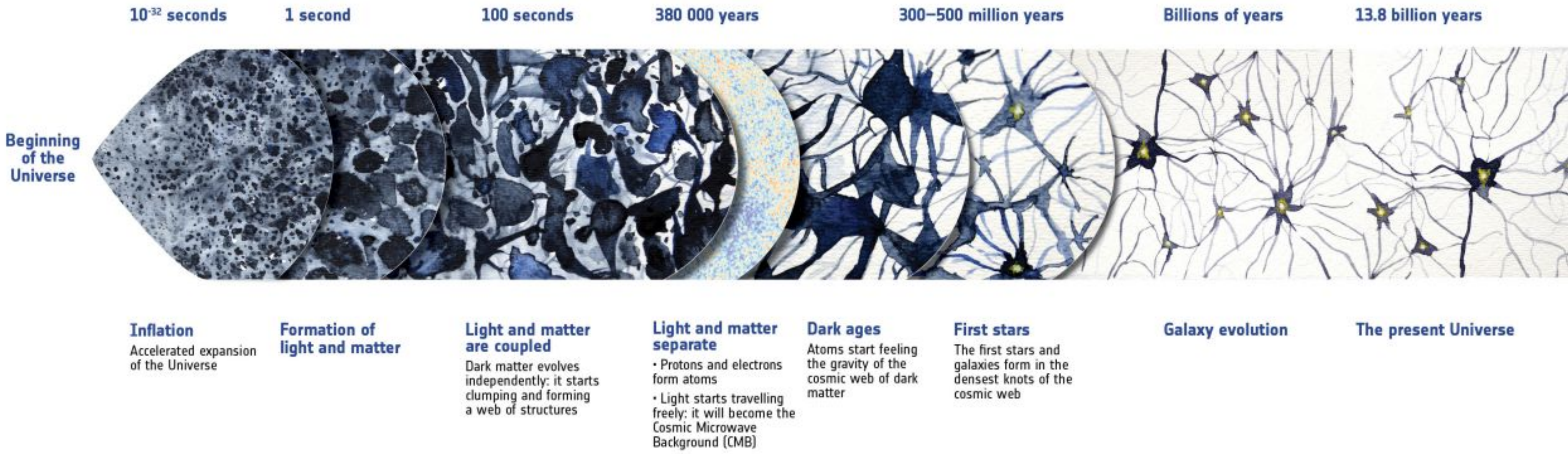
Machine learning boosted cosmological inference

Guilhem Lavaux (IAP/CNRS)

with Aquila consortium & Learning the Universe collaboration

Department of Astronomy, University of Helsinki, Finland (February 23, 2024)

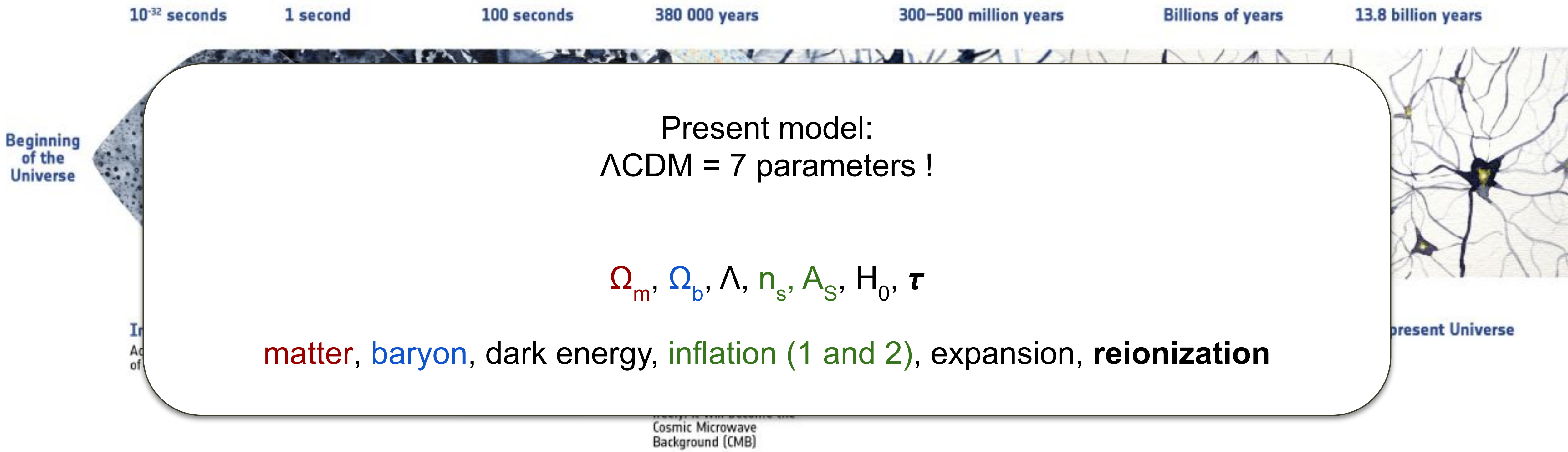
Cosmological context: current paradigm



Dynamical evolution of the universe from first instant to present time

Image credit: Planck collaboration

Cosmological context: current paradigm



Present model:
ΛCDM = 7 parameters !

Ω_m, Ω_b, Λ, n_s, A_S, H_0, $\boldsymbol{\tau}$

matter, baryon, dark energy, inflation (1 and 2), expansion, **reionization**

Dynamical evolution of the universe from first instant to present time

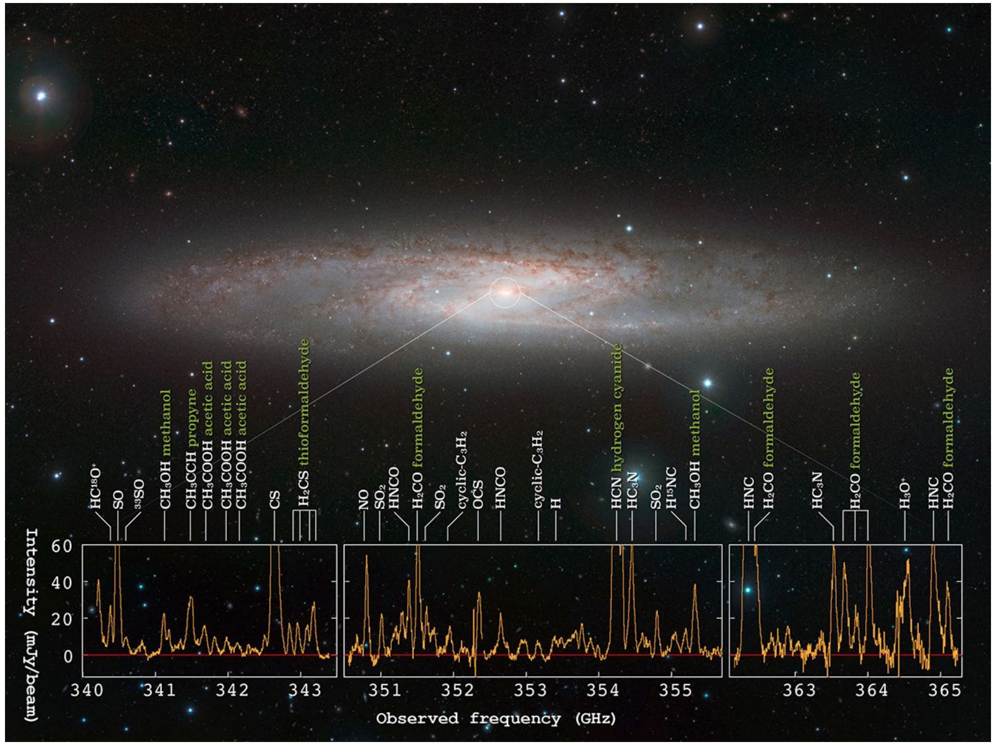
Image credit: Planck collaboration

Observations of large scale structures of the Universe

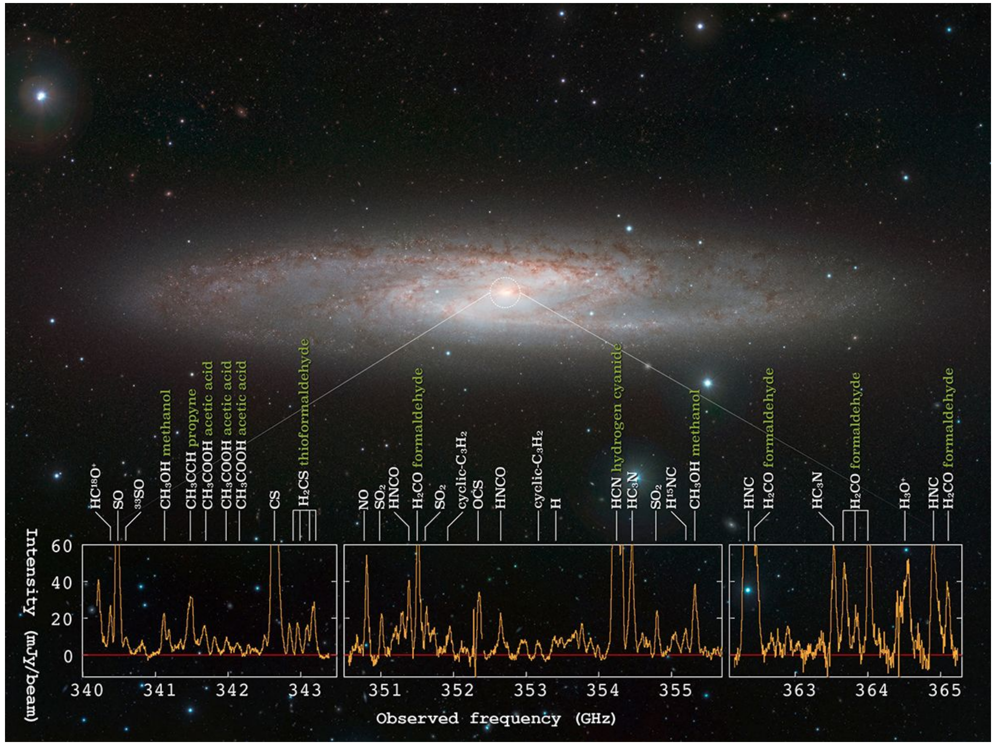
©CFHT/Coelum/Terapix/AstrOmatic

Mapping the Universe

+



Photometry



Spectrometry

Redshift z

Angular position

"3D Position"

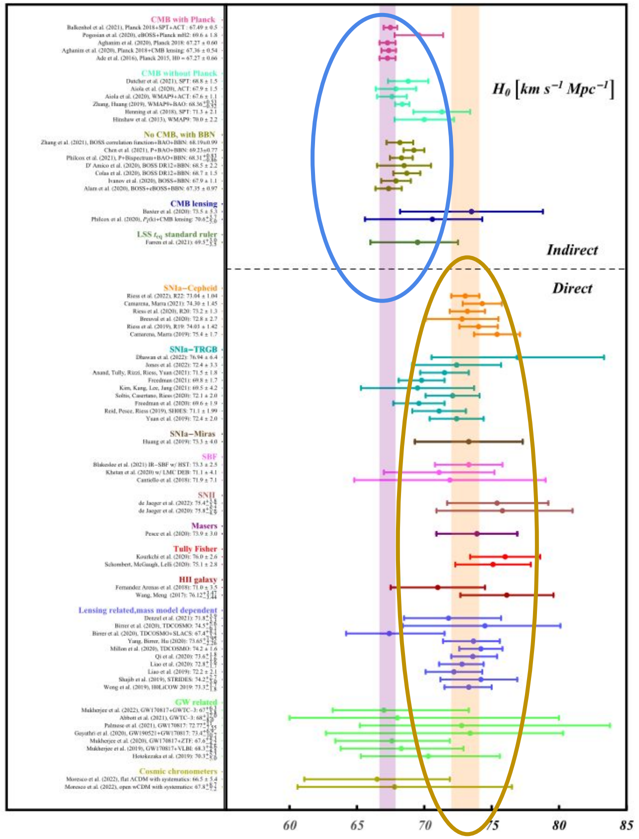
Distance

A special time with potential of discovery

ΛCDM model at the basis of the present paradigm is under tension

… and only using 2 or 3-point statistics ! What lurks beyond?

Local Expansion of the Universe (H_0)



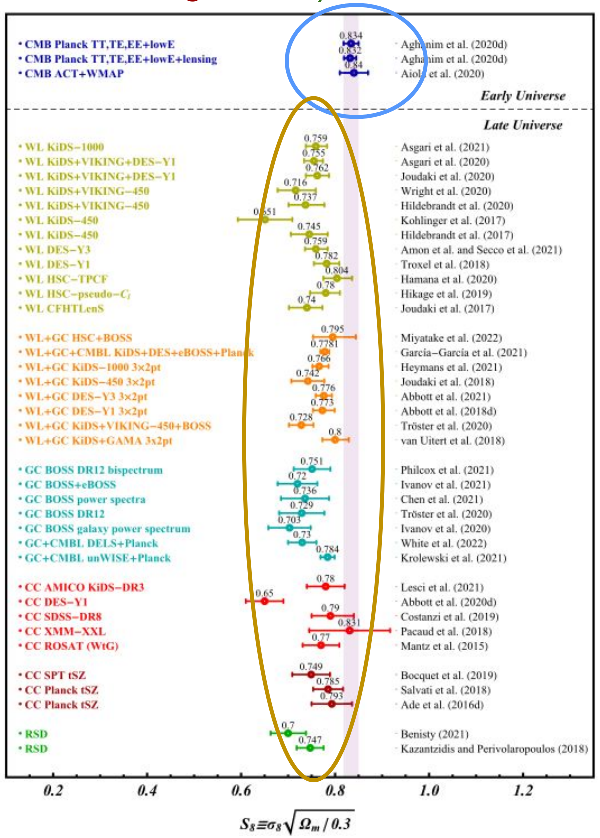
Abdalla et al. (2022, JHEA)

Early Universe

Late Universe

Small scale dynamics

$S_8 = f(\Omega_m, \sigma_8)$



A special time with potential of discovery

ΛCDM model at the basis of the present paradigm is under tension

Opportunities / Problems / Objectives

Delivery of massive new datasets

Absolute volume of observable universe is limited

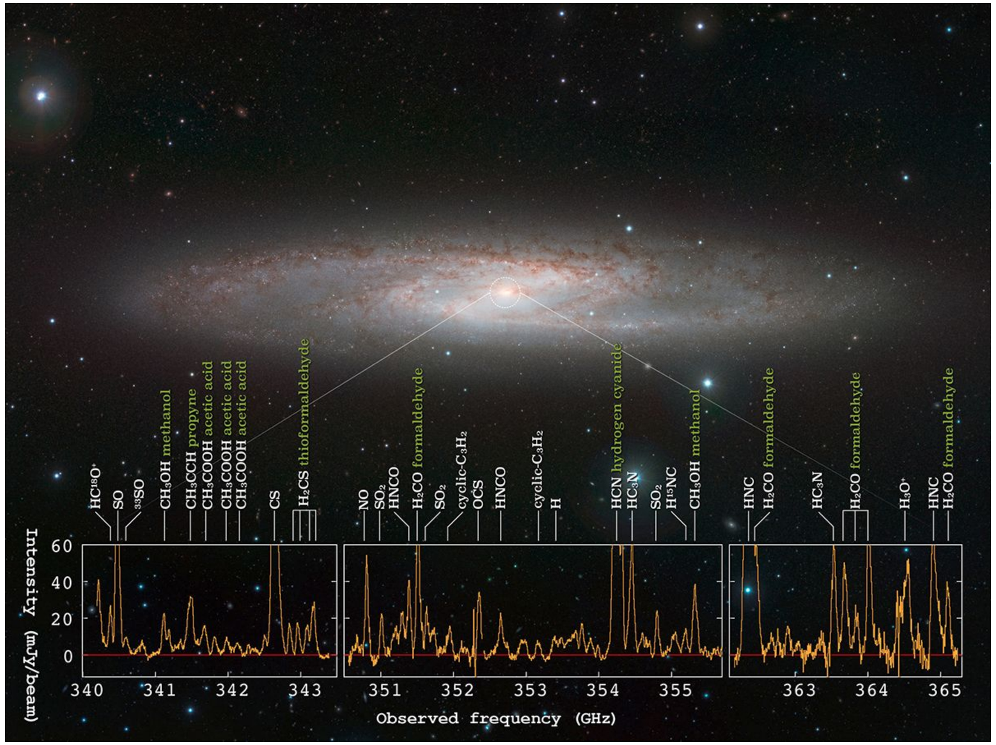
Direct information on primordial universe is low

Do better than 2- or 3- points statistics with modern data assimilation techniques

Potential for discovery of new physics

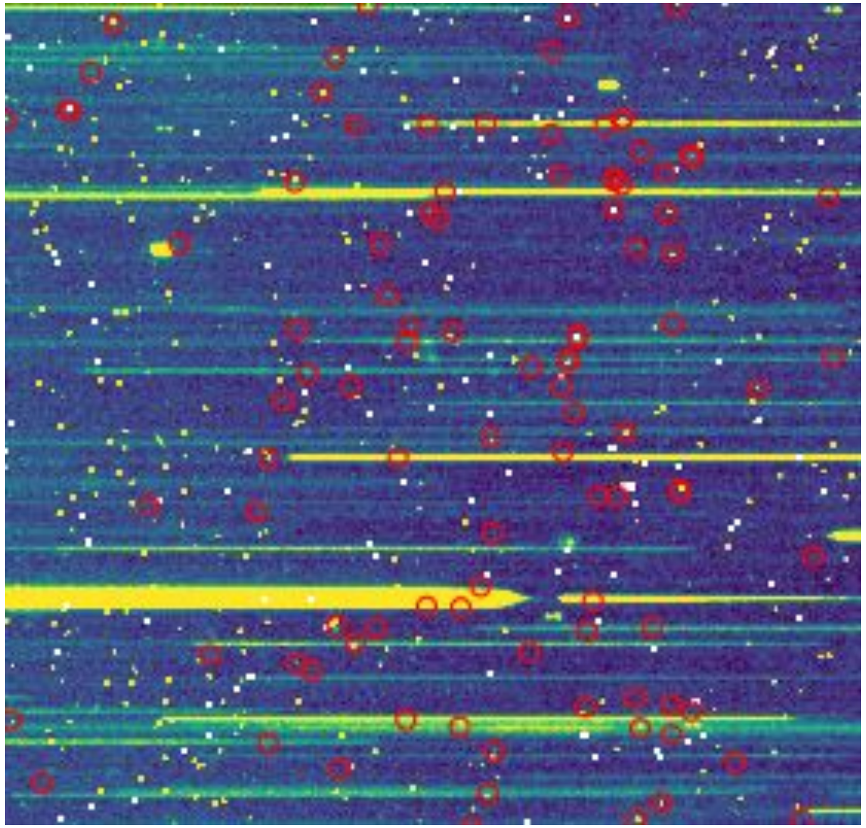
Survey challenges: spectroscopy

- New surveys = more complicated data processing, e.g. slitless spectroscopy



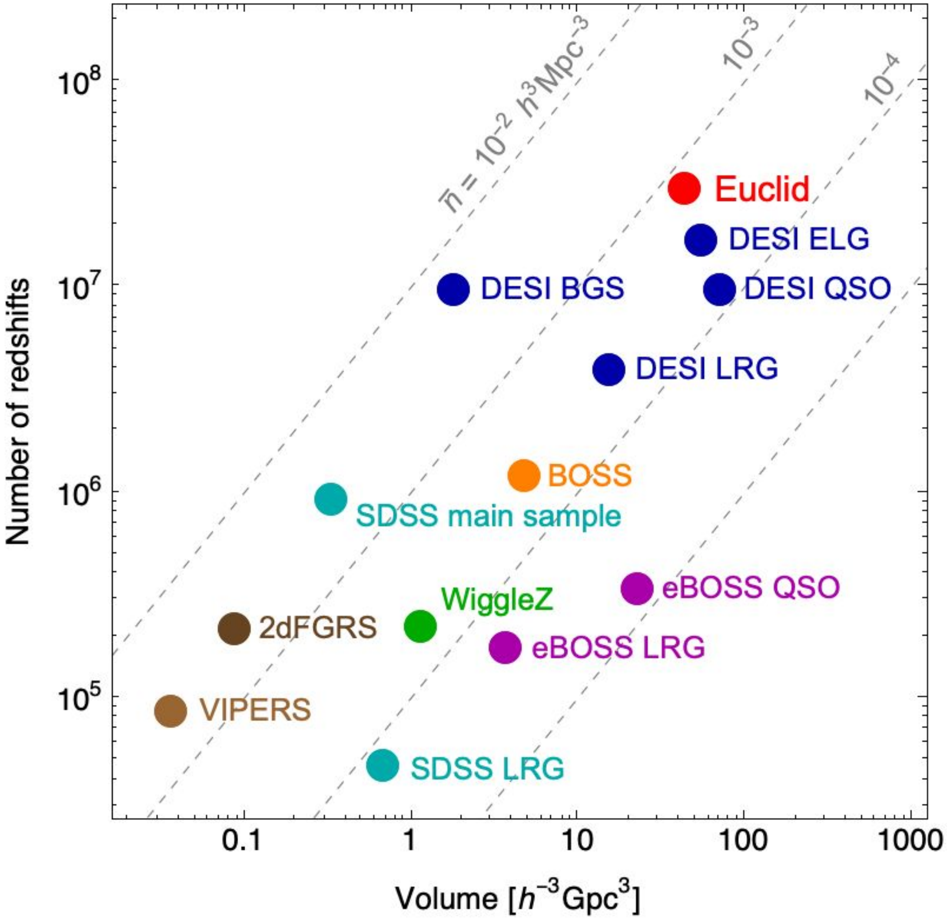
Example: a good galaxy spectrum

- New surveys = more complicated data processing, e.g. slitless spectroscopy

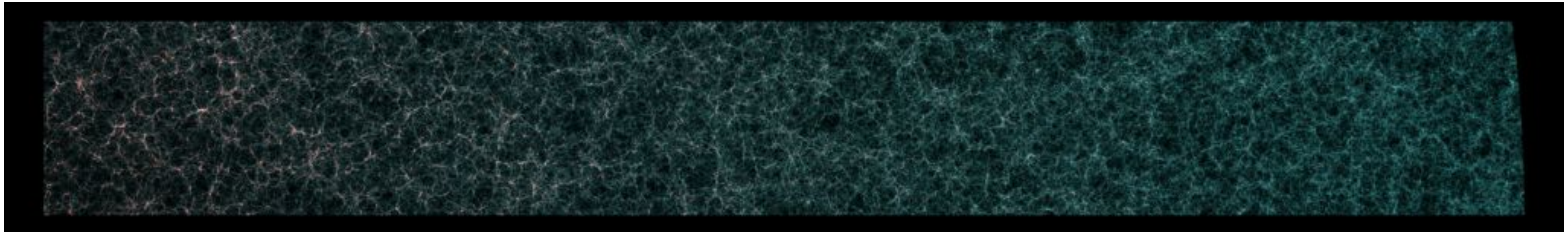


Euclid NISP-S simulated exposure, with H_a lines marked (B. Granett & e2e group)

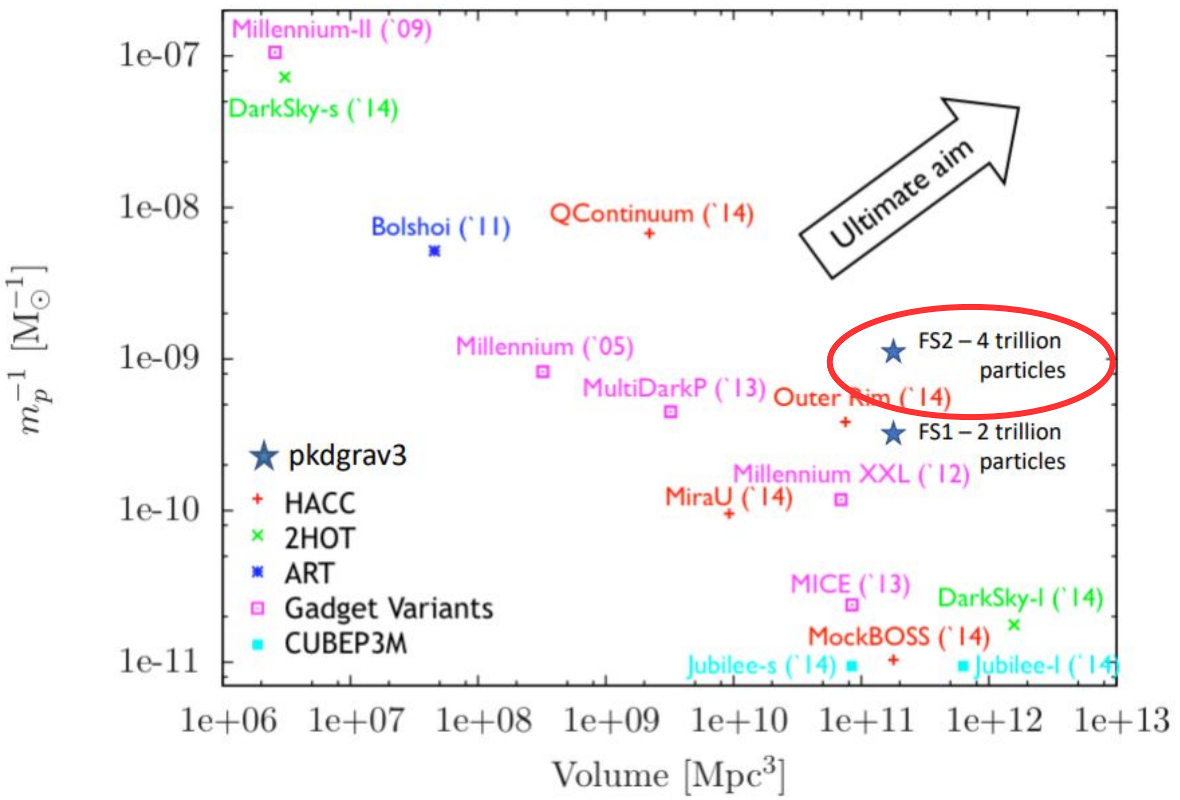
Survey challenges: huge data volume

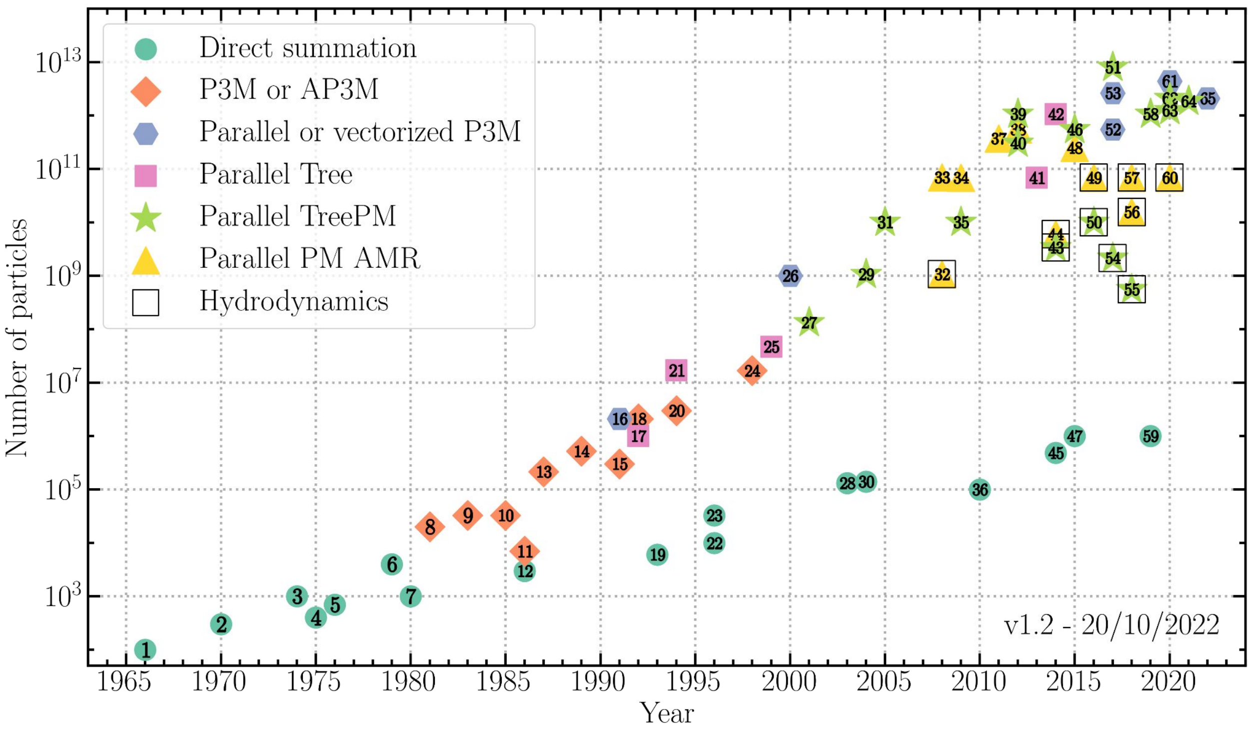


Flagship 2 simulation

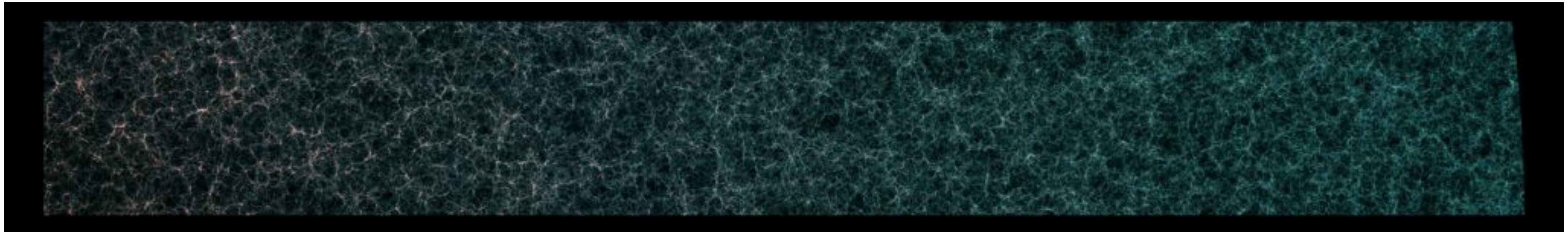


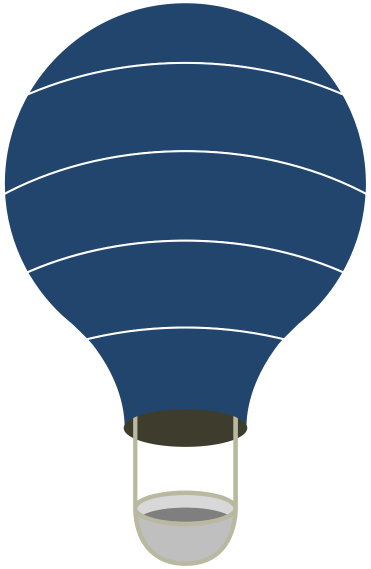
Survey challenges: computational power is not scaling as fast...





Flagship 2 simulation





How to address this challenge ?

- Emulation technologies
- Better inference techniques

- **Emulators (simulator accelerated with ML):**
  - Lyman alpha forest baryon
  - LPT + ML with displacement
  - BAM, PineTree, and CHARM
  - + lots of others at level of summaries (CosmoPower, BACCO, …)

- **Inferers (inference accelerated with ML):**
  - SELFI
  - ILI

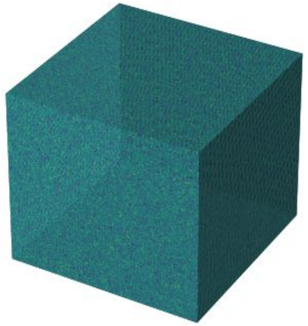
1

Neural Field emulator

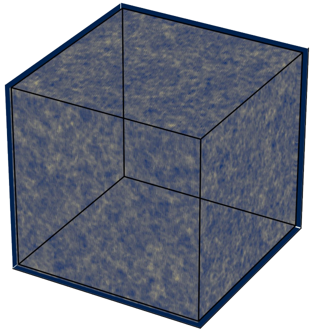
super cheap
high resolution
dark matter simulation

ABC of running N-body simulation

Prior Model

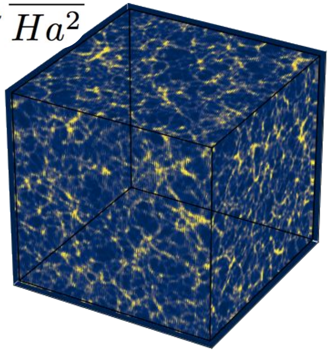


Structure Formation Model



$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}a} = \frac{\vec{p}}{\dot{a}a^2}$$
$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}a} = -\frac{3}{2}H_0^2\Omega_m\frac{\nabla^2\Phi}{Ha^2}$$

$$\vec{x} = \vec{q} + \vec{\nabla}\Delta^{-1}\delta_{\mathrm{init}}$$
$$\vec{p} \propto \dot{D}\vec{\nabla}\Delta^{-1}\delta_{\mathrm{init}}$$



How to get N-body simulations without paying the cost?

Idea: make an expansion of particle displacement

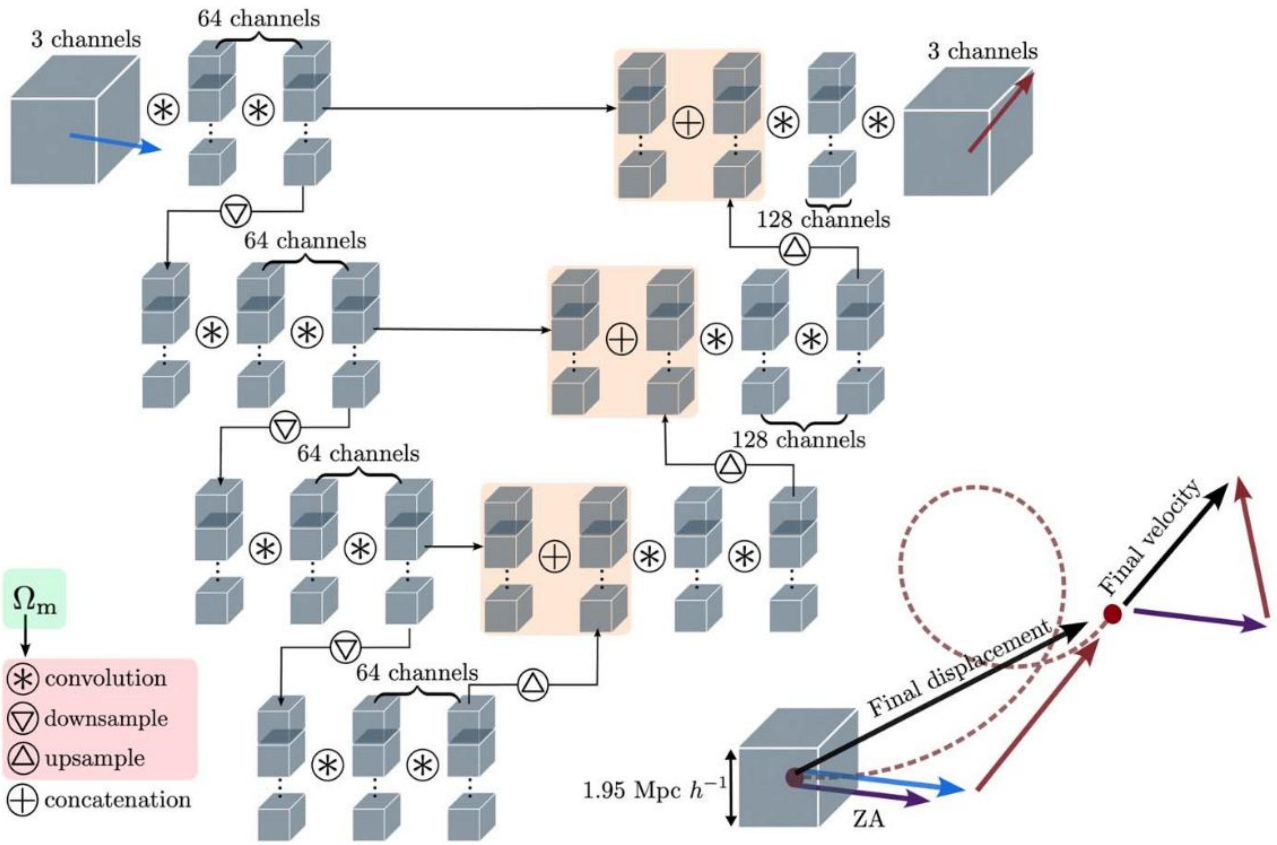
Final Position =
Initial **+** Analytic **+** Neural network

Two examples:

- LPT+NN
- NECOLA (tCOLA+NN)

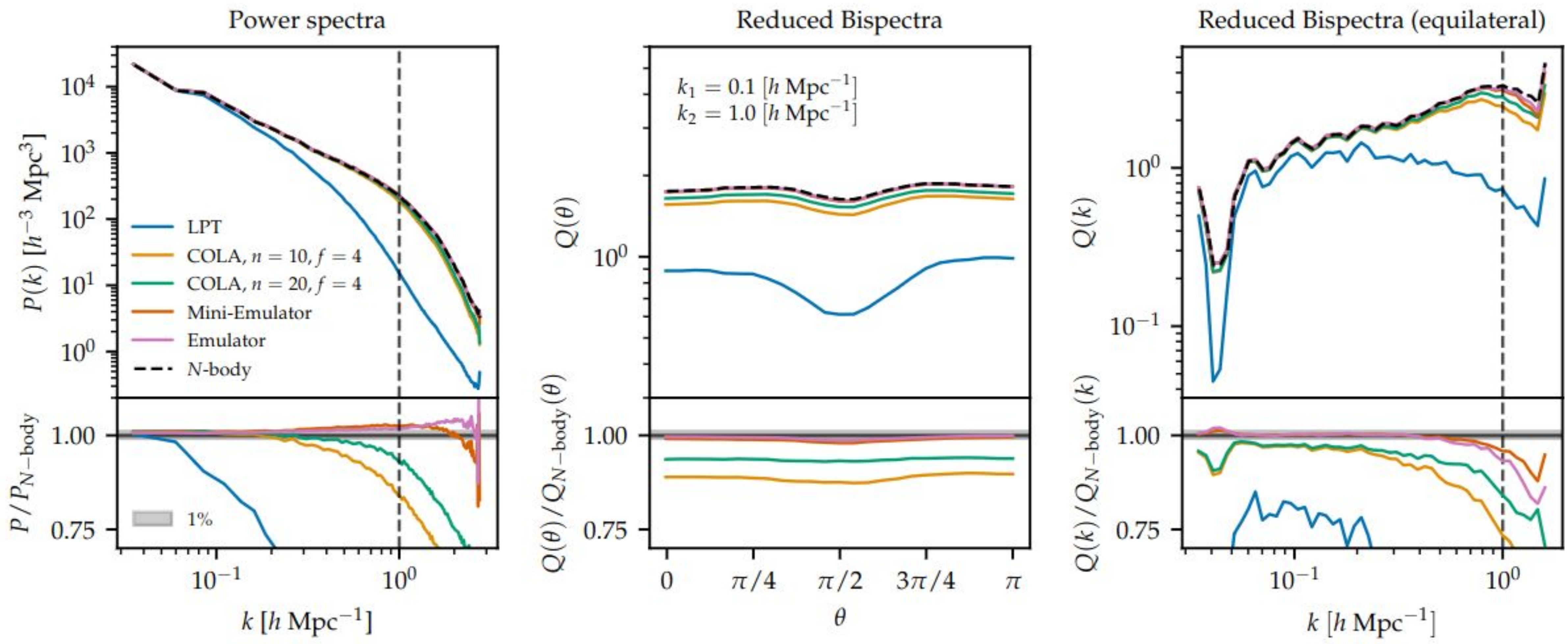
LPT+NN emulator: concept and architecture

- **Analytic displacement** = Lagrangian displacement field (= Zel'dovich approximation)
- **Residuals** are trained on Quijote N-body simulations (i.e. ~Gadget)
- **Advantages:**
  - super-fast: > 100x a PM simulation
  - GPU ready

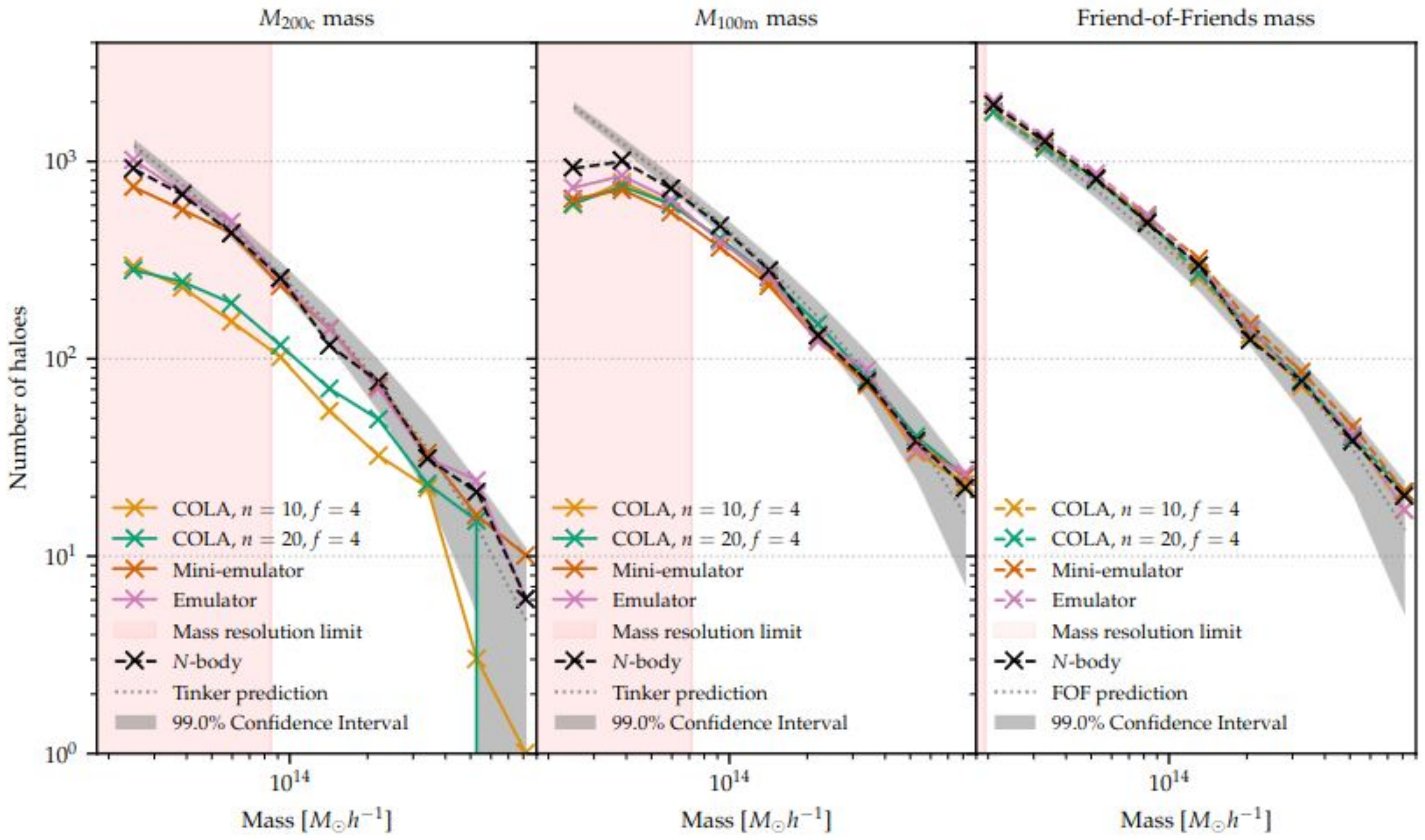


Doeser et al. (2023), Jamieson et al. (2023), de Oliveira et al. (2020)

Two and Three point statistics for emulator and other solvers

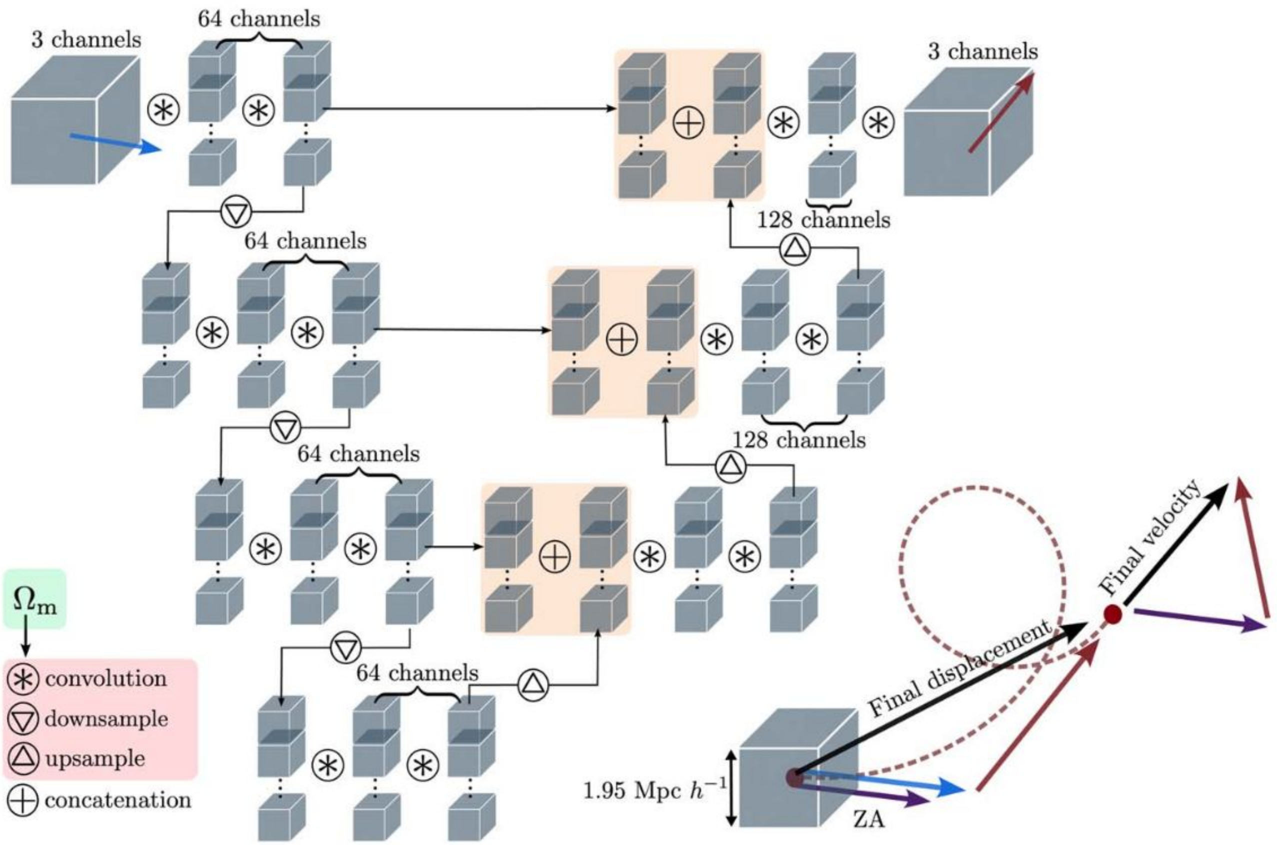


Mass function with emulators vs other solvers



LPT+NN emulator: concept and architecture

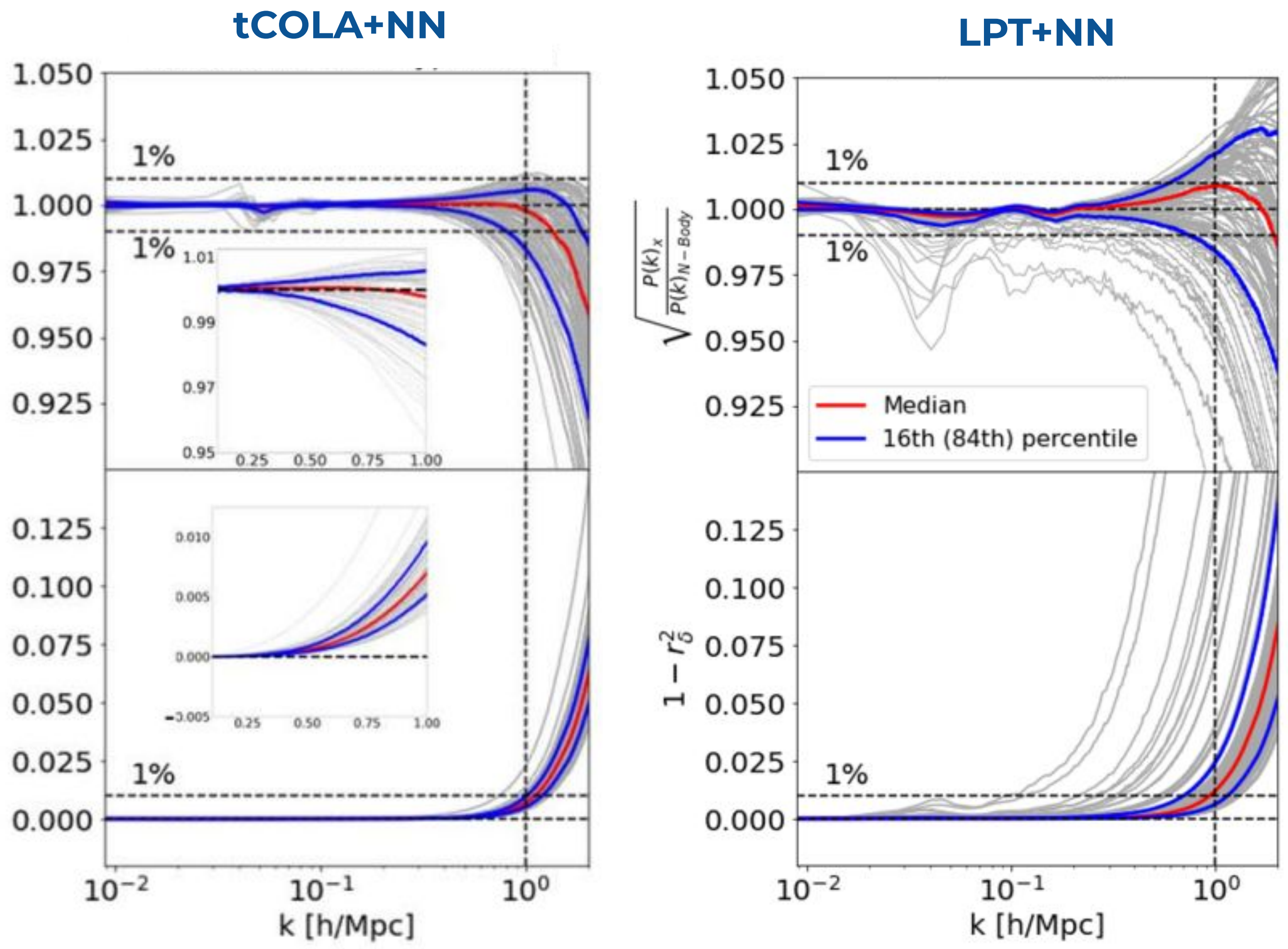
- Analytic displacement = Lagrangian displacement field (= Zel'dovich approximation)
- Residual is trained on Quijote N-body simulations (i.e. ~Gadget)
- **Advantages:**
  - super-fast: > 100x a PM simulation
  - GPU ready
  - **Accuracy!**
- **Disadvantages:**
  - large convolutional kernel (128^3+46 for padding), thus large GPU memory requirements
  - styled with a single cosmological parameter (Ω_m)
  - not completely explainable
- **Other works?**



Doeser et al. (2023), Jamieson et al. (2023)

Other works: NECOLA

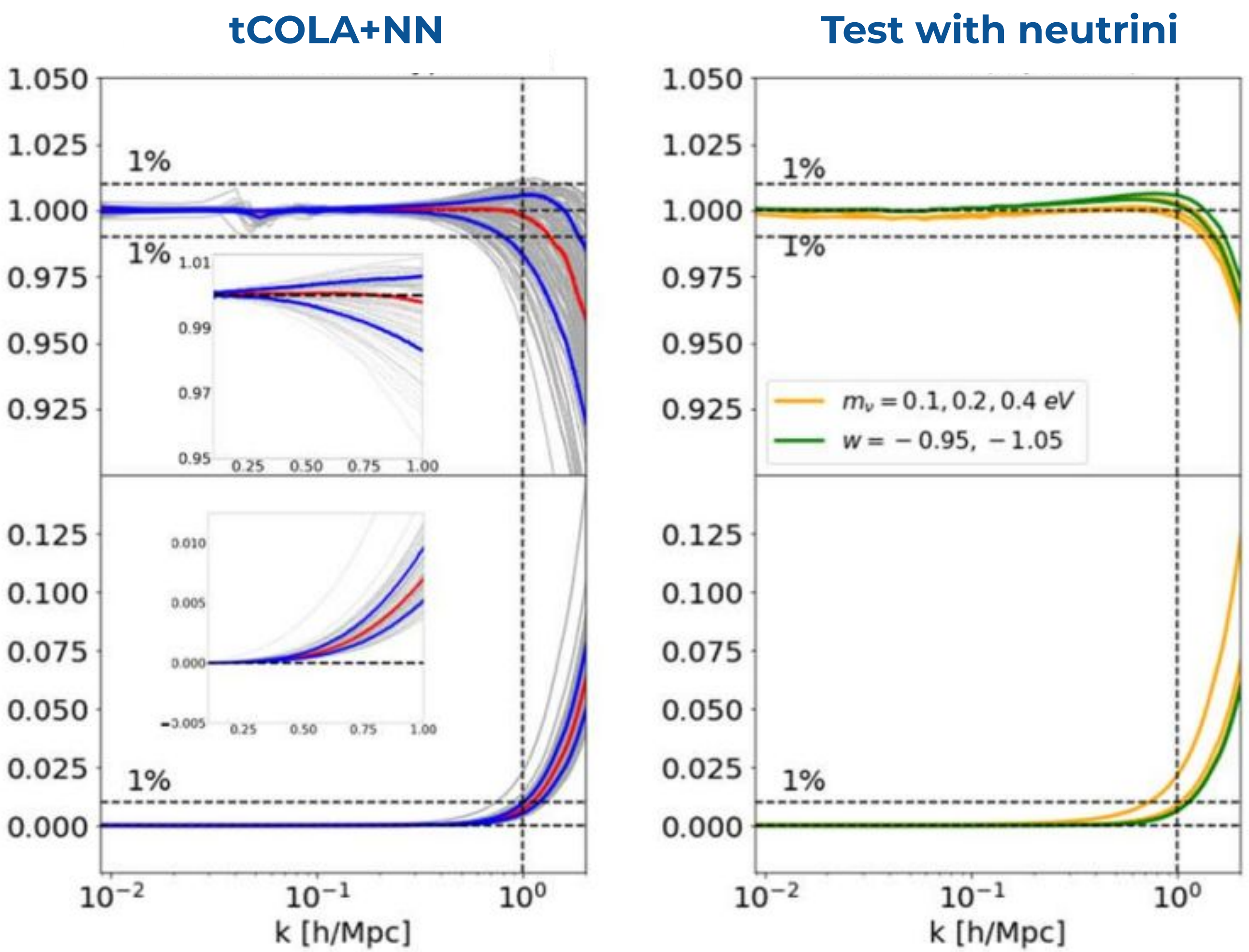
- Analytic displacement = tCOLA
- Residuals trained again on QUIJOTE set of simulations
- Advantages:
  - less cosmology dependent
- Disadvantages:
  - require a costly PM run



NECOLA (Kaushal et al. 2022)

Other works: NECOLA

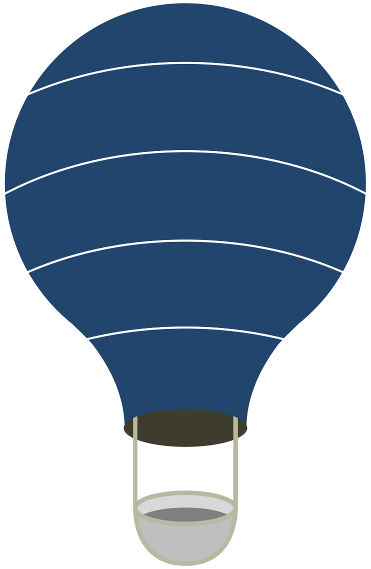
- Analytic displacement = tCOLA
- Residuals trained again on QUIJOTE set of simulations
- Advantages:
    - less cosmology dependent
- Disadvantages:
    - require a costly PM run



NECOLA (Kaushal et al. 2022)

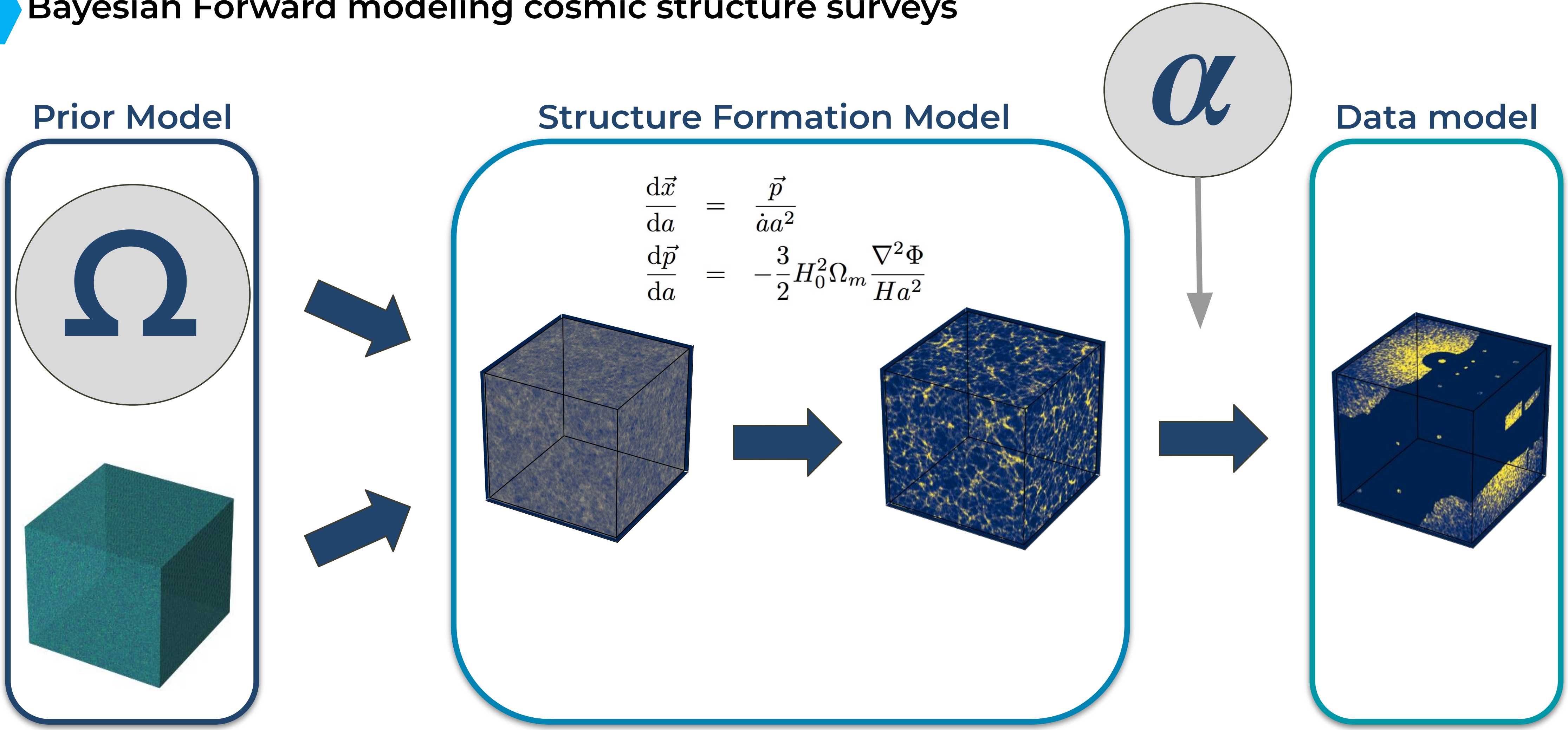
Take home message

- Accuracy – higher than current forward models in **`BORG`**; percent-level diff with N-body
- Speed – 100x faster than N-body
- Will likely unlock needed simulations for future survyes



Application: information content of Large scale structures using BORG

Bayesian Forward modeling cosmic structure surveys

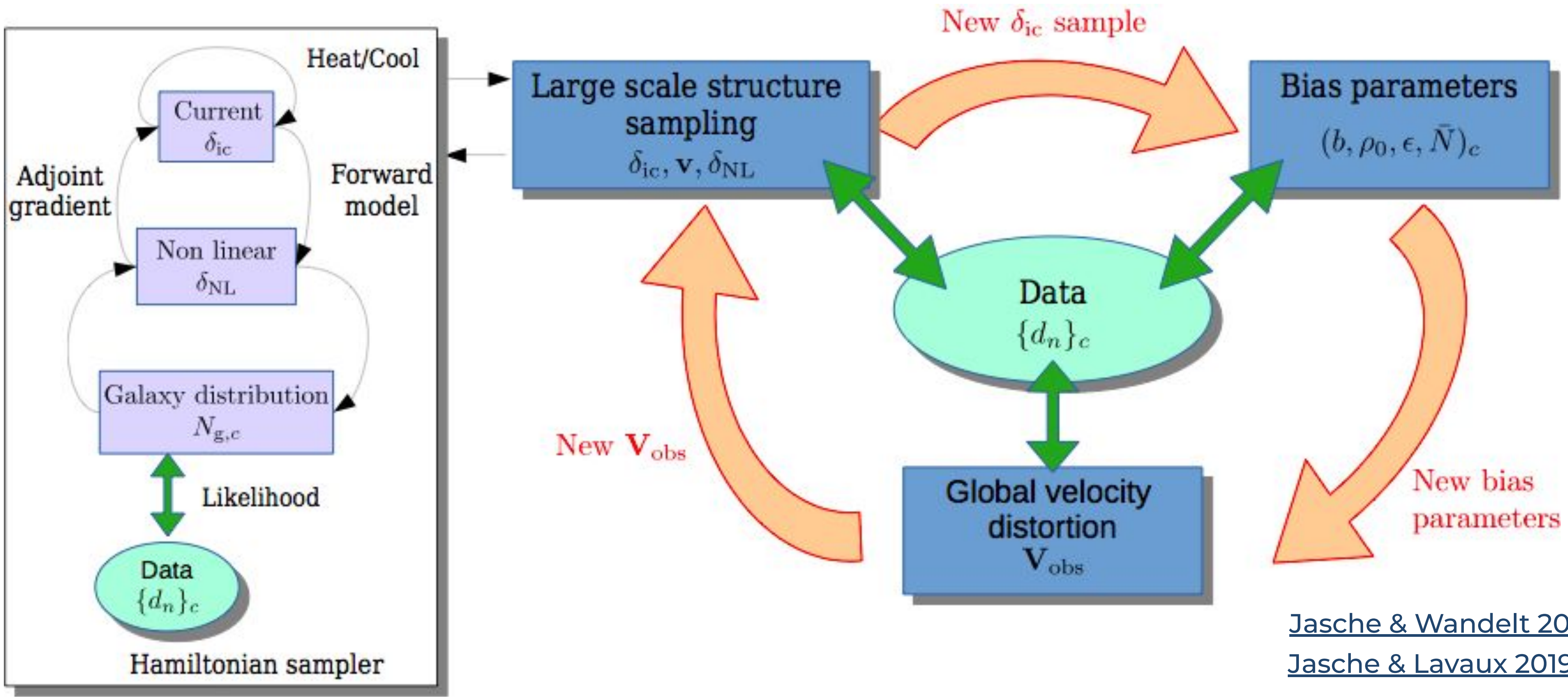


$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}a} = \frac{\vec{p}}{\dot{a}a^2}$$
$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}a} = -\frac{3}{2}H_0^2\Omega_m\frac{\nabla^2\Phi}{Ha^2}$$

$$\pi\left(\mathbf{x},\boldsymbol{\Omega}\right) \qquad \pi\left(\rho_{\mathbf{m}}|\mathbf{x},\boldsymbol{\Omega}\right) \qquad \pi\left(\mathbf{N_g}|\rho_{\mathbf{m}},\alpha,\boldsymbol{\Omega}\right)$$

BORG: A large scale MCMC framework

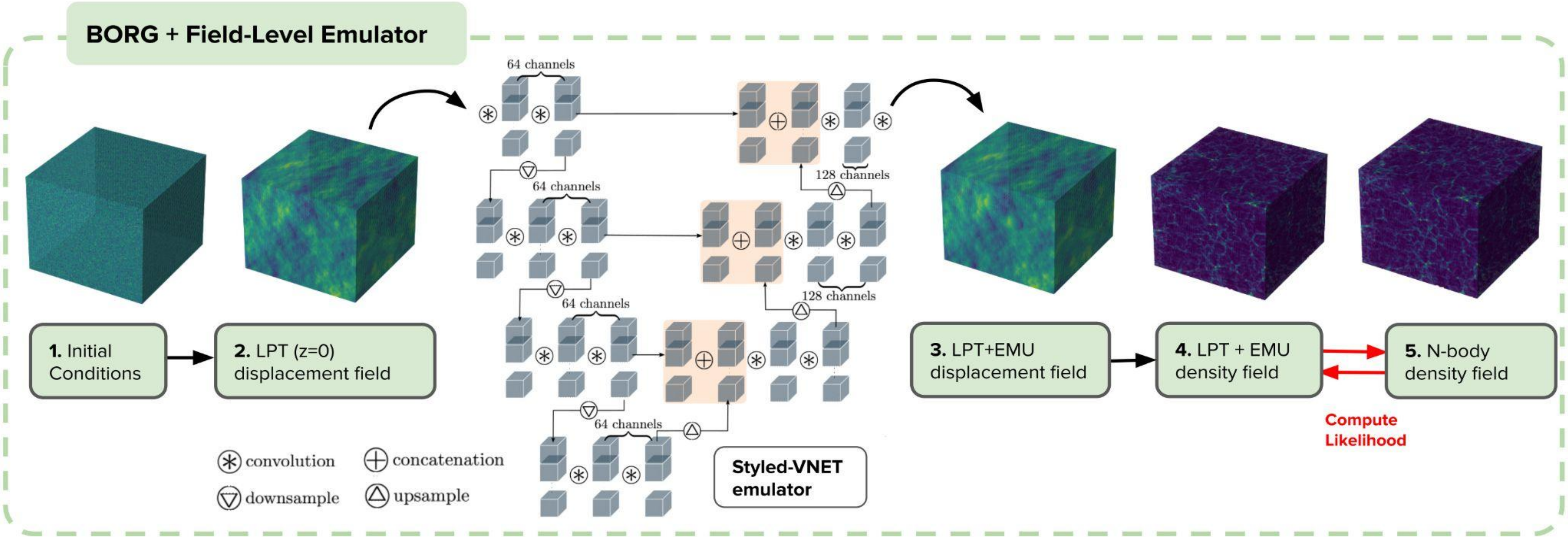
- **BORG's MCMC framework allows building flexible data models**
  - Hierarchical Bayes and block sampling
  - Efficient Hamiltonian Monte Carlo **(HMC)** technique
  - **Fully differentiable physics forward model**



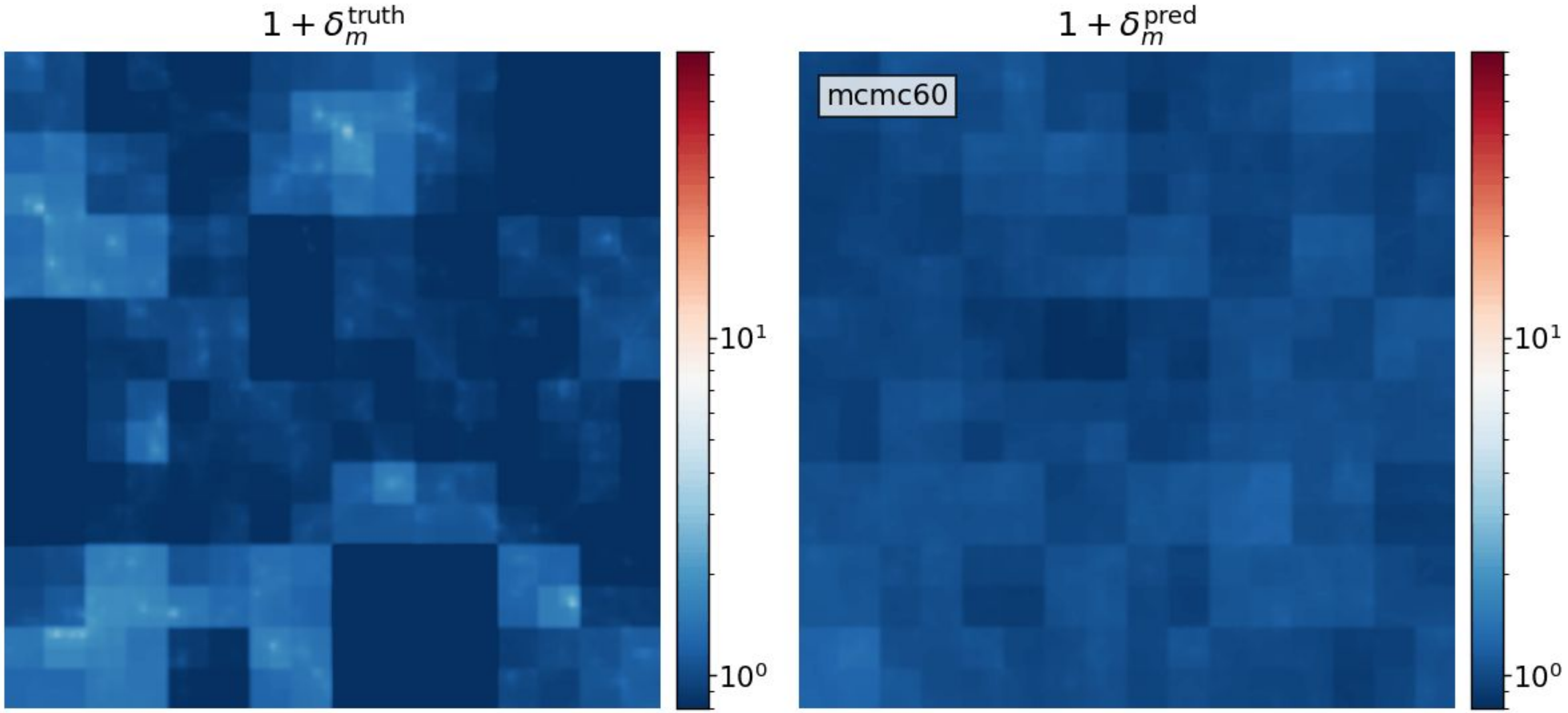
Jasche & Wandelt 2014
Jasche & Lavaux 2019

Neural Field-Level Emulator (Ludvig Doeser, Drew Jamieson)

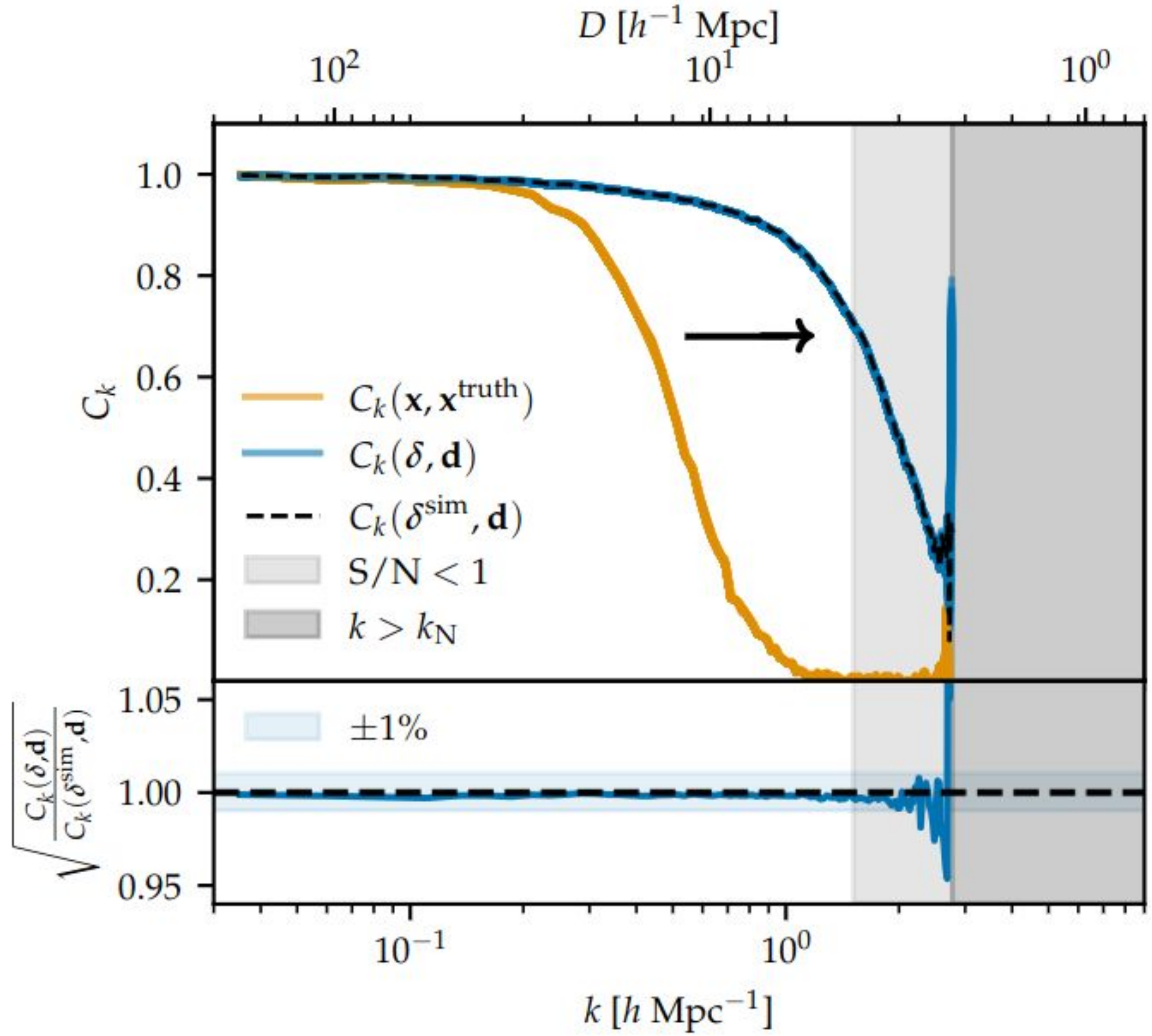
- Translate approximate LPT displacements to N-body-like displacements
- Differentiability – through autograd (`PyTorch`)



Neural Field-Level Emulator (Ludvig Doeser, Drew Jamieson)



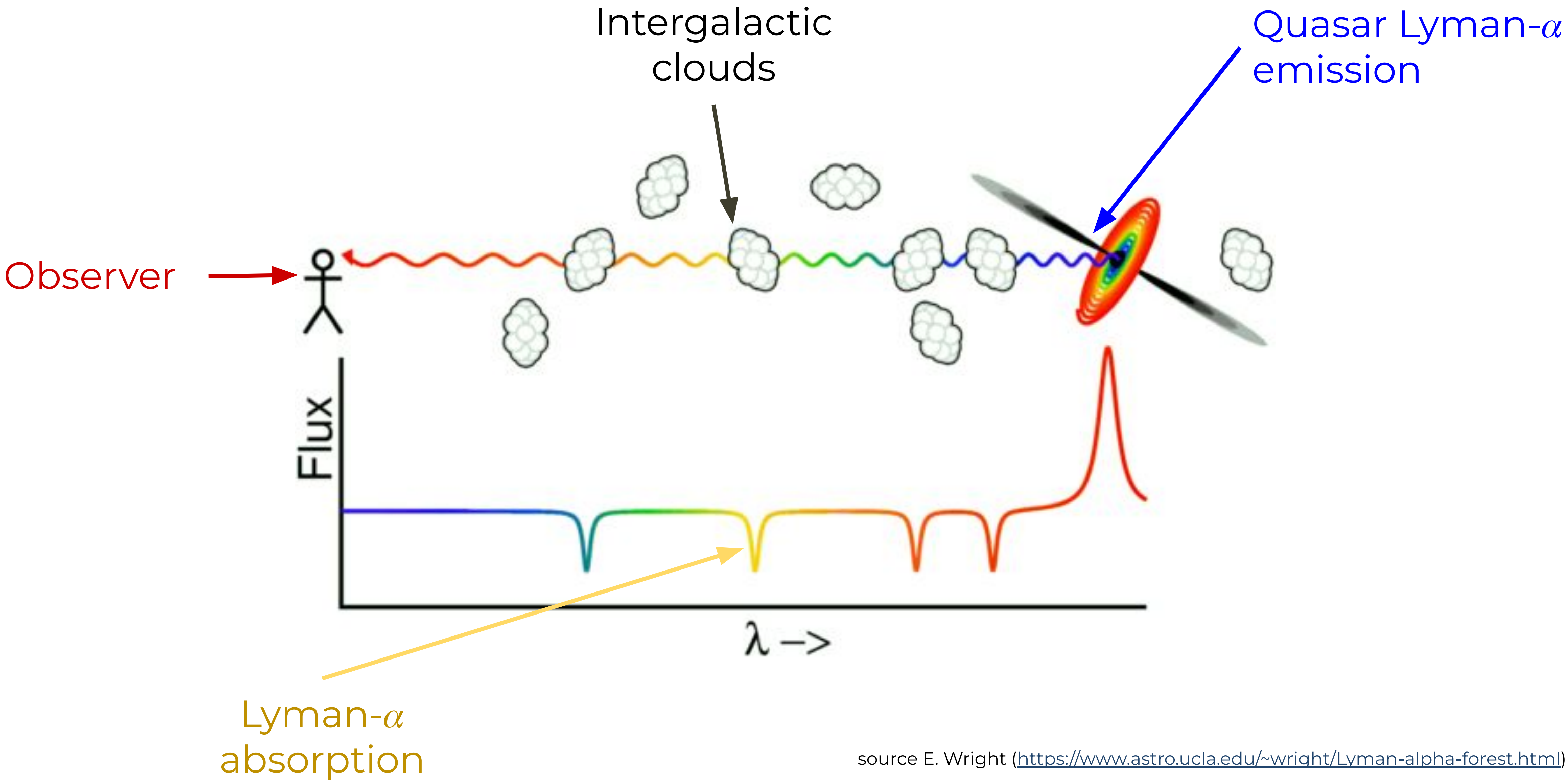
Information recovery on initial conditions



2

Baryon field emulator: application to Lyman alpha forest

Reminder: what is Lyman-α forest?



source E. Wright (https://www.astro.ucla.edu/~wright/Lyman-alpha-forest.html)

Pros/Cons of Lyman-α forest

- **Pros:**
  - More "direct" image of baryon density (wrt Galaxies)
  - Cosmological information
  - Higher redshift = easier to model physics
- **Cons:**
  - need to model baryon physics
  - non-linear signal
  - bunch of skewers, getting 3d information needs statistical work

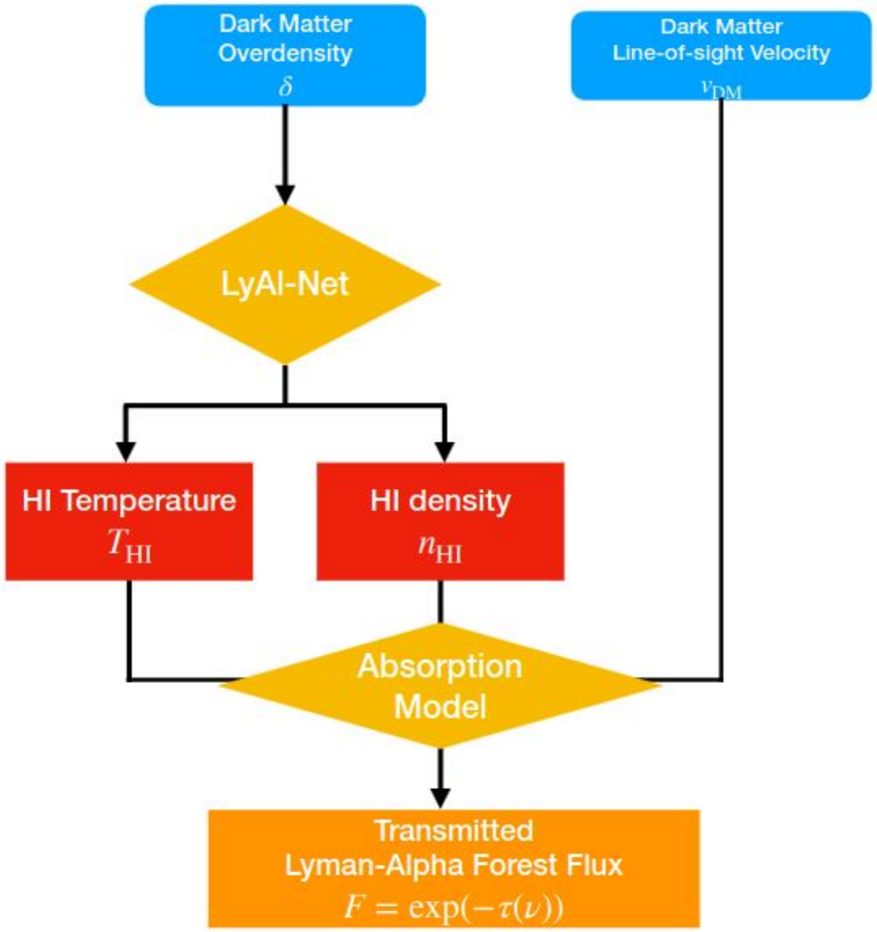
Building Ly-α model from the diffuse IGM

log absorption:

$$\tau(D_{\rm QSO}, \hat{u}, \nu_{\rm obs}) = \int_0^{s_{\rm QSO}(D_{\rm QSO})} n_{\rm HI}(s, \hat{u})\sigma_{\rm HI}(\nu_{\rm obs}, s, \hat{u})\, \mathrm{d}s$$

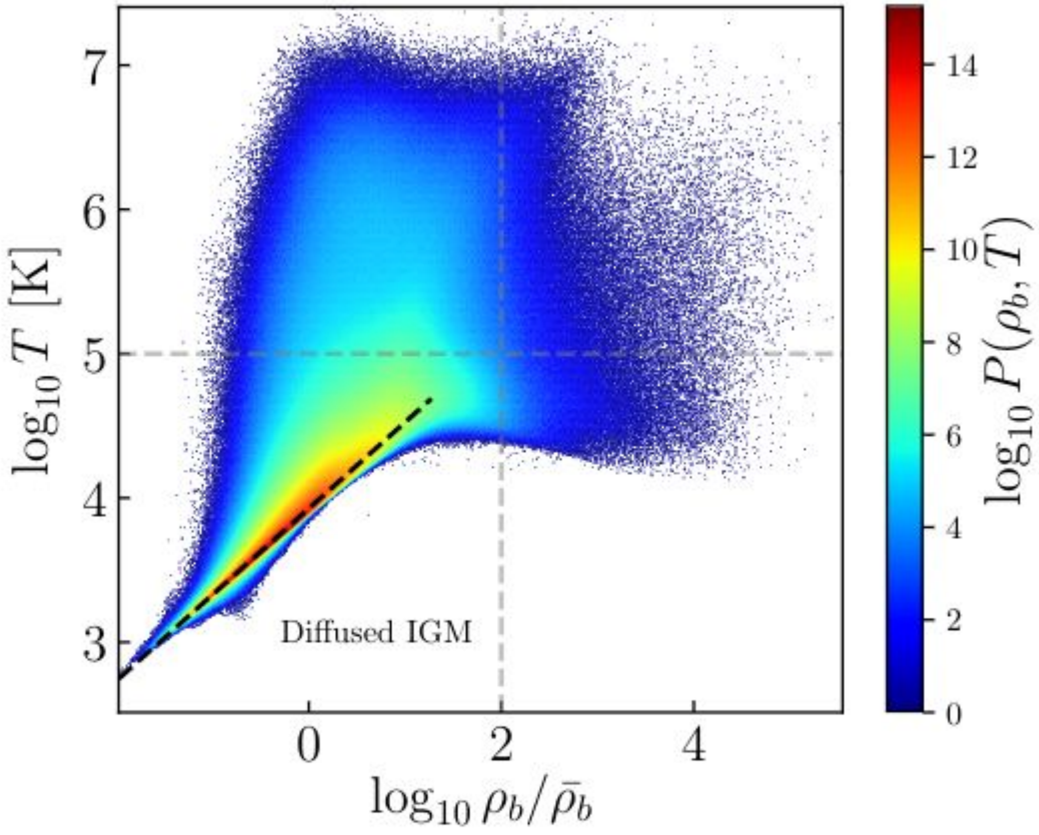
cross-section:

$$\sigma_{\rm HI}(\nu) = \frac{\pi e^2}{m_e c} f_{lu} L(\nu) = \frac{\pi e^2}{m_e c} f_{lu} \frac{\Gamma_{ul}/(4\pi^2)}{(\nu - \nu_{lu})^2 + (\Gamma_{ul}/(4\pi)^2)}$$

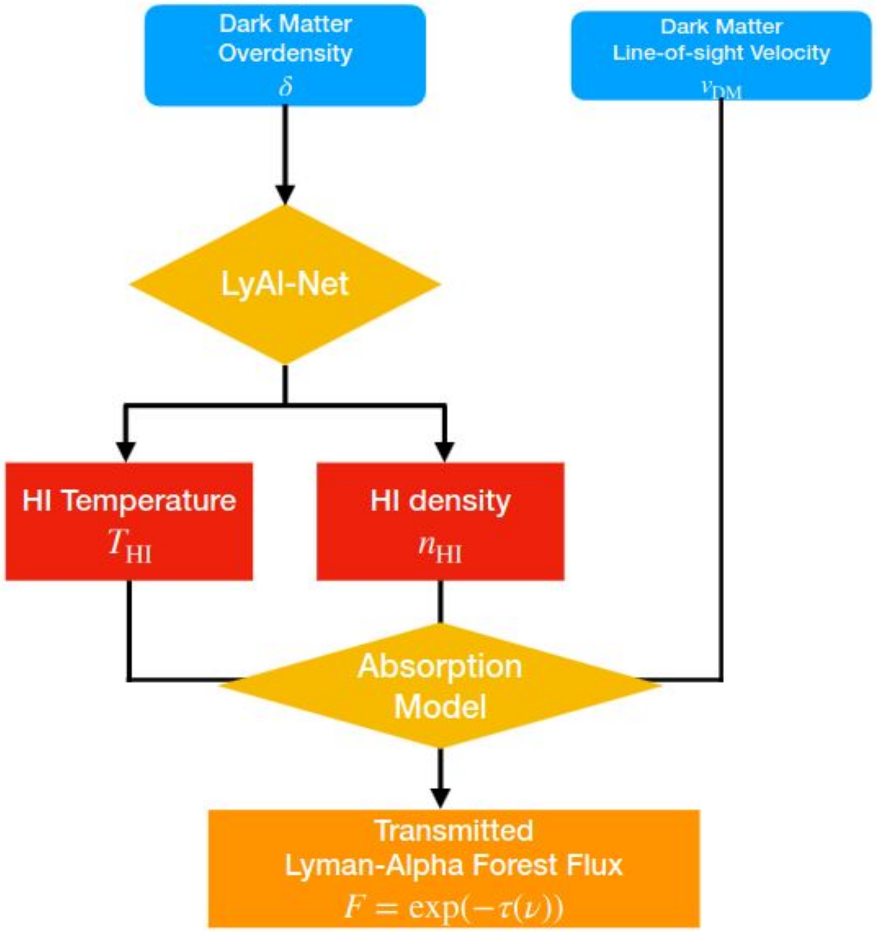


Gunn & Peterson (1965), Weinberg et al. (1997), Boonkongkird et al. (2023)

Building Ly-α model from the diffuse IGM



Equation of state of the IGM



Weinberg et al. (1997), Boonkongkird et al. (2023)

Emulator 1: Lymas2, absorption flux emulation through linear filtering

Matter field

↓

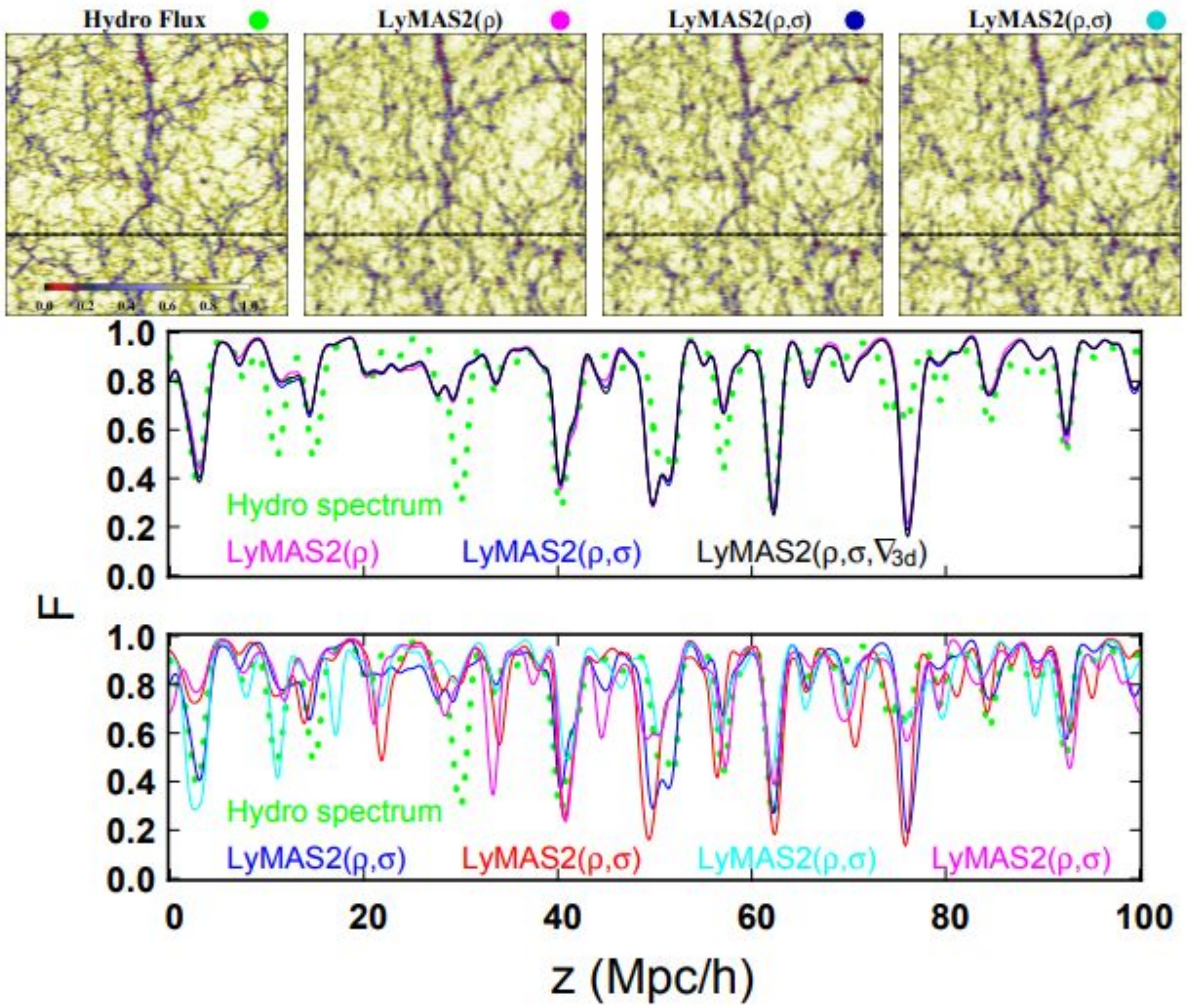
Gaussianize

↓

Fourier filter

↓

Flux

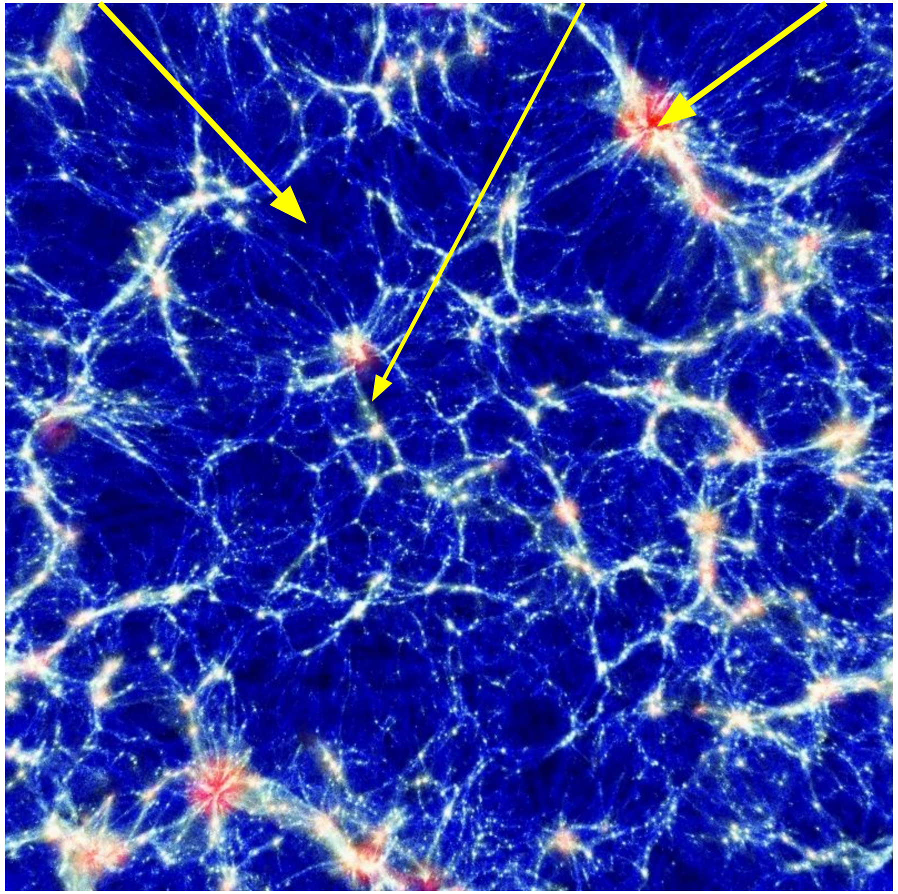


Peirani et al. (2022)

Emulator 2: Non-Local Fluctuating Gunn-Peterson Approx. (w/ Cosmic Web)

i =

voids pancakes filaments knots

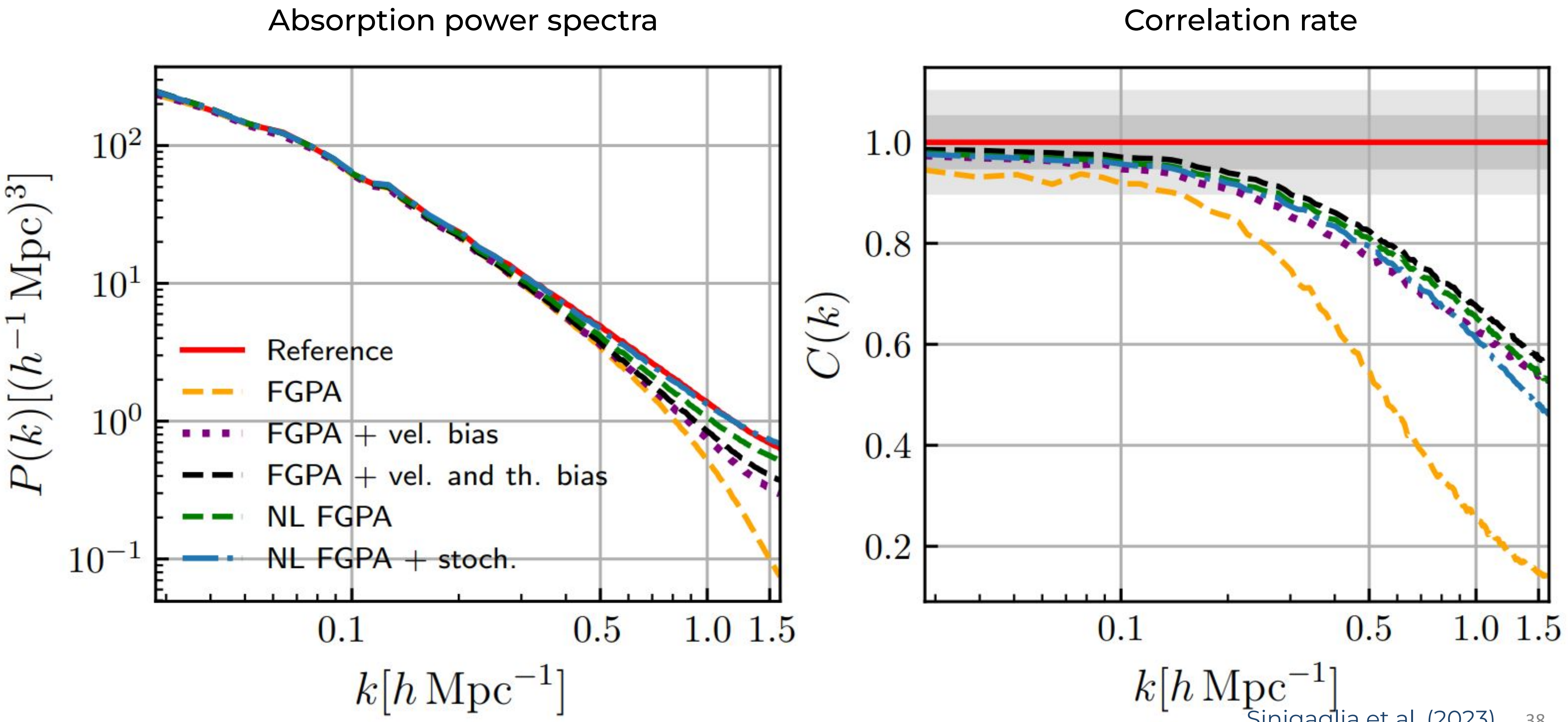


A_i, α_i, $\delta_{[1,2]i}$ = F(cosmic web class of i)

$$\tau = A_i(1+\delta)^{\alpha_i} \exp\left(-\frac{\delta}{\delta^*_{1,i}}\right) \exp\left(\frac{\delta}{\delta^*_{2,i}}\right) + \epsilon_i$$

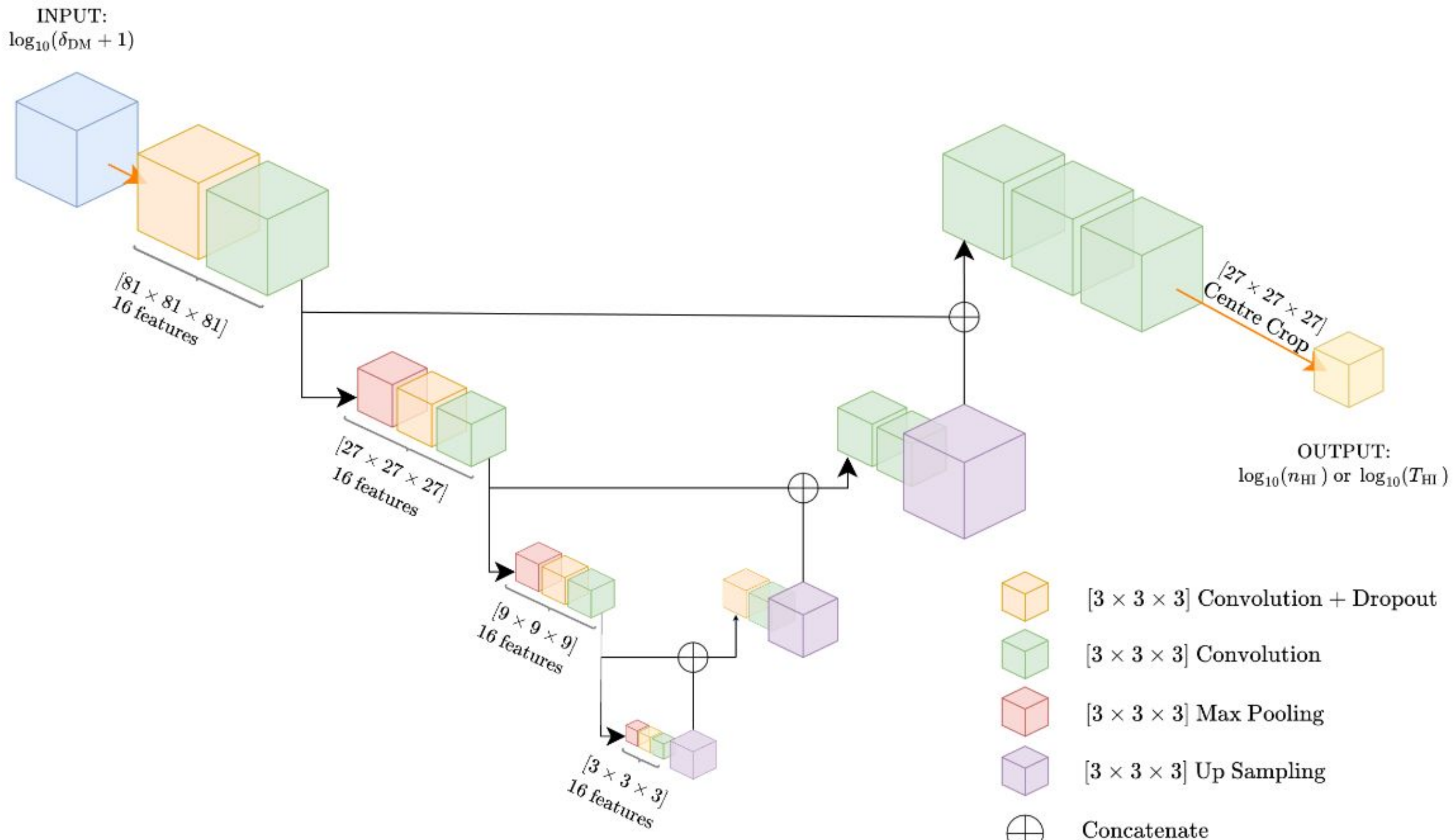
Sinigaglia et al. (2023)

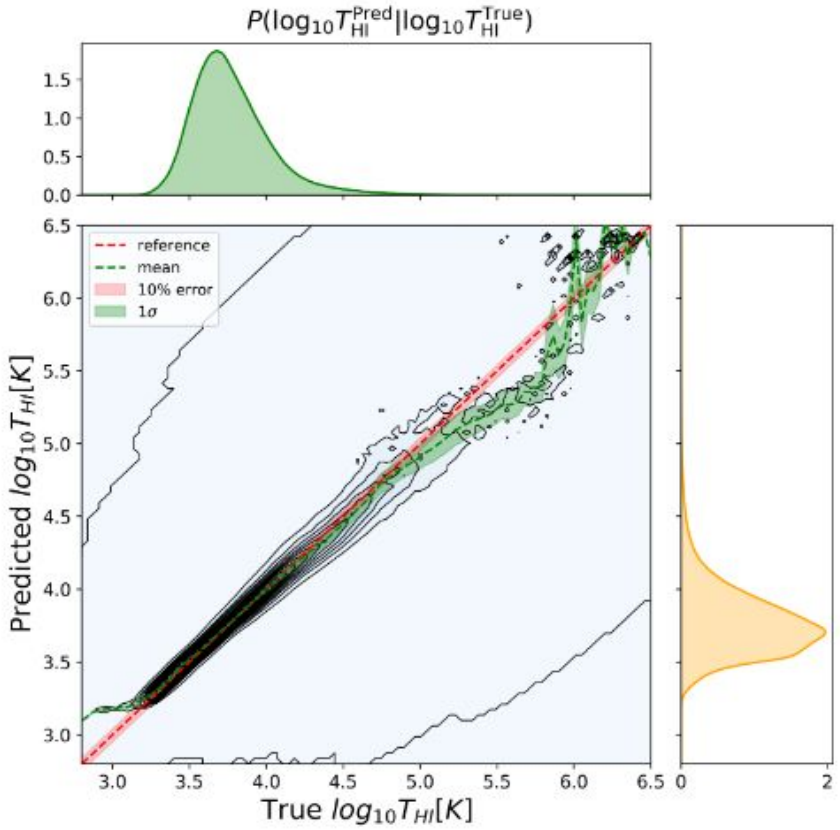
Emulator 2: Non-Local Fluctuating Gunn-Peterson Approx. (w/ Cosmic Web)



Sinigaglia et al. (2023)

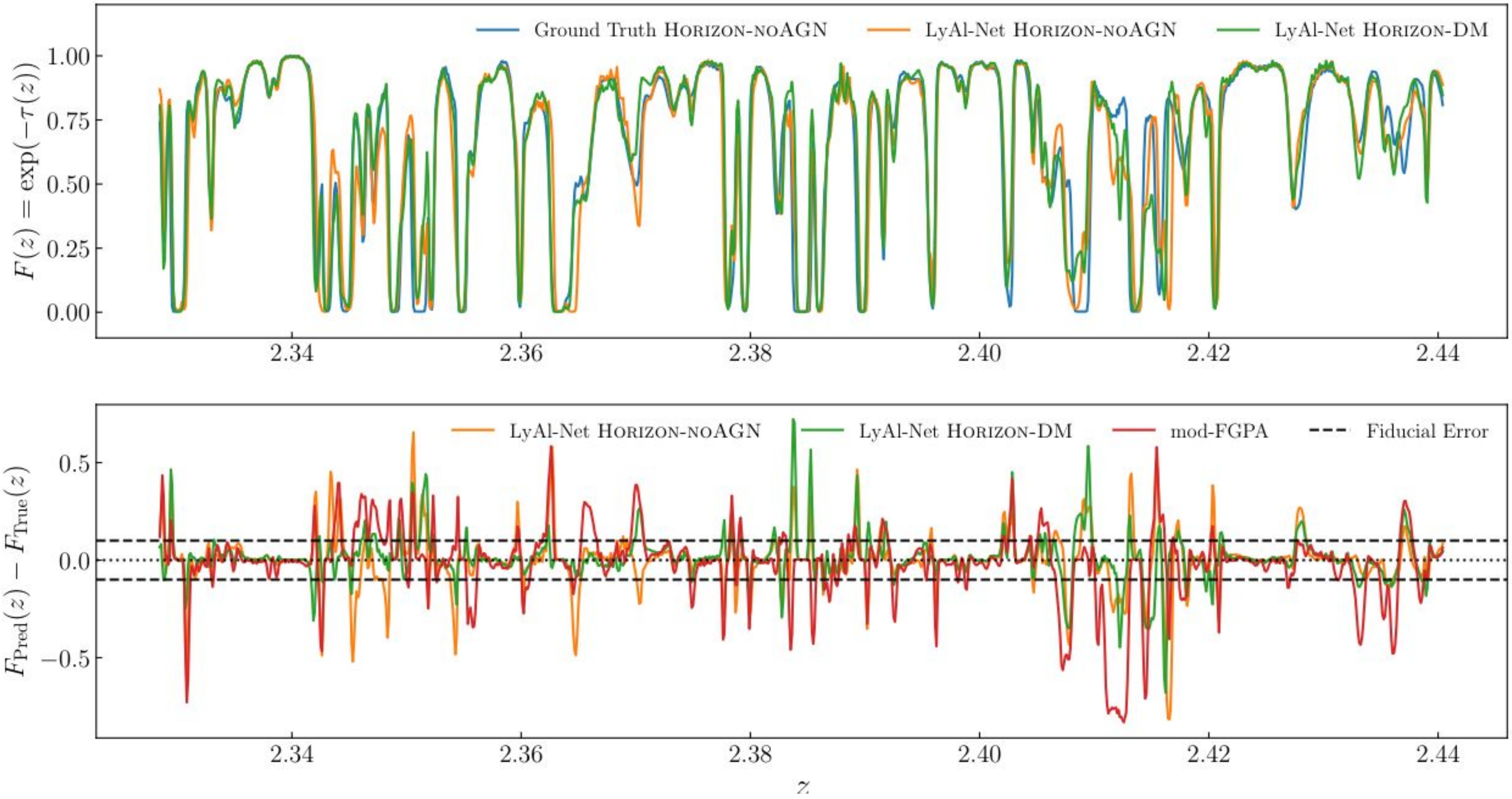
Emulator 3: LyAl-Net, emulation through non-linear convolution



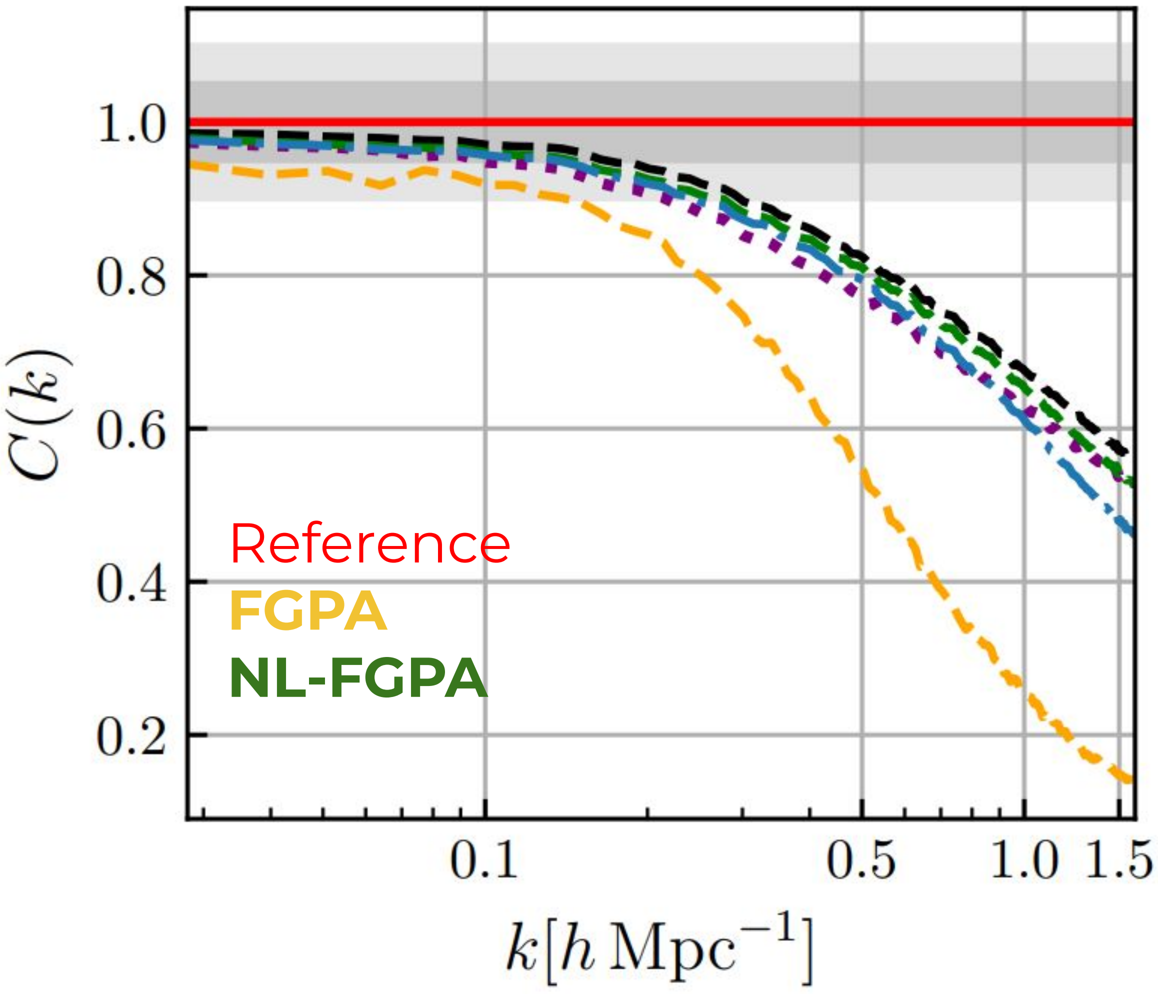


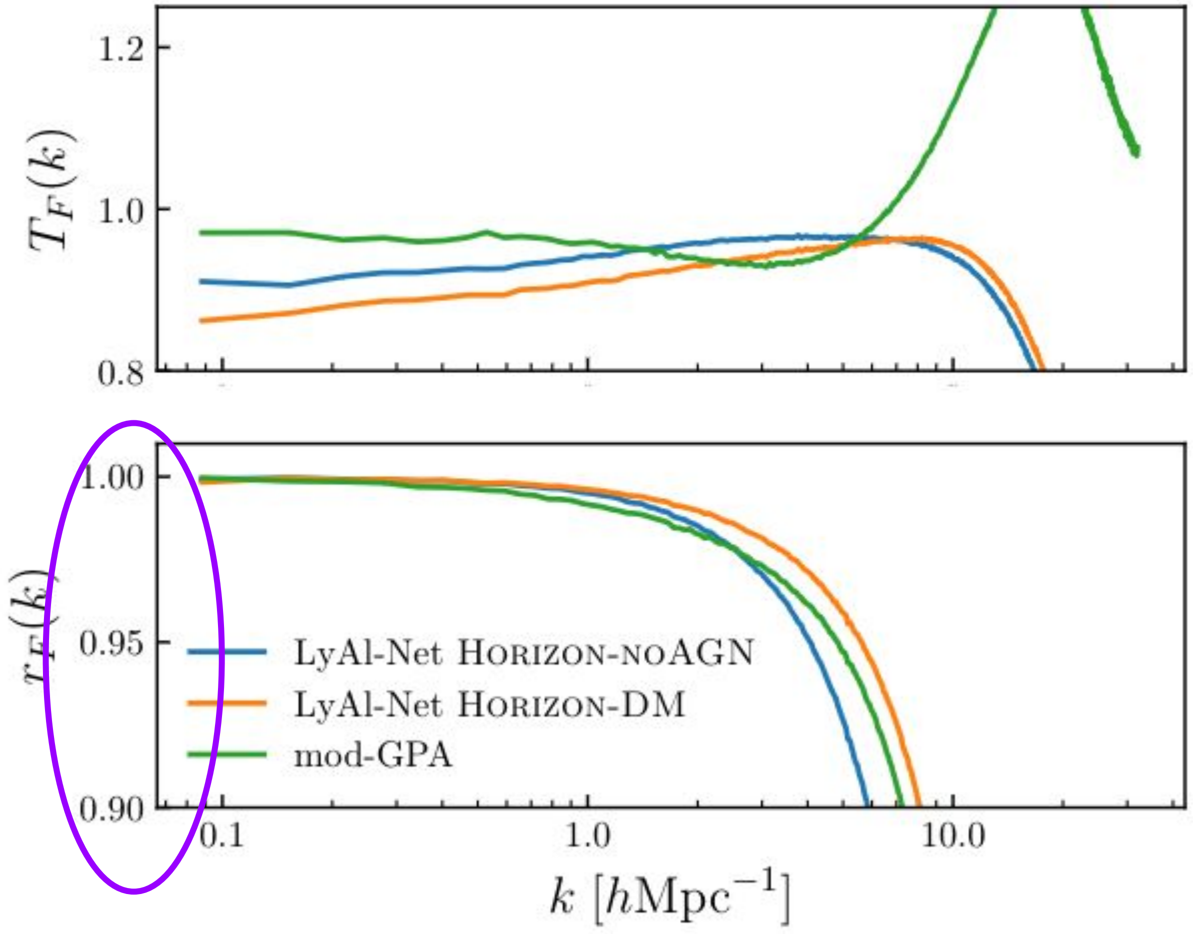
Boonkongkird et al. (2023)

Emulator 3: LyAl-Net, absorption prediction performance



Emulator 3: LyAl-Net, 2 point statistics performance





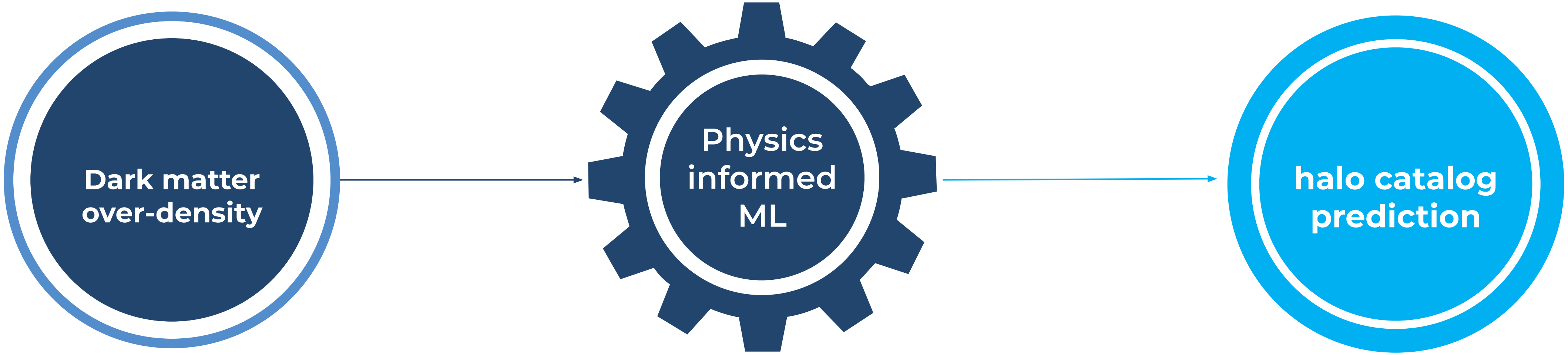
Take home message

- Deep Learning techniques becoming competitive:
  - can cover cosmological scales
  - LyAl-Net is resilient to change of baryonic physics
  - General resilience to change of cosmology
  - Need work on redshift dependence
- Accuracy:
  - tend to favor big networks
  - physics intuition can push down (i.e. use cosmic-web)
- Application to new surveys (e.g. SDSS4-QSO, DESI)

3

Populating mock universes with halos/galaxies

PineTree & CHARM (ex-NPE = Neural Physical Engine)
(S. Ding, S. Pandey, T. Charnock)



From approximate simulators
(e.g. 2LPT)

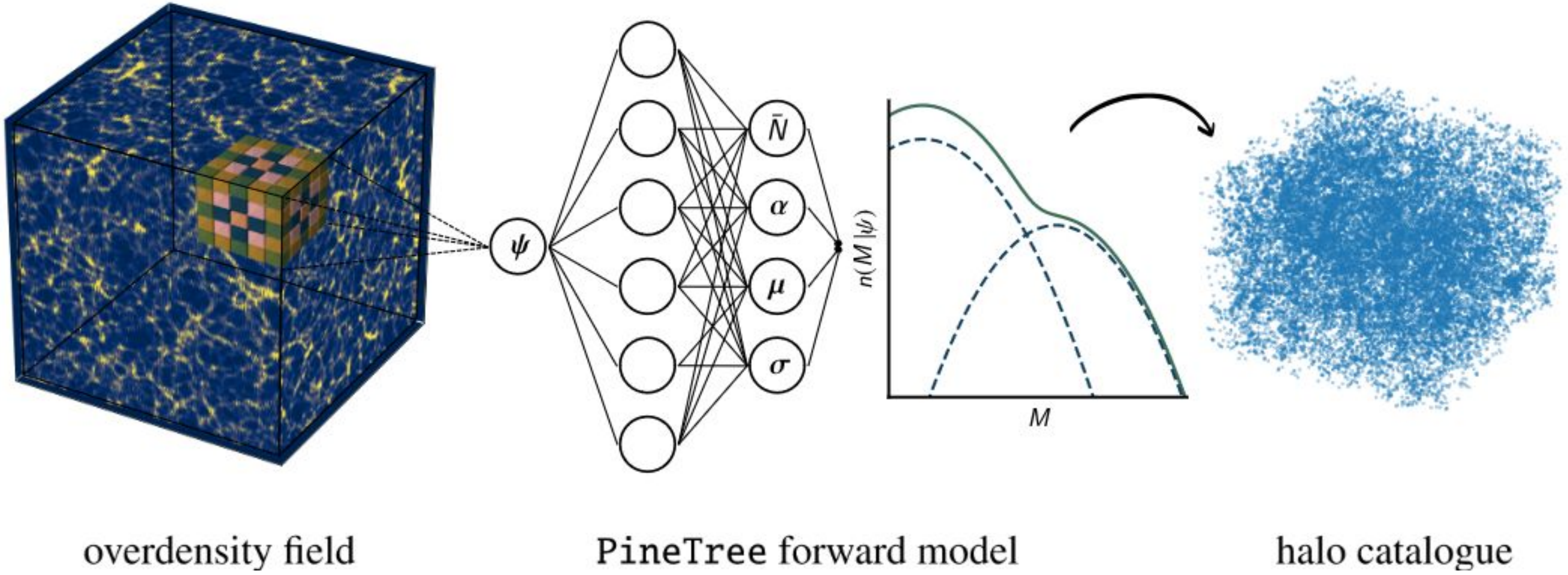
- Fast & Differentiable
- Stochastic
- Explainable
- 17-32 parameters

Validation:
- 1pt
- 2pt
- field-level

Ding et al (in prep), Pandey et al. (2024), Charnock et al 2020

Pinetree: Physical and Interpretable NEtworks for TRacer Estimation/Emulation

Preliminary



Ding et al. (in prep. 2024)

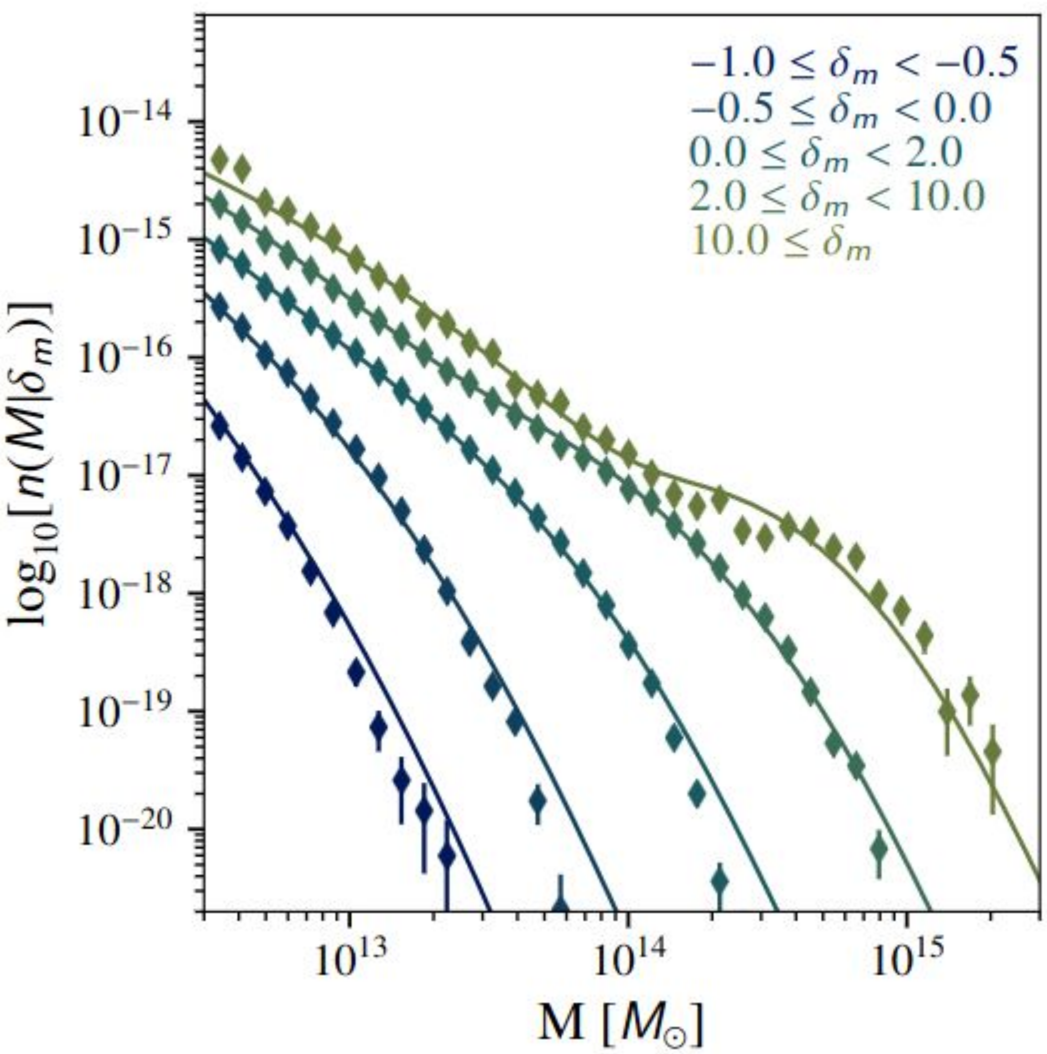
Preliminary

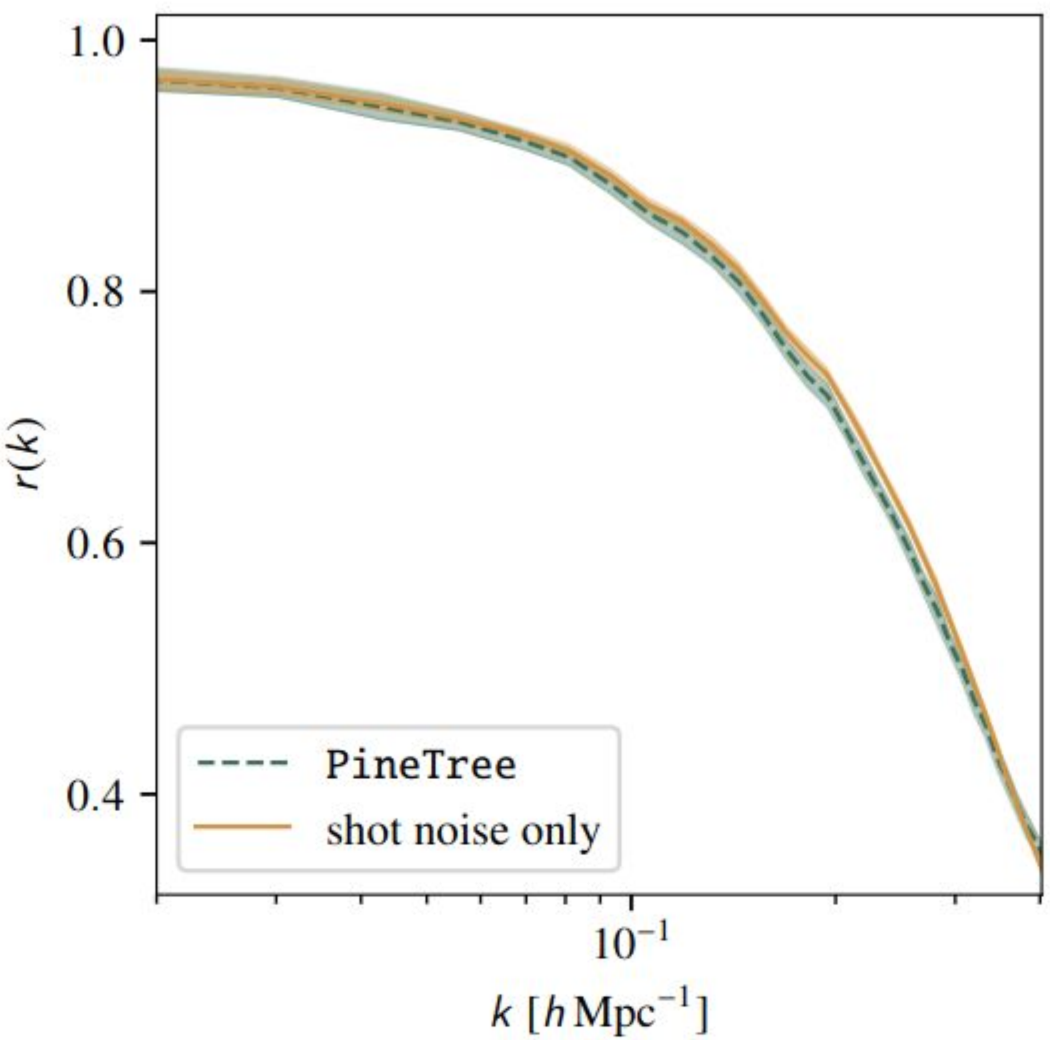
- Computed 40 N-body simulations
  - 500 Mpc/h, 512^3 particles
  - $m_p = 3 \times 10^{12}\ M_\odot$
- Training on:
  - baseline: one simulation
  - extended: 10 for training and 30 for validation
- Ideally: no training at all!

Ding et al. (in prep. 2024)

Preliminary

First look: mass function and halo field correlation

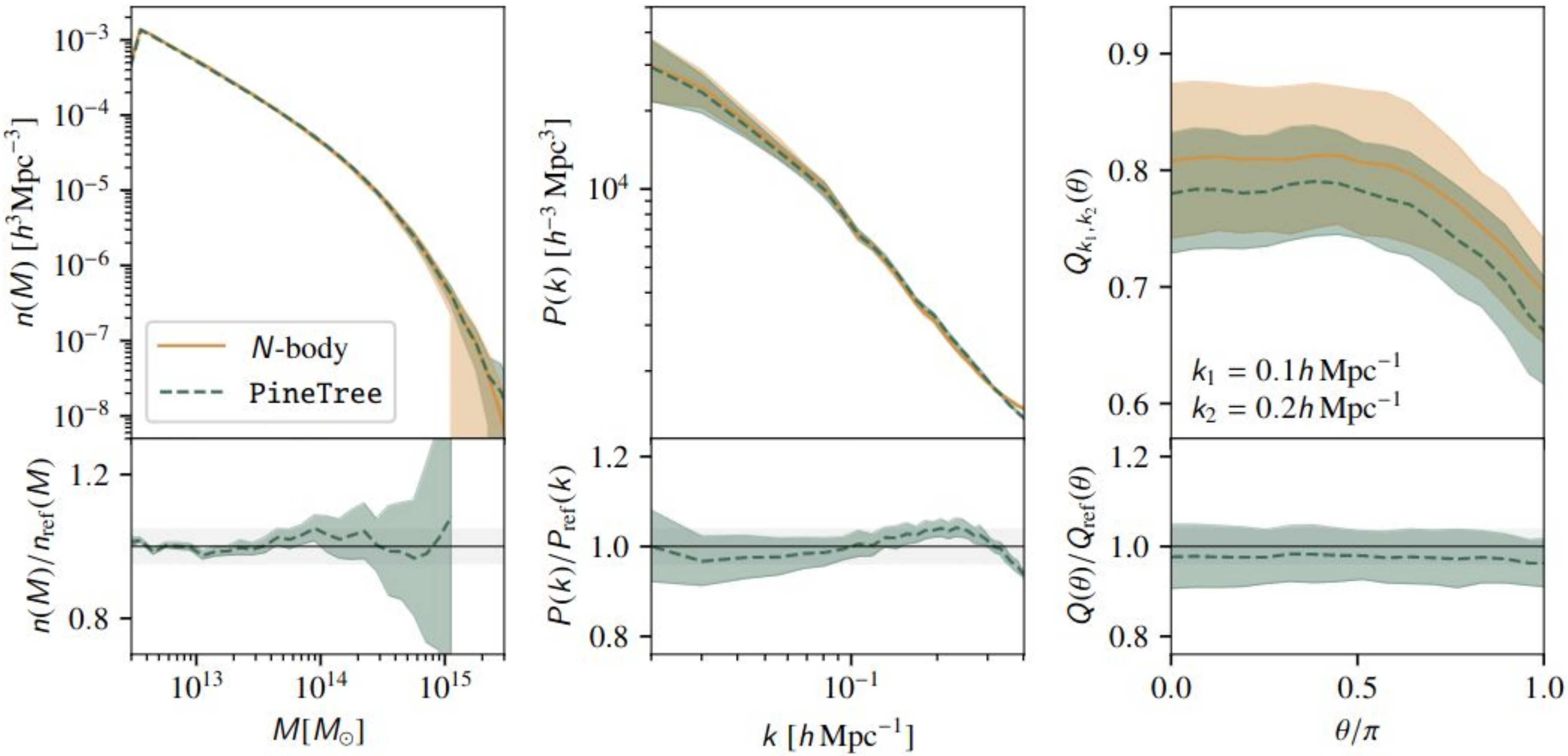




Ding et al. (in prep. 2024)

Second look: n(M), power spectra, bi-spectra

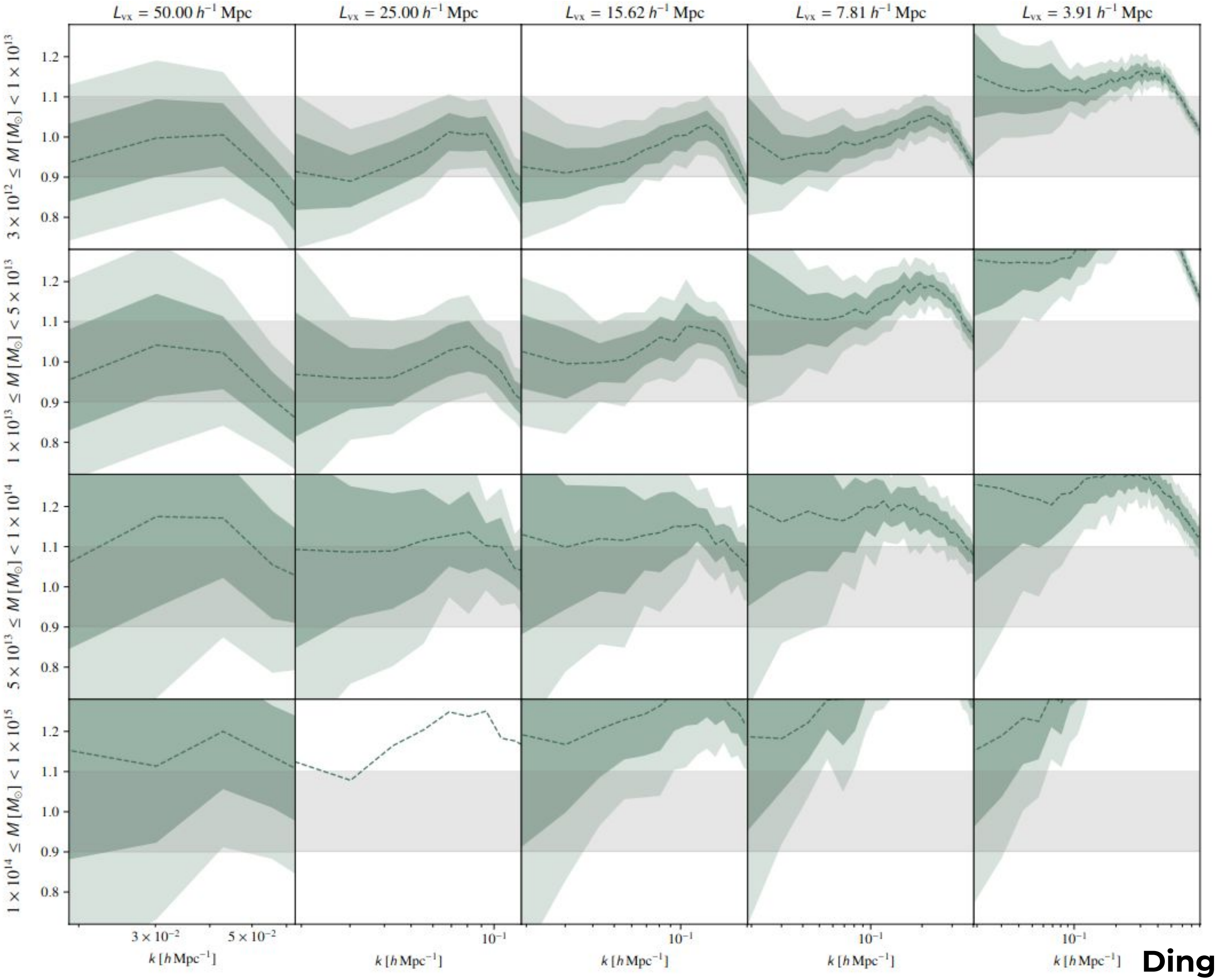
Preliminary



Ding et al. (in prep. 2024)

Effect of resolution

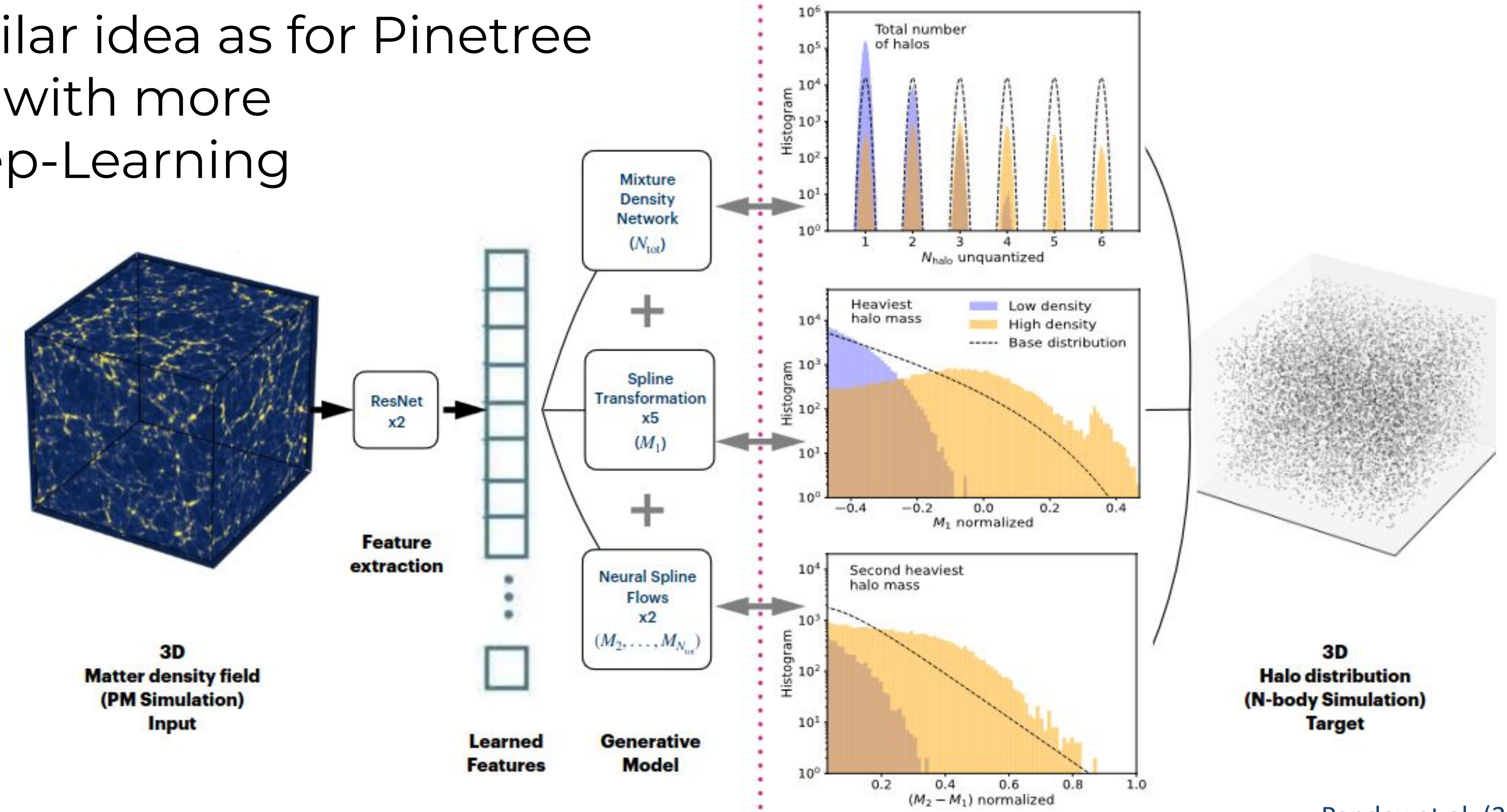
Preliminary



Ding et al. (in prep. 2024)

CHARM: Creating Halos with Auto-Regressive Multi-stage networks

Similar idea as for Pinetree but with more Deep-Learning



Pandey et al. (2024)

Take home message

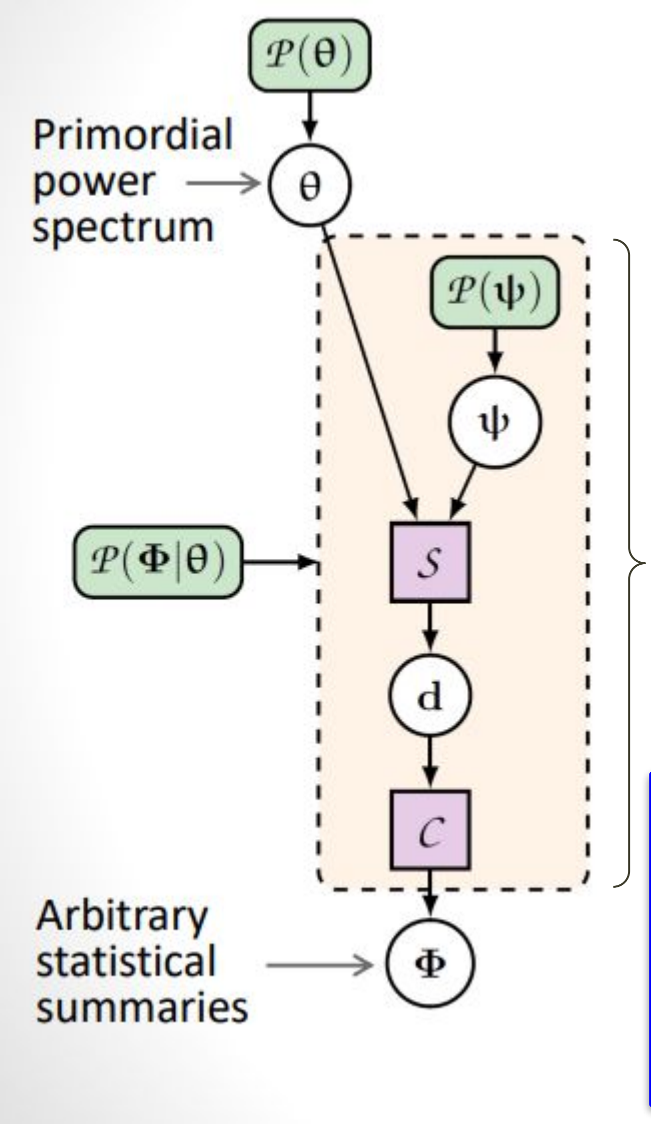
- Possible to generate large halo mock catalogs from rough simulations
- Statistics well understood for PineTree
- Scaling possible by going full Machine Learning with CHARM

4

Running cosmological inferences with ML

- **Different model variant:**
  - MOPED: massive data compression (expansion of log-likelihood)
  - SELFI: simulator expansion for LFI (expansion of the simulator)
  - BOLFI: Bayesian optimisation for LFI
  - ILI-LTU: Parameter density estimators through LFI/ILI
- **Motivations:**
  - Purely based on simulation
  - May fold model as complex as needed
- **Challenges:**
  - training data
  - robustness
  - parameter space
  - model misspecifications

SELFI: Simulator Expansion for Likelihood Free Inference



$$\Phi = f(\theta) + \epsilon \longrightarrow \hat{\mathbf{\Phi}}_{\theta} \approx \mathbf{f}_0 + \nabla \mathbf{f}_0 \cdot (\theta - \theta_0)$$

at primordial P(k) expansion point θ_0
+ covariance noise C_0 on final summaries

New distribution, Gaussian, for θ:

mean = $\theta_0 + \mathbf{\Gamma}\,(\nabla \mathbf{f}_0)^\intercal\, \mathbf{C}_0^{-1}(\mathbf{\Phi}_{\mathrm{O}} - \mathbf{f}_0)$

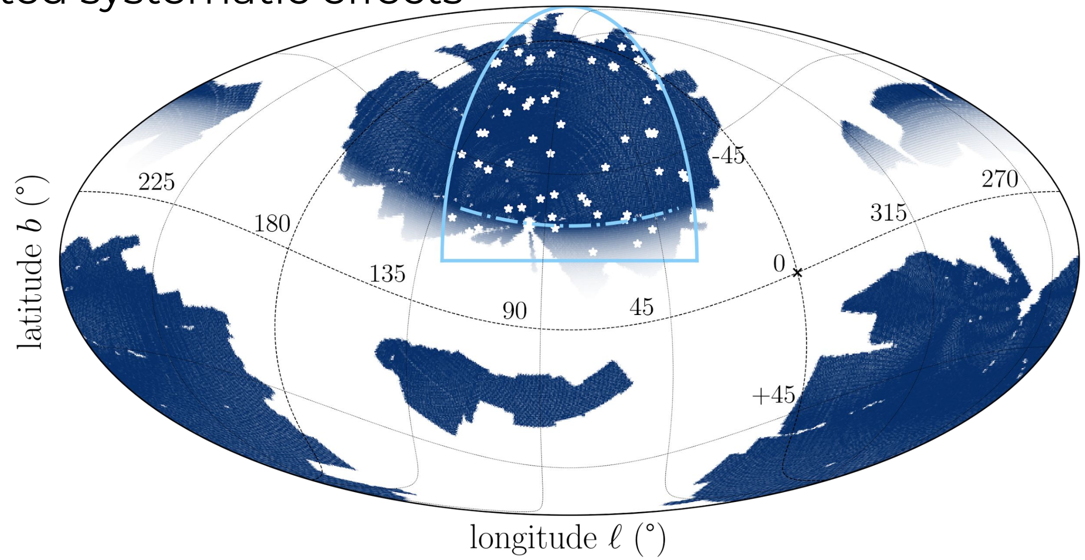
covariance = $\left[(\nabla \mathbf{f}_0)^\intercal\, \mathbf{C}_0^{-1} \nabla \mathbf{f}_0 + \mathbf{S}^{-1}\right]^{-1}$

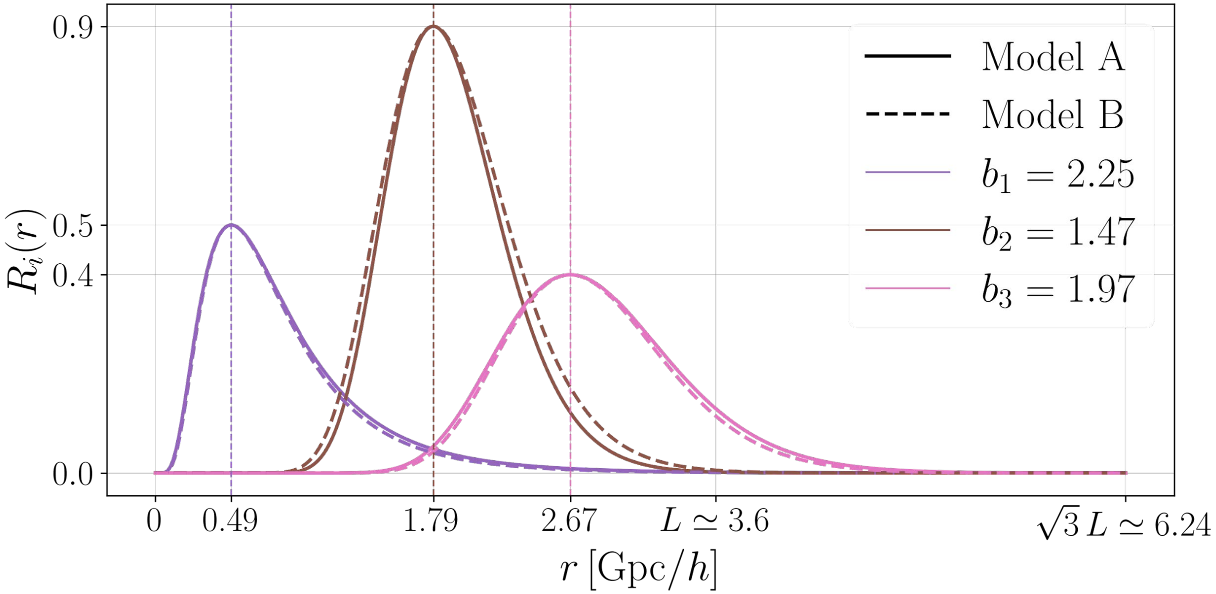
https://florent-leclercq.eu/talks/2019-11-11_Cambridge_DAMTP.pdf

Leclercq et al. (2019), **Hoellinger & Leclercq (in prep. 2024)**

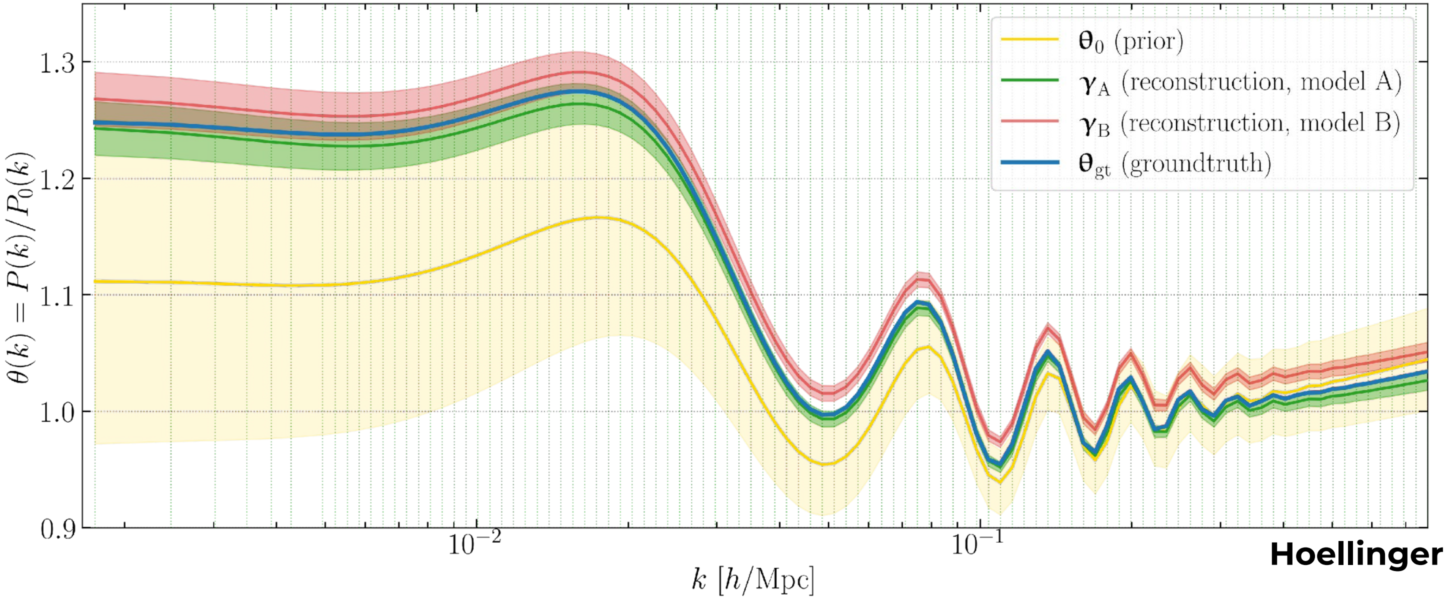
SELFI: Systematic diagnoser from summaries

Injected systematic effects



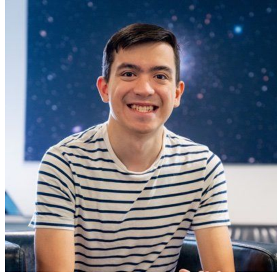


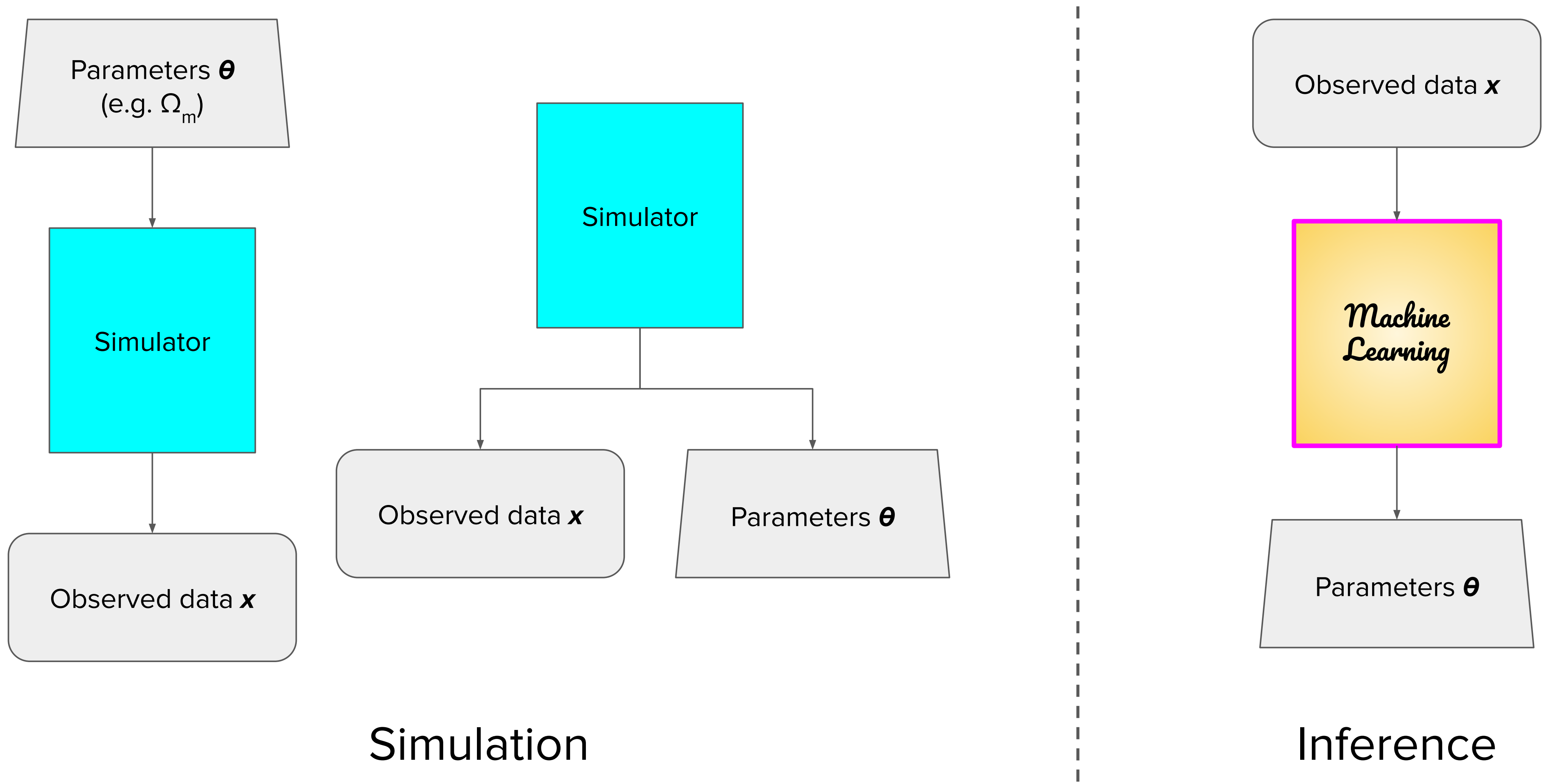
Results on parameter



Hoellinger & Leclercq (in prep. 2024)

Implicit Likelihood Inference (ILI, aka "SBI" & Likelihood Free Inference)





LtU-ILI ~ 3 algorithms: SNLE, SNPE, NRE

[Ho et al. (2024)](#)

inspired from Matt Ho presentation

Concept of ILI inference

$\mathbf{x}$ θ

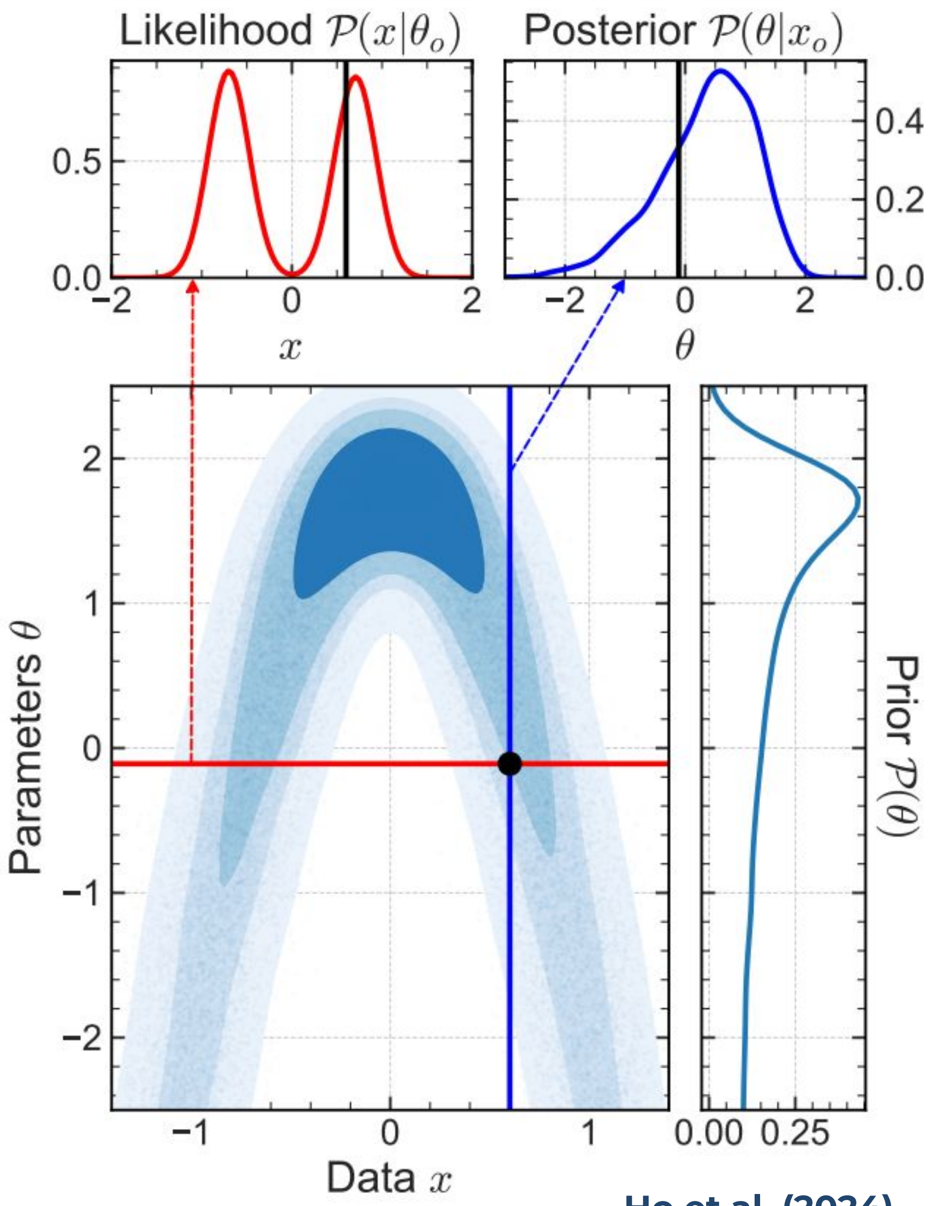
- Assuming we have a <u>perfect simulator</u>, we have (data, parameter) pairs. How do we do inference?

$$\underbrace{P(\boldsymbol{\theta}|\mathbf{x})}_{\text{“Posterior”}} \propto \underbrace{P(\mathbf{x}|\boldsymbol{\theta})}_{\text{“Likelihood”}} \underbrace{P(\boldsymbol{\theta})}_{\text{“Prior”}}$$

Neural **P**osterior **E**stimation

Neural **L**ikelihood **E**stimation

Neural **R**atio **E**stimation



Ho et al. (2024)

Concept of ILI inference: NLE

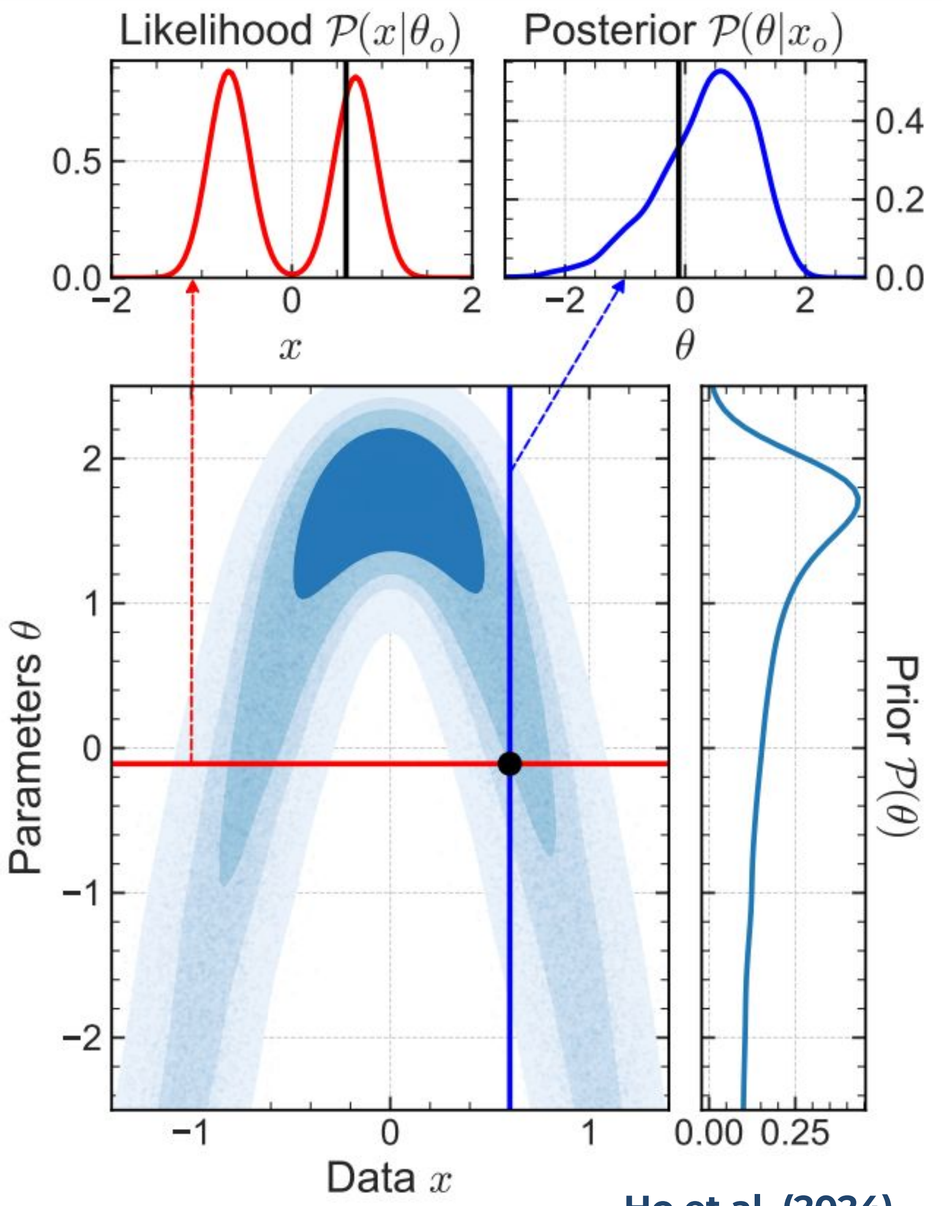
$\mathbf{x}$ θ

- Assuming we have a perfect simulator, we have (data, parameter) pairs. How do we do inference?

$$P(\boldsymbol{\theta}|\mathbf{x}) \propto \boxed{P(\mathbf{x}|\boldsymbol{\theta})}\ P(\boldsymbol{\theta})$$

Neural **L**ikelihood **E**stimation

- Fit a model for the likelihood given (data, parameters)
- Train only one model, and evaluate posterior given a prior at the cost of additional sampling (e.g. MCMC, VI...)



Ho et al. (2024)

Concept of ILI inference: NPE

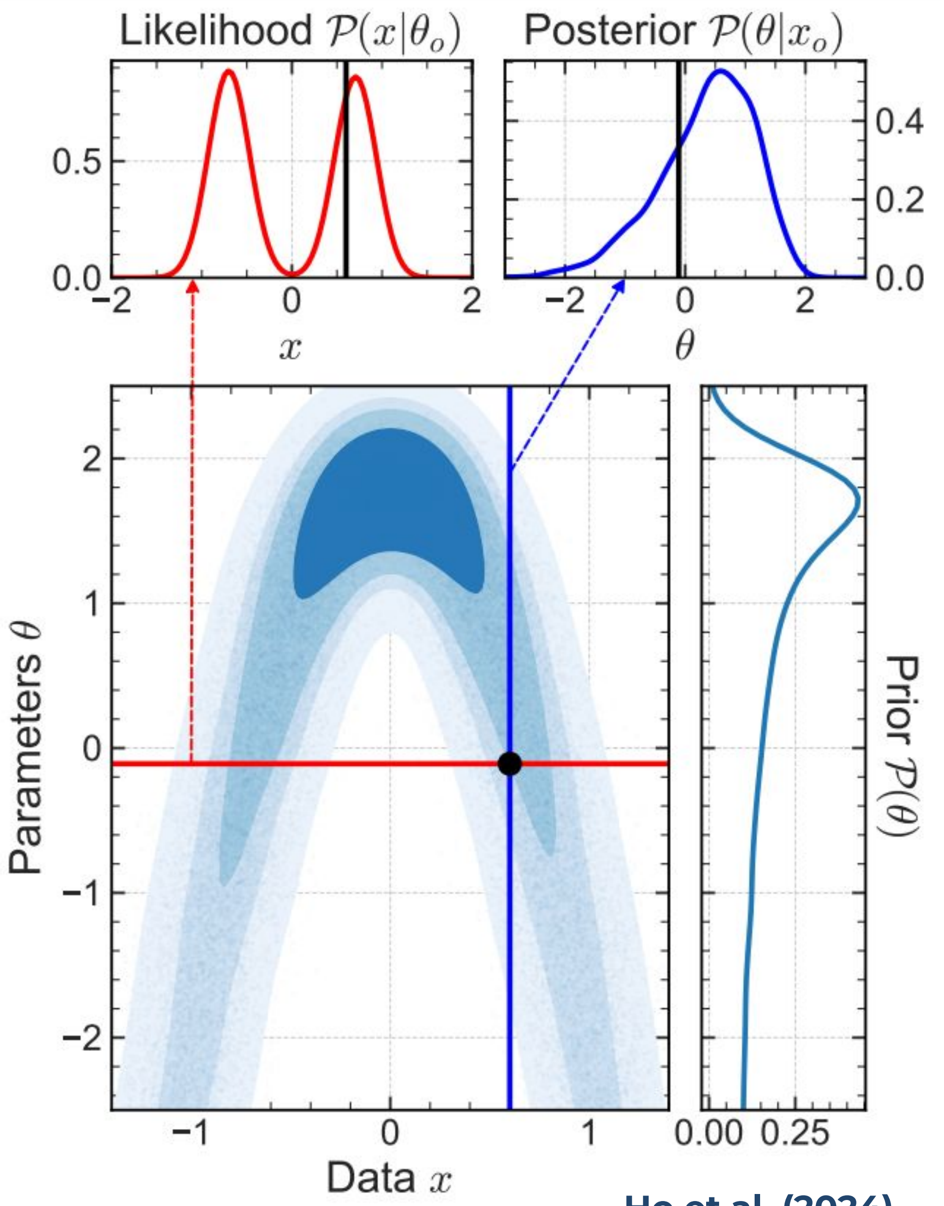
$\mathbf{x}$ θ

- Assuming we have a perfect simulator, we have (data, parameter) pairs. How do we do inference?

$$\boxed{P(\boldsymbol{\theta}|\mathbf{x})} \propto P(\mathbf{x}|\boldsymbol{\theta})\ P(\boldsymbol{\theta})$$

Neural **P**osterior **E**stimation

- Fit a model for the posterior distribution given (data, parameter) pairs.
- Directly outputs posterior to compute validation metrics (one model trained per prior)



Ho et al. (2024)

Concept of ILI inference: NRE

$\mathbf{x}$ θ

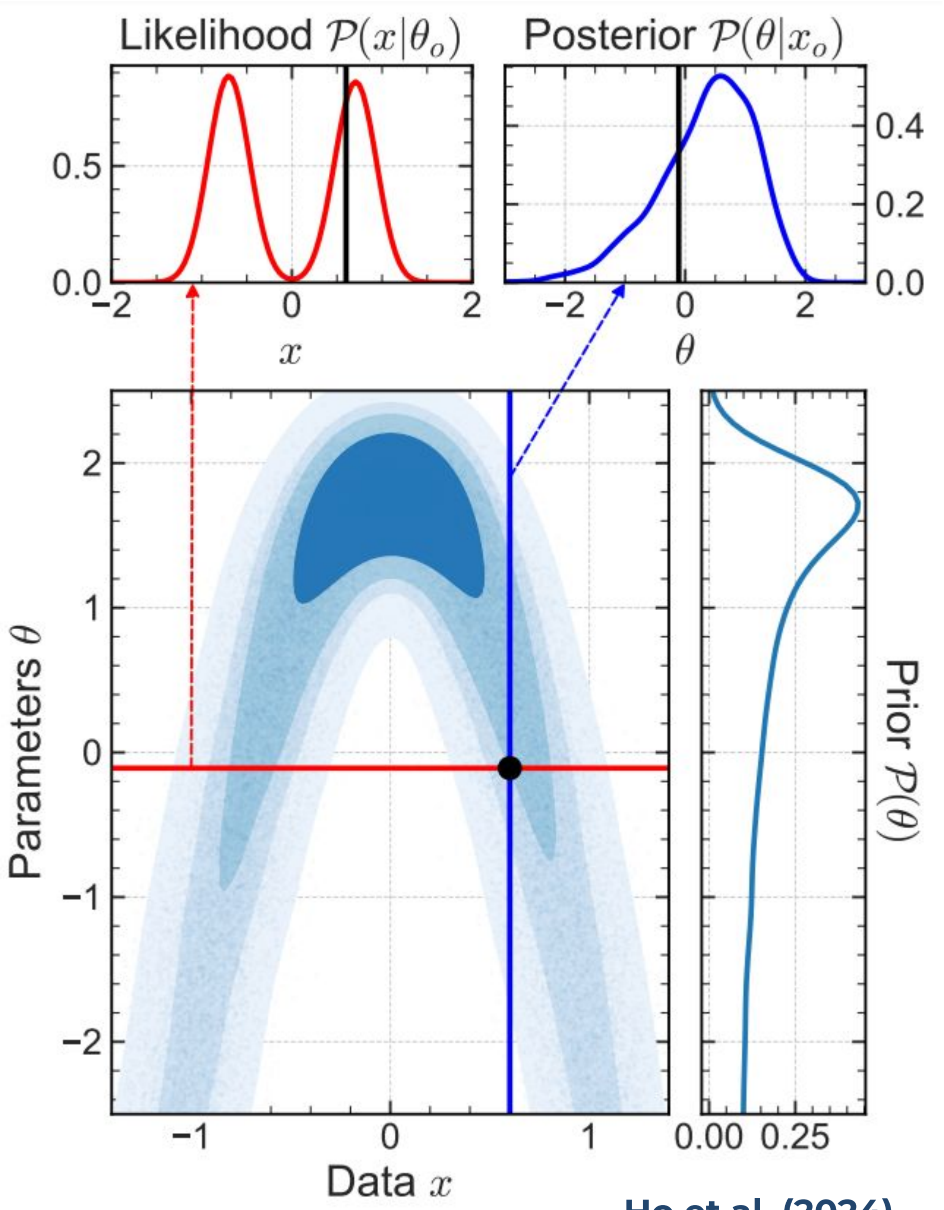
- Assuming we have a perfect simulator, we have (data, parameter) pairs. How do we do inference?

$$P(\boldsymbol{\theta}|\mathbf{x}) \propto P(\mathbf{x}|\boldsymbol{\theta})\ P(\boldsymbol{\theta})$$

Neural **R**atio **E**stimation

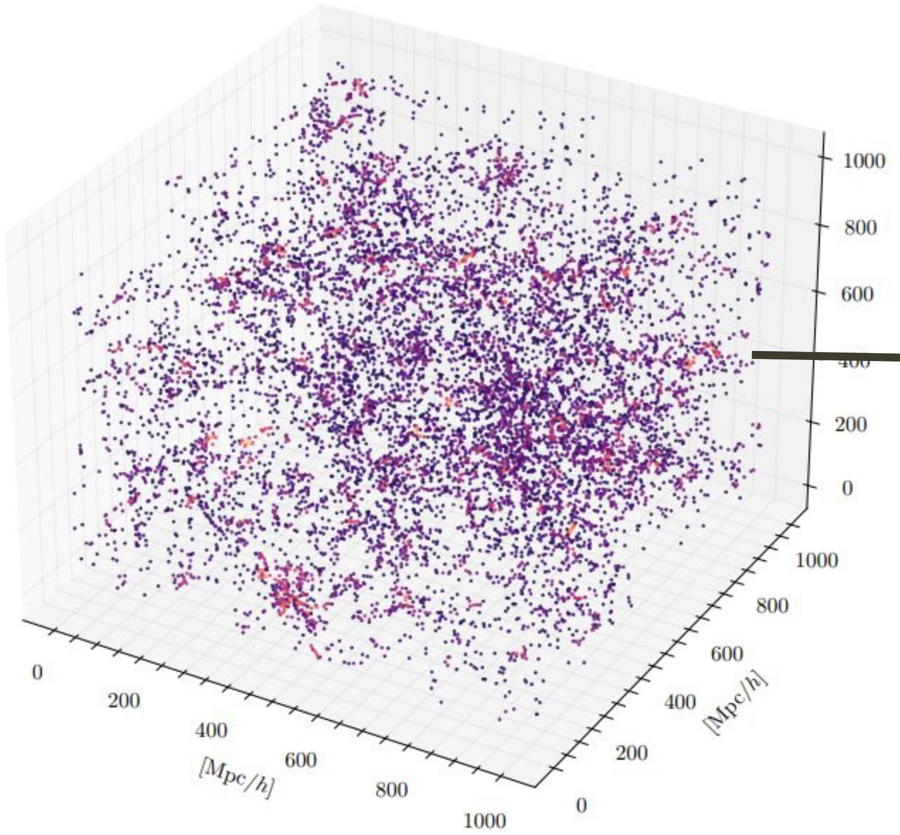
$$\alpha = \boxed{\frac{P(\theta_1|\mathbf{x})}{P(\theta_0|\mathbf{x})}}$$

Acceptance Ratio



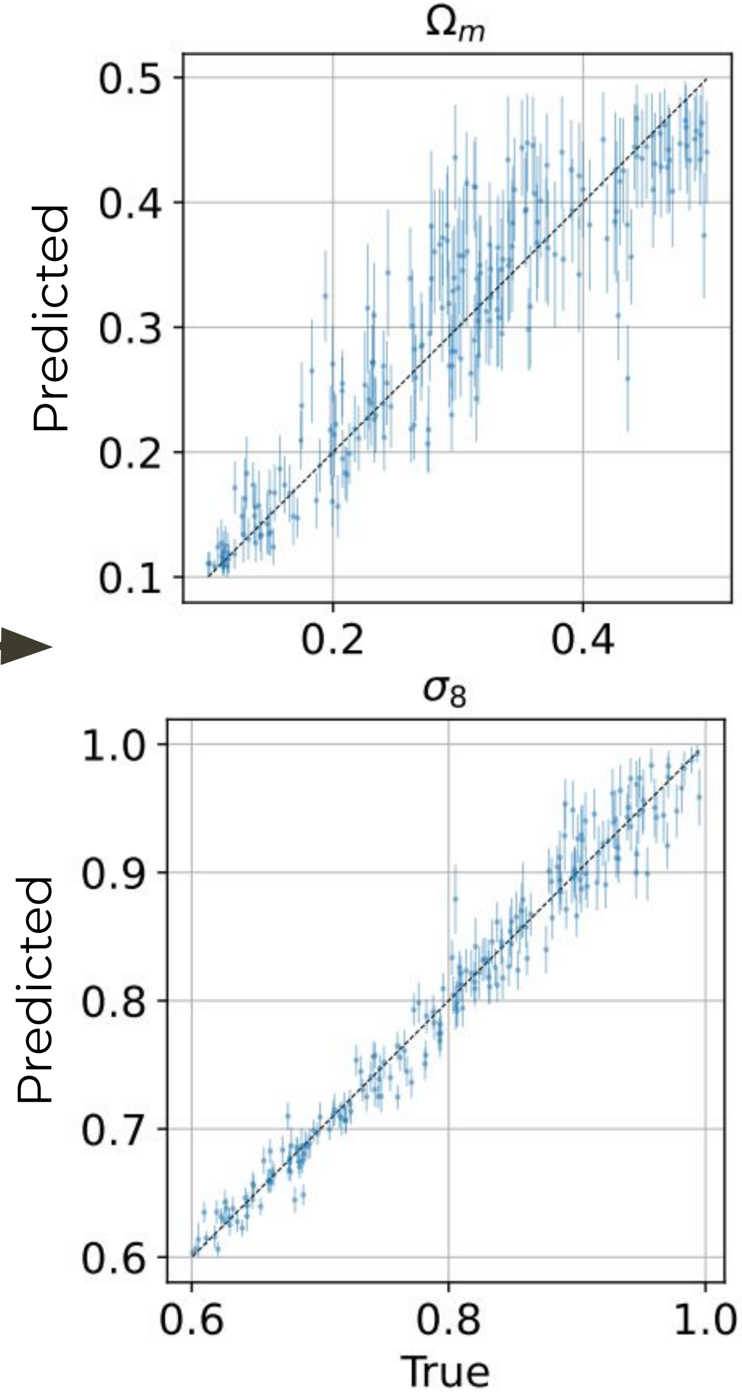
Ho et al. (2024)

Example of an LtU-ILI to cosmology



Halo power spectrum multipole

ILI



Ho et al. (2024)

5

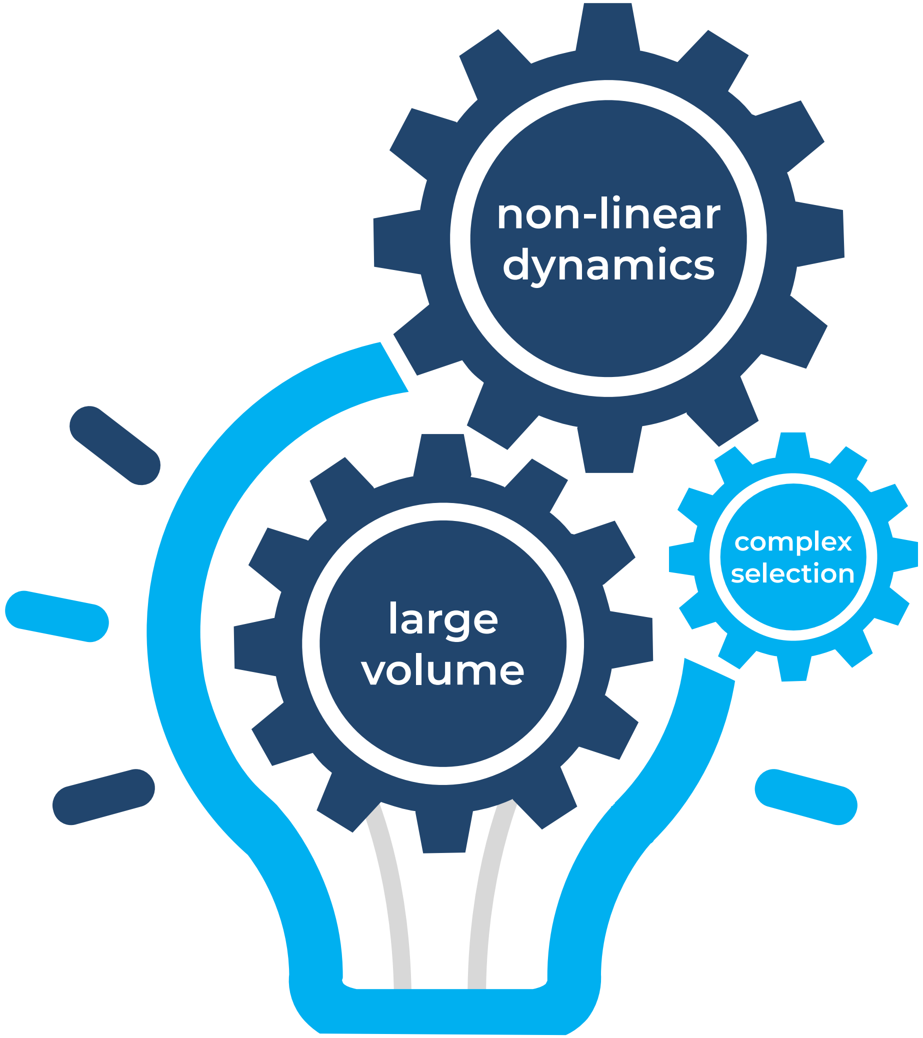
Summary & Outlook

Rise of the machines

Large cosmological surveys = very complex to analyse

Rise of the machines is inevitable to continue progressing

Unlocked by GPU hardware with large memory



Rise of the machines

Full panorama of ML in cosmology is difficult
(last conference attracted ~400 people)

https://indico.iap.fr/e/ml-2023

Emulation:

- Models validated on large datasets
- Exhibit interesting generalization

Statistical techniques based on ML showing increasing robustness for inference

Limits are:

- validity of simulations
- Resilience to unknown systematics

Opportunities by choosing carefully crafted I/O to neural networks