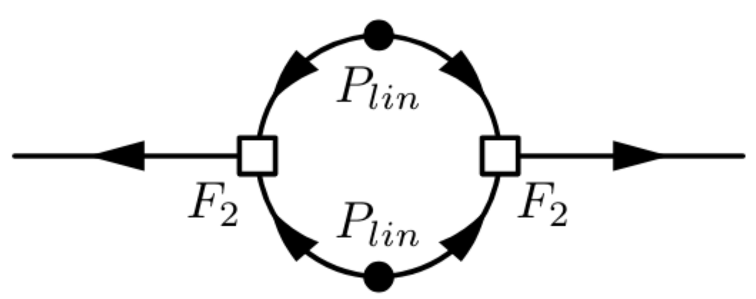
...

- For given set of parameters $H_0, \Omega_m, \sigma_8 \ldots$, compute $\langle \delta_g \delta_g \rangle, \langle \delta_g \delta_g \delta_g \rangle, \ldots$

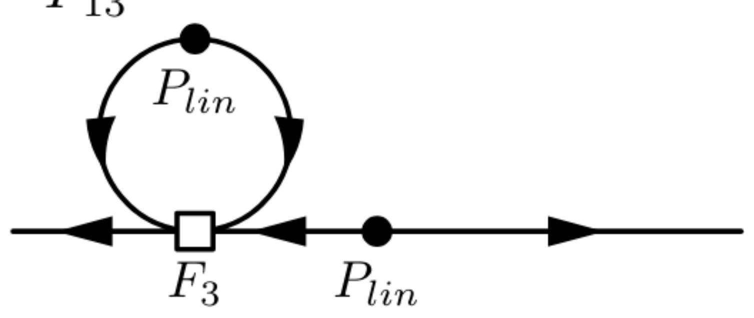
- Power spectrum at NLO: $\delta_g = \delta_g^{(1)} + \delta_g^{(2)} + \delta_g^{(3)}$

$$\langle \delta_g \delta_g \rangle = \langle \delta_g^{(1)} \delta_g^{(1)} \rangle + \langle \delta_g^{(2)} \delta_g^{(2)} \rangle + \langle \delta_g^{(1)} \delta_g^{(3)} \rangle + \langle \delta_g^{(3)} \delta_g^{(1)} \rangle$$

P_{22}



P_{13}



Chudaykin, Ivanov (2023)

Simon++ (2023)

Cabass++ (2022)

Simonovic++ (2018)

Dataset	$\omega_{\rm cdm}$	h	$\ln(10^{10} A_s)$	n_s	Ω_m	σ_8
$P_\ell(k)$	$0.139^{+0.011}_{-0.015}$	$0.699^{+0.015}_{-0.017}$	$2.63^{+0.15}_{-0.16}$	$0.883^{+0.076}_{-0.072}$	$0.333^{+0.019}_{-0.020}$	$0.704^{+0.044}_{-0.049}$

- Few % constraints on H_0, Ω_m, σ_8 + also potential for beyond ΛCDM physics

- 10 parameters, give or take

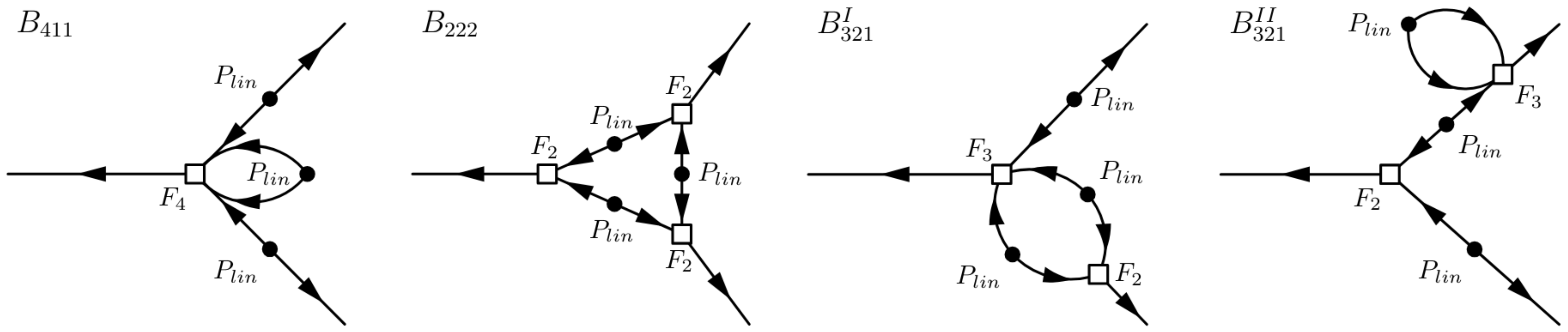
Philcox, Ivanov (2022)

Current state-of-the-art (2)

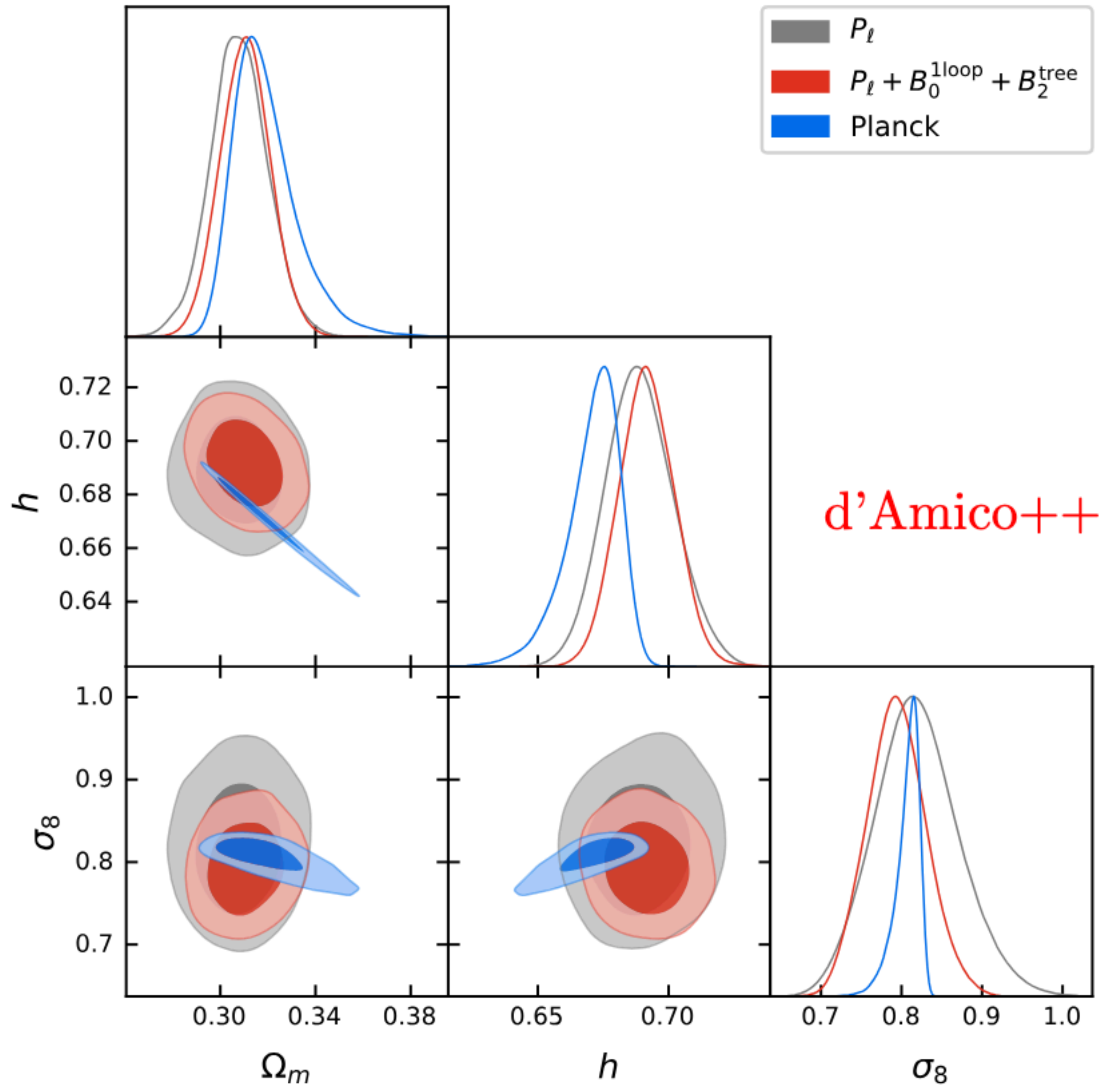
- Bispectrum at tree level or **one loop** (4th order PT)

Philcox++ (2022)

- $\delta_g = \delta_g^{(1)} + \delta_g^{(2)} + \delta_g^{(3)} + \delta_g^{(4)}$



Simonovic++ (2018)



d'Amico++ (2022)

Lu++ (2025)

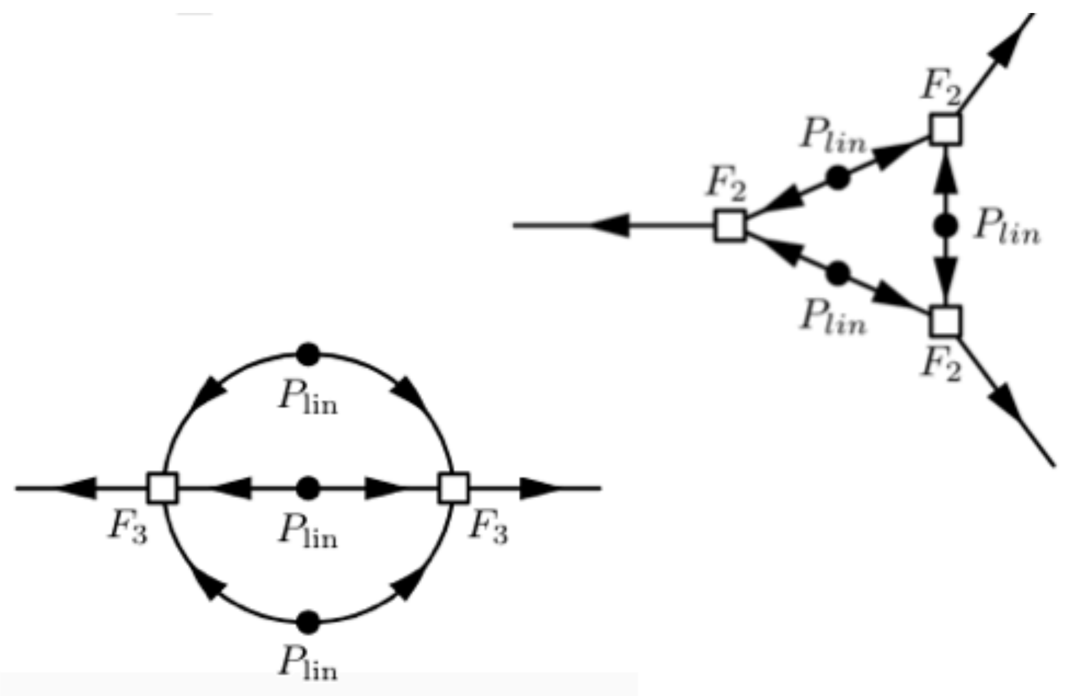
Racco++ (2024)

- Moderate reduction in error on σ_8, i.e. 30 percent, other parameters less
- Marginalize analytically over many nuisance parameters (impose Gaussian priors)
- $\sim$40 extra parameters (but not really degrees of freedom), hundreds of extra data points

Challenges

- Needs to be practically feasible in MCMC (preferably << 1s, *the faster the better*)

- <u>Theory predictions are difficult to evaluate:</u> integrals over $P_{\text{lin}}(q) = \langle \delta^{(1)} \delta^{(1)} \rangle'$



$$\equiv \quad 8 \int_{\mathbf{q}} F_2(\mathbf{q}, \mathbf{k}_1 - \mathbf{q}) F_2(\mathbf{k}_1 - \mathbf{q}, \mathbf{k}_2 + \mathbf{q}) F_2(\mathbf{k}_2 + \mathbf{q}, -\mathbf{q}) P_{\text{lin}}(q) P_{\text{lin}}(|\mathbf{k}_1 - \mathbf{q}|) P_{\text{lin}}(|\mathbf{k}_2 + \mathbf{q}|)$$

$$\equiv \quad 6 \int_{\boldsymbol{q}} \int_{\boldsymbol{p}} F_3(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{k} - \boldsymbol{q} - \boldsymbol{p}) F_3(-\boldsymbol{q}, -\boldsymbol{p}, \boldsymbol{q} + \boldsymbol{p} - \boldsymbol{k}) P_{\text{lin}}(q) P_{\text{lin}}(p) P_{\text{lin}}(|\boldsymbol{k} - \boldsymbol{q} - \boldsymbol{p}|)$$

Simonovic++ (2018)

$$F_2(\boldsymbol{q}, \boldsymbol{k} - \boldsymbol{q}) = \frac{5}{14} + \frac{3k^2}{28q^2} + \frac{3k^2}{28|\boldsymbol{k} - \boldsymbol{q}|^2} - \frac{5q^2}{28|\boldsymbol{k} - \boldsymbol{q}|^2} - \frac{5|\boldsymbol{k} - \boldsymbol{q}|^2}{28q^2} + \frac{k^4}{14|\boldsymbol{k} - \boldsymbol{q}|^2 q^2}.$$

- Linear power spectrum has a large dynamical range

- Integrals are e.g. 3D for one-loop bispectrum, or 5D for two-loop power spectrum

Carrasco++ (2013)

Blas++ (2013)

Efficient theory model evaluation (3)

McEwen++ (FAST-PT) (2016)

Schmittfull, Vlah (FFT-PT) (2017)

- *Cosmology dependence in the loop integral is only through $P_L(k)$*
- Try to find a decomposition where we *separate scale dependence from cosmology*

Anastasiou++ (2022)

$$P_L^{\Theta}(k) = \sum_{i=1}^{N_b} w_i(\Theta) v_i(k)$$

$$\begin{aligned}\mathcal{P}^{\Theta}_{\ell\text{-loop}} &= \text{const.} + \mathcal{S}^l[P_L^{\Theta}] \\ &+ \mathcal{S}^q[P_L^{\Theta}, P_L^{\Theta}] + \mathcal{S}^c[P_L^{\Theta}, P_L^{\Theta}, P_L^{\Theta}] \\ &+ \ldots\end{aligned}$$

$$\mathcal{S}_i^l = \mathcal{S}^l[v_i]$$

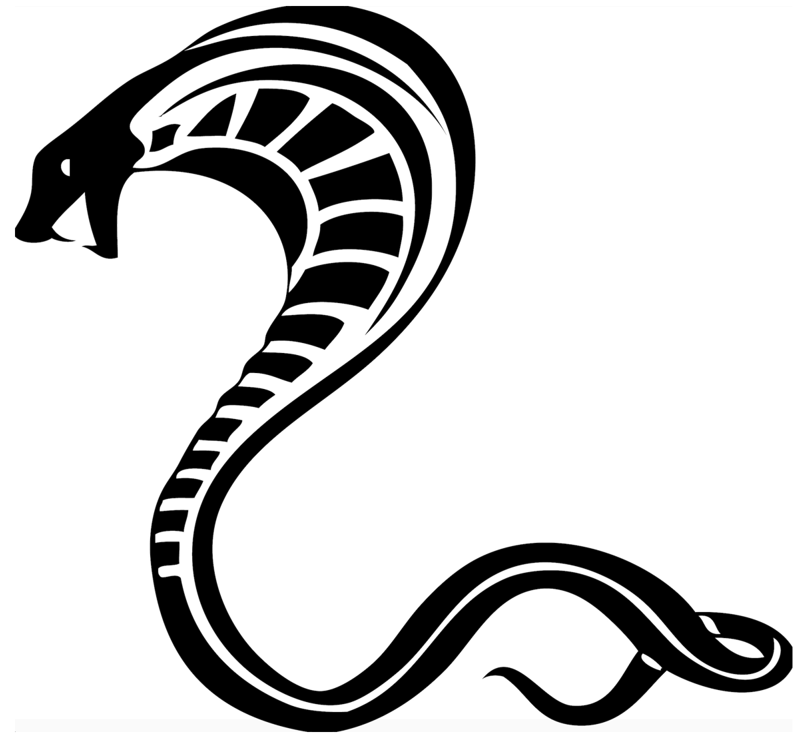
$$\mathcal{S}_{ij}^q = \mathcal{S}^q[v_i, v_j]$$

$$\mathcal{S}_{ijk}^c = \mathcal{S}^c[v_i, v_j, v_k]$$

- Pre-computation of integrals happens only once, **doesn't matter if it is numerically done**
- Works for *any N-point function at any order!* Precompute some small $\mathcal{S}_{ij}$, $\mathcal{S}_{ijk}$, … tensors
- So… what is the *minimal* way to separate scale and cosmology dependence?

Cosmology with Optimally compressed Basis for Rapid Approximation

COBRA (1)



TB, Vlah, Chisari (2024)

- For given range of cosmological parameters, want a *compressed* basis

$$P_L^{\Theta}(k) = \sum_{i=1}^{N_b} w_i(\Theta) v_i(k)$$

- For given N_b, what is the best possible choice for $v_i(k)$?
- Singular Value Decomposition (SVD) on templates in cosmological parameter space
- Use a regular grid for cosmology, log-binning in wavenumber, normalise by some fiducial $\bar{P}$

$$\hat{P}_L^{\Theta}(k_m) = P_L^{\Theta}(k_m)/\bar{P}(k_m)$$

$$\hat{P} \approx \hat{U}\Sigma\hat{V}^T$$

$$\hat{v}_i(k_m) = v_i(k_m)/\bar{P}(k_m)$$

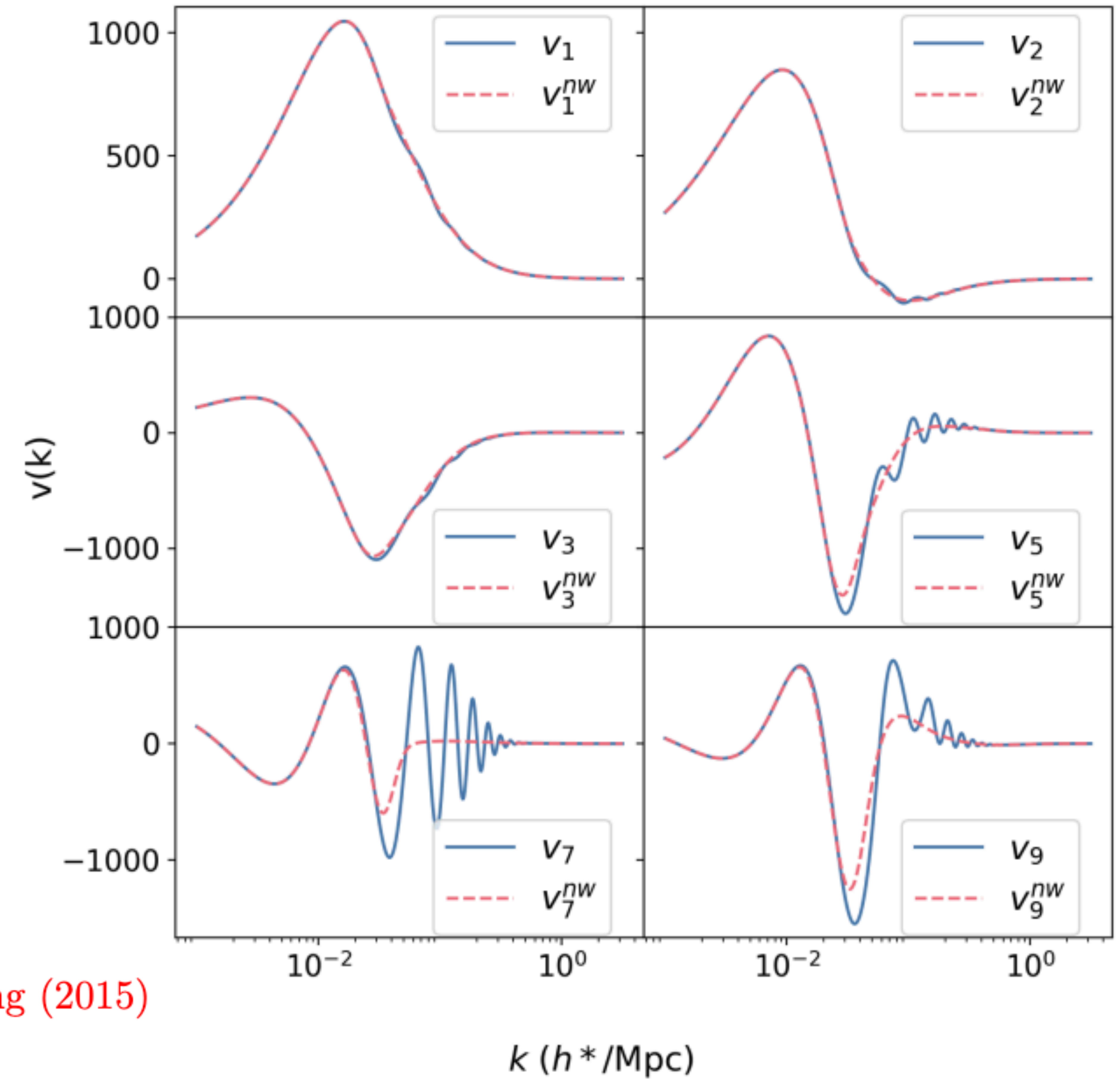
$$w_i(\Theta) = \sum_{m=1}^{N_k} \hat{v}_i(k_m)\hat{P}_L^{\Theta}(k_m).$$

- SVD: finds vectors such that projection into their subspace yields minimal mean square error

COBRA (2)

- Need <10 basis functions for very high precision
- First vector $v_1(k)$ is like Planck
- More complicated broadband, more wiggles for $i > 1$
- Use wiggle-no-wiggle-split for IR resummation
- Do smoothing for for each $v_i(k)$ individually

$$P_{\rm nw}^{\Theta}(k) = P_{\rm EH}^{*}(k)\frac{1}{\lambda(k)\sqrt{2\pi}}\int \mathrm{d}\log q\,\frac{P_L^{\Theta}(10^{\log q})}{P_{\rm EH}^{*}(10^{\log q})} \times \exp\left(-\frac{1}{2\lambda(k)^2}(\log q - \log k)^2\right).$$

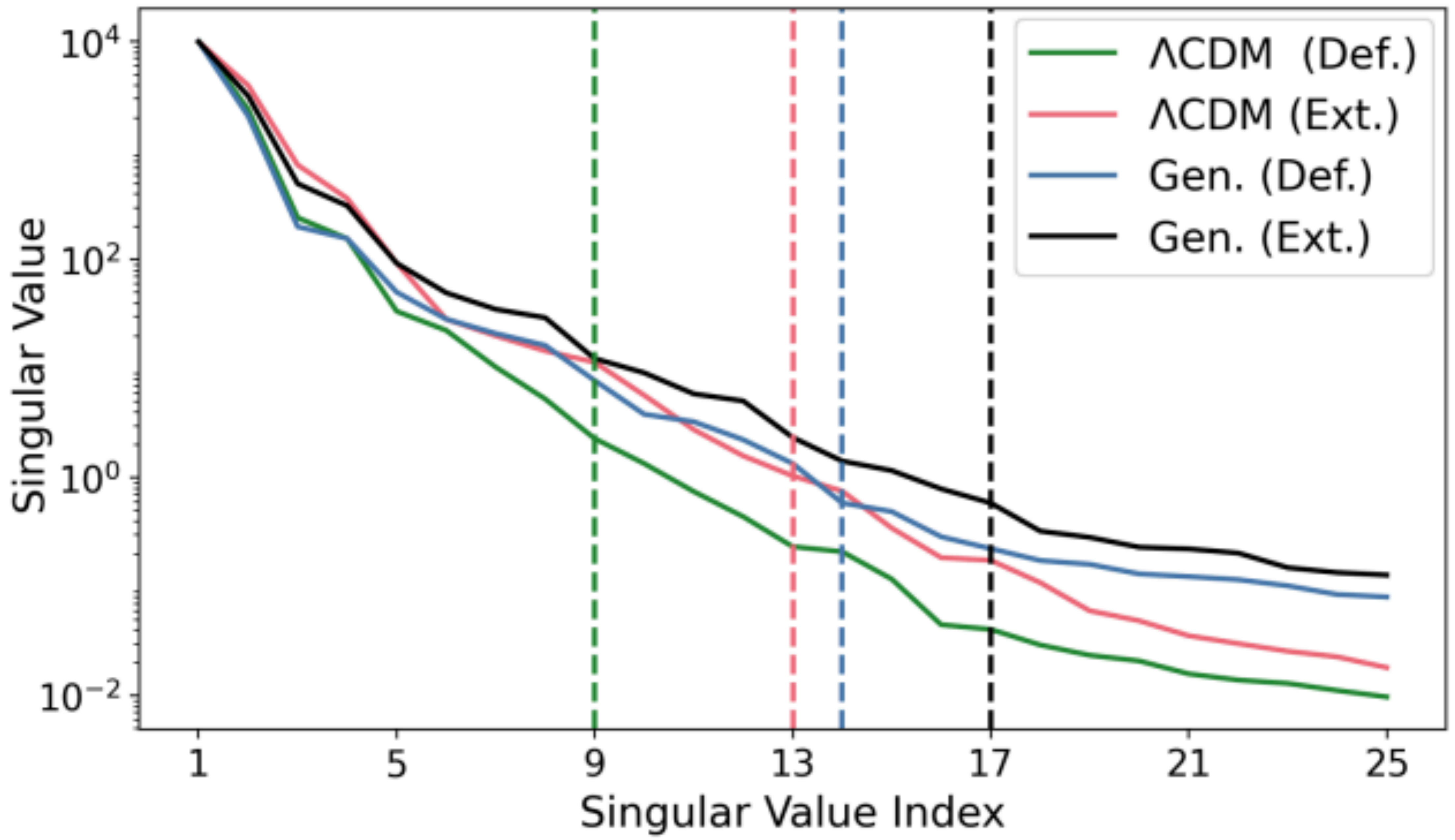


Vlah, Seljak, Chu, Feng (2015)

COBRA (3)

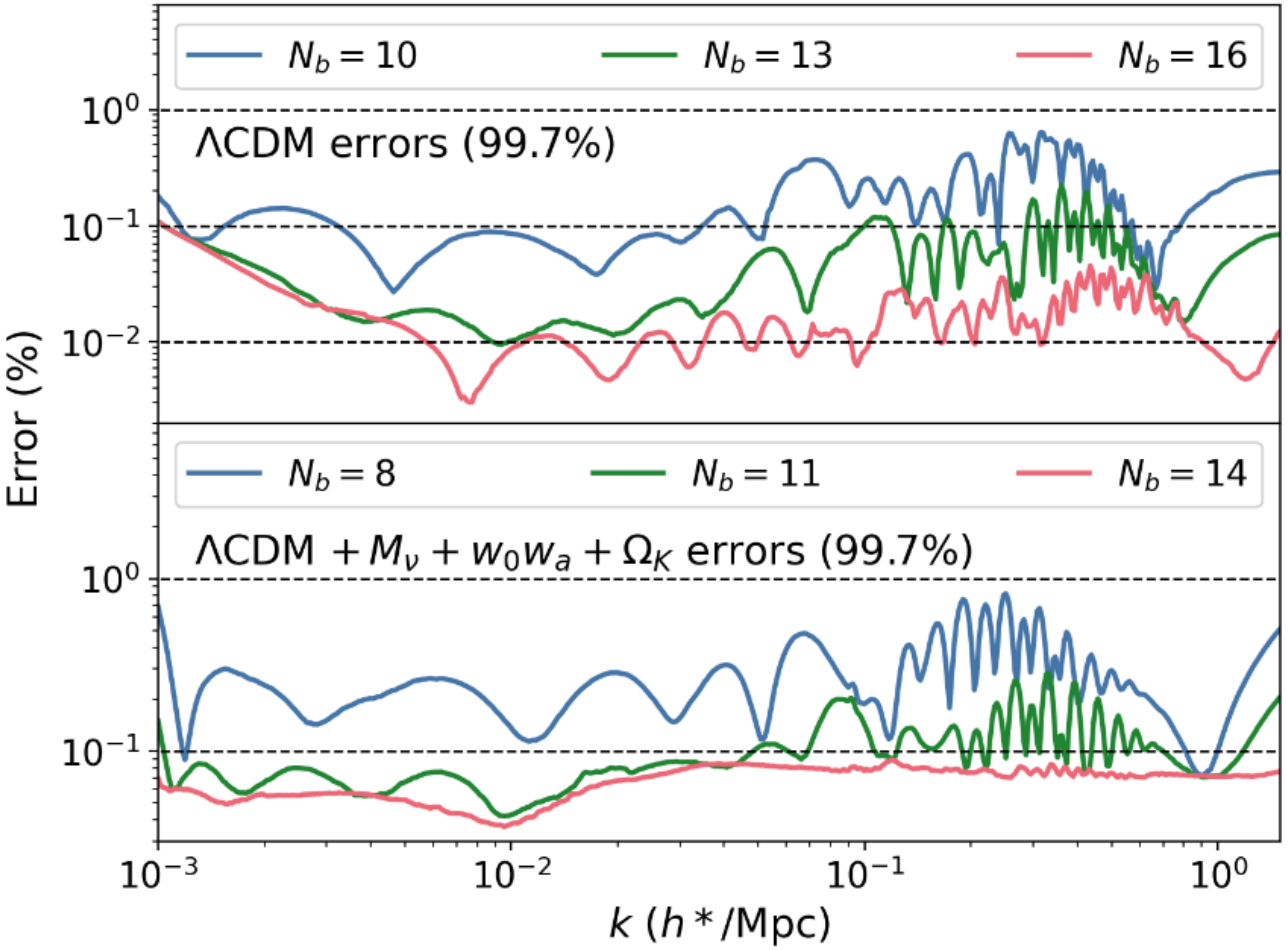
- Higher basis functions rapidly decrease in importance
- Loops with higher basis functions are less relevant

	Default		Extended	
Θ	Range	Grid size	Range	Grid size
ω_c	[0.095,0.145]	27	[0.08,0.175]	40
ω_b	[0.0202,0.0238]	18	[0.020,0.025]	20
n_s	[0.91,1.01]	12	[0.8,1.2]	20
$10^9 A_s$	-	$10^9 A_s^* = 2$	-	$10^9 A_s^* = 2$
h	[0.55,0.8]	$h^* = 0.7$	[0.5,0.9]	$h^* = 0.7$
z	[0.1,3]	$z^* = 0$	[0.1,3]	$z^* = 0$



	Default		Extended	
Θ	Range	Grid size	Range	Grid size
ω_c	[0.095,0.145]	20	[0.082,0.153]	30
ω_b	[0.0202,0.0238]	12	[0.020,0.025]	12
Ω_K	[-0.12,0.12]	12	[-0.2,0.2]	12
h	[0.55,0.8]	$h^* = 0.7$	[0.51,0.89]	$h^* = 0.7$
n_s	[0.9,1.02]	8	[0.81,1.09]	15
M_ν [eV]	[0,0.6]	12	[0,0.95]	15
w_0	[-1.25,-0.75]	12	[-1.38,-0.62]	15
w_a/w_+	[-0.3,0.3]	$w_a^* = 0$	[-1.78,-0.42]	$w_+^* = -1$
z	[0.1,3]	$z^* = 0$	[0.1,3]	$z^* = 0$
$10^9 A_s$	-	$10^9 A_s^* = 2$	-	$10^9 A_s^* = 2$

Linear power spectrum



Θ	Range
ω_c	[0.08,0.175]
ω_b	[0.020,0.025]
n_s	[0.8,1.2]
$10^9 A_s$	-
h	[0.5,0.9]
z	[0.1,3]

Θ	Range
ω_c	[0.095,0.145]
ω_b	[0.0202,0.0238]
Ω_K	[-0.12,0.12]
h	[0.55,0.8]
n_s	[0.9,1.02]
M_ν [eV]	[0,0.6]
w_0	[-1.25,-0.75]
w_a/w_+	[-0.3,0.3]
z	[0.1,3]
$10^9 A_s$	-

- Vectorised code, do 250 evaluations in 4ms for LCDM, or 30ms for the larger space

Emulating weights (1)

- Weights are given by orthogonal projection:

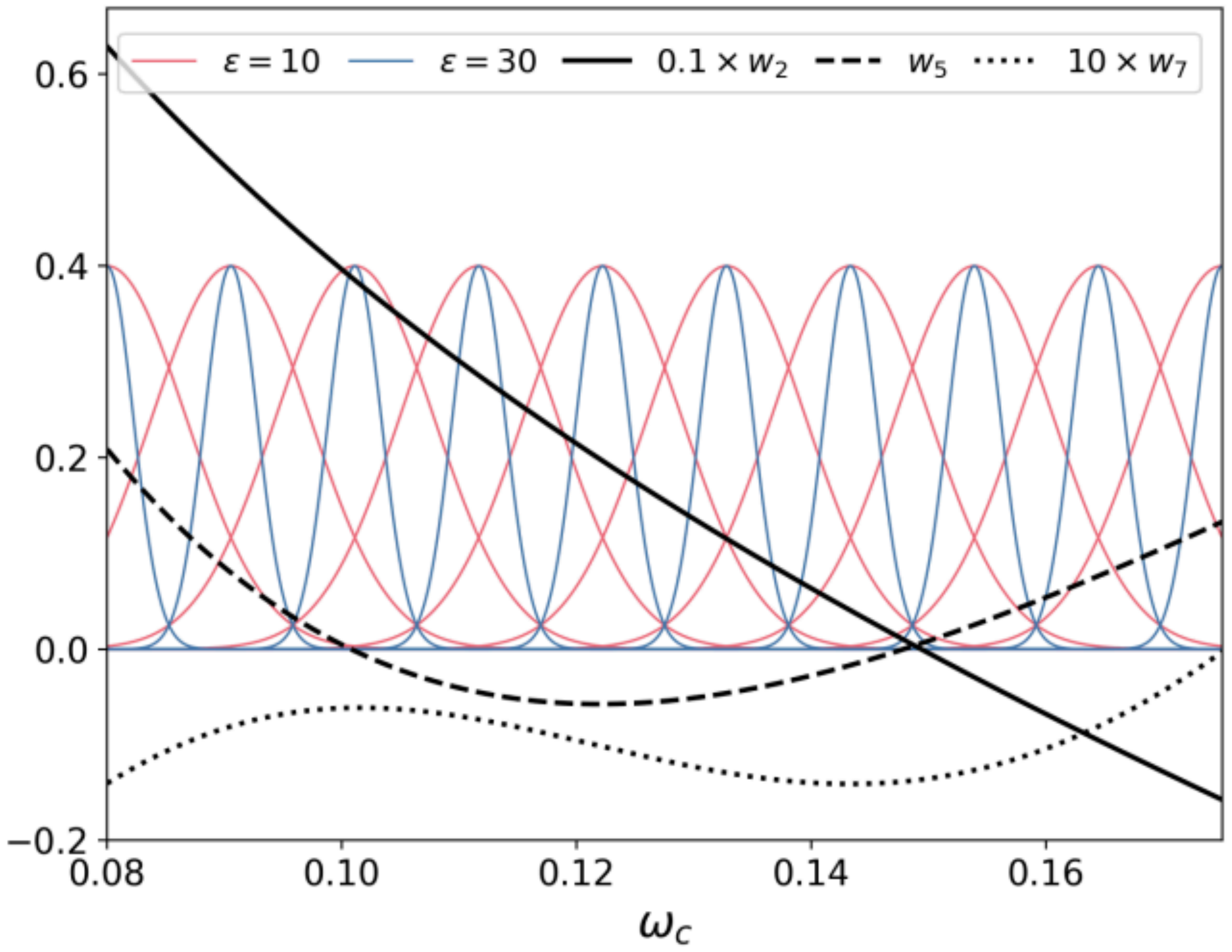
$$w_i(\Theta) = \sum_{m=1}^{N_k} \hat{v}_i(k_m)\hat{P}_L^{\Theta}(k_m).$$

- Would like to compute without calling Boltzmann solver
- In principle many possibilities, e.g. neural net, Gaussian process
- Here I used *radial basis functions:*

$$g(\Theta) = \sum_{b=1}^{N_n} c_b\phi_b(\Theta).$$

$$\phi_b(\Theta) = \phi(|\Theta - \Theta_b|)$$

$$\phi^{\epsilon}(\theta) = \exp(-\epsilon^2\theta^2).$$



Emulating weights (2)

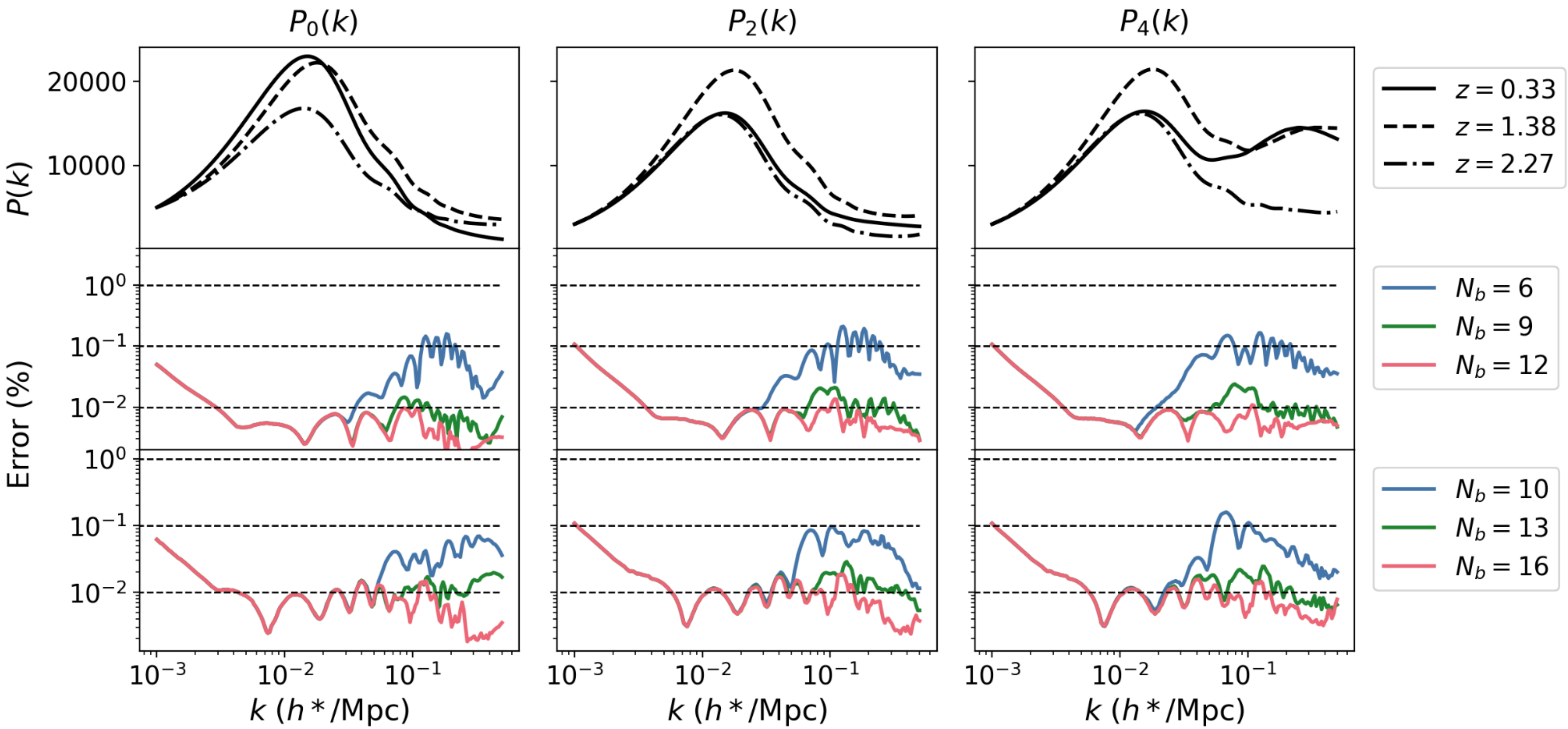
- If ϵ is large, peaks are narrow - need many points for interpolation
- Want to decrease ϵ, but when ϵ to zero, run into numerical instabilities (e.g $\epsilon = 0.1$)
- This can be circumvented… see paper for more math / details   Fasshauer, McCourt (2012)
- End result: can cover 9D parameter space with only 20,000 pts for < 0.1 percent errors
- Use ‘maximum-discrepancy’ covering set, e.g. Halton nodes (like Latin hypercube)
- Potentially useful when training points are expensive, i.e. nonlinear spectrum?

Galaxy power spectrum at NLO

Bonici, d'Amico, Bel, Carbone (2025)

Reeves, Zhang, Zheng (2025)

- 4000 evals / s, $\sim 10^{-4}$ precision (speedup $> 100\times$ vs e.g. velocileptors)

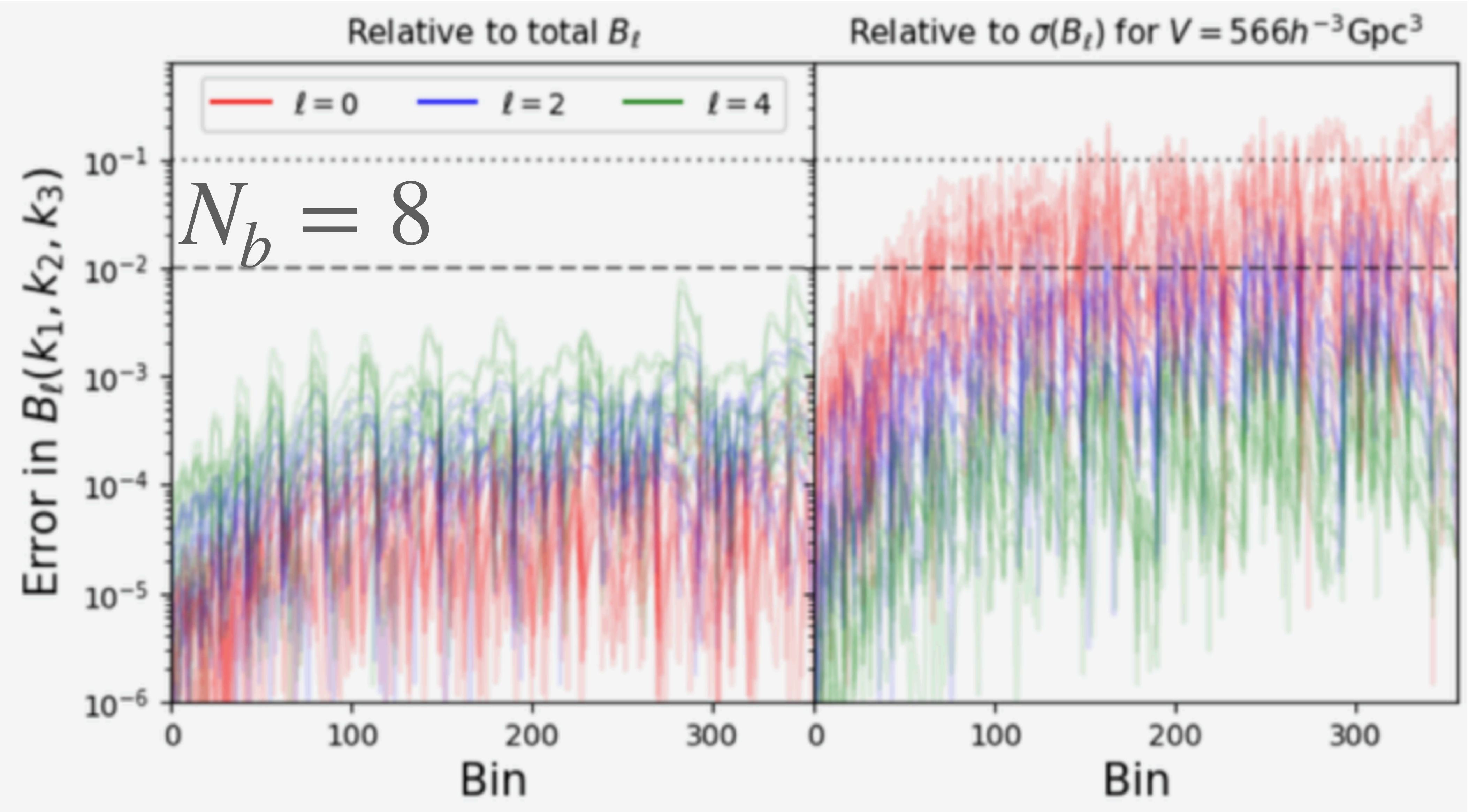


Universiteit Utrecht

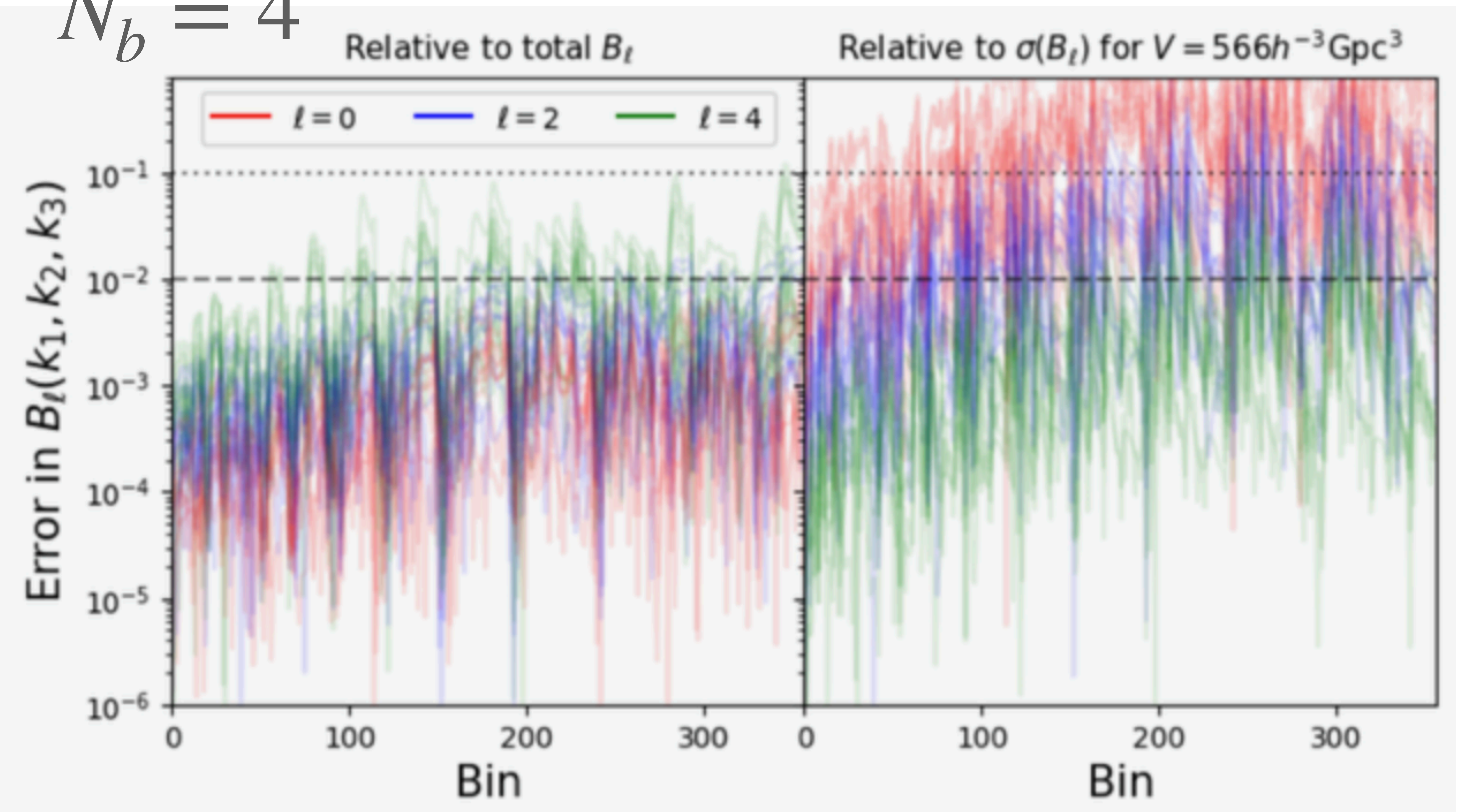
Galaxy bispectrum at NLO (1)

- Re-analyse PT challenge data varying H_0, Ω_m, σ_8
- AP distortions and IR resummation in the loops
- Need only $N_b = 4$ for good accuracy
- Compute bispectrum tensors $\mathcal{S}_{ijk}$ with FFTlog
- Not strictly necessary, but definitely convenient
- Evaluating one-loop B: under 1 second
- Increases cubically with N_b

TB, Ivanov, Philcox, Vlah (2025)



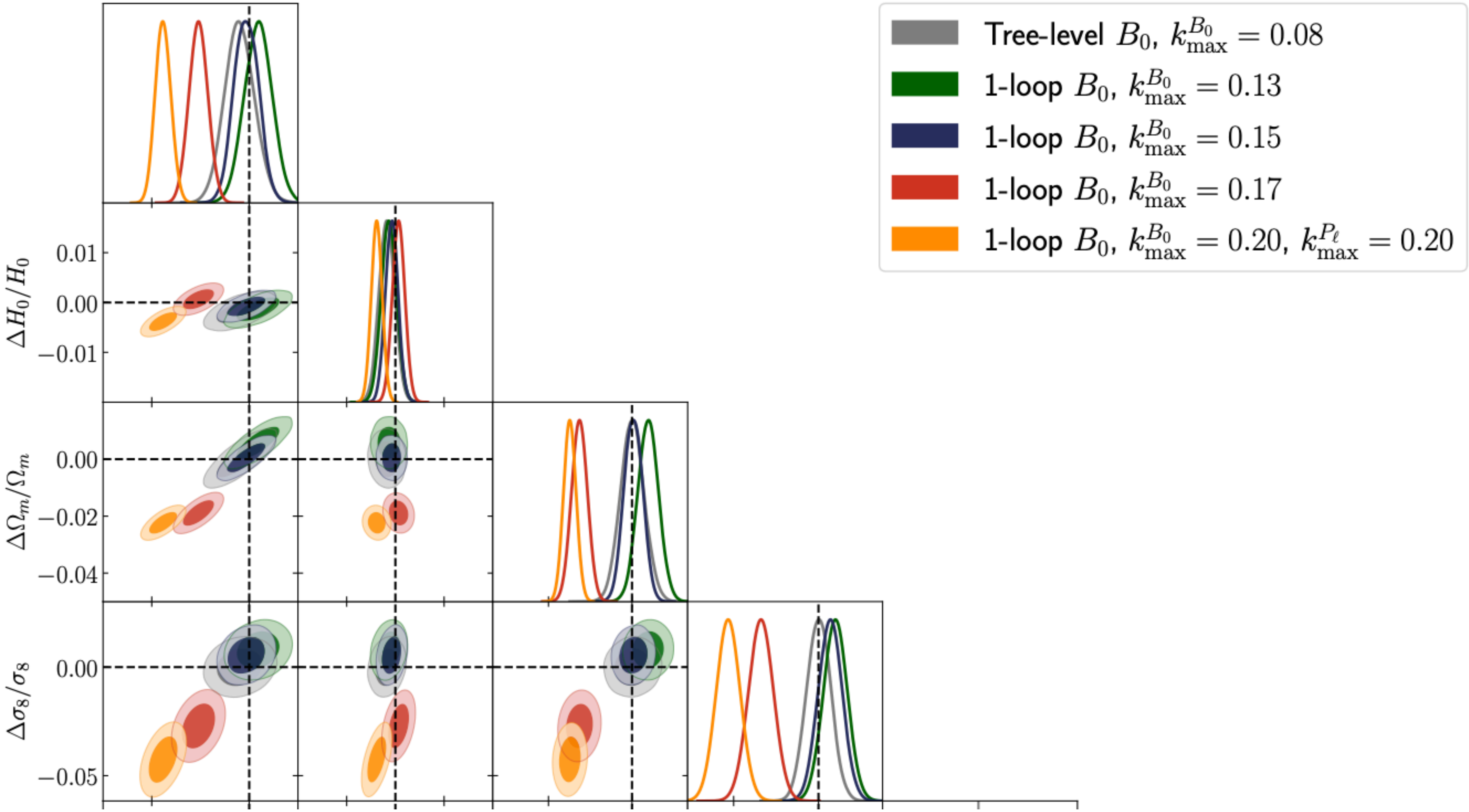
$N_b = 4$



Universiteit Utrecht

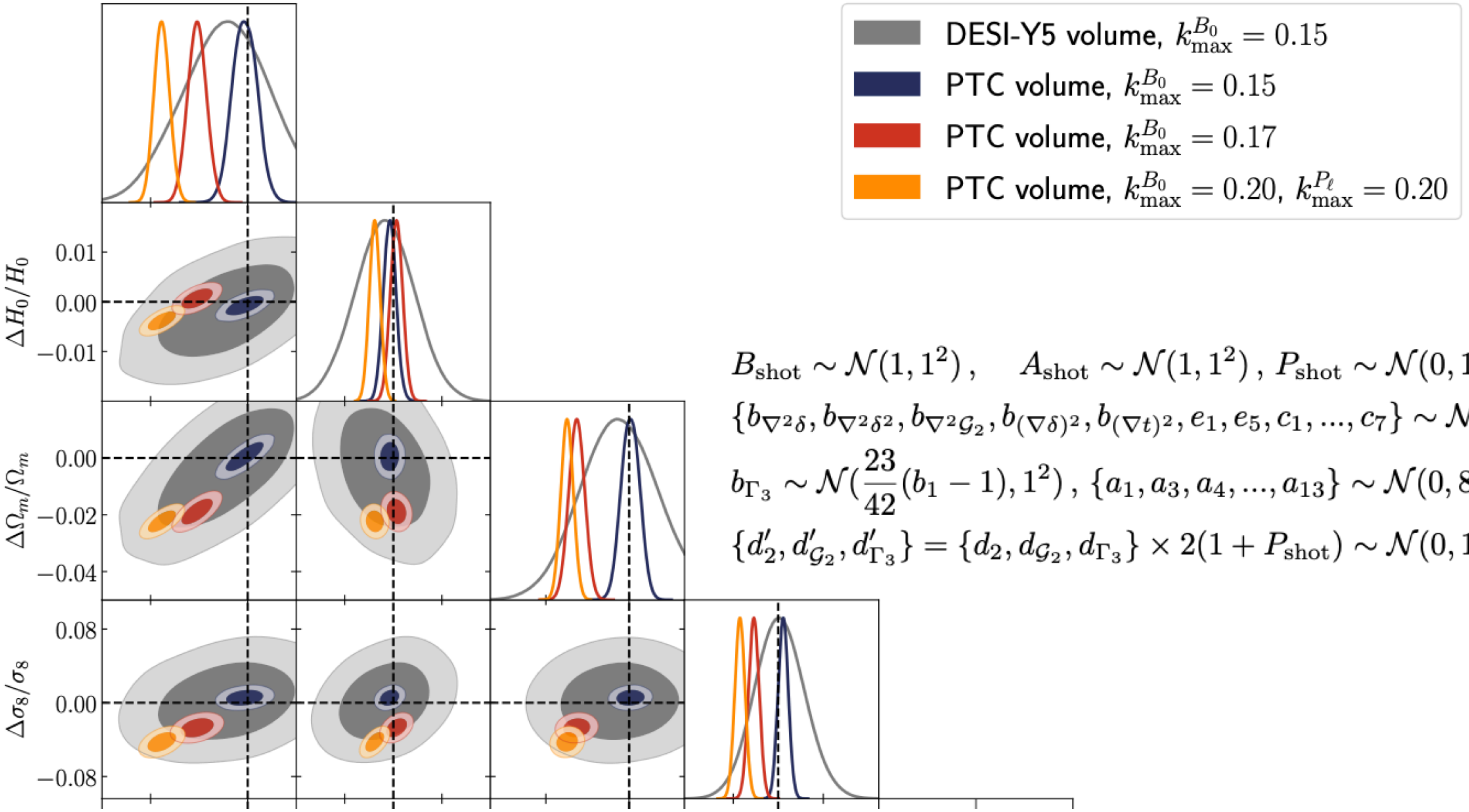
Galaxy bispectrum at NLO (2)

25



Universiteit Utrecht

Galaxy bispectrum at NLO (3)

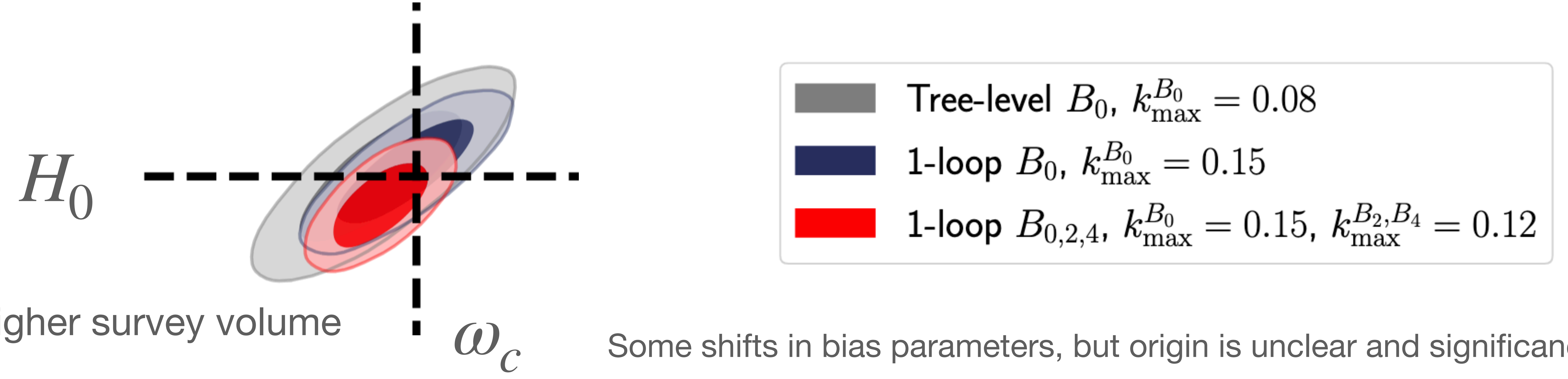


$$B_{\rm shot} \sim \mathcal{N}(1,1^2)\,, \quad A_{\rm shot} \sim \mathcal{N}(1,1^2)\,, P_{\rm shot} \sim \mathcal{N}(0,1^2)\,, a_{0,2} \sim \mathcal{N}(0,1^2)\,, b_1 \in [0,4]\,,$$
$$\{b_{\nabla^2\delta}, b_{\nabla^2\delta^2}, b_{\nabla^2\mathcal{G}_2}, b_{(\nabla\delta)^2}, b_{(\nabla t)^2}, e_1, e_5, c_1, ..., c_7\} \sim \mathcal{N}(0,8^2)\,, \{b_2, b_{\mathcal{G}_2}\} \sim \mathcal{N}(0,1^2)\,,$$
$$b_{\Gamma_3} \sim \mathcal{N}(\frac{23}{42}(b_1-1),1^2)\,, \{a_1,a_3,a_4,...,a_{13}\} \sim \mathcal{N}(0,8^2)\,, \gamma_{\mathcal{O}} \sim \mathcal{N}(0,10^2)\,,$$
$$\{d_2', d_{\mathcal{G}_2}', d_{\Gamma_3}'\} = \{d_2, d_{\mathcal{G}_2}, d_{\Gamma_3}\} \times 2(1+P_{\rm shot}) \sim \mathcal{N}(0,10^2)\ ,$$

Galaxy bispectrum at NLO (4)

- Add new contributions Misha found: $\langle \epsilon(k_1)\delta(k_2)(\epsilon_\delta \delta)(k_3)\rangle \implies \frac{1}{\bar{n}}(P_{22} + P_{13})$ at NLO
- Unbiased cosmology for $k = 0.15\, h/\mathrm{Mpc}$ with B_0
- Add B_2, B_4 with $k = 0.12\, h/\mathrm{Mpc}$
- Noticeable improvements in cosmology results (combined with two-point)

TB, Ivanov, Philcox, Vlah (2025)



2x higher survey volume

Some shifts in bias parameters, but origin is unclear and significance low for DESI

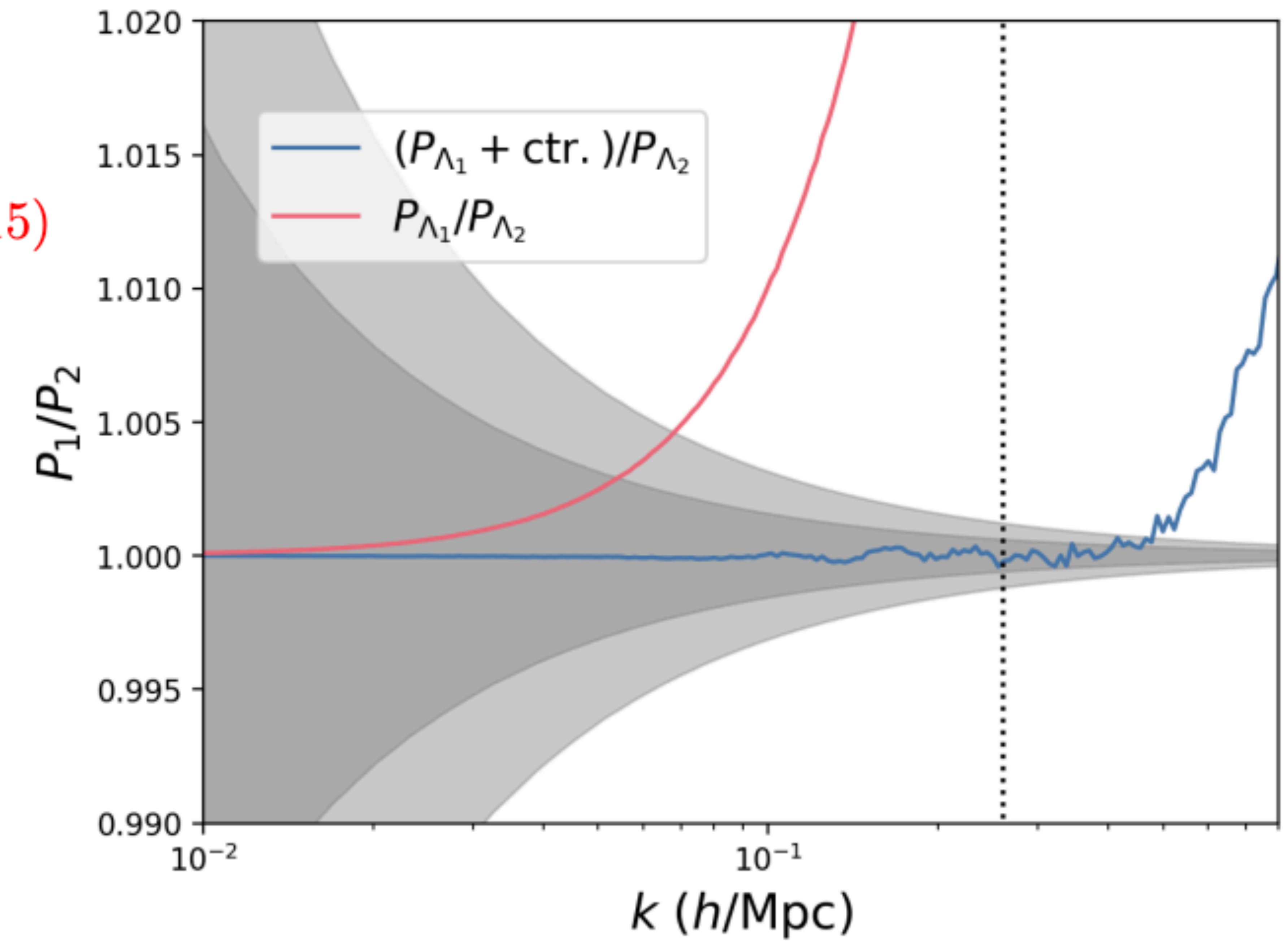
Matter power spectrum at NNLO (1)

- No fancy analytics available for 2-loop power spectrum (at the moment…)
- Some partial results available in literature, e.g. Slepian (2015) Simonovic++ (2018)
- See also Matt's talk for more discussion on counterterms Anastasiou++ (2025)
- We use four-parameter prescription from earlier work

Foreman++ (2015)

$$P_{\mathrm{EFT},21} = P_L + P_{\mathrm{SPT},11} + P_{\mathrm{SPT},21} - 2(2\pi)c_{\mathrm{s}}^2\frac{k^2}{k_{\mathrm{nl}}^2}P_L$$
$$-2(2\pi)c_{\mathrm{s},1}^2\frac{k^2}{k_{\mathrm{nl}}^2}P_{\mathrm{SPT},11} - 2(2\pi)c_{\mathrm{quad}}\frac{k^2}{k_{\mathrm{nl}}^2}P_{\mathrm{quad}}$$
$$+(2\pi)^2[(c_{\mathrm{s},1}^2)^2 - 2c_4]\frac{k^4}{k_{\mathrm{nl}}^4}P_L + \cancel{(2\pi)^2 c_\epsilon\frac{k^4}{k_{\mathrm{nl}}^4}\frac{1}{k_{\mathrm{nl}}^3}}\,, \qquad (6)$$

TB, Rubira, Chisari, Vlah (2025)



Matter power spectrum at NNLO (2)

- What is minimal number of pre-computations to be done?
- Rank-k symmetric tensor has $\binom{N_b + k - 1}{k - 1}$ elements where N_b is dimension
- $N_b = 6$, k = 3 yields only 56 integrals!
- Even better, we can write *all* off-diagonal terms in terms of *only* ‘auto-correlations’

One loop PS $:\propto P_L P_L = (\sum v_i)^2$ use $2v_1v_2 = (v_1 + v_2)^2 - v_1^2 - v_2^2$

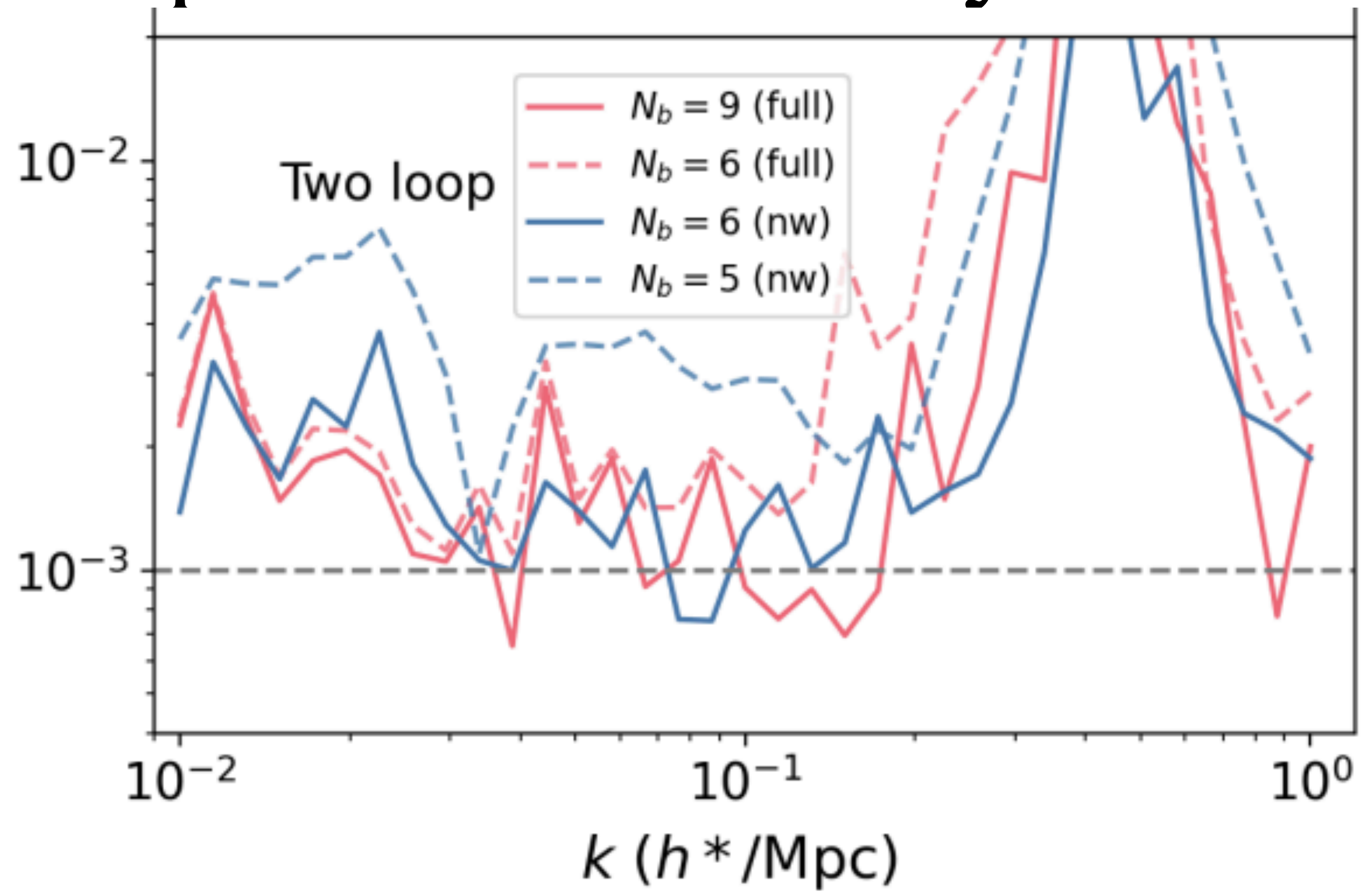
Two loop PS $:\propto P_L P_L P_L = (\sum v_i)^3$ use $6v_1v_2v_3 = (v_1 + v_2 + v_3)^2 - \frac{1}{6}\left((v_1 + 2v_2)^3 - \ldots\right)$

- Generalizes to all ranks - no need to modify anyone’s code, more numerically stable

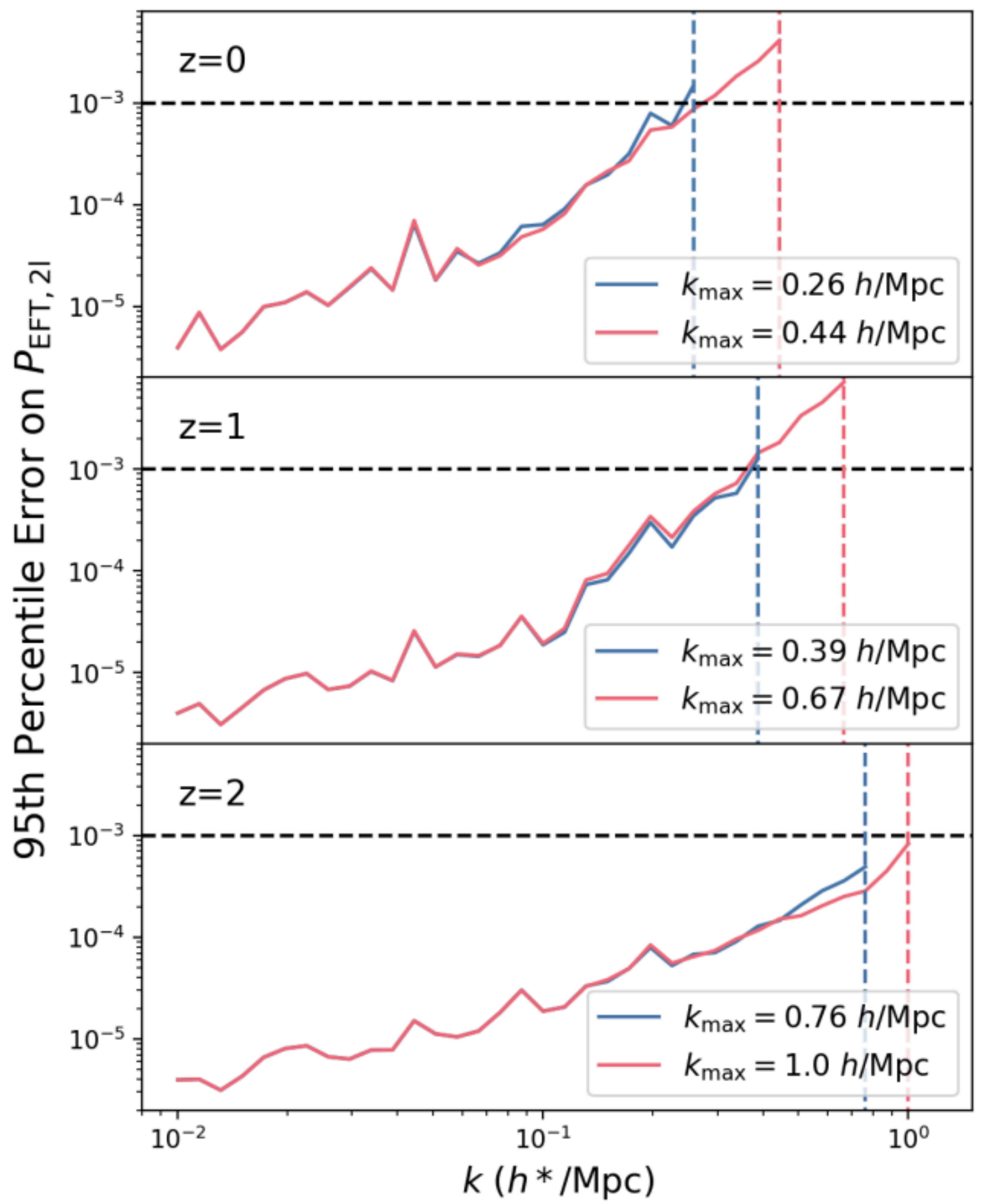
TB, Rubira, Chisari, Vlah (2025)

Matter power spectrum at NNLO (3)

- Go up to $N_b = 9$ for two-loop SPT piece
- $0.01 < k < 1$, more bins between 0.05 and 0.5
- Recover full power spectrum to per-mille accuracy
- Fewer N for no-wiggle
- Higher z is better
- Counterterms from HaloFit
- 1ms evaluation on CPU



5



TB, Rubira, Chisari, Vlah (2025)

Universiteit Utrecht

Matter power spectrum at NNLO (4)

TB, Rubira, Chisari, Vlah (2025)

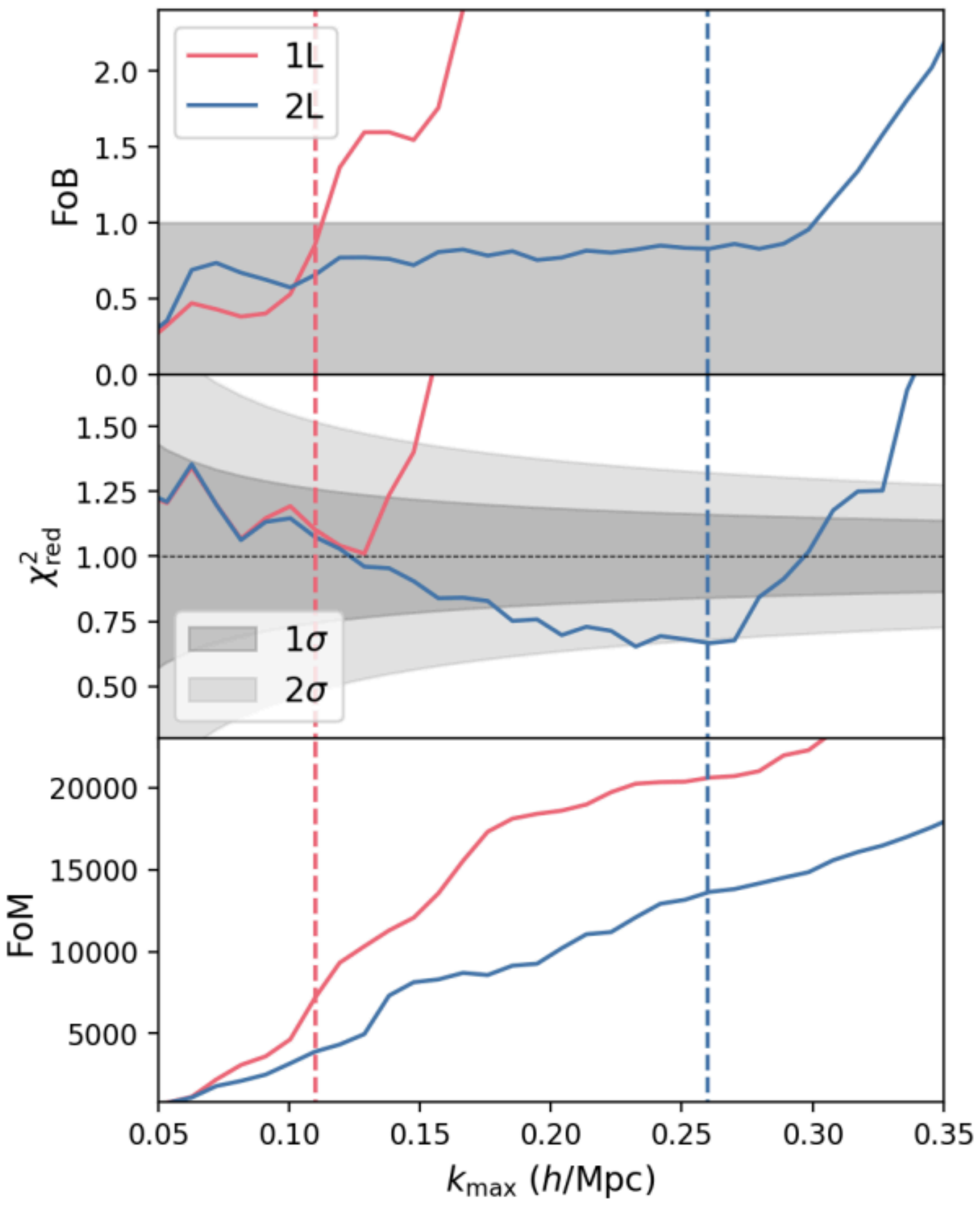
- MCMC: Use Dark Sky at z=0, rescale to $V = 25\,(\mathrm{Gpc}/h)^3$
  Skillman++ (2014)
- Consistently marginalise over baryon density, slope
- $\omega_b \sim \mathcal{N}(0.0221, 0.00055^2),\, n_s \in \mathcal{U}(0.91, 1.01)$
- Compare one-loop EFT with two-loop EFT
- FoB on Ω_m, A_s, H_0

$$\mathrm{FoB}(k_{\mathrm{max}}) = \left(\sum_{\alpha,\beta} (\theta^\alpha - \theta^\alpha_{\mathrm{fid}}) S^{-1}_{\alpha\beta} (\theta^\beta - \theta^\beta_{\mathrm{fid}})\right)^{1/2}$$

- FoM on Ω_m, A_s, H_0

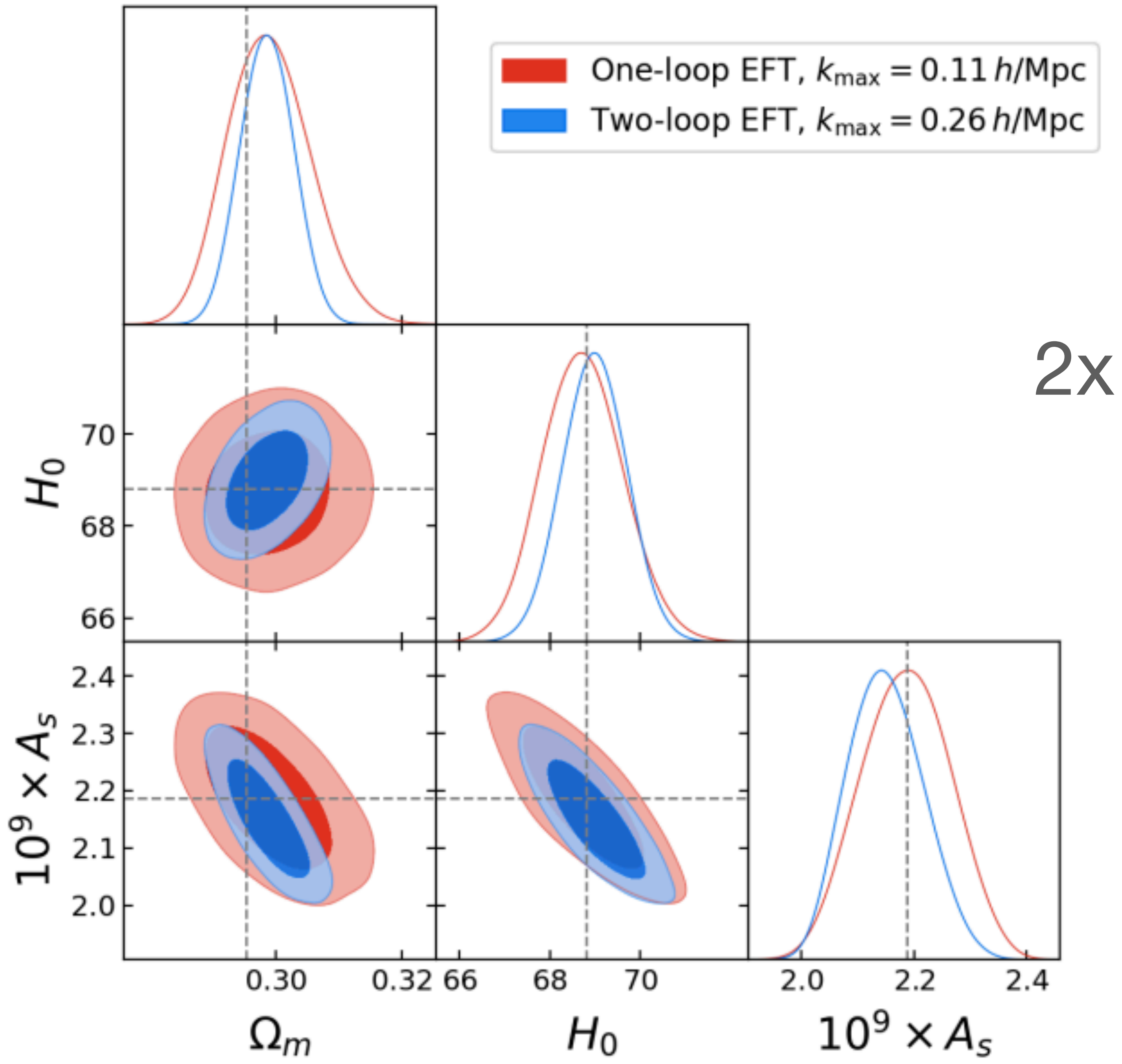
$$\mathrm{FoM}(k_{\mathrm{max}}) = \frac{1}{\left(\det(S_{\alpha\beta})/\theta^\alpha_{\mathrm{fid}}\theta^\beta_{\mathrm{fid}}\right)^{1/2}},$$

- Reduced chi-squared (goodness of fit) at full sim volume
- Also consider running of counterterms



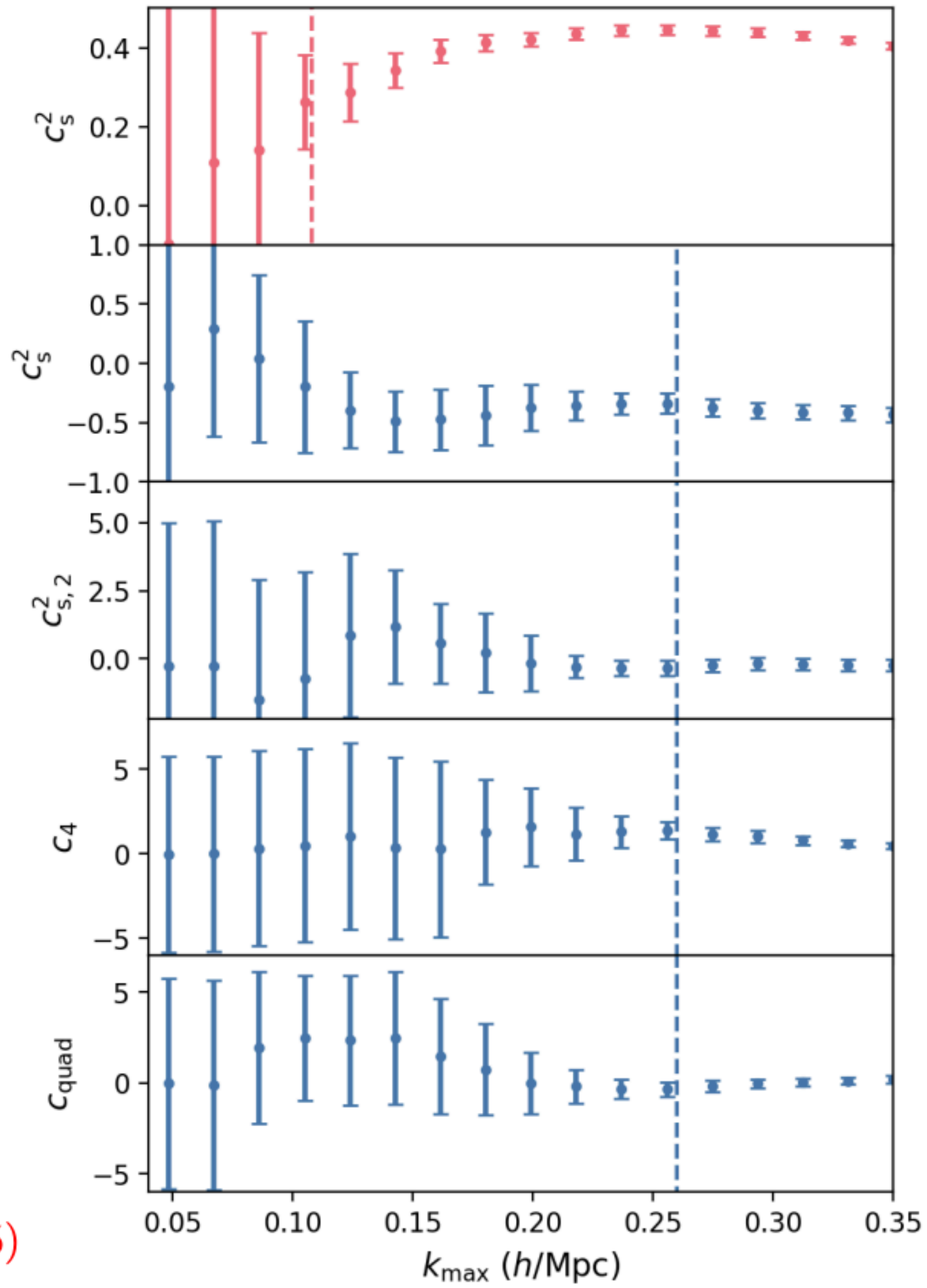
Matter power spectrum at NNLO (5)

- Scale cuts: $k_{1l} = 0.11\,h/\text{Mpc}$, $k_{2l} = 0.26\,h/\text{Mpc}$ (mainly FoB < 1)



2x higher survey volume

Ω_m 35 % ↓



- Avoid dumping information into (well-constrained) counterterms?
- See also Mathias' talk

TB, Rubira, Chisari, Vlah (2025)

Universiteit Utrecht

Beyond matter

TB, Garny, Rubira, Vlah (2025)

- Some results on renormalized galaxy bias at 2 loops, see Henrique Rubira's talk
- Remove cutoff-dependence of 2-loop integrals (so far leading order in gradients)
- Need (??) bias parameters after removing (??) *exact* / approximate degeneracies…
- *Definitely numerically feasible with COBRA*
- Maybe reduce N_b further by up-/down-weighting some scales in $P_L(k)$?
- Numerical 'preconditioning' by subtracting UV sensitive parts?
- Stay tuned!

The road ahead

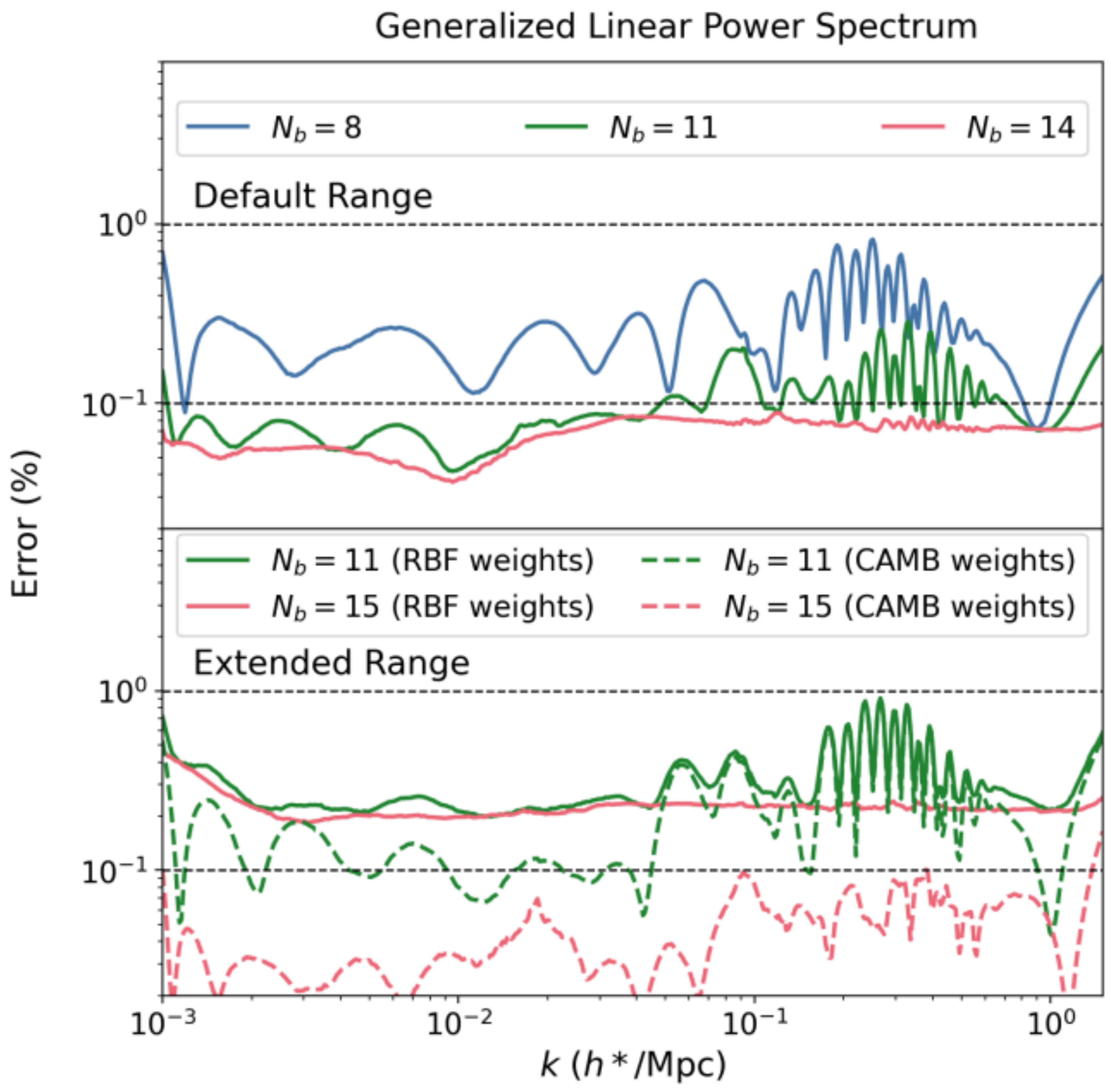
TB, Vlah, Chisari (2024)

- **Publicly available package!** `pip install cobra`
- Perturbation theory implementations beyond $P_{1\text{-loop}}$ collected in one place?
- 'Derived' statistics like skew spectra, marked spectra?
- Questions?

Backup 1

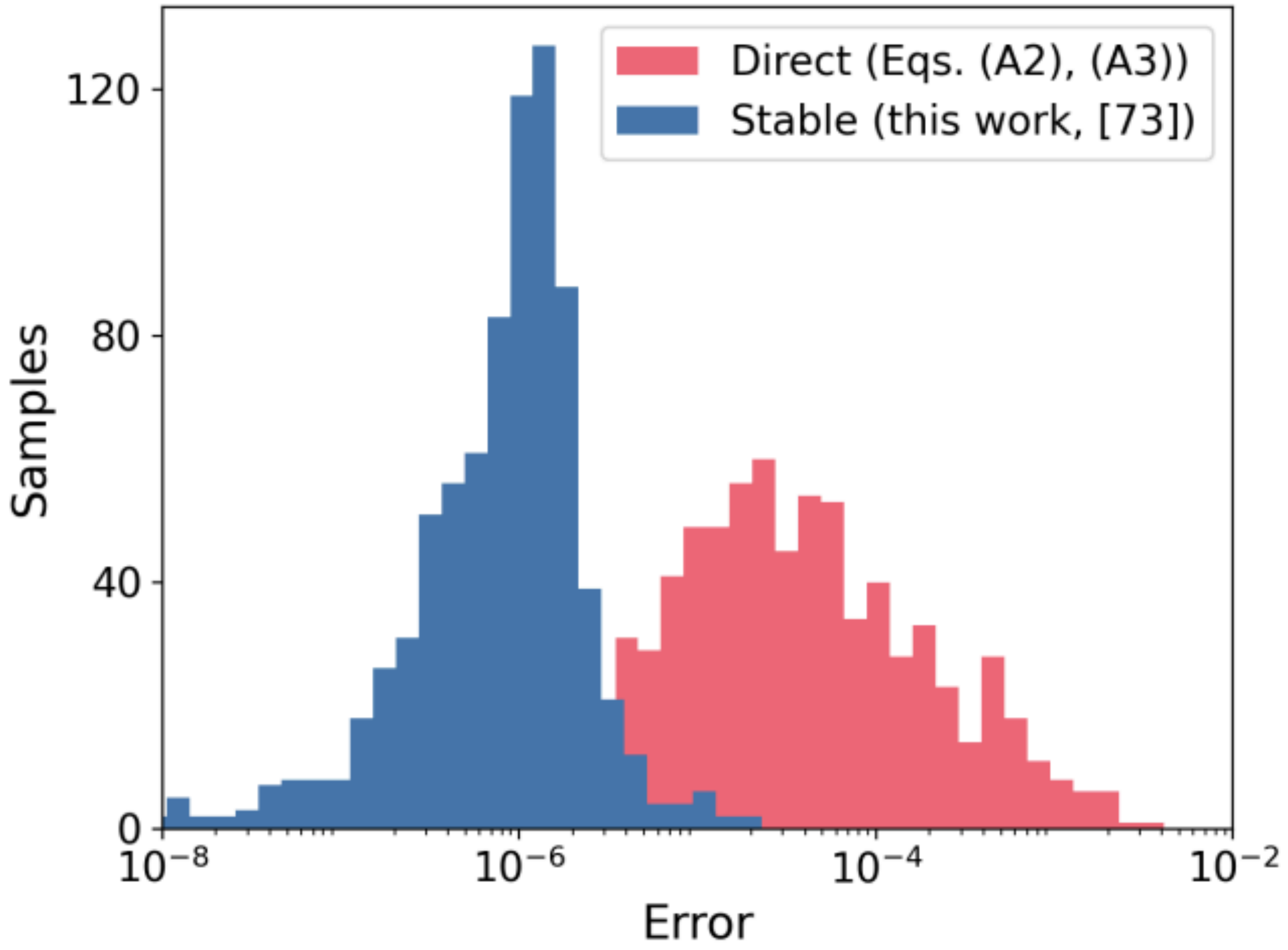
3

	Default		Extended	
Θ	Range	Grid size	Range	Grid size
ω_c	[0.095,0.145]	20	[0.082,0.153]	30
ω_b	[0.0202,0.0238]	12	[0.020,0.025]	12
Ω_K	[-0.12,0.12]	12	[-0.2,0.2]	12
h	[0.55,0.8]	$h^* = 0.7$	[0.51,0.89]	$h^* = 0.7$
n_s	[0.9,1.02]	8	[0.81,1.09]	15
M_ν [eV]	[0,0.6]	12	[0,0.95]	15
w_0	[-1.25,-0.75]	12	[-1.38,-0.62]	15
w_a/w_+	[-0.3,0.3]	$w_a^* = 0$	[-1.78,-0.42]	$w_+^* = -1$
z	[0.1,3]	$z^* = 0$	[0.1,3]	$z^* = 0$
$10^9 A_s$	-	$10^9 A_s^* = 2$	-	$10^9 A_s^* = 2$



Universiteit Utrecht

Backup



Universiteit Utrecht