Full-catalog field-level inference in 2.5 dimensions for imaging surveys

The Non-Gaussian Universe
Heraklion, Crete
17 June 2026

Work in progress by
University of Pennsylvania crew:

Supranta Boruah
(→IIT Bombay late 2026)

Vernon Wetzell
(GS, → final yr)

Gary Bernstein
(presenting today)

Alex Tong
(GS →3rd yr)

Megan Zhao
(GS →3rd yr)

Maximal Information

Maximal Information Astrophysics

…to simplify and accelerate analysis of imaging survey data

- Uses both the shear on and the spatial density of the galaxies (incl. magnification).

- Uses available information from ***each*** galaxy.

- Compress galaxies to feature vector with calculable selection/noise properties *p(measurement | truth)* - observations in the likelihood, simulations.

- Robust to the details of biasing, IA models, *e.g.* can retrieve correct cosmology from any sensible simulation.

- Minimize the number of “analysis choices” and their associated effort.

- Fast, sample-able without owning Nvidia.

Wide images

Here are many selection steps
….are they included in these models?

D

$\xi_{\pm ij}(\theta)$

$\gamma_{tij}(\theta)$

$w_i(\theta)$

NG stats...

Inference

$p(\Omega \,|\, D)$

$\mathscr{L}(D \,|\, \Omega)$

$\bar{\xi}_{\pm ij}(\theta)$

$\bar{\gamma}_{tij}(\theta)$

$\bar{w}_i(\theta)$

NG stats...

Ω

$P_\delta(k, z)$

Baryon effects

$n_{\rm src}(z)$

shear cal

IA

$n_{\rm lens}(z)$

magnifi cation

galaxy bias

Wide images

Object catalog

selections

Lens bin 1

Lens bin 4

Lens bin 5

Lens bin 6

Src bin 1

Src bin 2

Src bin 3

Clusters

maps

$w_i(\theta)$

$\gamma_{tij}(\theta)$

$\xi_{\pm ij}(\theta)$

mass cal

$\{F_g\}, \quad g = 1,2,\ldots,10^9$

- *F*={flux, color, position, 2nd moments}
- Simple flux selection
- Noise/selection derivable from image noise+PSF.

Wide images

Object catalog

Inference

$p(\Omega \mid D)$

$\bar{D}$

$\bar{\xi}_{\pm ij}(\theta)$

$\bar{\gamma}_{tij}(\theta)$

$\bar{w}_i(\theta)$

$+\mathrm{Cov}(D)$

Ω

(k, z)

Baryon effects

$n_{\mathrm{src}}(z)$

shear cal

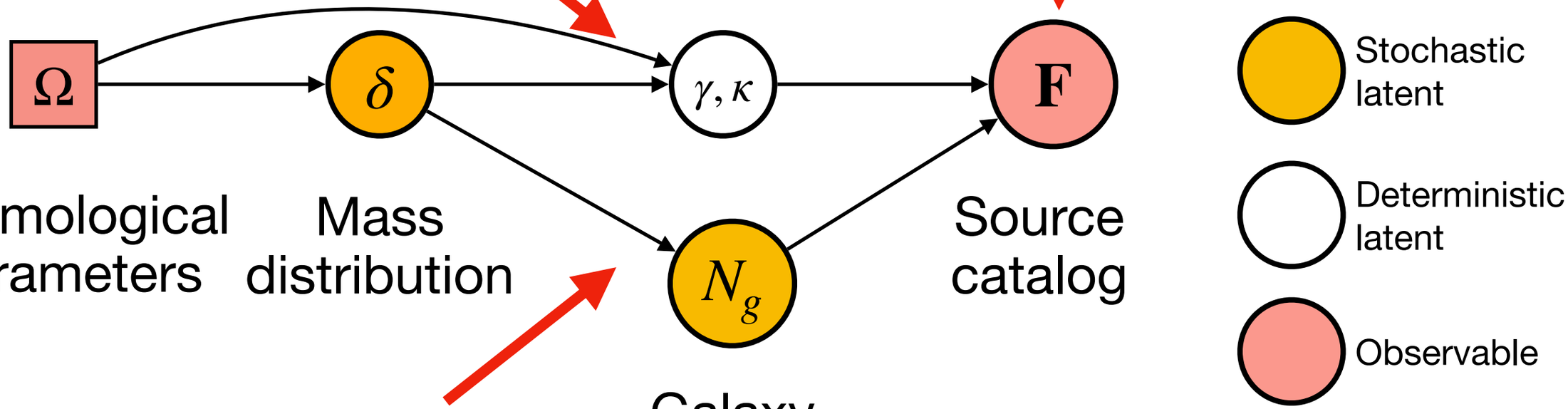
IA

$n_{\mathrm{lens}}(z)$

magnifi cation

Deterministic relation here is the strength of weak lensing

***Every** galaxy is informing us of …*
local density
applied WL
its own z
…to varying extents.
No bins - “catalog level”



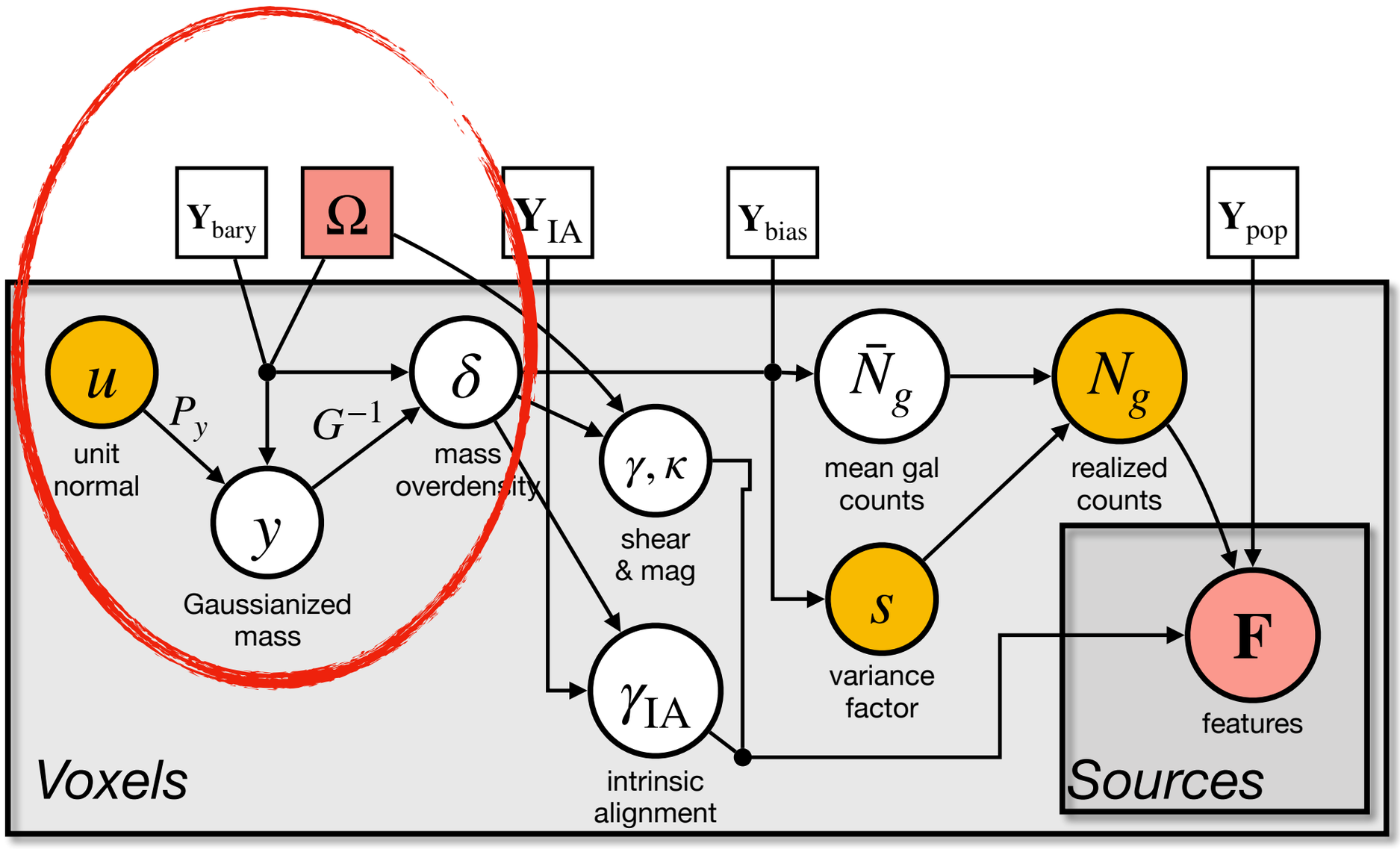
Higher S/N per voxel, better LOS localization.

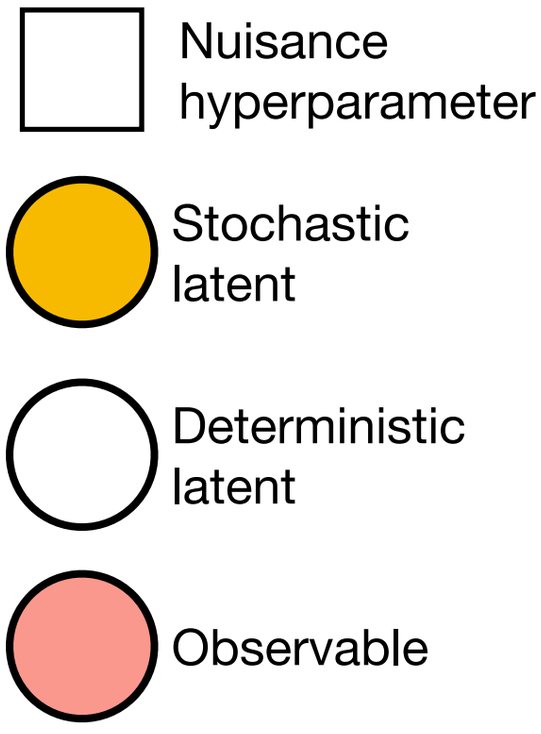
BUT $p(N_g \mid \delta)$ is highly non-linear and stochastic in a manner beyond theoretical reach

Field level is the natural domain for models of the local, non-linear galaxy process.

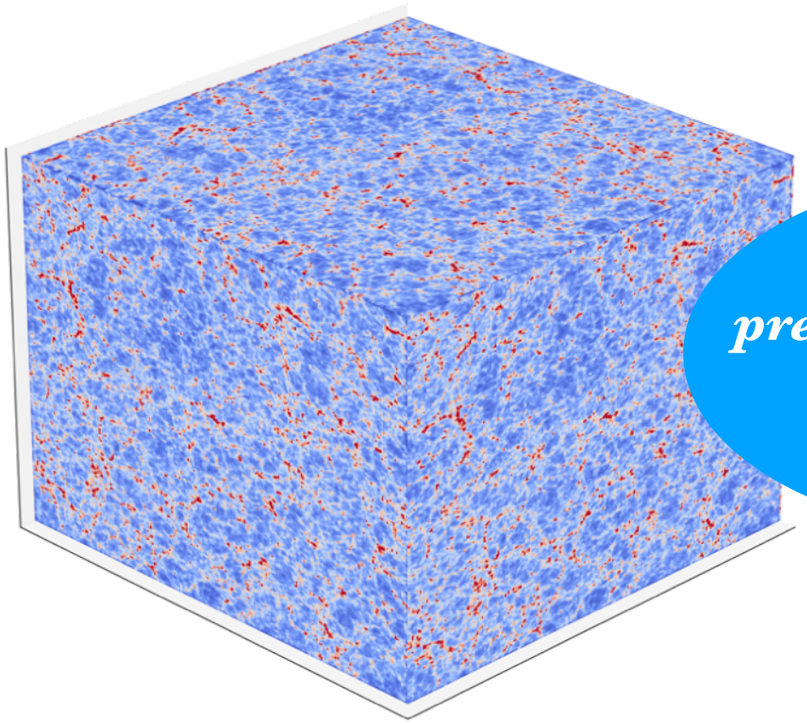
WL signals calibrate parameters of this process to transfer galaxy densities to mass.

Mass Model

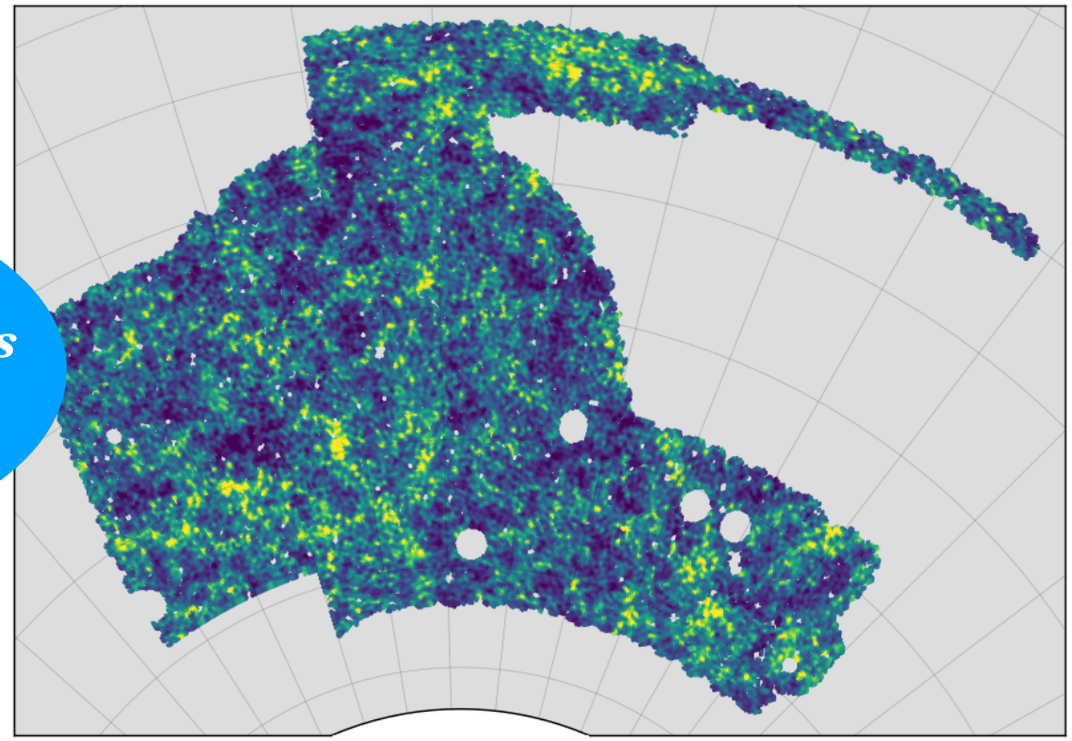

$\mathbf{Y}_{\text{bary}}$
Ω
$\mathbf{Y}_{\text{IA}}$
$\mathbf{Y}_{\text{bias}}$
$\mathbf{Y}_{\text{pop}}$
u
unit normal
P_y
y
Gaussianized mass
G^{-1}
δ
mass overdensity
γ, κ
shear & mag
γ_{IA}
intrinsic alignment
$\bar{N}_g$
mean gal counts
s
variance factor
N_g
realized counts
$\mathbf{F}$
features
Voxels
Sources


Nuisance hyperparameter
Stochastic latent
Deterministic latent
Observable

Strategy 1: 3D field modeling from initial conditions



Strategy 2: Direct 2D modeling of projected fields



How do we preserve the strengths of both these approaches?

✚ Very well known, fast prior on initial conditions

✚ Good physical models for mass evolution (with free parameters for baryons)

✚ Halo, EFT, other plausible models for galaxy occupation, with quite a few nuisance parameters.

✚ Can reveal inflationary signals in initial field.

✚ Very high mode count for good stats

– Very expensive (even with acceleration from ML)

– Enormous number of free parameters (>10^9 for translinear scales)

– >90% of these parameters are unconstrained by WL or photo-z data → poor chain convergence

Good for spec-z surveys; some WL work

✚ Many fewer DOF for ~5 bin tomography

✚ Requires empirical models from N-body

✚ Very fast, already feasible for Stage III WL

✚ Captures most of the WL information (but not all?)

– Galaxy density information not used to date, WL only. **Poor usage of galaxy density info.**

– Does not yield initial field - but mode count is not huge compared to CMB anyway.

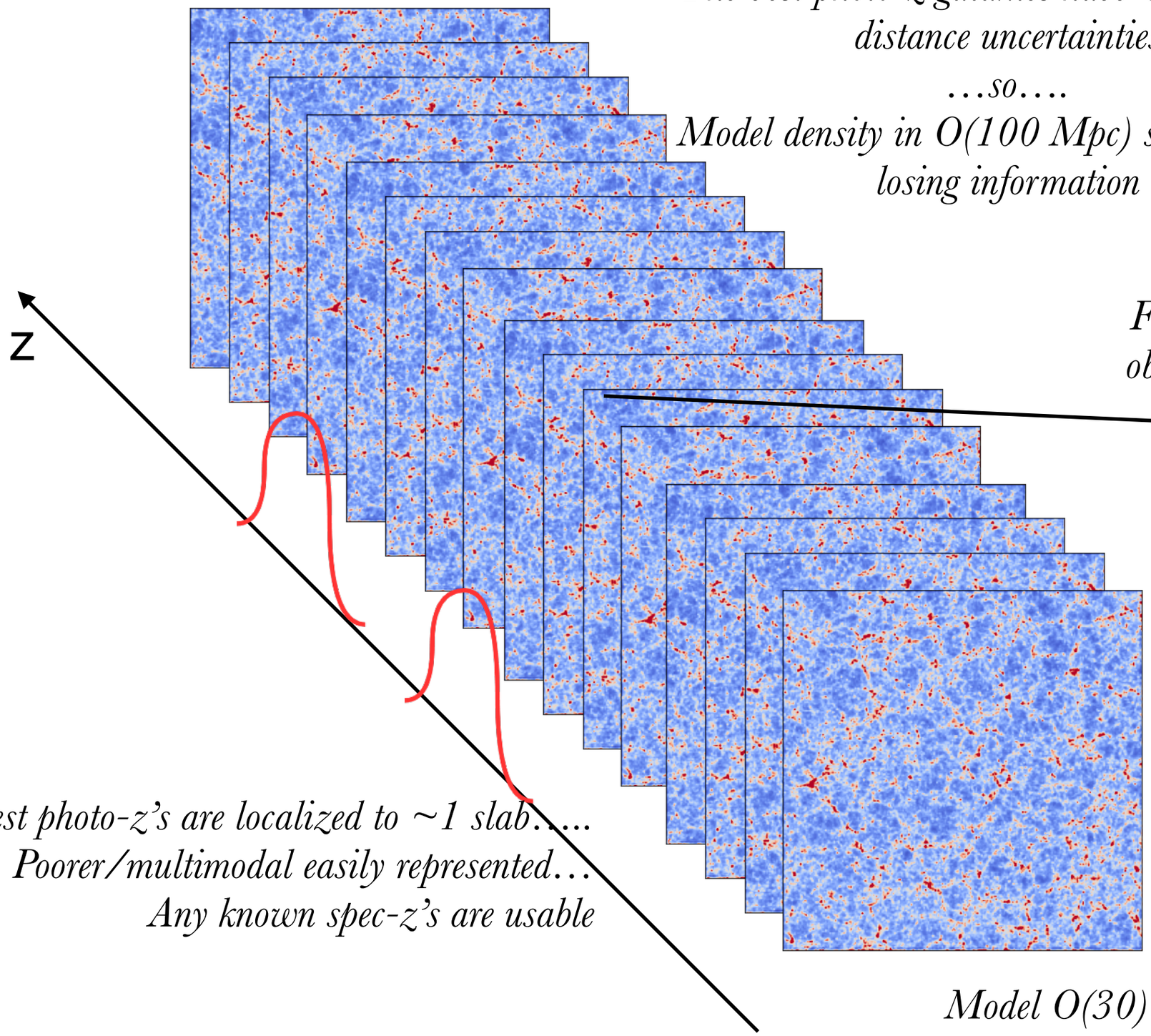
Good for tomographic WL (usually after some form of map-making from shear)

Slab-based (or 2.5 dimensional) field-level inference

The best photo-z galaxies have O(100 Mpc) distance uncertainties

…so….

Model density in O(100 Mpc) slabs witho losing information

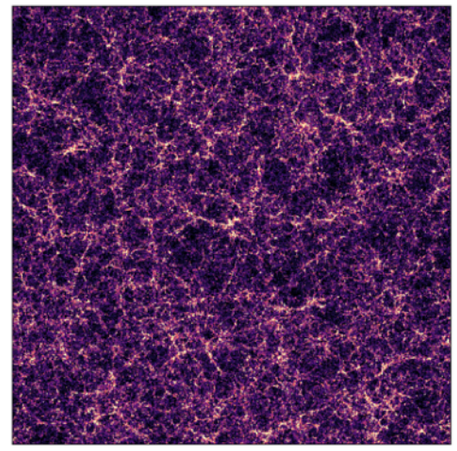


Field-level observables

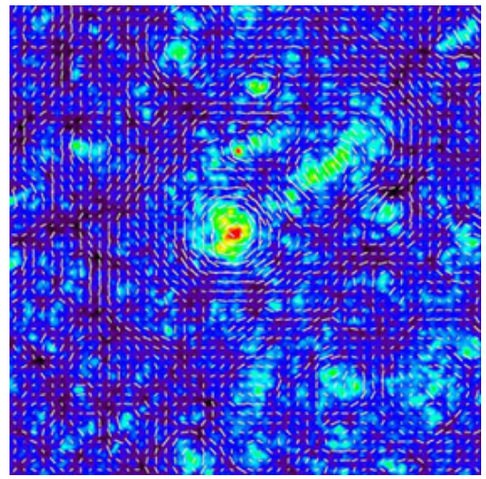
Best photo-z's are localized to ~1 slab….
Poorer/multimodal easily represented…
Any known spec-z's are usable

Model O(30) density slabs

Galaxy number count



Shear data

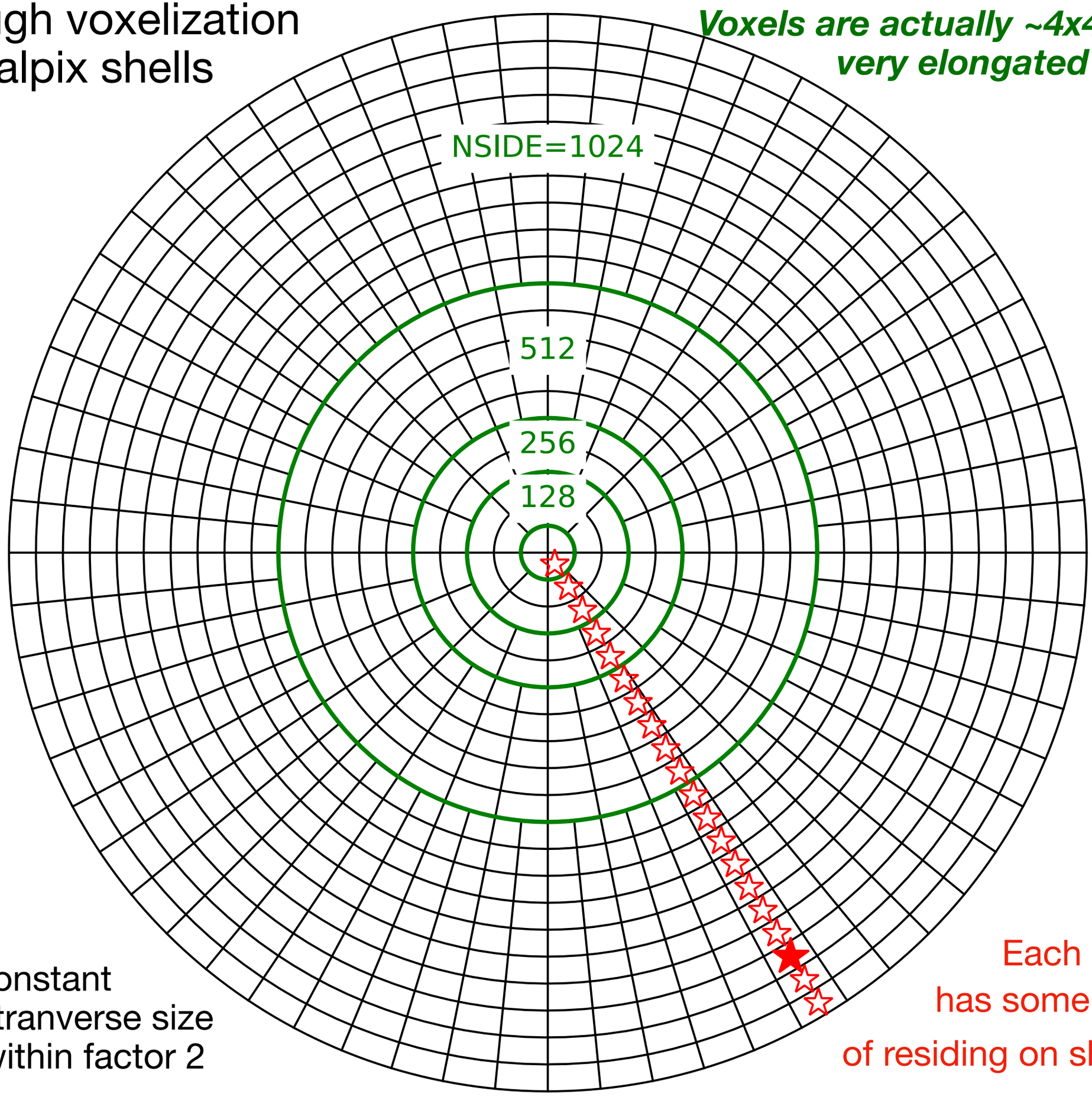


Joint shear-galaxy likelihoods!

The unmodelled line-of-sight structure within each slab becomes a source of stochasticity in the astrophysical models for galaxy occupancy and IA.

Slice through voxelization
Nested healpix shells

Voxels are actually ~4x4x150 Mpc, very elongated along LOS



Maintain constant
comoving tranverse size
of voxels within factor 2

Each detected g
has some prob $p_g(z)$
of residing on shell index z

Generalized point-transformed Gaussian (GPTG)

- Reproducing the one-point distribution (PDF) $p(\delta_m \,|\, \Omega, z)$ as the most important aspect of non-Gaussian behavior?

- Need to define a smoothing filter $W(\theta) \Leftrightarrow \tilde{W}(\ell)$ to measure a PDF.

- There is alway a "normalizing" function $W(\theta) * \delta_m(\theta) = G[y(\theta)]$, $G(\delta_m) : y = G(\delta_m) \sim \mathcal{N}(0,1)$ (unit normal).

- We ***assume*** that $y(\theta)$ is a zero-mean Gaussian random field specified by a power spectrum $P_y(\ell), \ell < \ell_{\rm Nyq}$.

- One can pick a $P_y(\ell)$ to recreate almost any chosen

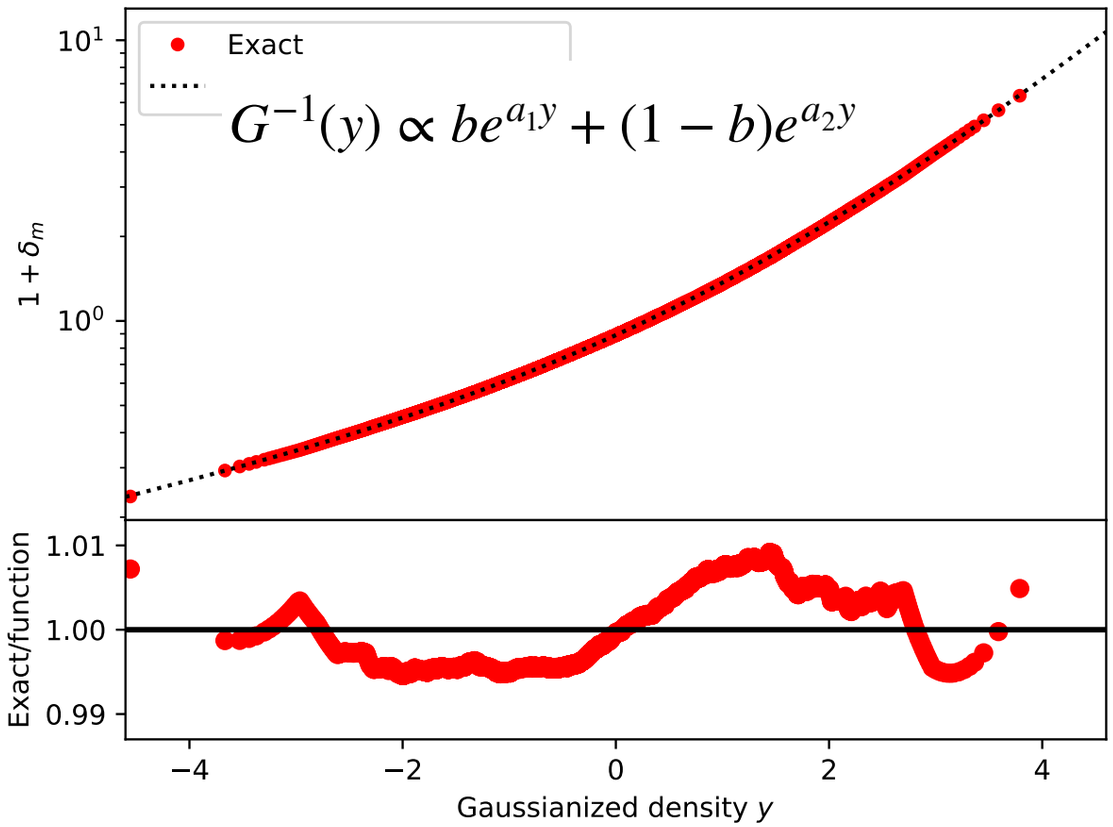
- Currently using square pixels for *W* function; size ul

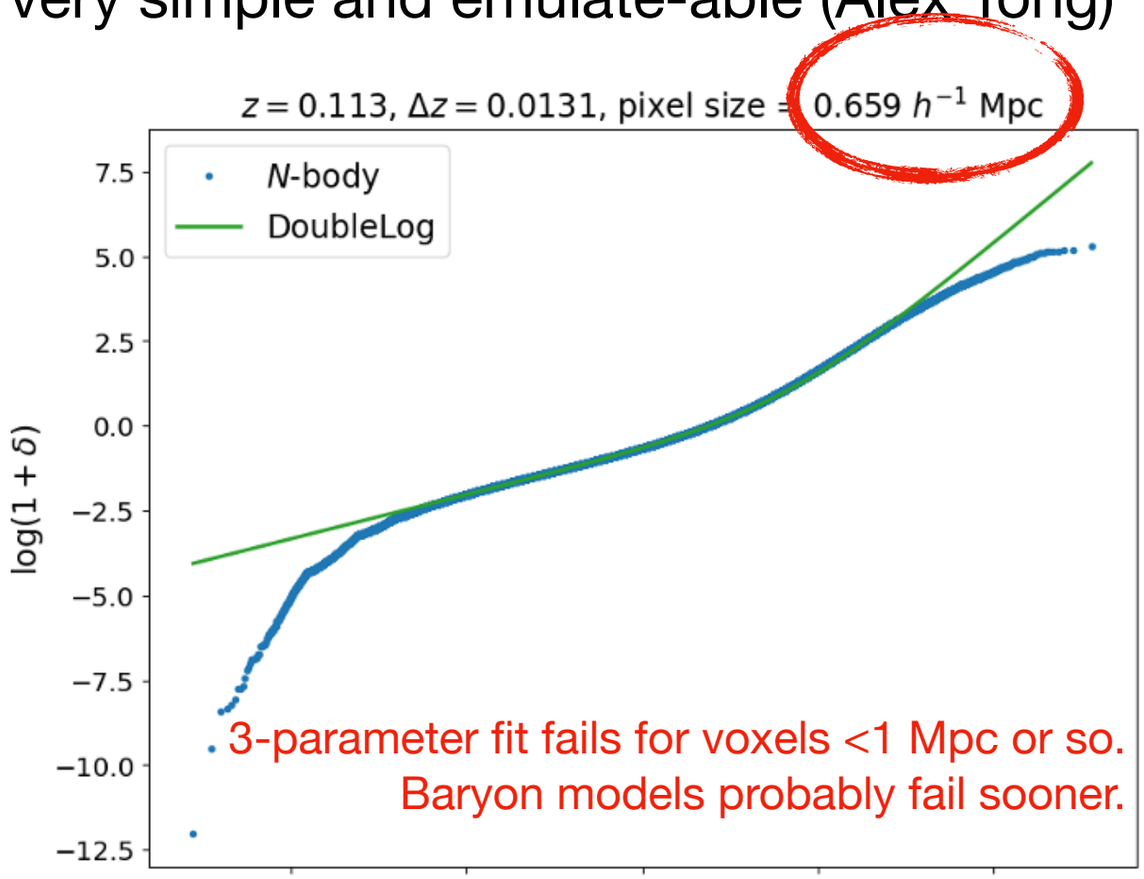
  - G, P_y derived from simulations are stable or easily p a variety of baryon treatments

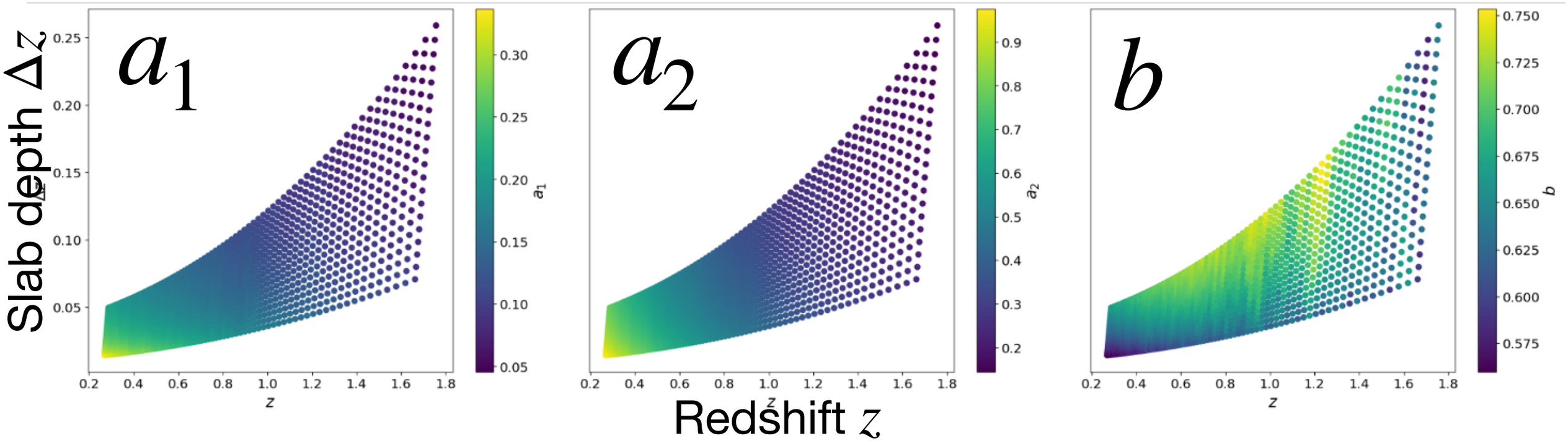
  - Generative galaxy model (described later) spans all reasonable simulations' results.

Question:
Are there better operators than linear filters to "censor" the baryonic effects in FLI?

The Gaussianizing functions are, empirically, very simple and emulate-able (Alex Tong)

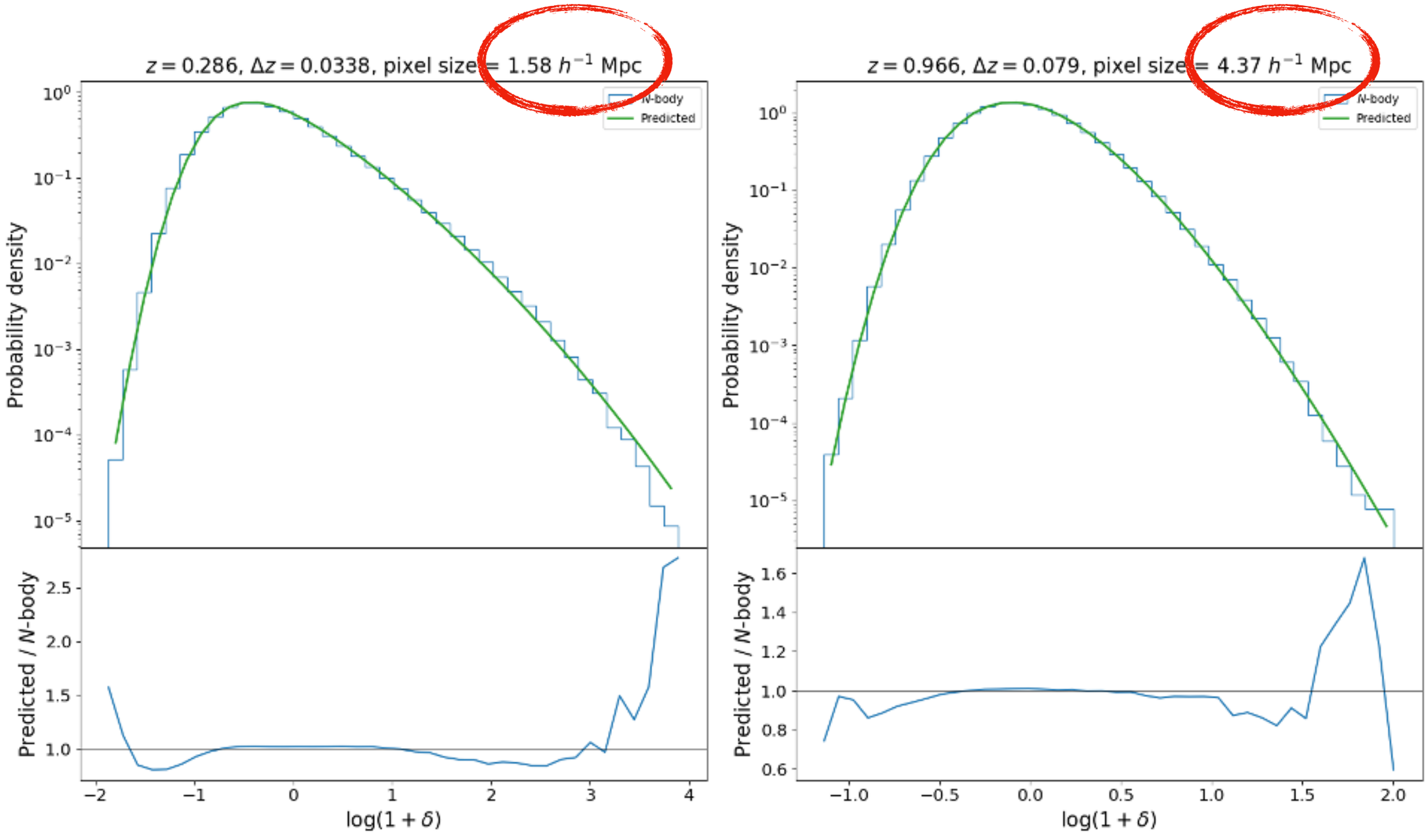


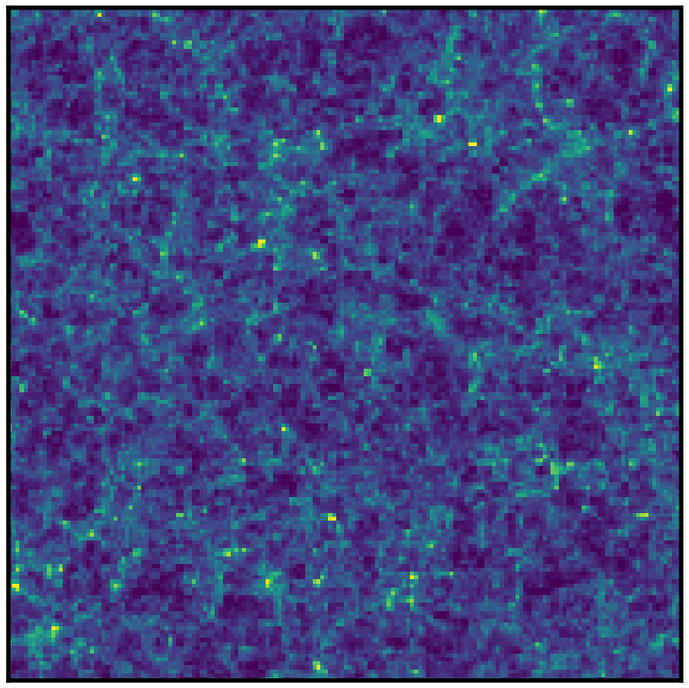


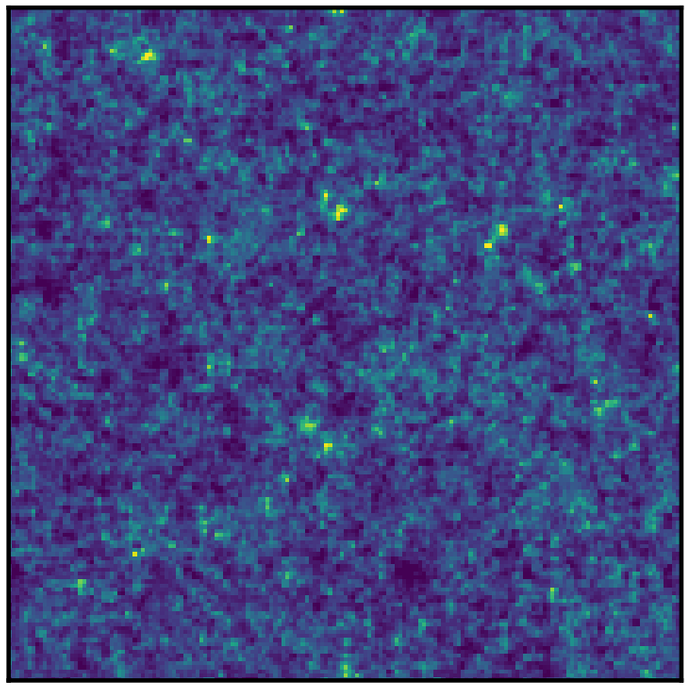


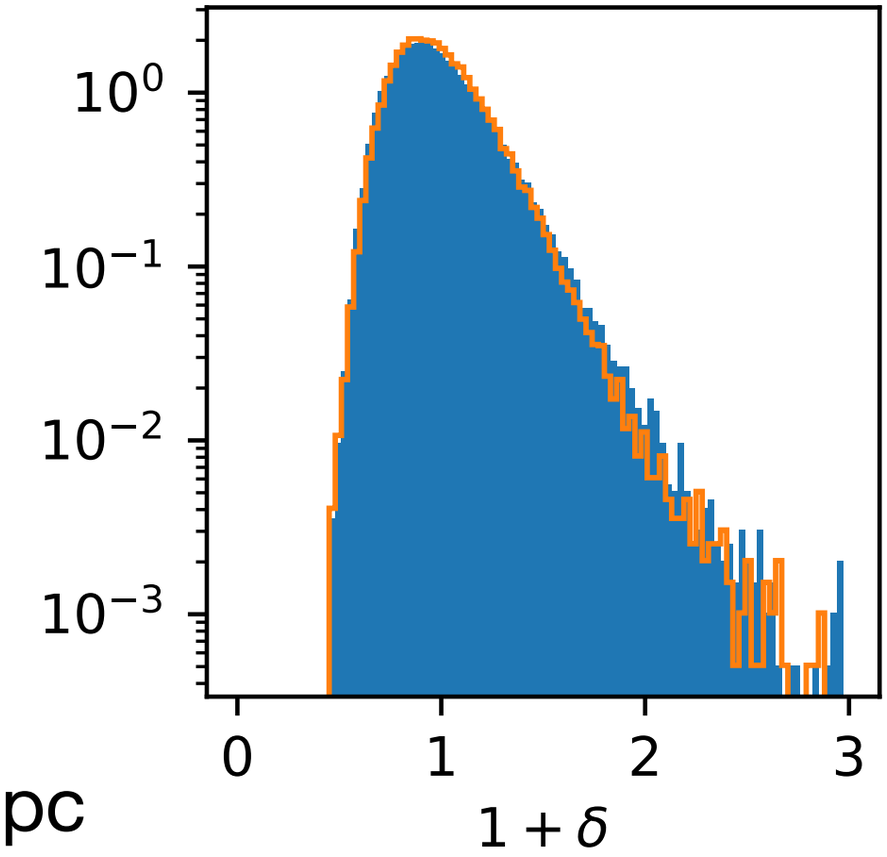
Emulation across Ω_m, σ_8, w using 38 PKDGRAV3 runs

- Fit (a_1, a_2, b) with simple polynomials across cosmology
- …and across slab central *z* and depth
- Leave-one-out validation results at 2 redshifts









1 Gpc x 1 Gpc x 500 Mpc

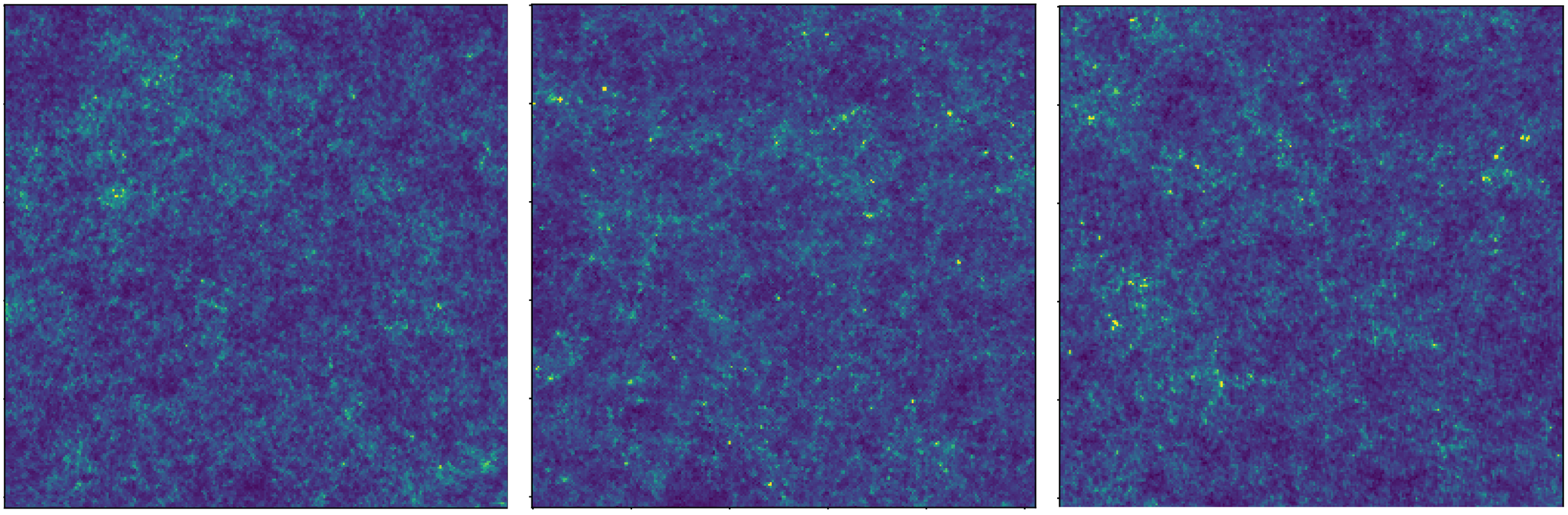
GPTG is obviously not a correct description of LSS.

But does it capture **most** of the non-Gaussianity that carries cosmological information? (halo collapse)

Neyrinck (2011) showed strong information recovery in power spectra from Gaussianization

…if not, we can switch the generative slab mass model to something more sophisticated!

GPTG κ map emulation from Zhong et al (2025)
1 is N-body, 1 is log-normal, 1 GPTG

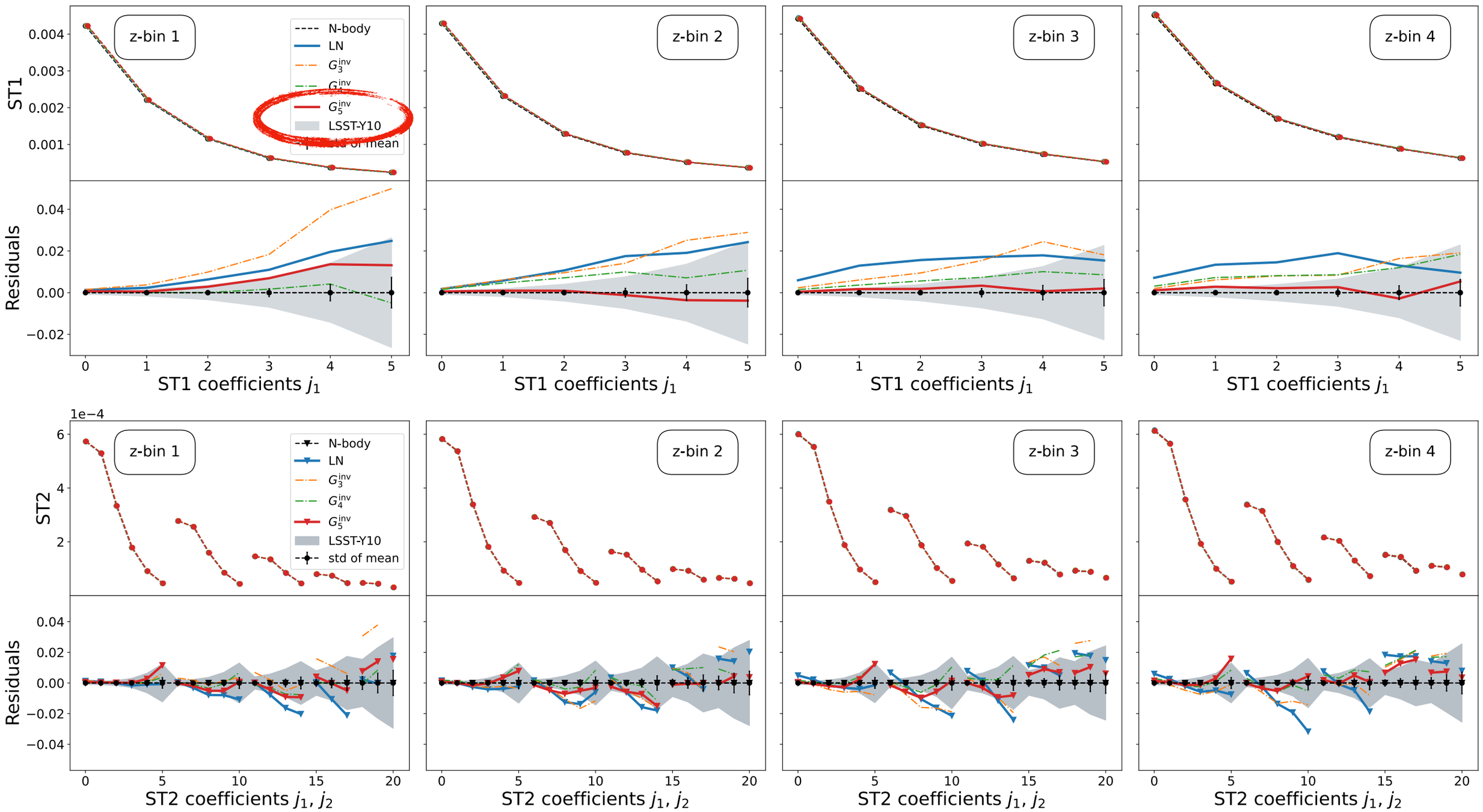


Lognormal

N-body

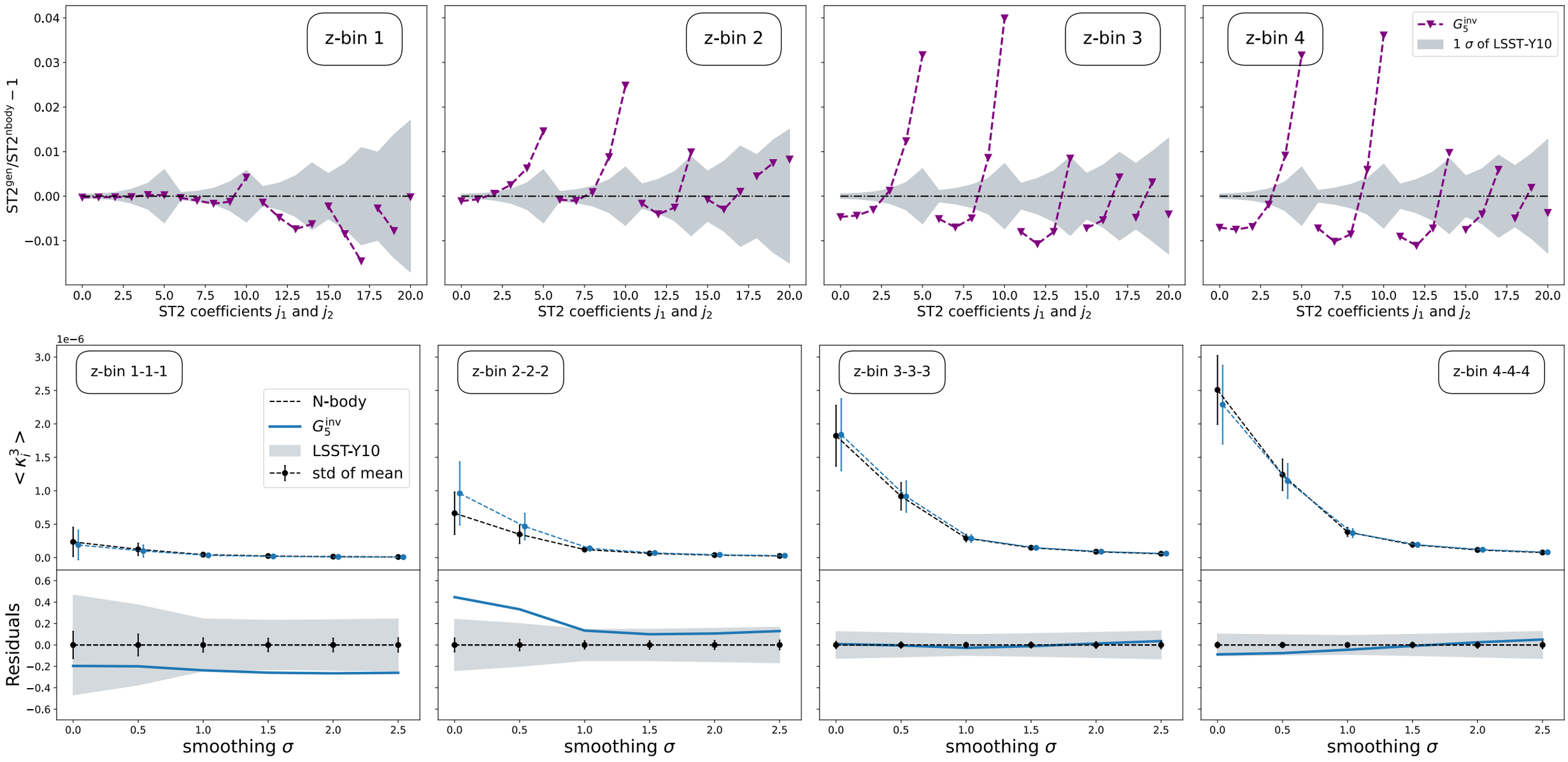
GPTG

Scattering Wavelet Transform stats for GPTG vs N-body

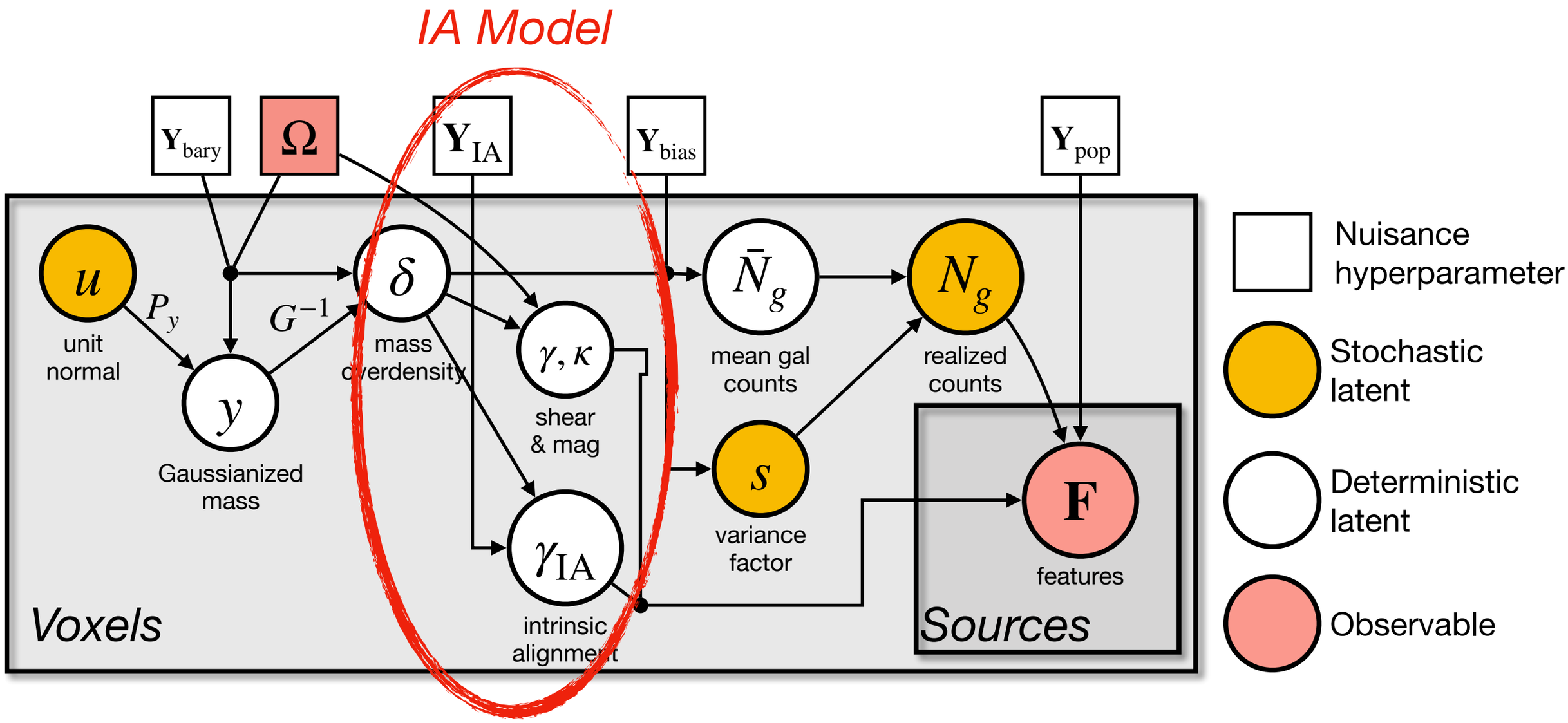


7’ pixels

SWT & 3rd moments for GPTG vs N-body



3.4’ pixels



- State of the art for IA models is *already* a simple symmetry-constrained local model.
- We easily implement (projected) tidal term (NLA) and $\delta_m s$ terms into forward models.
- We can make coefficients dependent upon galaxy *z* and type.
- We do not have to propagate the varying IA coefficients into summary statistics.
- 2.5d, FLI should be better than tomographic power spectra at measuring nonlinear IA terms

Initial tests from Supranta on IA in 2.5d FLI

$$\bar{\gamma}_{ij}^{\rm IA} = A_1 s_{ij} + A_{1\delta}\delta s_{ij} + A_2 \sum_k s_{ik}s_{kj} + \cdots,$$

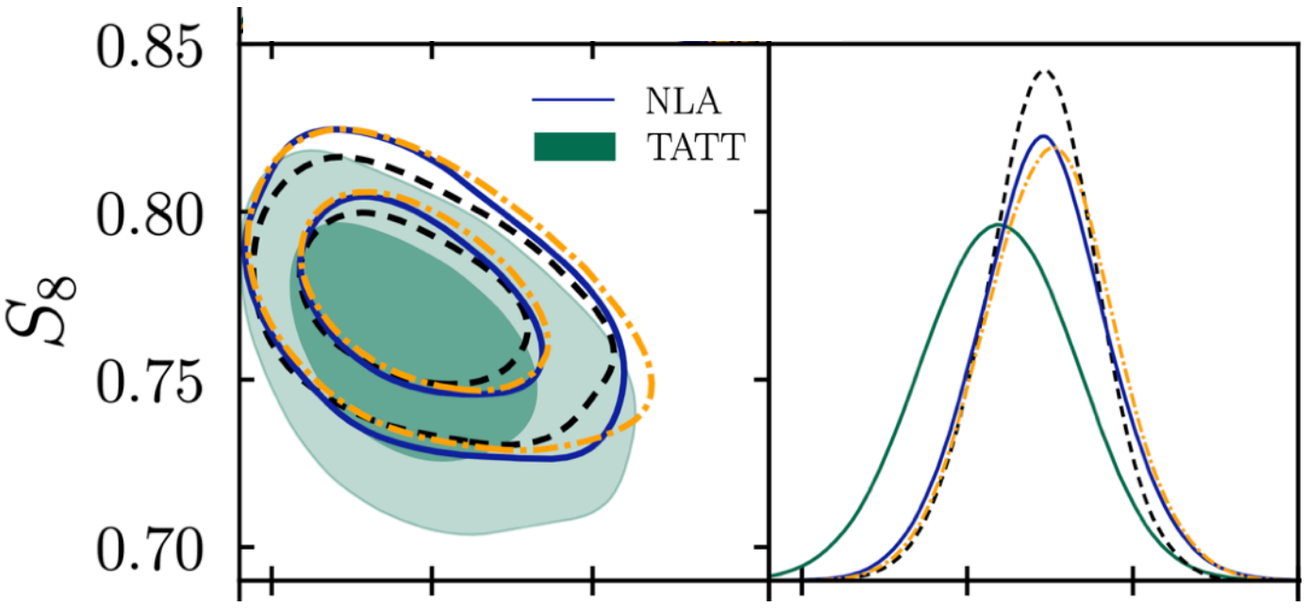
Alignment with local tidal fields

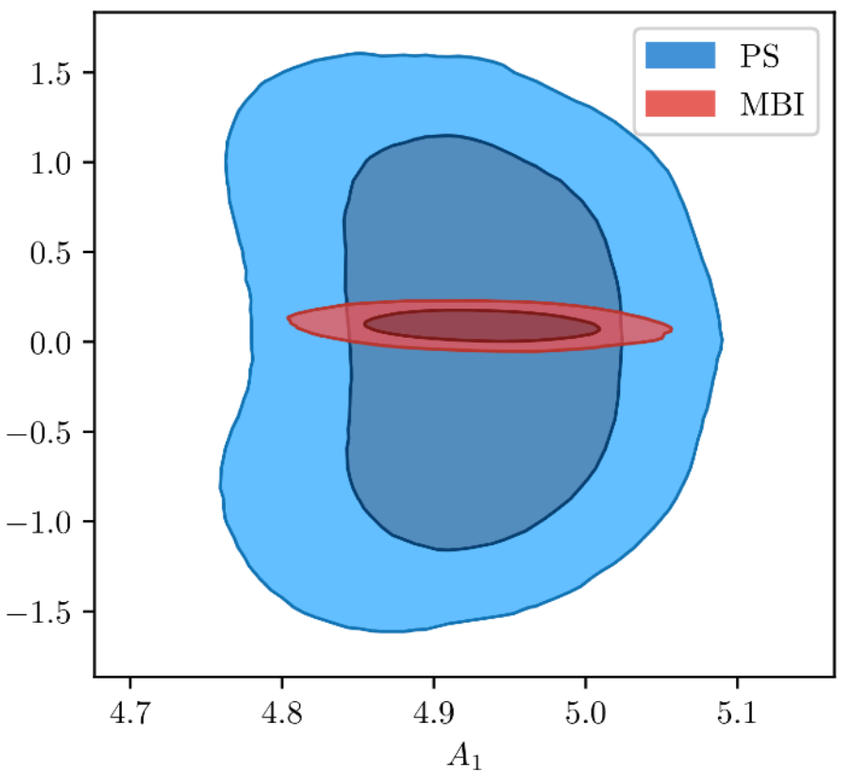
Higher order non-linear effects

Intrinsic alignment is related to the tidal fields

FLI puts much stronger constraints on the non-linear terms

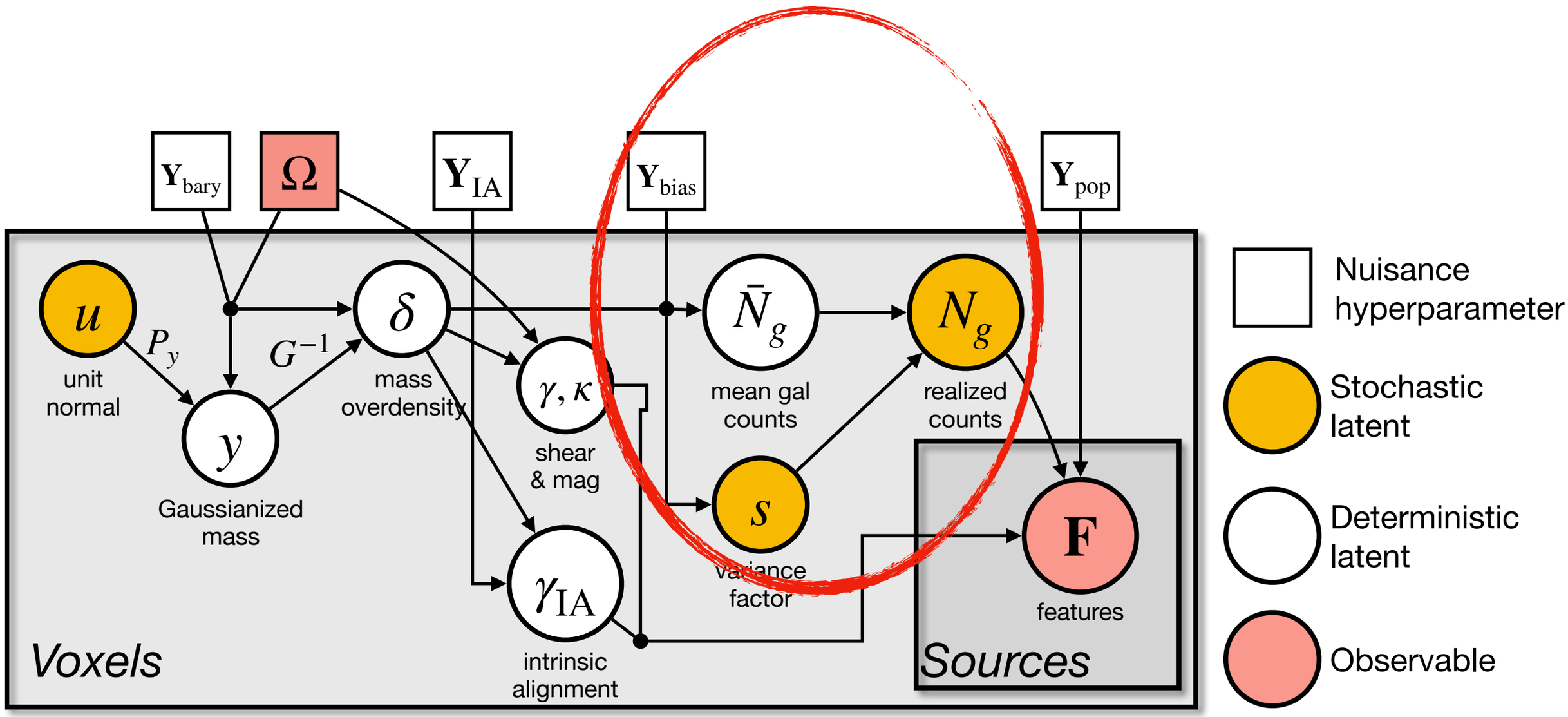
$> 25\,\%$ degradation on S_8 in DES-Y3 because of these non-linear alignment terms





FLI providing much more information on (non-linear) astrophysical systematics — making the analyses more robust (or sensitive) to these systematics

Galaxy Bias Model



Are galaxy counts really linearly (or quadratically) related to δ_m?

Is $p(N_g \mid \delta_m)$ really a Poisson distribution?

Will simulations/theory ever give us <1% accuracy for $p(N_g \mid \delta_m)$ without fitting to real-sky data?

Should we skip the theory part and just let the data tell us the appropriate functions? (a.k.a. Effective Field ~~Theory~~)

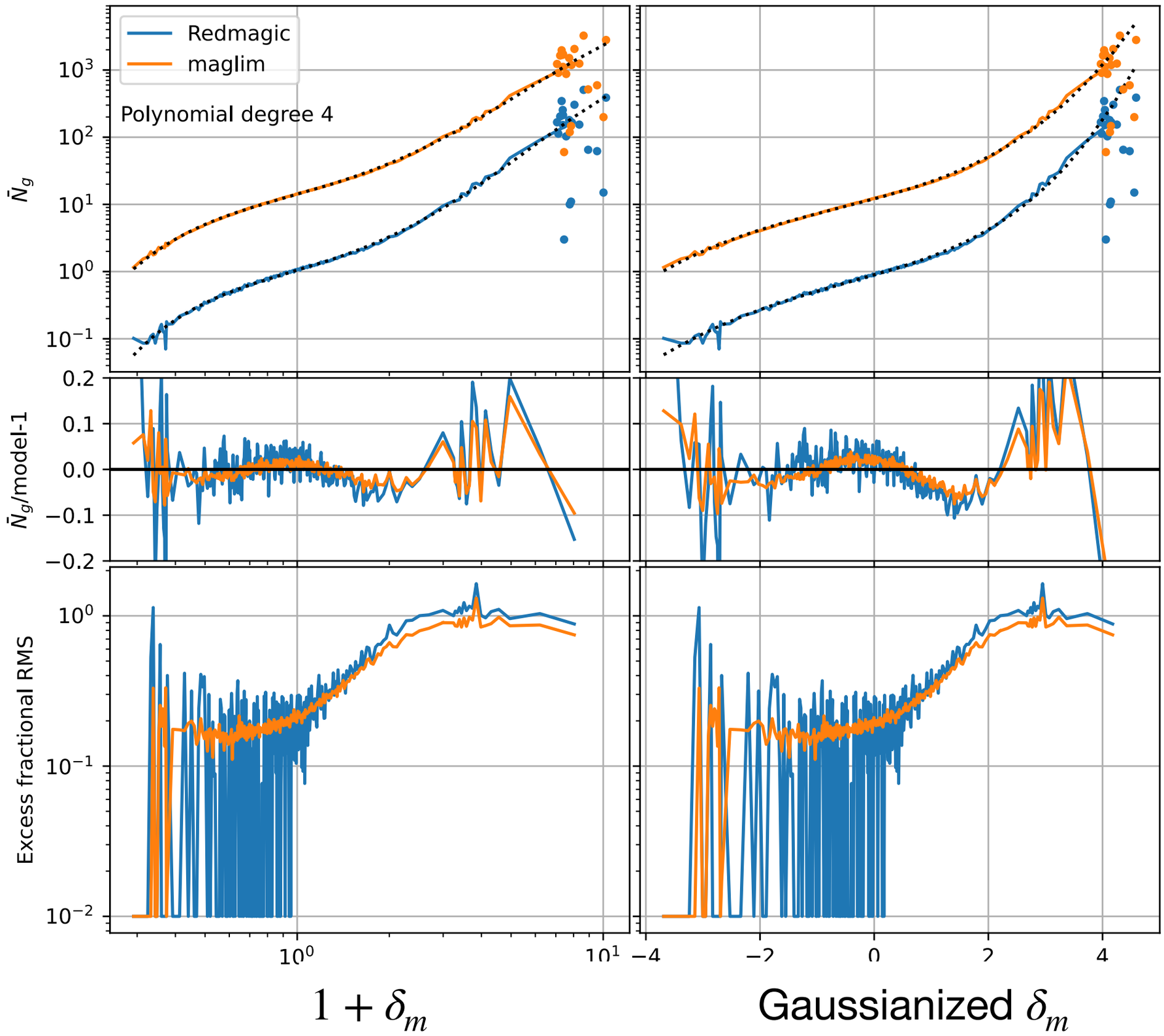
"Real" $p(N_g \mid \delta_m)$ in 4x4x125 Mpc voxels

(galaxies produced from DES HOD models)

Mean bias $\bar{N}_g(\delta_m)$ is nothing like linear or quadratic.

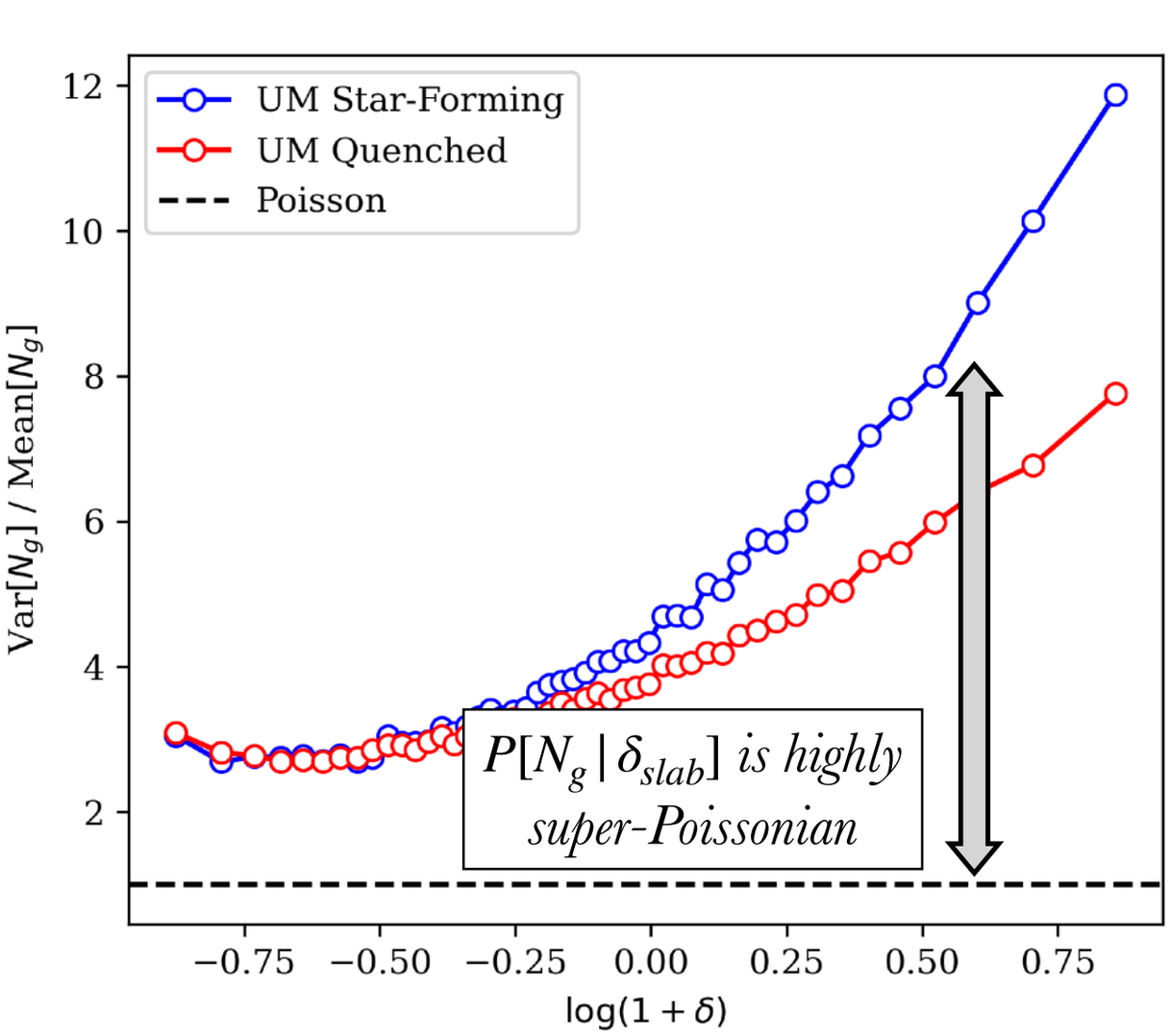
But smooth, <4 DOF?

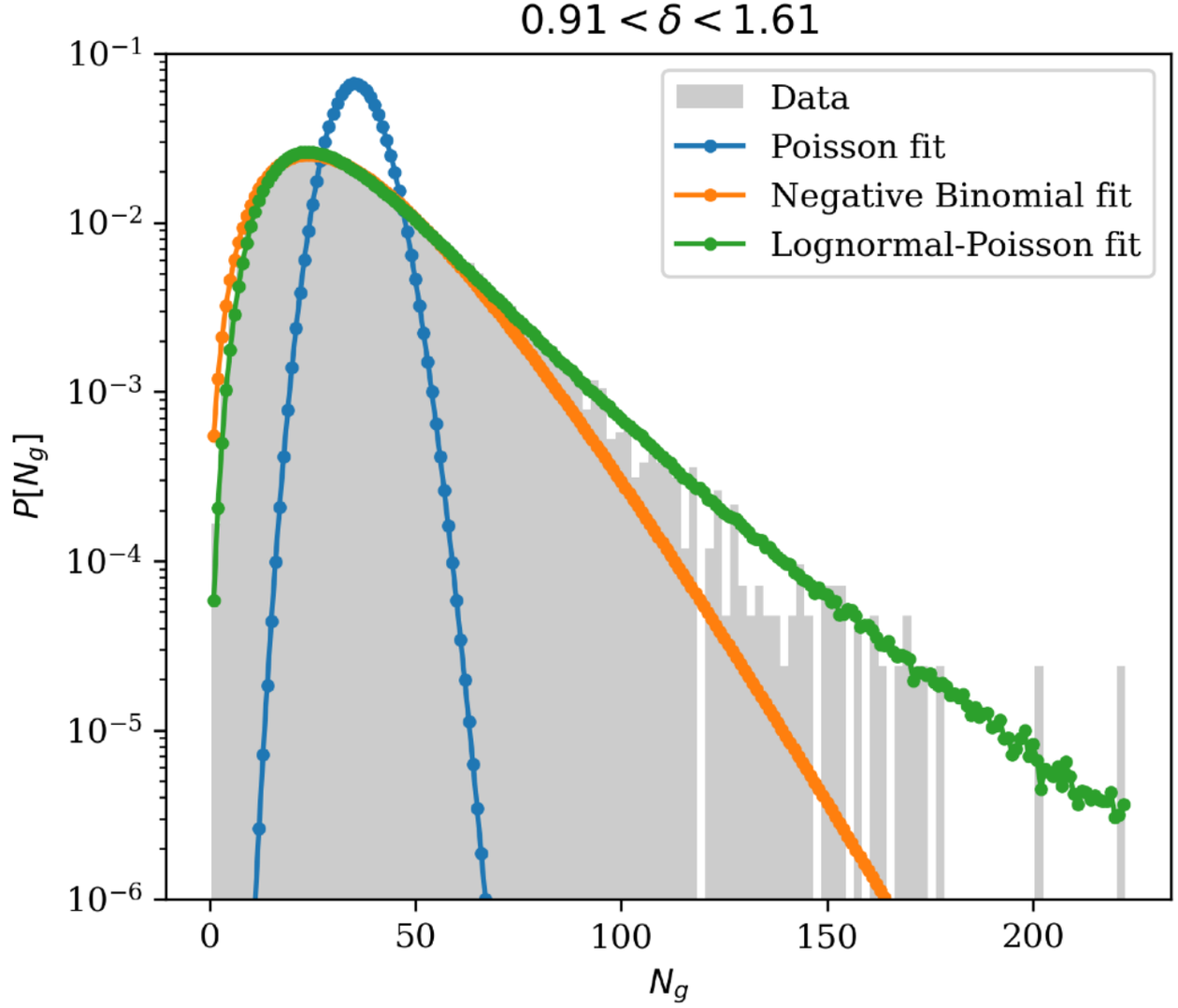
Var $N_g(\delta_m)$ does not agree with Poisson, never mind full $p(N_g \mid \delta_m)$.



Galaxy distribution in slabs — non-Poissonianity

We investigated the distribution of UniverseMachine galaxies to model $P[N_g \mid \delta_{slab}]$





- *Model the non-Poissonian galaxy distribution with a Lognormal-Poisson mixture model.*
- *Two-step sampling process*: *1.* $\lambda_i = \mu[\delta_i]\exp[\sigma(\delta_i)s_i]$, *where,* $s_i \sim \mathcal{N}(0,1)$,
  2. $N_g^i \sim \text{Poisson}[\lambda_i]$
- One latent *s* variable per voxel, i.e, need a new 'stochasticity' map

Type-dependent galaxy distribution in slabs

- *Model the joint distribution using the Lognormal-Poisson approach, but with a shared latent map approach,*

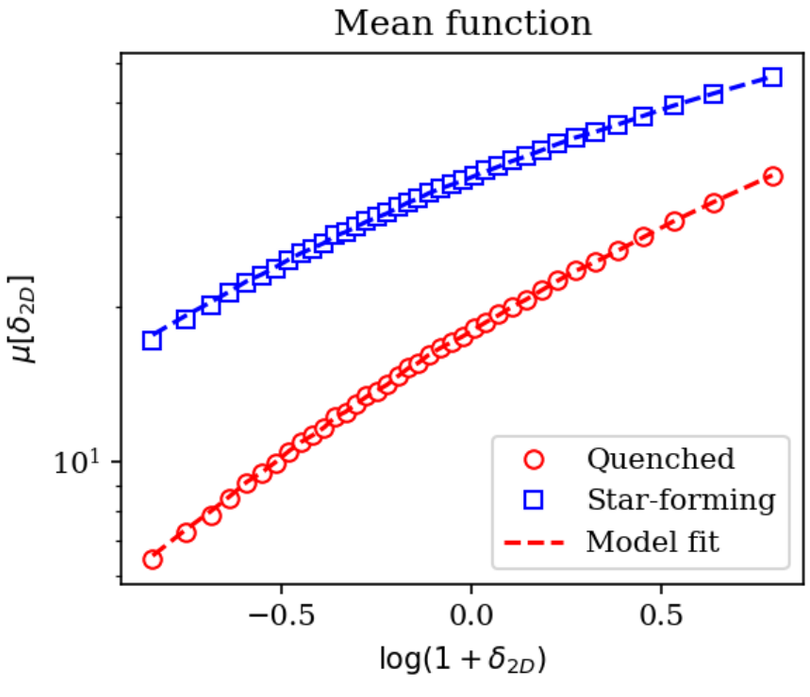
$z_i \sim \mathcal{N}[0,1]$

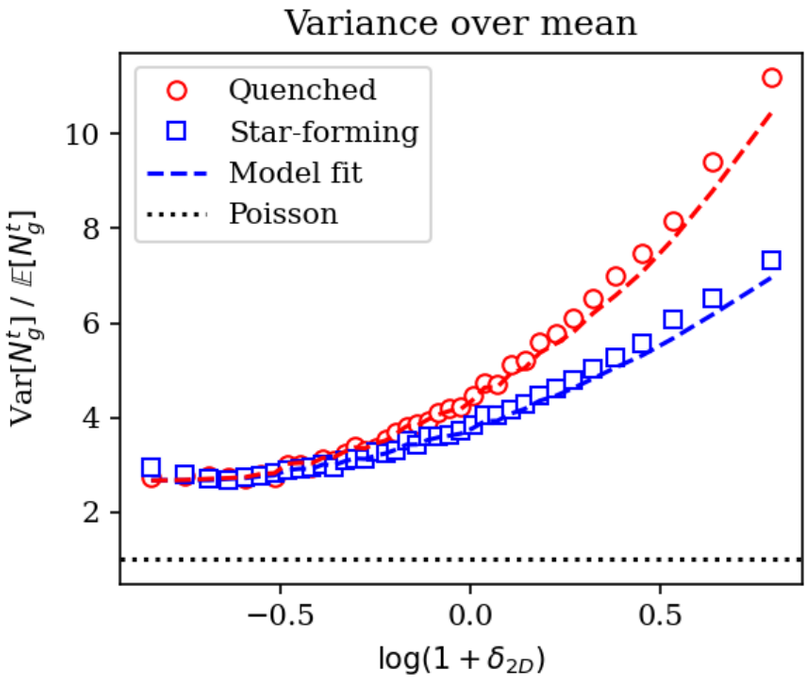
$N^i_{g,SF} \sim \text{Poisson}[\lambda^{SF}_i]$, *where,* $\lambda^{SF}_i = \mu_{SF}[\delta_i]\exp[\sigma_{SF}(\delta_i)z_i - \sigma^2_{SF}/2]$.

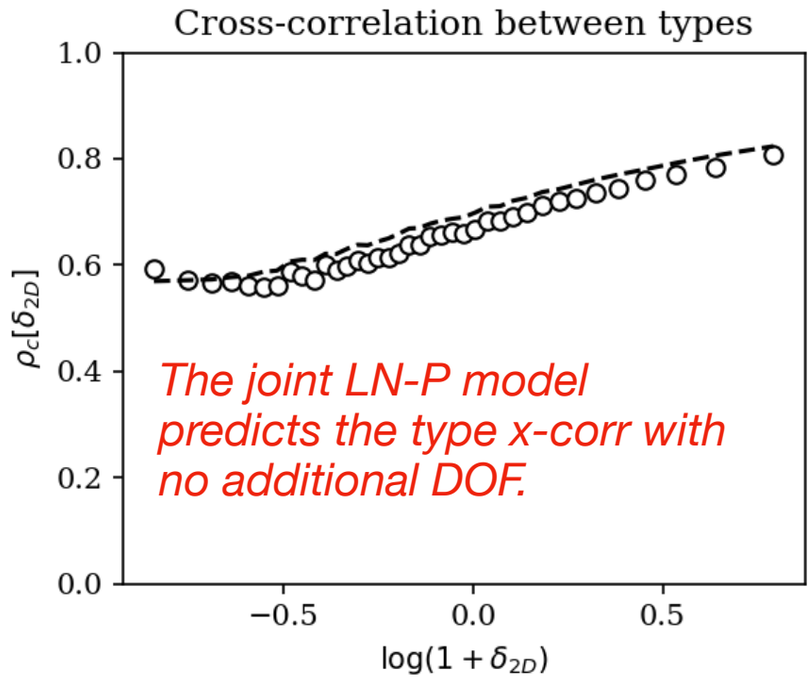
$N^i_{g,Q} \sim \text{Poisson}[\lambda^{Q}_i]$, *where,* $\lambda^{Q}_i = \mu_{Q}[\delta_i]\exp[\sigma_{Q}(\delta_i)z_i - \sigma^2_{Q}/2]$.

- *Now, different types of galaxies share the same noise map, z.*

We can model the joint, non-Poissonian distribution of red/blue galaxies in Universe Machine







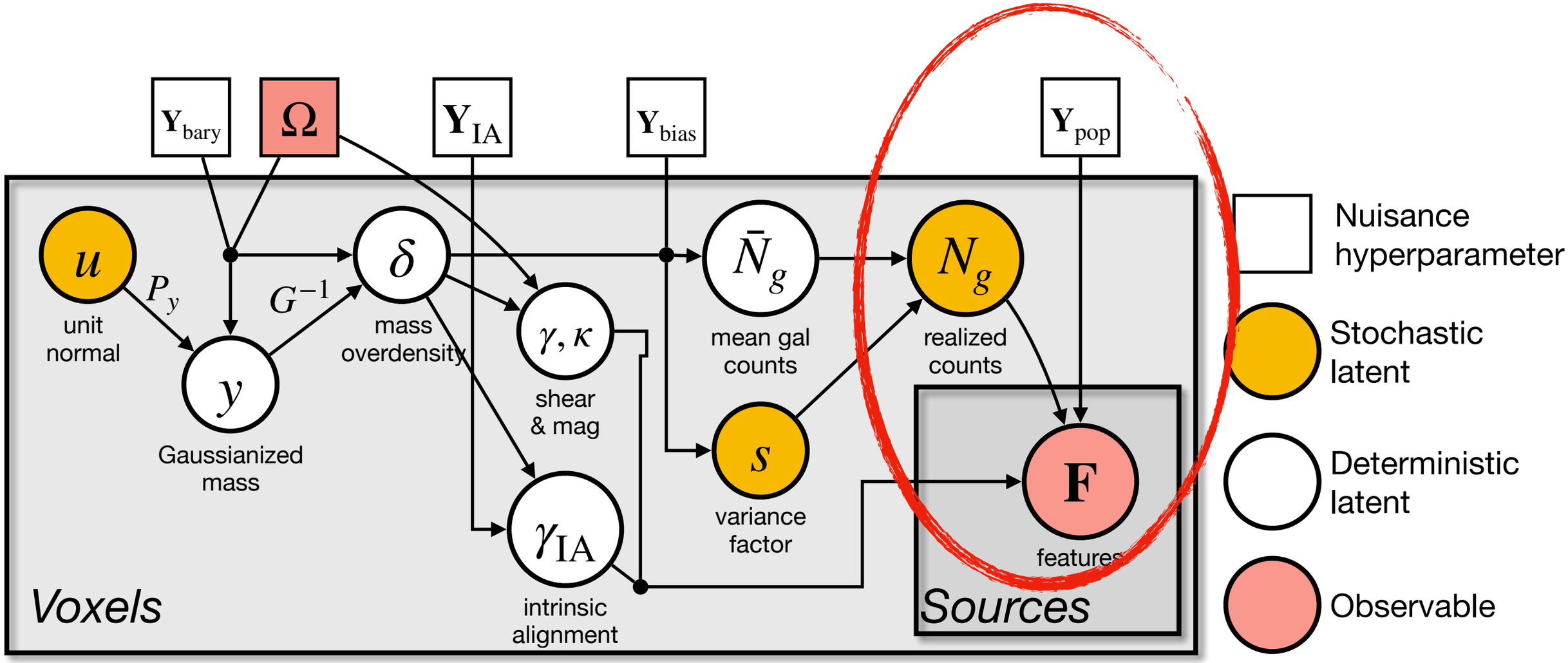
Summary of the galaxy occupation model

- $p(N_g \mid \delta_m, \text{type}, z)$ is specified by two parametric functions and one latent variable *s* per voxel:

- $\mu(\delta_m, \text{type}, z)$ gives mean value

- $\sigma(\delta_m, \text{type}, z)$ gives amount of super-Poisson variance to apply from s value.

- Use a variety of galaxy-emplacement simulations (HODs, Universe Machine / SHAM, hydro) to determine an envelope of parameteric functions that should be allowed on real data.



I don't expect we will ever have an *a priori* model for the distribution of galaxies in (color, shape/size, redshift) space.

Fortunately there are a large number of galaxies on the sky from which we can learn this distribution.

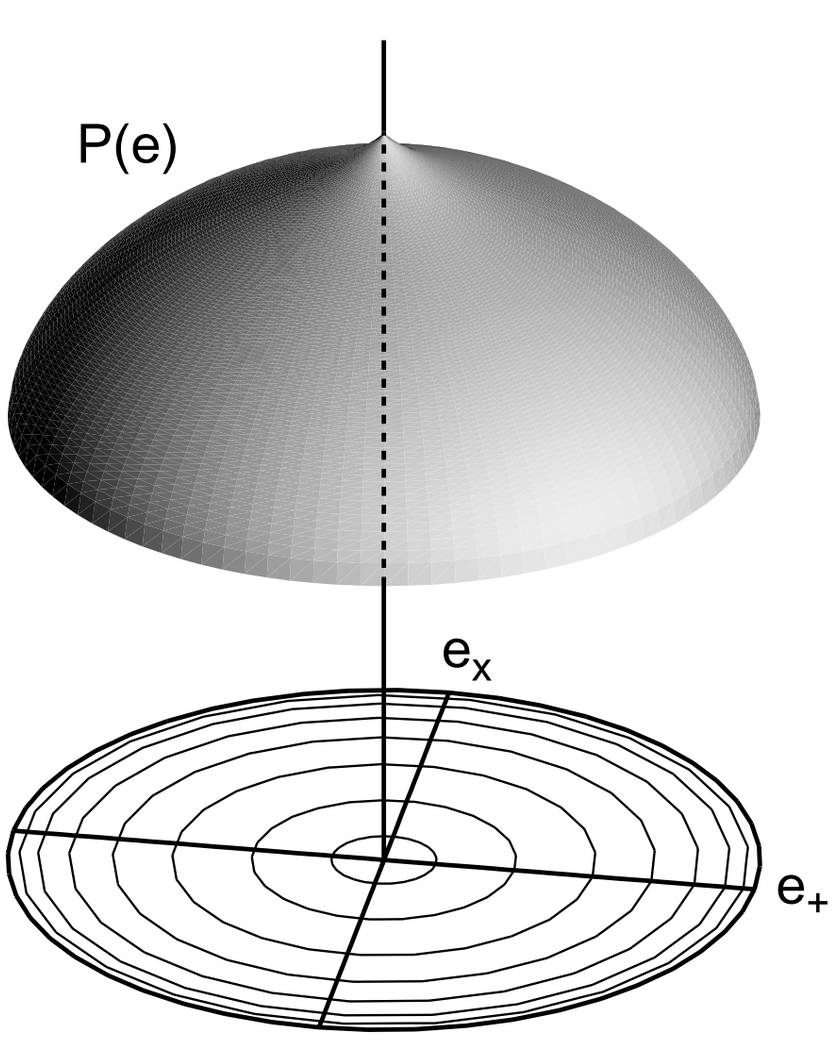
All of the high-accuracy shear measurement algorithms extract the dependence of the galaxy population on applied shear/ magnification from manipulation of the real population.

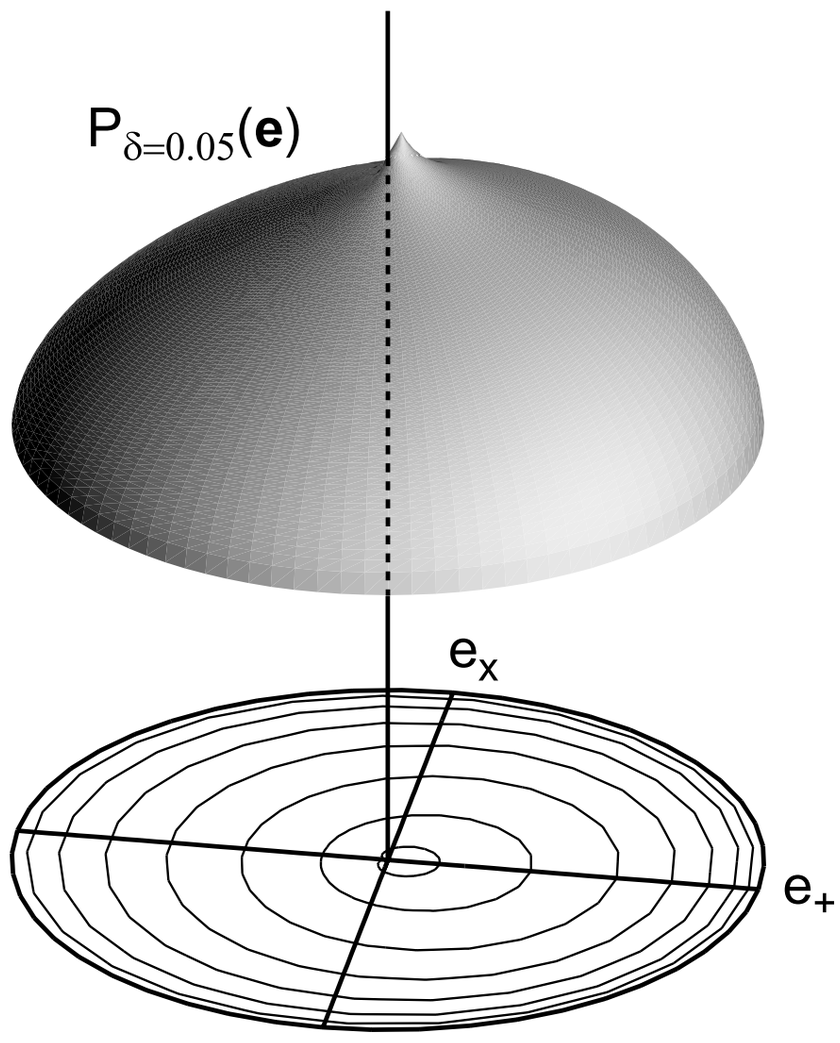
Summary of the galaxy feature model

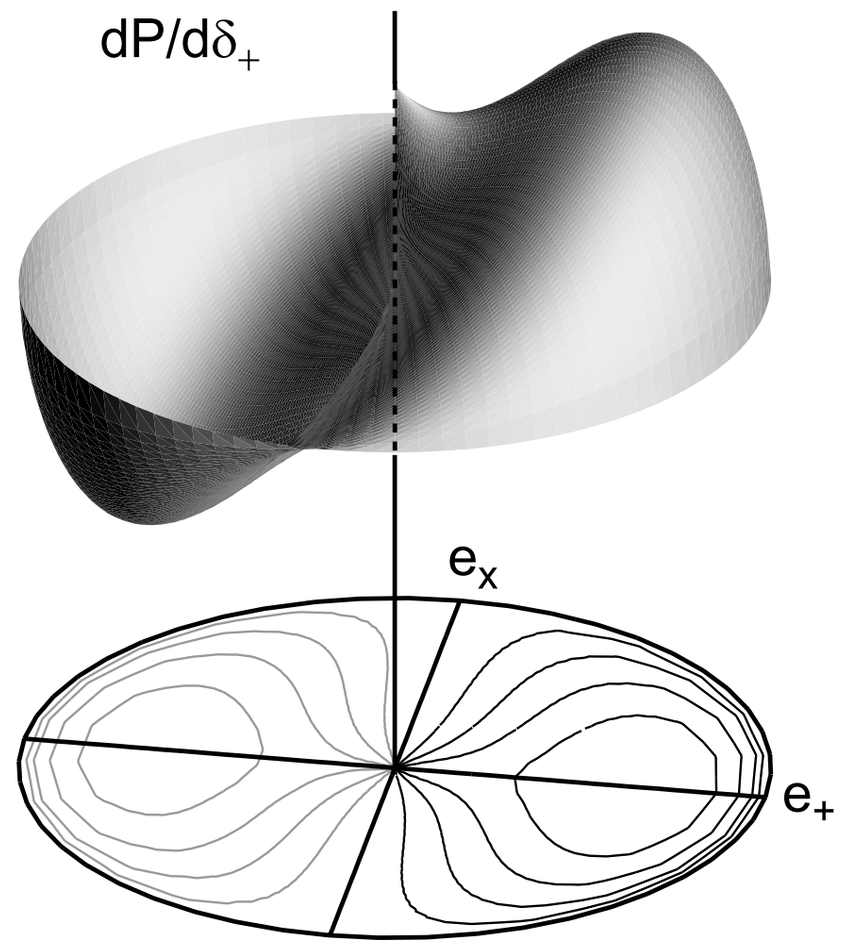
- Adopt Bayesian Fourier Domain methods: compress each galaxy g's data into a short feature vector

$$F_g = \{\text{broadband fluxes } f_g, \text{ PSF-corrected 2nd moments } M_g\}$$

- This feature set is chosen so that observational noise can be explicitly propagated into selection probability and feature noise, knowing image noise level & PSF. *Simulations for validation, not for calibration - far lower computation budget!*
- Learn sky density $p_g(F_g \mid z, \gamma, \kappa)$ of features from (artificially lensed) millions high-S/N galaxy images.
- $p(f_g \mid z_g, \text{type})$ derived from dimensional-reduction methods in many-color space (SOM used in DES, KIDS; UMAP looking great!)
- New work by Vernon Wetzell: conditional normalizing flow fits to $p(M_g \mid \gamma, \kappa, \text{type})$





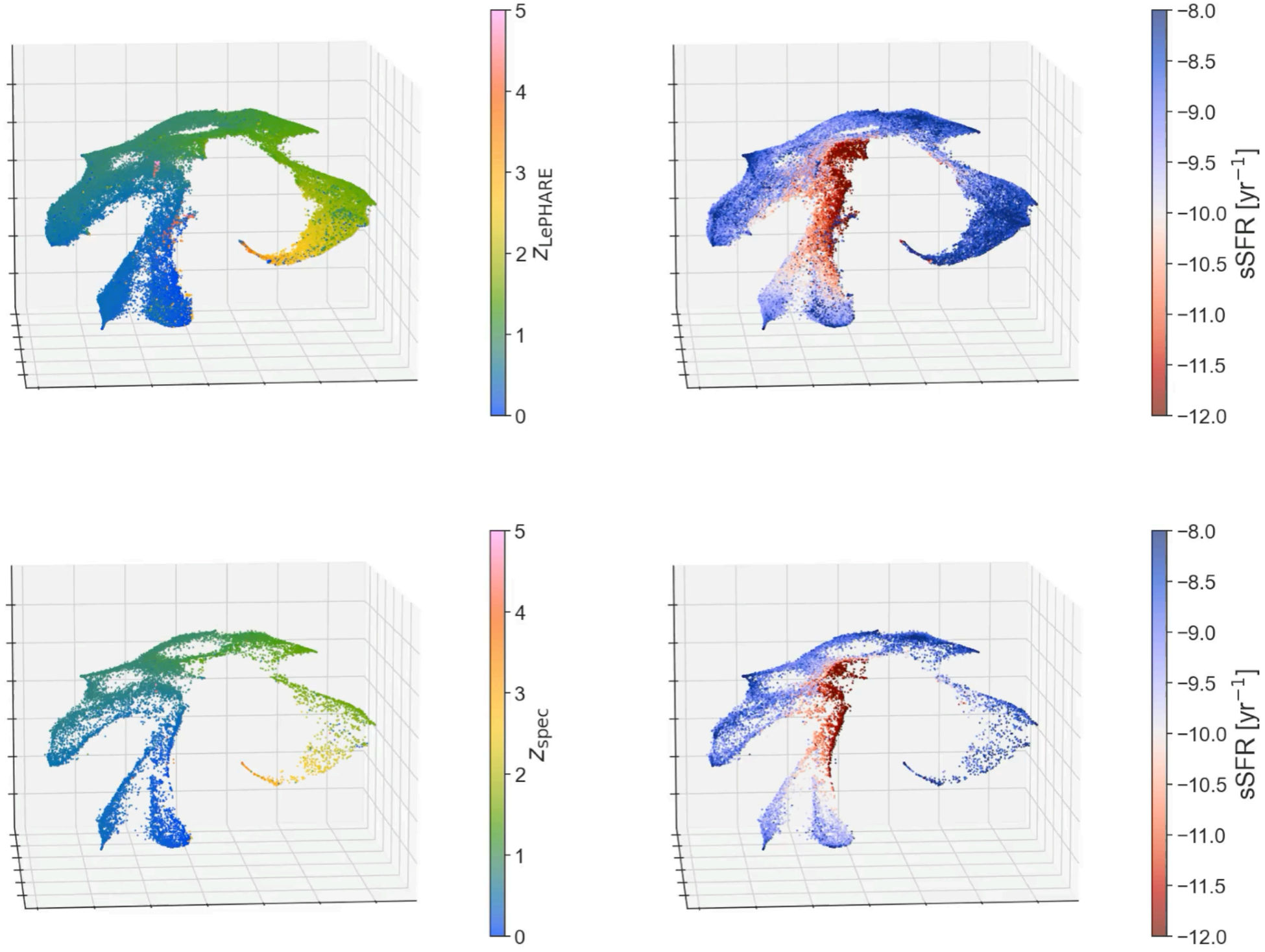


This is what we need for WL likelihood of a field proposal!

PZ UMAP from Ashmead *et al.* (2025)

Segment this 2d manifold into z/type regions

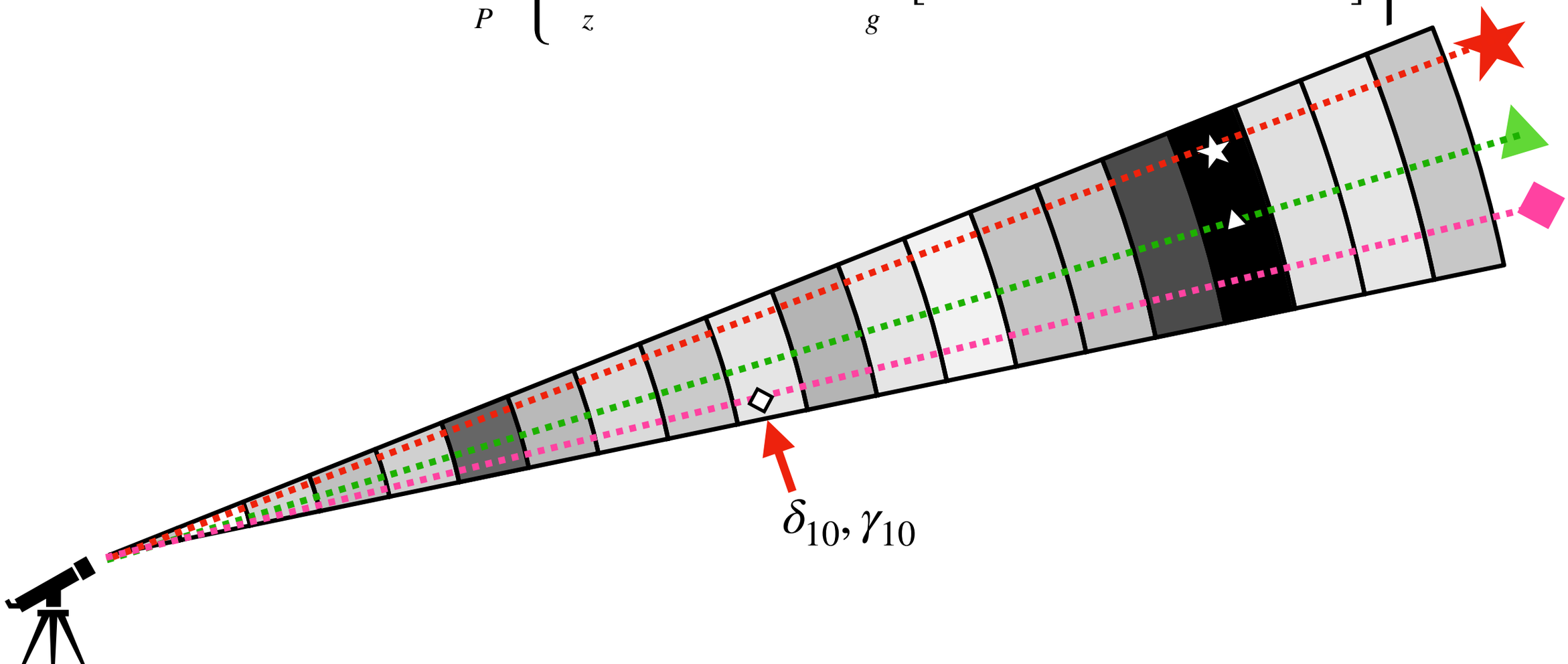
Integrate each observed source's fluxes/errors over the manifold to get its $p(f_g \mid z_g, \text{type}_g)$



The galaxy likelihood function in a given healpixel

- The voxel contains selected galaxies $g \in 1,2,\ldots N_g$ with observed flux and moment vectors f_g, M_g and latent redshift, type indices z_g, t_g .
- The sampler has proposed cosmological/baryon parameters and a realization of the density field that yield mass overdensities δ_z and applied lensing shear, magnifcation γ_z, κ_z to each voxel along the healpixel's line of sight.
- We will need to marginalize over every possible partition $P = \{z_g, t_g\}$ of the galaxies into slabs/types. A partition yields counts N_{zt} of galaxies in each slab/type combination. The likelihood of obtaining the observed catalog for this healpixel is

$$p(\{f_g, M_g\} \mid \{\delta_z, \gamma_z\}) = \sum_P \left\{ \prod_z \left[p(\{N_{zt}\} \mid \delta_z) \right] \prod_g \left[p(f_g \mid z_g, t_g) p(M_g \mid \gamma_{z_g}, z_g, t_g) \right] \right\}$$



The galaxy likelihood function in a given healpixel

- The voxel contains selected galaxies $g \in 1,2,\ldots N_g$ with observed flux and moment vectors f_g, M_g and latent redshift, type indices z_g, t_g .
- The sampler has proposed cosmological/baryon parameters and a realization of the density field that yield mass overdensities δ_z and applied lensing shear, magnifcation γ_z, κ_z to each voxel along the healpixel's line of sight.
- We will need to marginalize over every possible partition $P = \{z_g, t_g\}$ of the galaxies into slabs/types. A partition yields counts N_{zt} of galaxies in each slab/type combination. The likelihood of obtaining the observed catalog for this healpixel is

$$p(\{f_g, M_g\} \mid \{\delta_z, \gamma_z\}) = \sum_P \left\{ \prod_z \left[p(\{N_{zt}\} \mid \delta_z) \right] \prod_g \left[p(f_g \mid z_g, t_g) p(M_g \mid \gamma_{z_g}, z_g, t_g) \right] \right\}$$

- Good news: this expression contains all joint information available from the colors and shapes of every galaxy in the catalog, including redshift distribution information both from the galaxies' colors *and* the information normally extracted from "clustering redshifts"
- Bad news: completely unevaluable in practice, # of partitions is exponentially in galaxy count.

The galaxy likelihood function in a given healpixel

$$p(\{f_g, M_g\} \,|\, \{\delta_z, \gamma_z\}) = \sum_P \left\{ \prod_z \left[p(\{N_{zt}\} \,|\, \delta_z) \right] \prod_g \left[p(f_g \,|\, z_g, t_g) p(M_g \,|\, \gamma_{z_g}, z_g, t_g) \right] \right\}$$

- Let's introduce now our preferred form for $p(\{N_{tz}\} \,|\, \delta_z)$, using the latent stochasticity realization s_z also assigned by the sampler to each voxel. The first term in the sum becomes a product of Poisson distributions, with means λ_{zt}:

$$\log p(\{f_g, M_g\} \,|\, \{\delta_z, \gamma_z, s_z\}) = \sum_g \log \left[\sum_{z_g, t_g} \lambda_{z_g, t_g}(s_z)\, p(f_g \,|\, z_g, t_g)\, p(M_g \,|\, \gamma_{z_g}, z_g, t_g) \right] - \sum_{z,t} \lambda_{zt}(s_z)$$

- This expression is *much* faster to evaluate, scaling linearly with the number of galaxies and redshift bins.
- This expression is also manifestly parallelizable over healpixels, so even with 10^9 galaxies *it becomes possible to marginalize over all the latent galaxy redshifts (and types) on the fly for every sample of the field.*

Summary (definitely not conclusions!)

- We believe it's feasible to write a robust, rapidly evaluable likelihood for an entire survey catalog arising from a given 2.5d mass proposal.

- We have candidate generic local models for galaxy biasing, IA that circumscribe models / simulations to date.

- Inference code is a few hundred lines of NumPyro

- GPTG mass emulator almost done - will need to introduce baryons into the simulation space and determine scale cutoff, then see if more sophisticated ML slab model needs to be trained on simulations. Should be easier than extensive work done to date in 3d.

- Still need mass/baryon model, and a good initial *p(z | color)* map, but many other steps in current analyses are avoided.

- *What would it take for you to believe these inferences?*