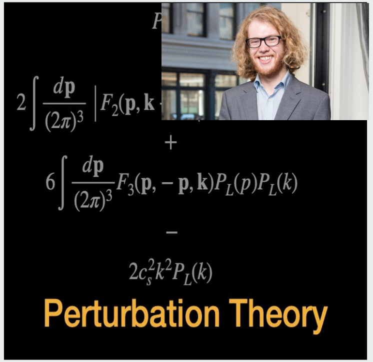
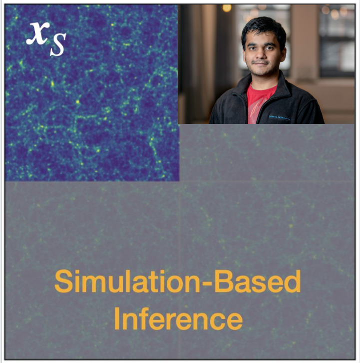
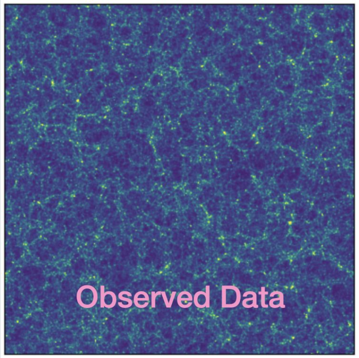
FLATIRON INSTITUTE
Center for Computational Astrophysics

Hybrid SBI or How I learned to stop worrying and learn the likelihood







Chirag Modi
Center for Computational Astrophysics
Center for Computational Mathematics
Flatiron Institute

with
Oliver Philcox

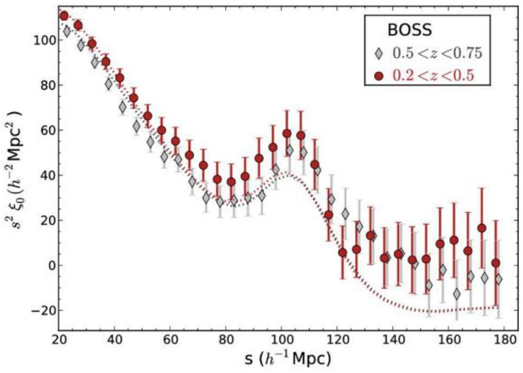
arXiv: 2309.10270

Cosmological analysis until now

Workhorse tool: analytical methods such as perturbation theory (PT / EFT of LSS)

Highly successful for-

- low-order clustering statistics such as the two- and three-point functions
- linear and quasi-linear scales.



$$
\begin{aligned}
P_g(k,\mu) =& Z_1^2(\mathbf{k})P_{\rm lin}(k) + 2\int_{\mathbf{q}} Z_2^2(\mathbf{q},\mathbf{k}-\mathbf{q})P_{\rm lin}(|\mathbf{k}-\mathbf{q}|)P_{\rm lin}(q) \\
&+ 6Z_1(\mathbf{k})P_{\rm lin}(k)\int_{\mathbf{q}} Z_3(\mathbf{q},-\mathbf{q},\mathbf{k})P_{\rm lin}(q) \\
&- 2\tilde{c}_0 k^2 P_{\rm lin}(k) - 2\tilde{c}_2 f\mu^2 k^2 P_{\rm lin}(k) - 2\tilde{c}_4 f^2\mu^4 k^2 P_{\rm lin}(k)\,, \\
&- \tilde{c} f^4 \mu^4 k^4 (b_1 + f\mu)^2 P_{\rm lin}(k) + P_{\rm shot}\,,
\end{aligned}
$$

Analytic model based on renormalized PT loop integrals

$$
\ln \mathcal{L}(\mathcal{D}|\mathbf{p}) = -\frac{1}{2}\sum_{\rm samp}\sum_{\ell,\ell'}\sum_{i,j}^{k_{\rm max}} \left[P_\ell^{\mathcal{D}}(k_i) - P_\ell(k_i;\mathbf{p})\right] \times \mathrm{Cov}^{-1}\left[P_\ell(k_i), P_{\ell'}(k_j)\right]\left[P_{\ell'}^{\mathcal{D}}(k_j) - P_{\ell'}(k_j;\mathbf{p})\right], \quad (19)
$$

Gaussian Likelihood

Ivanov et.al. 2020, Philcox & Ivanov 2022, D'Amico et.a.l. 2022 etc.

Why simulation-based inference?

We would like to -

- go beyond 2/3-point analysis (higher order statistics, learnt neural statistics)
- push to smaller scales

Standard analysis is challenging

- Need theoretical models and analytic likelihood distribution.
- PT/EFT breaks down on small scales
- Including survey systematics is difficult

Computational modeling is easier, and can be more accurate, so we would like to use simulations.

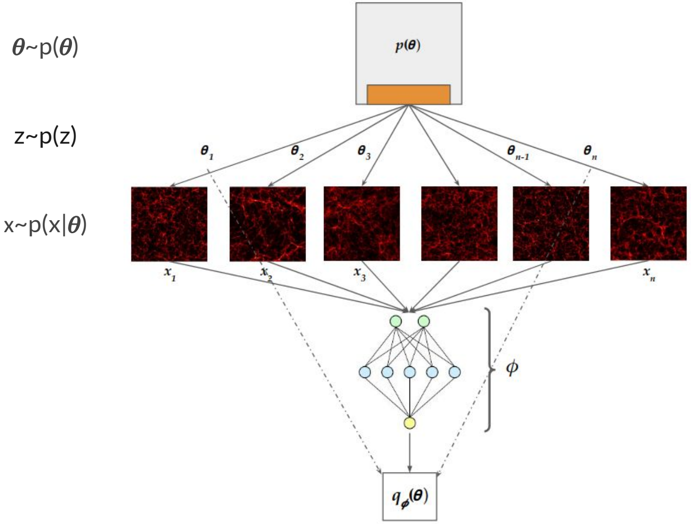
Simulation-based inference

Generating simulations is equivalent to sampling from the joint distribution

Training data = $\{x_i, \theta_i\} \sim p(x,\theta) = p(x|\theta) \times p(\theta)$

1. Neural likelihood estimation:
   Learn the likelihood function as a parametric distribution $q_\phi(x|\theta)$

2. Neural posterior estimation:
   Learn the posterior distribution as a parametric distribution $q_\phi(\theta|x)$



Flexible q ⇒ Normalizing flows

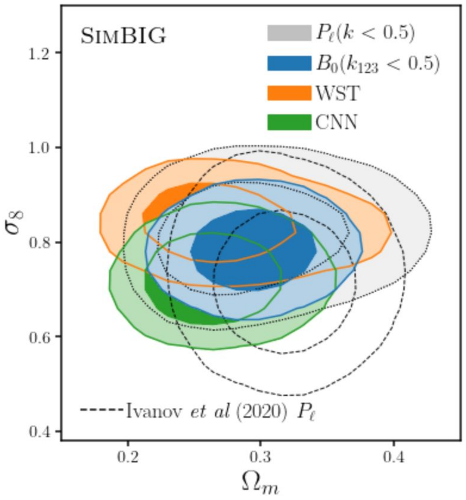
Application on data: SimBIG

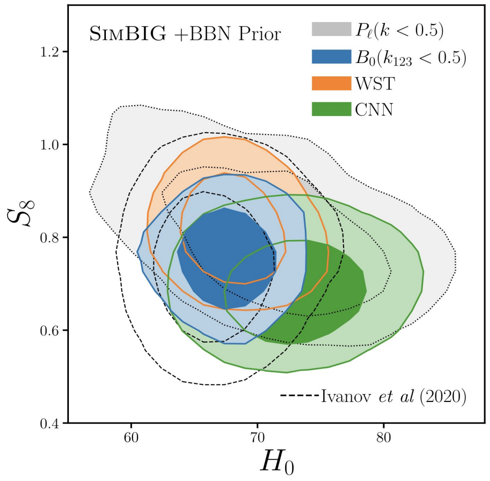
Data: 100,000 BOSS-SGC galaxies

Statistics:

- Power spectrum multipole
- Bispectrum
- Wavelet scattering transform
- CNN (field level)

Takeaway: Using higher order statistics & accessing data on the small scales improves constraints.





arXiv: 2211.00660
arXiv: 2211.00723
Credits: Changhoon Hahn, Pablo Lemos, Bruno Régaldo-Saint Blancard

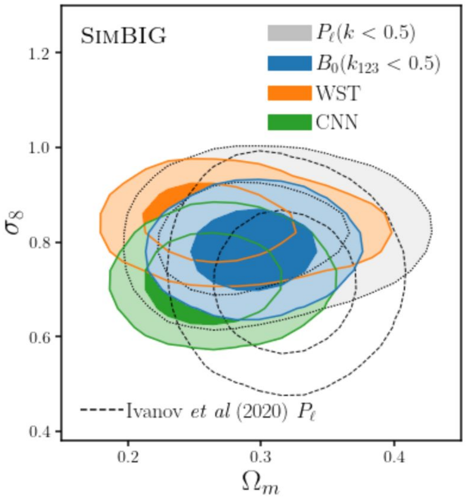
Application on data: SimBIG

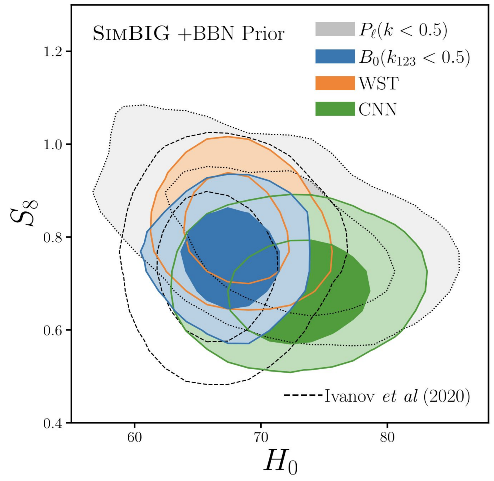
Data: 100,000 BOSS-SGC galaxies

Statistics:

- Power spectrum multipole
- Bispectrum
- Wavelet scattering transform
- CNN (field level)

Takeaway: Using higher order statistics & accessing data on the small scales improves constraints.





arXiv: 2211.00660
arXiv: 2211.00723
Credits: Changhoon Hahn, Pablo Lemos, Bruno Régaldo-Saint Blancard

Are we ready for the upcoming surveys?

Next generation of surveys

SBI for surveys like DESI, LSST, Euclid etc.

1) We will require simulations with larger volume, better mass resolution.
2) We need *increasingly accurate* forward models.

Current computational landscape: Quijote latin hypercube simulations

⟶ ~10 million CPU hours
⟶ Small volume: 1 Gpc/h - smaller than BOSS survey volume
⟶ Coarse resolution: 1 Mpc/h with 5 snapshots

Next generation of surveys

SBI for surveys like DESI, LSST, Euclid etc.

1) We will require simulations with larger volume, better mass resolution.
2) We need *increasingly accurate* forward models.

Current computational landscape: Quijote latin hypercube simulations

⟶ ~10 million CPU hours
⟶ Small volume: 1 Gpc/h - smaller than BOSS survey volume
⟶ Coarse resolution: 1 Mpc/h with 5 snapshots

Computationally prohibitive to scale to the next (current?) generation of cosmological surveys.*

How do we scale?

*for more introspection, consider the carbon cost shown in Rupert’s talk.

Recap: Motivation

SBI is needed to push to smaller scales with higher-order statistics

- higher-order statistics extract more information from non-Gaussian fields
- cannot be modeled analytically

Recap: Motivation

SBI is needed to push to smaller scales with higher-order statistics

- higher-order statistics extract more information from non-Gaussian fields
- cannot be modeled analytically

On the largest scales, *simulation-based approaches are not necessary.*

- the density field is close to Gaussian and can be modeled using PT
- traditional statistics like the P(k), B(k) are close to optimal (Cabass et al. 2023)

Recap: Motivation

SBI is needed to push to smaller scales with higher-order statistics

- higher-order statistics extract more information from non-Gaussian fields
- cannot be modeled analytically

On the largest scales, *simulation-based approaches are not necessary.*

- the density field is close to Gaussian and can be modeled using PT
- traditional statistics like the P(k), B(k) are close to optimal (Cabass et al. 2023)

Combine PT on large scales with SBI on small scales

Hybrid SBI (HySBI)

HySBI: formalism

Data-vector **x** can be split into two components $\mathbf{x}=\{x_L, x_S\}$ -- large scales $\mathbf{x}_L$, and small scales $\mathbf{x}_S$

$$p(x|\boldsymbol{\theta}) = p(x_L|\boldsymbol{\theta}) \times p(x_S|x_L, \boldsymbol{\theta})$$

HySBI: formalism

Data-vector **x** can be split into two components $\mathbf{x}=\{\mathbf{x}_L, \mathbf{x}_S\}$ -- large scales $\mathbf{x}_L$, and small scales $\mathbf{x}_S$

$$p(x|\boldsymbol{\theta}) = p(x_L|\boldsymbol{\theta}) \times p(x_S|x_L, \boldsymbol{\theta})$$

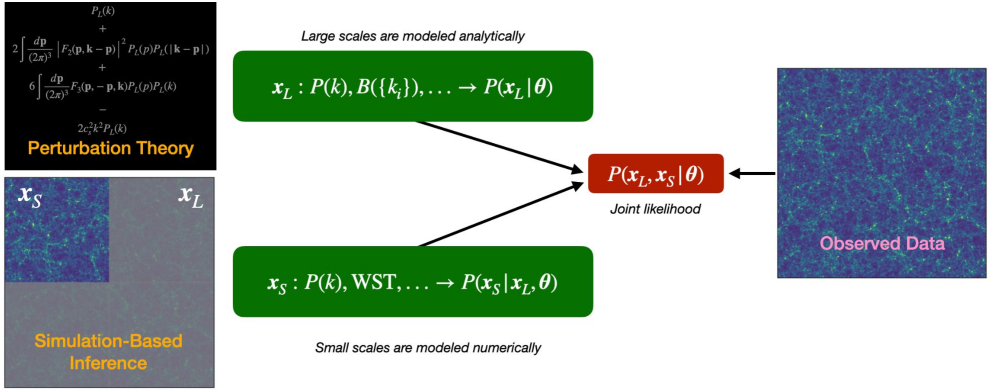
$p(x_L|\boldsymbol{\theta})$: model analytically with perturbation theory

x_L : classical statistics such as the P(k), B(k)

$p(x_s|x_L, \boldsymbol{\theta})$: learnt with SBI

simulating only a small sub-volume at high-fidelity, instead of the full survey volume

x_s : any statistic of choice (P(k), B(k), wavelets, neural statistics)

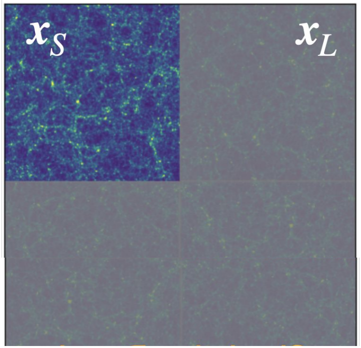


HySBI: no free lunch*

Two new issues:

(1) Learning $\mathbf{p(x_s | x_L, \boldsymbol{\theta})}$ requires new, customized simulations

- depends on $\mathbf{x_L}$
- need access to the correct large-scale statistics $\mathbf{x_L}$ corresponding to $\mathbf{x_s}$ …without simulating the entire volume at *full-fidelity*
- simulations with separate evolution on large and small scales, e.g., S-COLA, zoom-ins



*yes, I am aware that I am making this statement at the Flatiron Institute :)

HySBI: no free lunch*

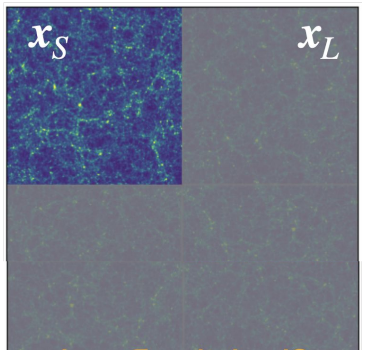
Two new issues:

(1) Learning $\mathbf{p(x_s|x_L,\boldsymbol{\theta})}$ requires new, customized simulations

- depends on $\mathbf{x_L}$
- need access to the correct large-scale statistics $\mathbf{x_L}$ corresponding to $\mathbf{x_s}$ …without simulating the entire volume at *full-fidelity*
- simulations with separate evolution on large and small scales, e.g., S-COLA, zoom-ins

(2) Super sample effects

- evolution in sub-volume is affected by large scale modes from the full box
- small scale statistics $\mathbf{x_s}$ are noisy



*yes, I am aware that I am making this statement at the Flatiron Institute :)

HySBI: proof-of-principle

Setup: Infer $\boldsymbol{\Omega}_m$ and $\boldsymbol{\sigma}_8$ from three-dimensional dark matter density field

$\mathbf{x}_L$: power spectrum (k < 0.15 h/Mpc)

p(x$_L$| $\boldsymbol{\theta}$):

$$-2\log p(\boldsymbol{x}_L|\boldsymbol{\theta}) = \sum_k \left[\frac{P_{\rm loop}(k) - 2c_s^2 P_{\rm ct}(k) - \hat{P}(k)}{\sigma_P(k)}\right]^2 \quad (2)$$

$\mathbf{x}_S$: power spectrum (0.15 < k < 0.5 h/Mpc), wavelet coefficients

p(x$_s$|x$_L$, $\boldsymbol{\theta}$): split 1 Gpc/h Quijote simulations into 8 sub-volumes
measure $\mathbf{x}_s$ in the sub-volumes *(only for training)*

HySBI: proof-of-principle

Setup: Infer $\boldsymbol{\Omega}_m$ and $\boldsymbol{\sigma}_8$ from three-dimensional dark matter density field

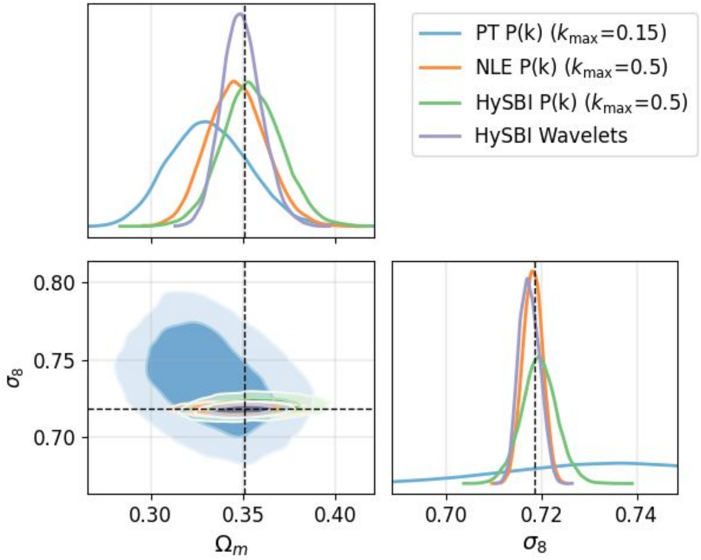
$\mathbf{x}_L$: power spectrum (k < 0.15 h/Mpc)

p($\mathbf{x}_L$ | $\boldsymbol{\theta}$):

$$-2\log p(\boldsymbol{x}_L|\boldsymbol{\theta}) = \sum_k \left[\frac{P_{\rm loop}(k) - 2c_s^2 P_{\rm ct}(k) - \hat{P}(k)}{\sigma_P(k)}\right]^2 \quad (2)$$

$\mathbf{x}_S$: power spectrum (0.15 < k < 0.5 h/Mpc), wavelet coefficients

p($\mathbf{x}_S$ | $\mathbf{x}_L$, $\boldsymbol{\theta}$): split 1 Gpc/h Quijote simulations into 8 sub-volumes
measure $\mathbf{x}_S$ in the sub-volumes *(only for training)*



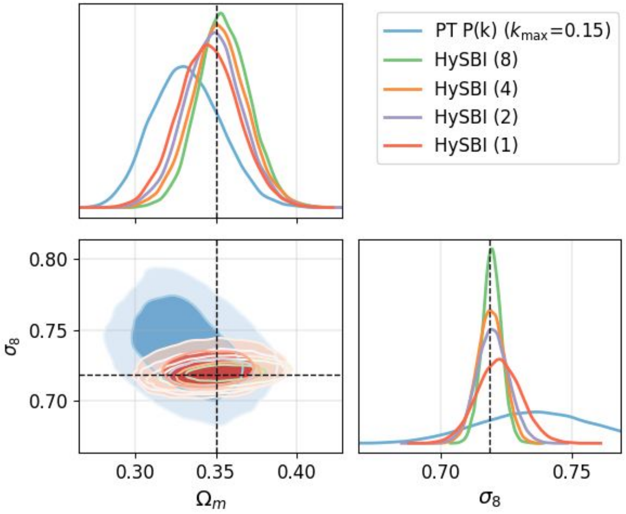
Results:

- HySBI outperforms traditional analysis
- Global NLE with P(k) better than HySBI because PT marginalizes over $\boldsymbol{c_s}$

HySBI: super-sample effects

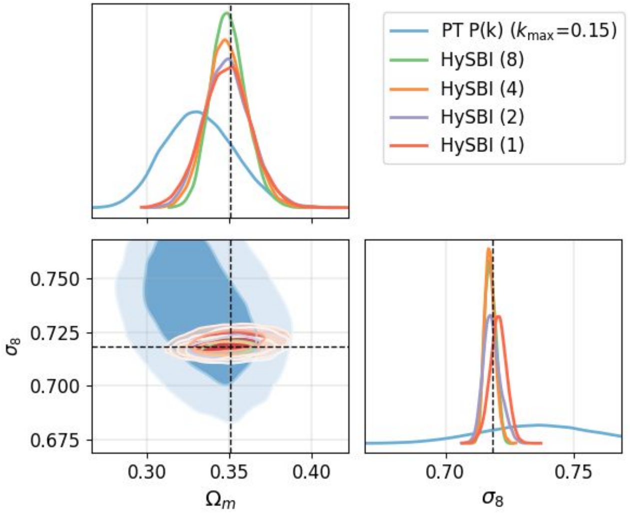
Significant gains with using only one-eighth of the simulation volume!

HySBI with power spectrum



uncertainties are inflated by
5 - 10% for $\boldsymbol{\Omega_m}$
40 - 120% for $\boldsymbol{\sigma_8}$
upon using 4, 2, 1 sub-volumes

HySBI with wavelets



uncertainties are inflated by
20 - 50% for $\boldsymbol{\Omega_m}$
40 - 100% for $\boldsymbol{\sigma_8}$
upon using 4, 2, 1 sub-volumes

Summary

- SBI is one the most promising techniques to go beyond current cosmological analyses

- We do not have the computational resources to generate training dataset for upcoming surveys

- **Hybrid SBI–** combine PT on large scales with SBI on small scales, trained on small sub-volumes
    - a realistic path for scaling SBI to large survey volumes

- Beyond proof of principle:
    - Customized simulations with approximate large-scale evolution & accurate small-scale simulations
        - multi-grid force computation (FlowPM), S-COLA, zoom-in simulations

    - Consistently treat nuisance parameters for observables like galaxies
        - Bias parameters and counter-terms in PT/EFT, and HOD parameters for SBI, led by Gemma Zhang

    - Correctly account for systematic effects like survey masks that mix small and large scales

Thank you!