A Hybrid Deep Learning Approach to Cosmological Constraints From Galaxy Redshift Surveys

Michelle Ntampaka
in collaboration with Daniel Eisenstein,
Sihan Yuan, and Lehman Garrison

HDSI | Harvard Data Science Initiative

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

ΛCDM Cosmological Model

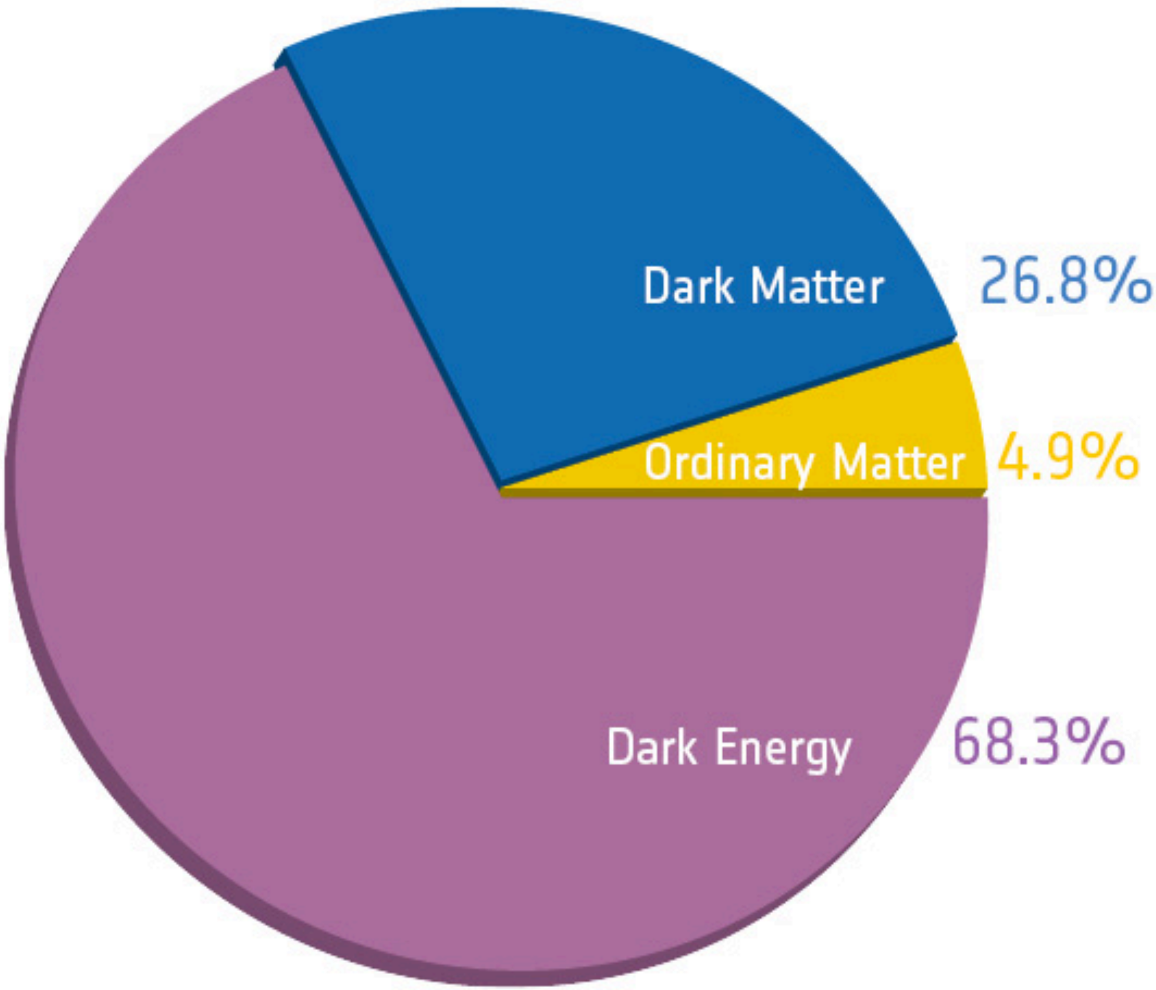
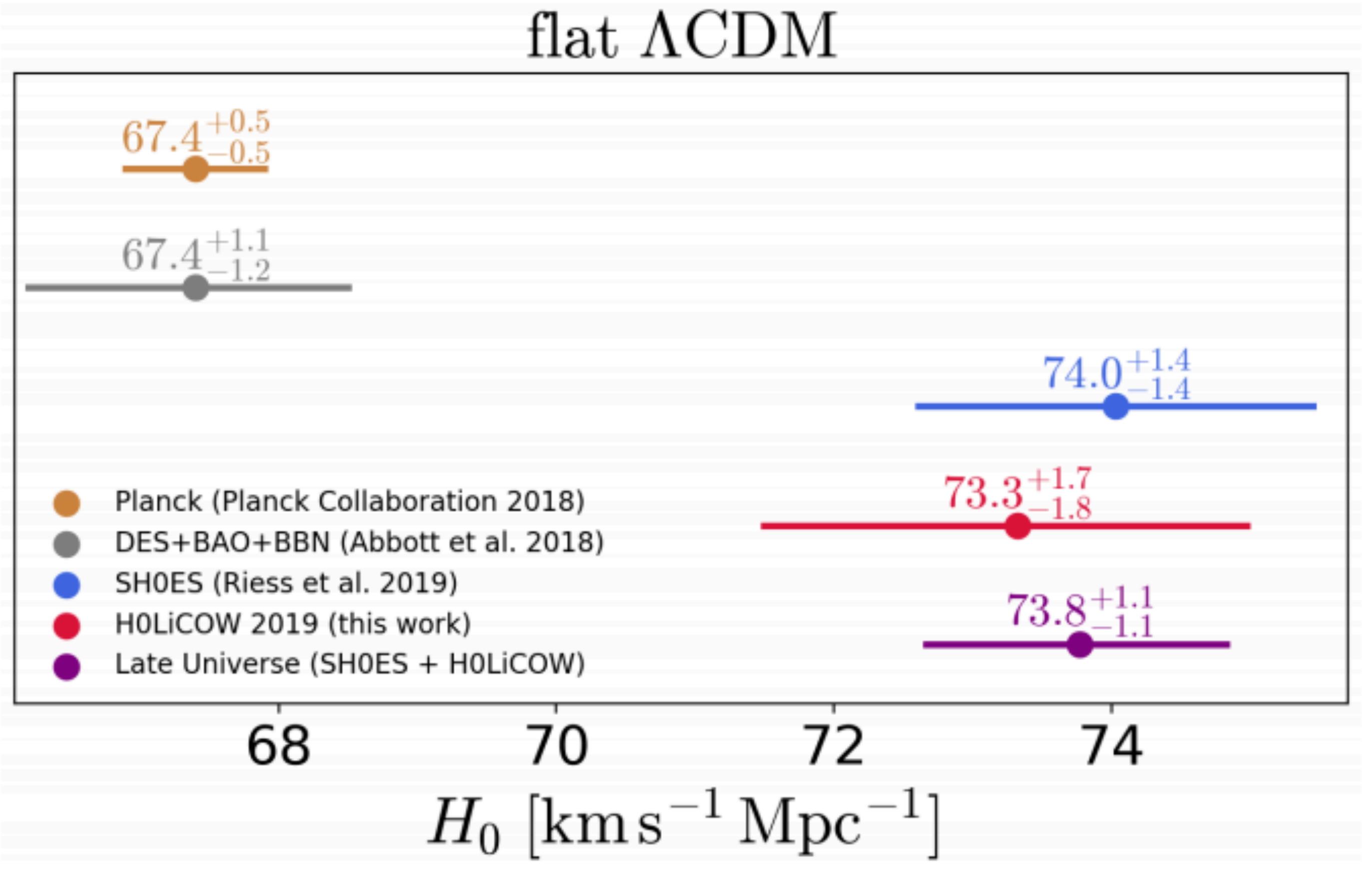


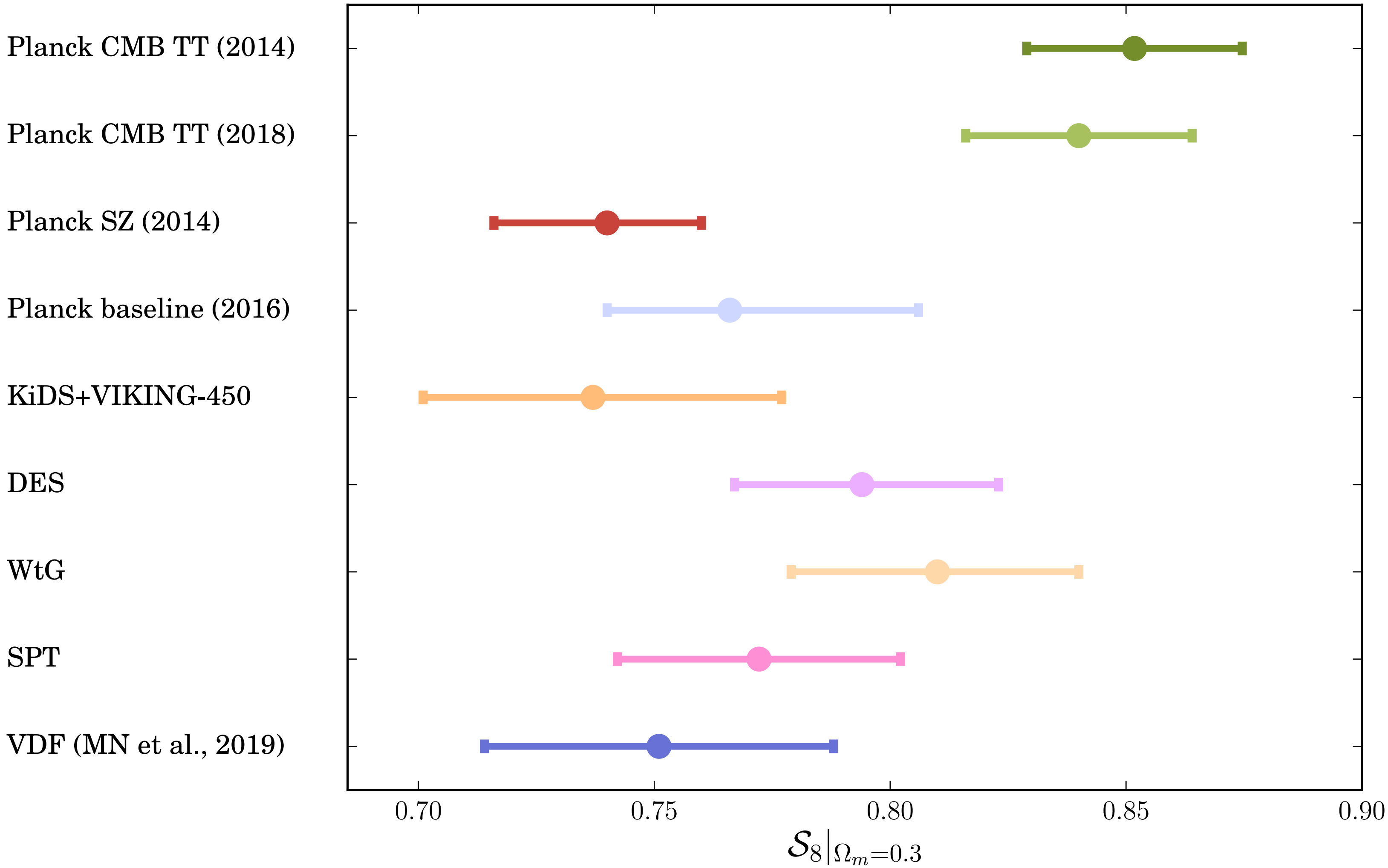
Image credit: ESA/Planck

Tensions in the current cosmological model: H_0 (early vs. late Universe)

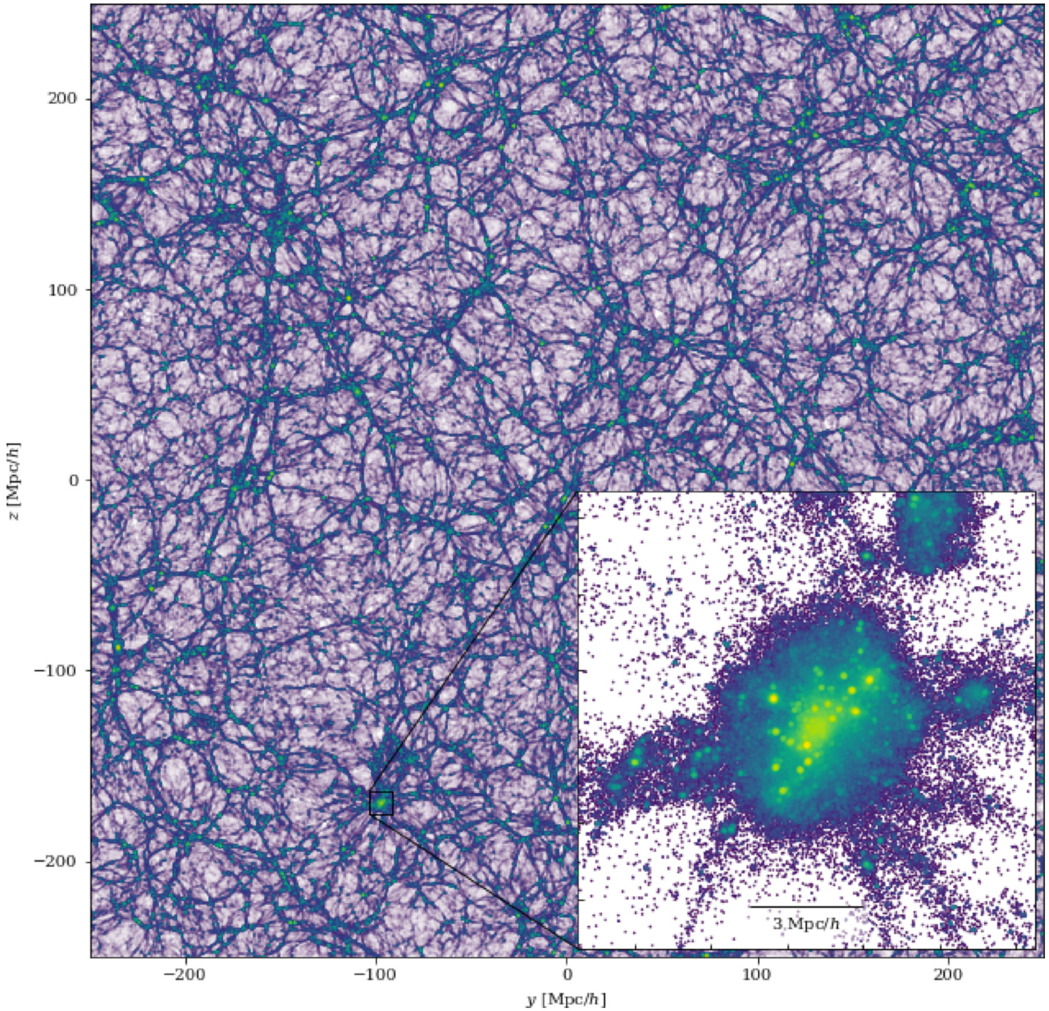


Wong+ 2019

Tensions in the current cosmological model: σ_8 (CMB vs. LSS)



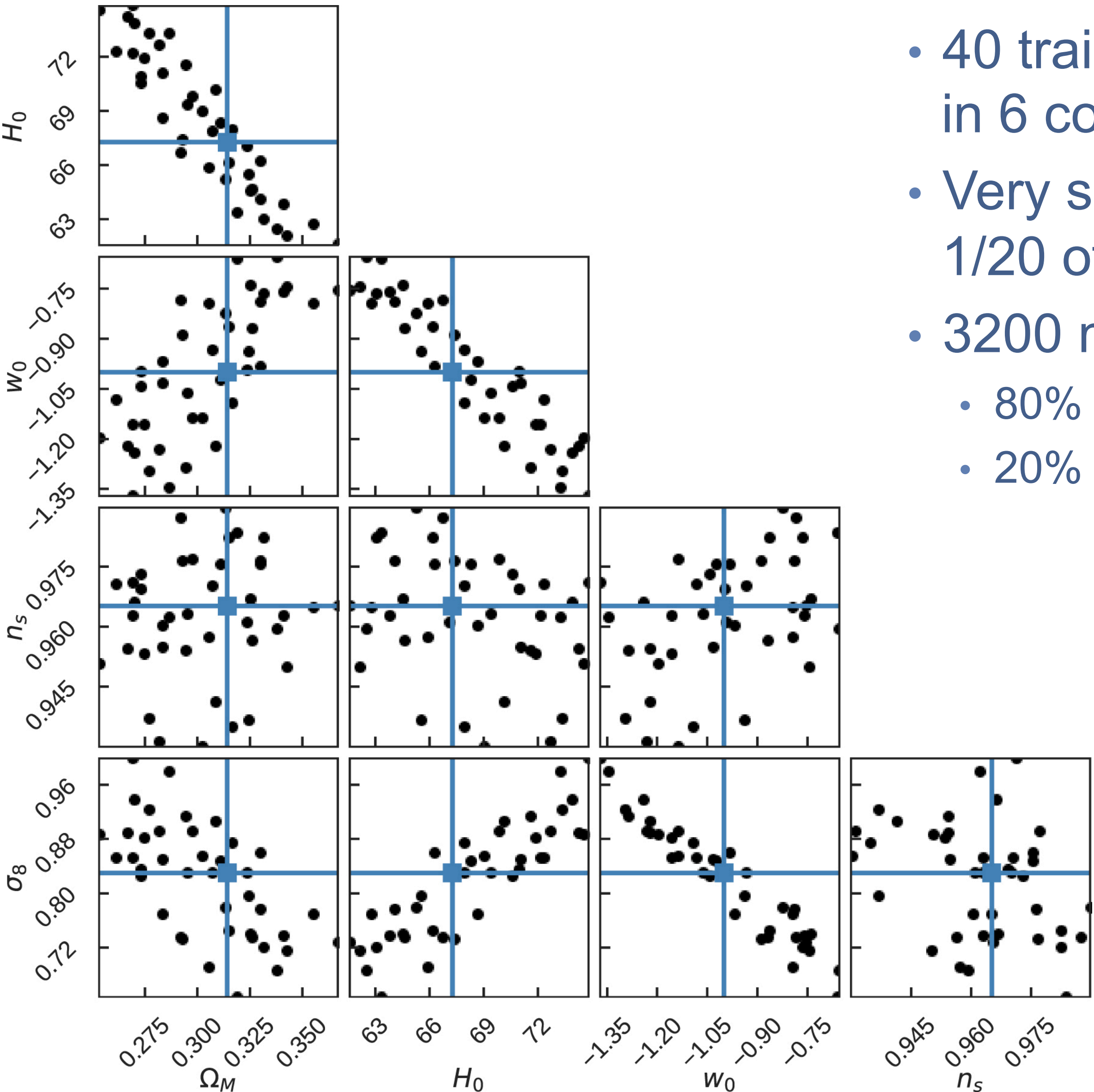
Constraining ΛCDM with Large Scale Structure



Garrison, Eisenstein, & Pinto 2019

- Spatial distribution and clustering of galaxies (via the power spectrum)
- Cosmic shear
- Baryon acoustic oscillations
- Abundance of clusters

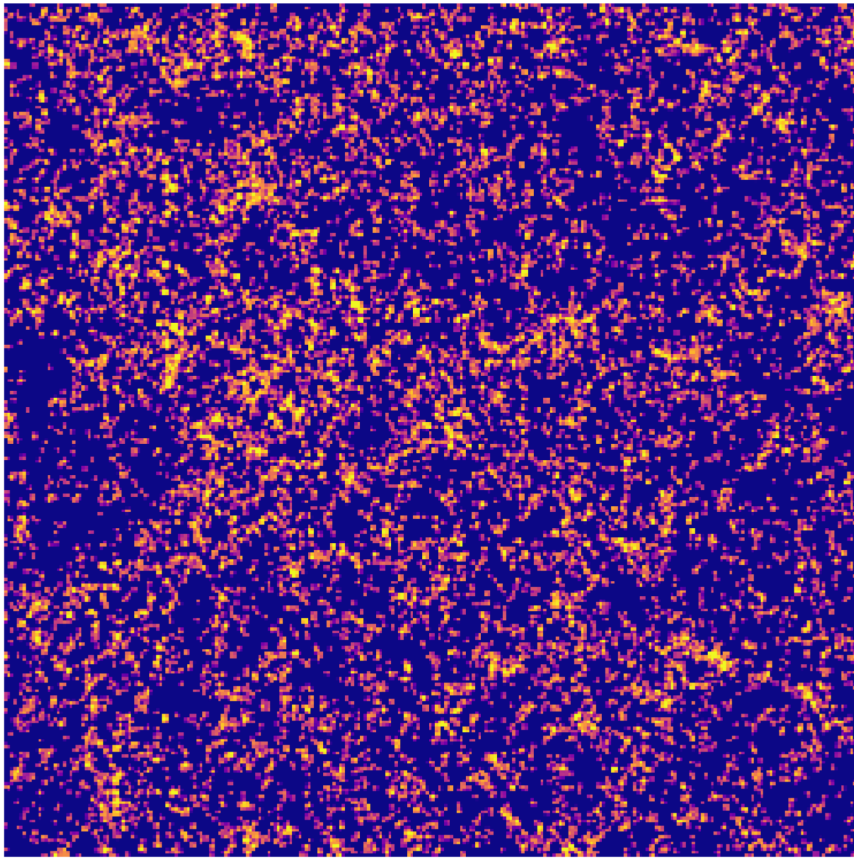
AbacusCosmos Suite of *N*-body Simulations



- 40 training simulations that vary in 6 cosmological parameters
- Very small mock observations, 1/20 of the ~Gpc^3 box
- 3200 mock observations
  - 80% for training
  - 20% for validation

Garrison+ 2018

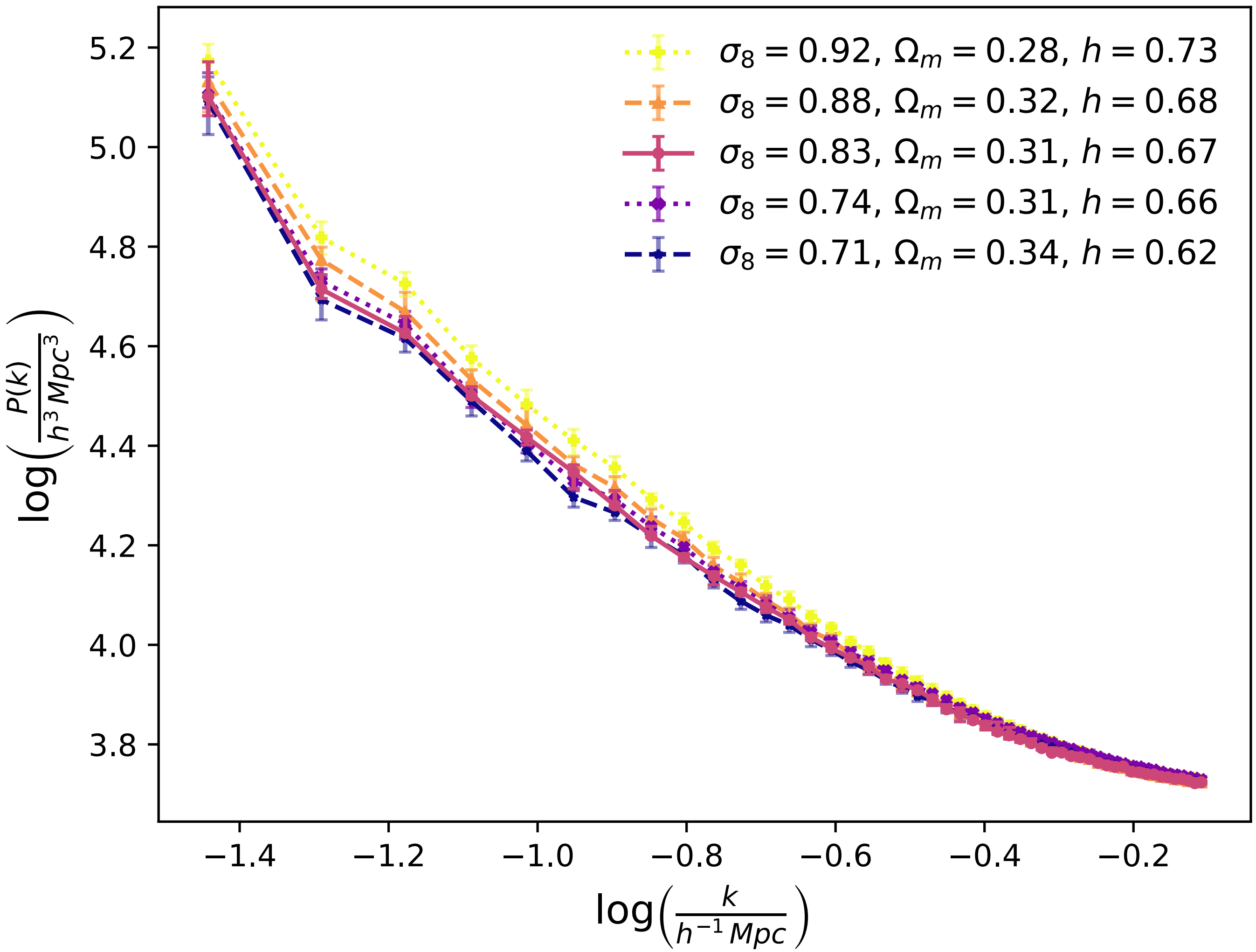
3D Mock Galaxy Catalogs



A 2D projection of
a sample 3D galaxy catalog.

- Halo catalogs are populated with galaxies according to a generalized Halo Occupation Distribution.
- 6 parameters to capture a range of galaxy formation models.

A Standard Approach: The Galaxy Power Spectrum



Galaxy Power Spectrum

The power spectrum does not tell the whole story!

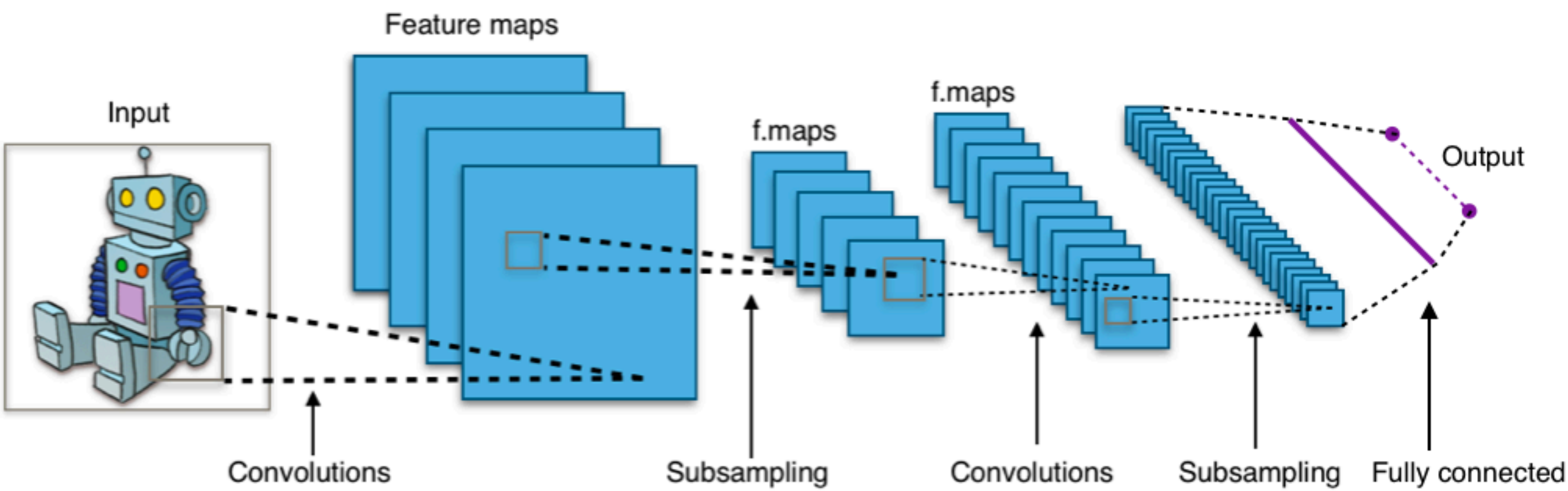
Other statistics are rich in complementary cosmological information:

- 3-point correlation function (Yuan et al. 2018).
- Redshift space power spectrum (Kobayashi et al. 2019).
- Counts-in-cylinders (Wang et al. 2019).

Can we use physics *plus* ML to improve constraints on cosmological parameters?

Can a deep ML method find meaningful patterns – beyond the power spectrum – that correlate with cosmology?

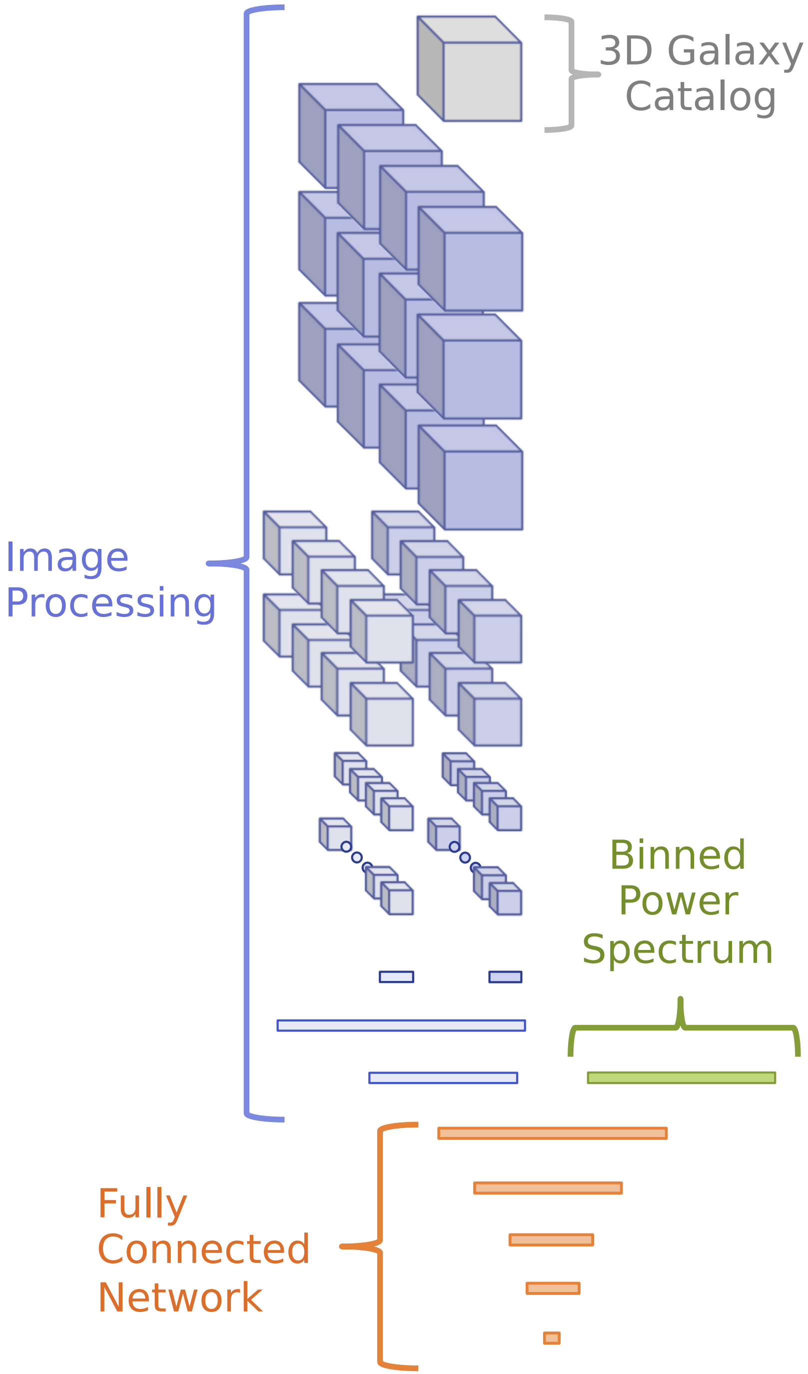
2D Convolutional Neural Network (CNN)



- Start with an input image
- Goal: predict a label (at the output neuron)
- Learn a system of convolutional filters to extract features (shapes, edges, textures, etc.) from the image
- Learn the weights and biases to use these features to predict answers and minimize loss.

3D hybrid CNN Architecture

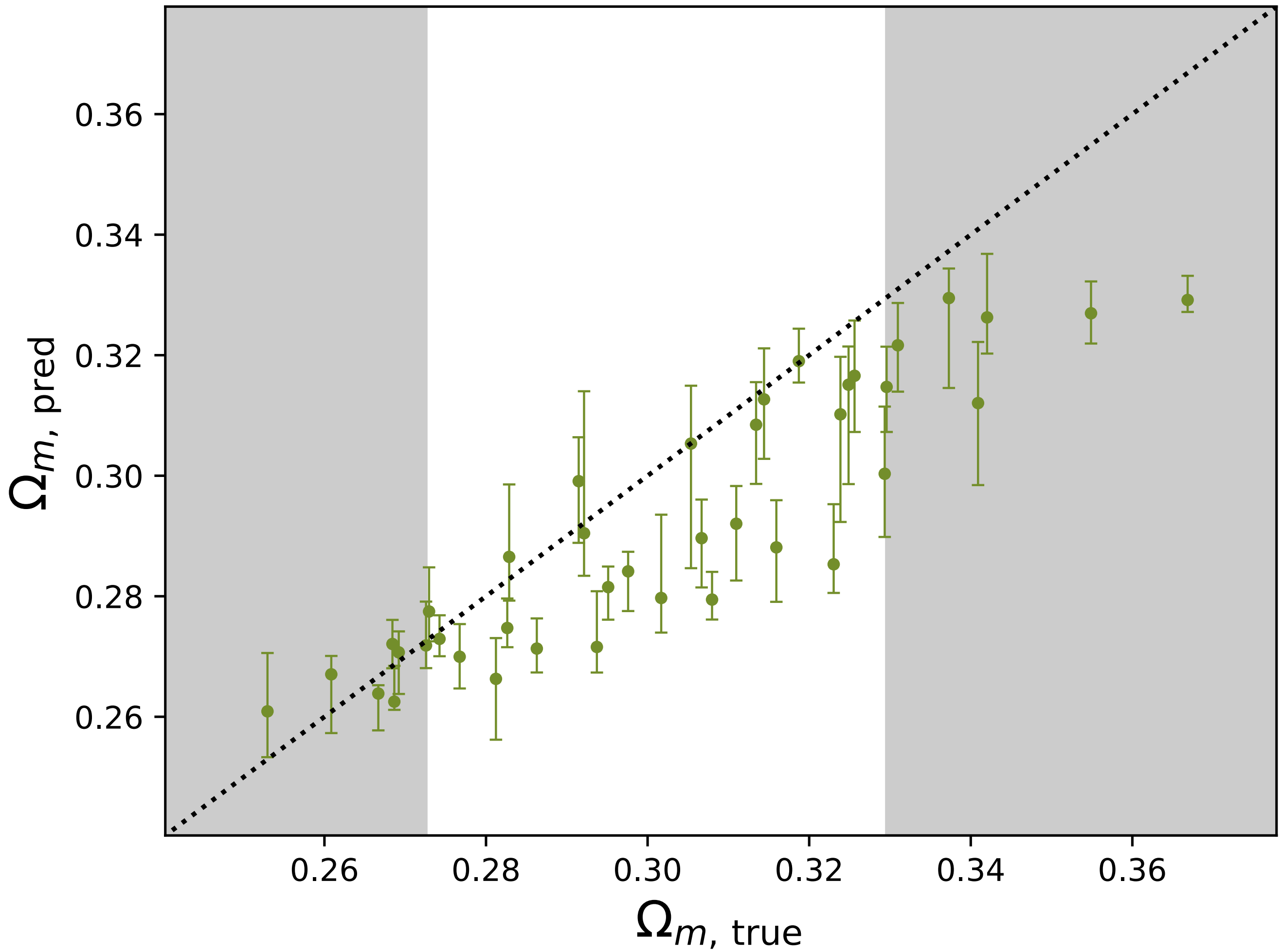
- Power Spectrum Neural Network input:
  - Binned Power Spectrum
- Hybrid CNN input:
  - 3D Galaxy Catalog
  - Binned Power Spectrum
- 2 cosmological parameters predicted (σ_8 and Ω_m)



Architecture Choices

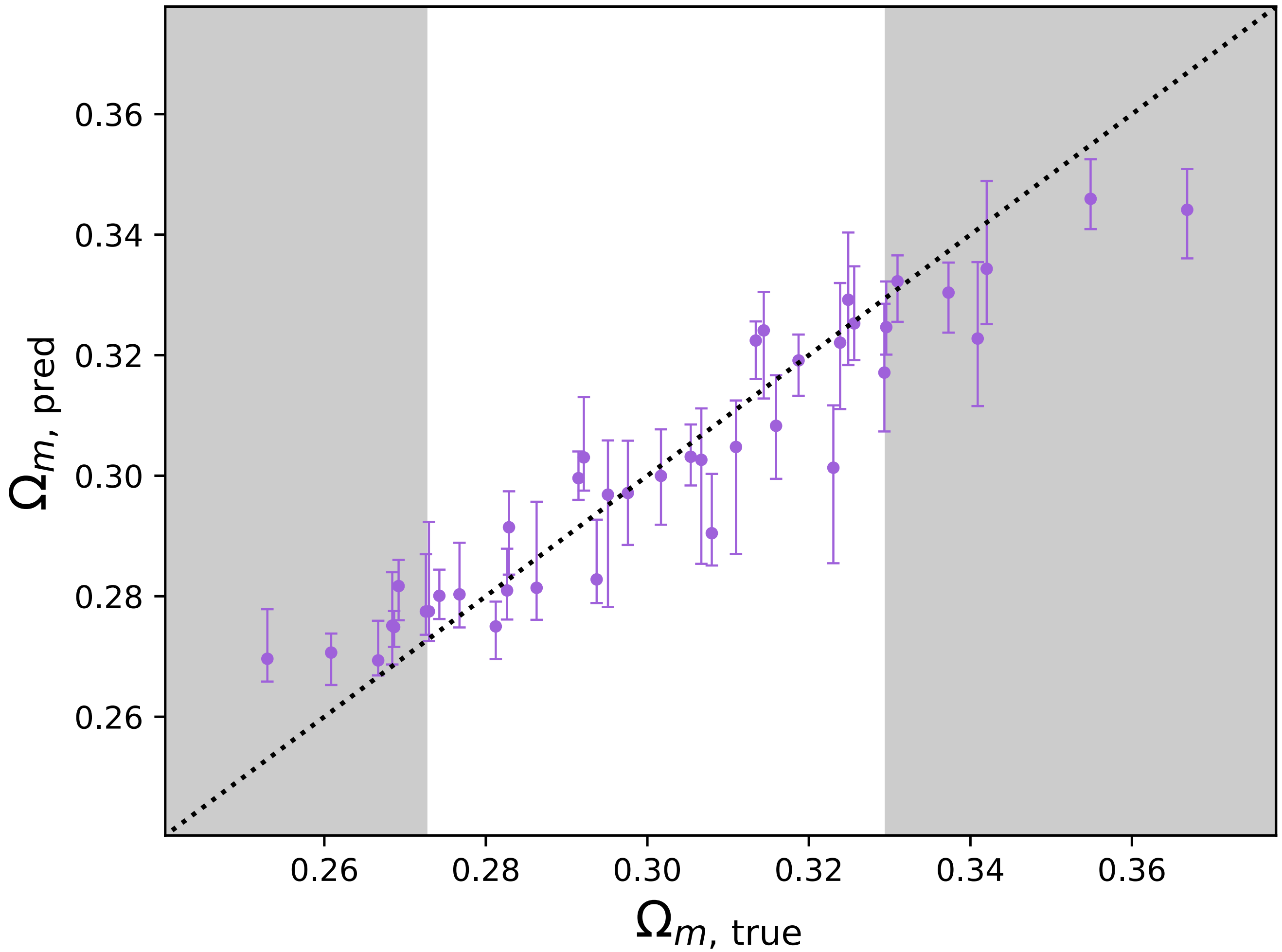
- Simulation length scales → Voxel size
- Matched phase simulations → Train/validate split
- Matched phase simulations → ML architecture with aggressive dropout
- Correlations in galaxy number count with cosmology → downsampling

Ω_m Constraints – Power Spectrum



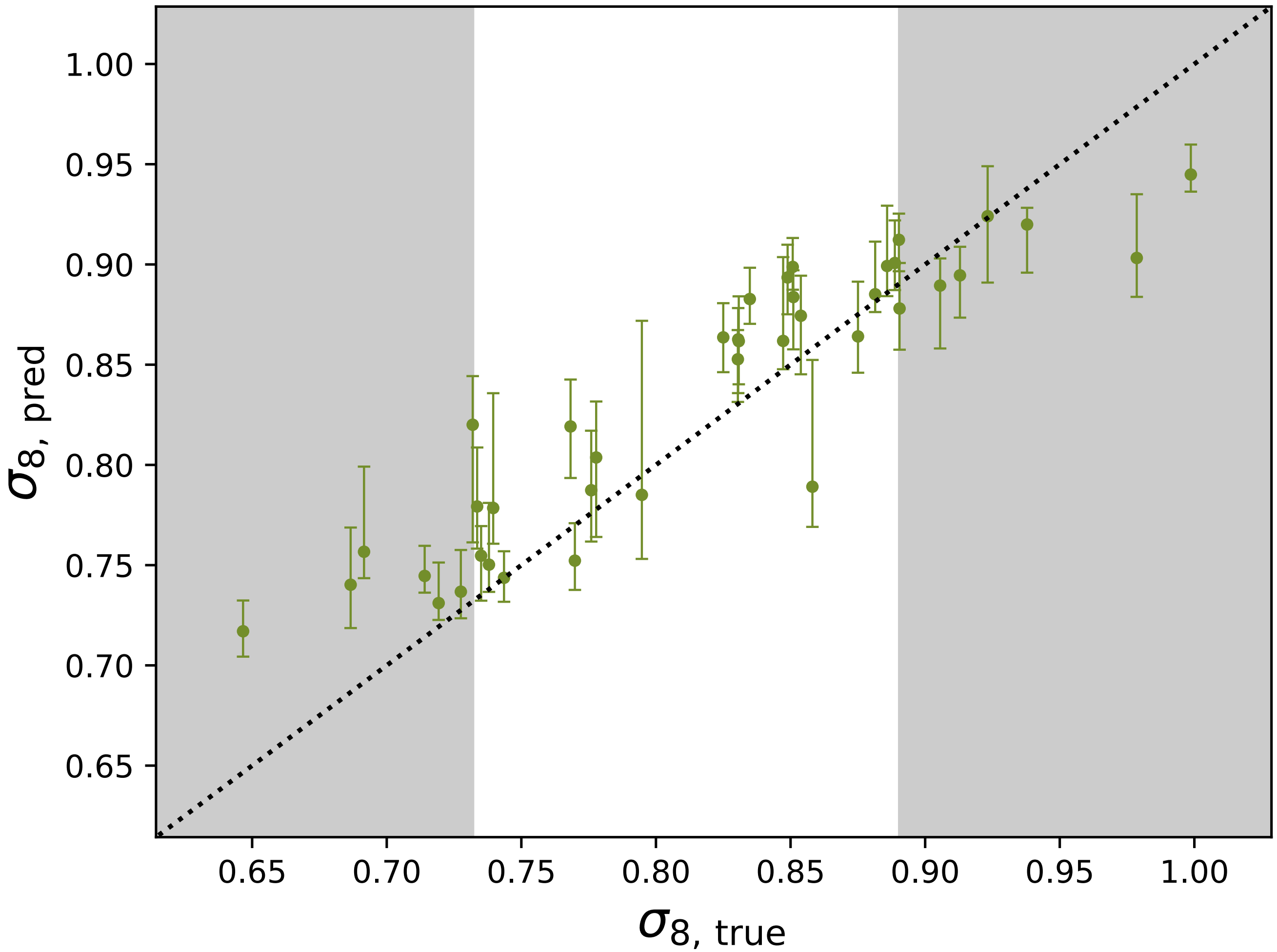
Ntampaka, Eisenstein, Yuan, & Garrison 2019 (1909.10527)

Ω_m Constraints – Hybrid CNN



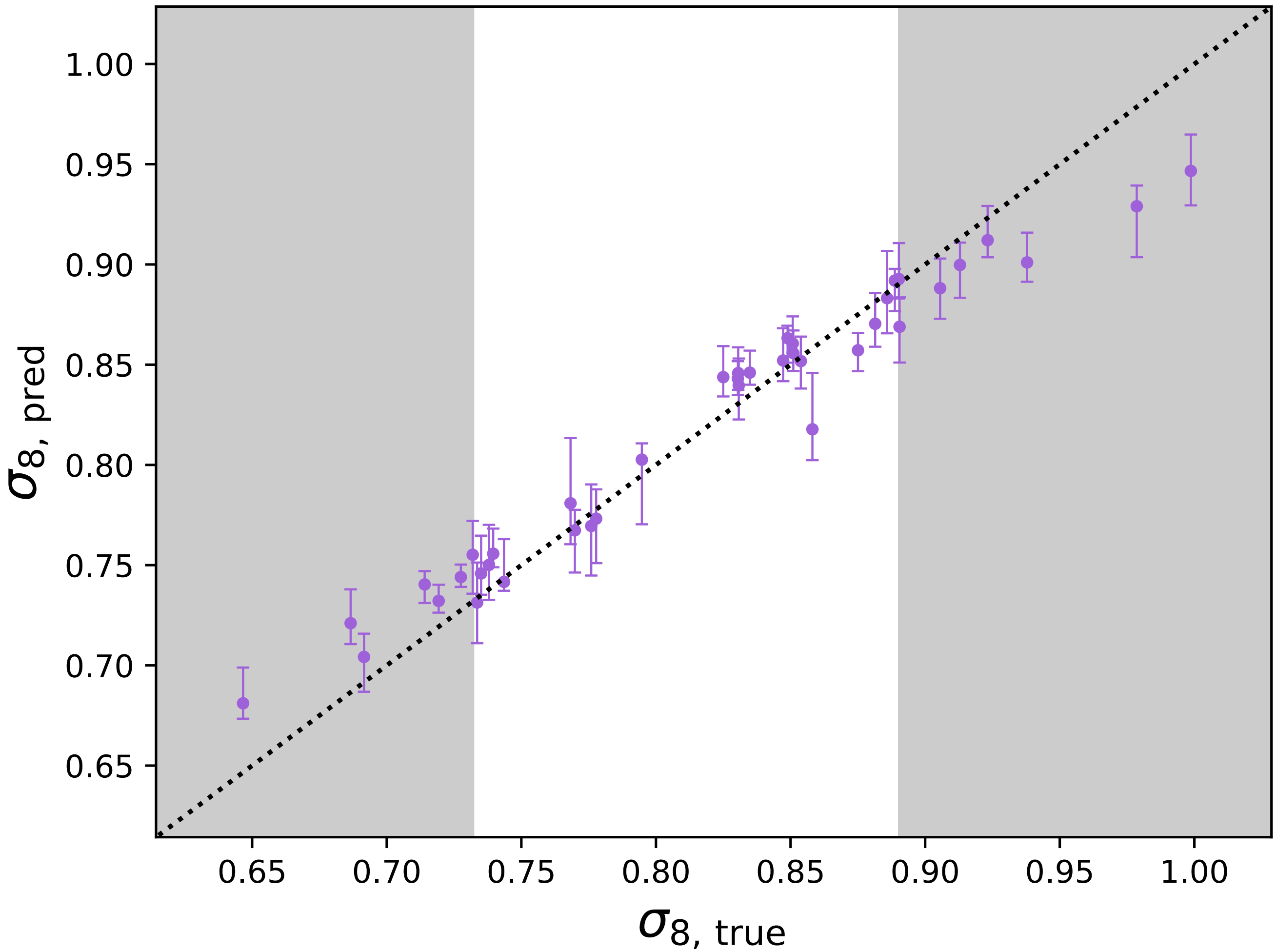
Ntampaka, Eisenstein, Yuan, & Garrison 2019 (1909.10527)

σ_8 Constraints – Power Spectrum



Ntampaka, Eisenstein, Yuan, & Garrison 2019 (1909.10527)

σ_8 Constraints – Hybrid CNN

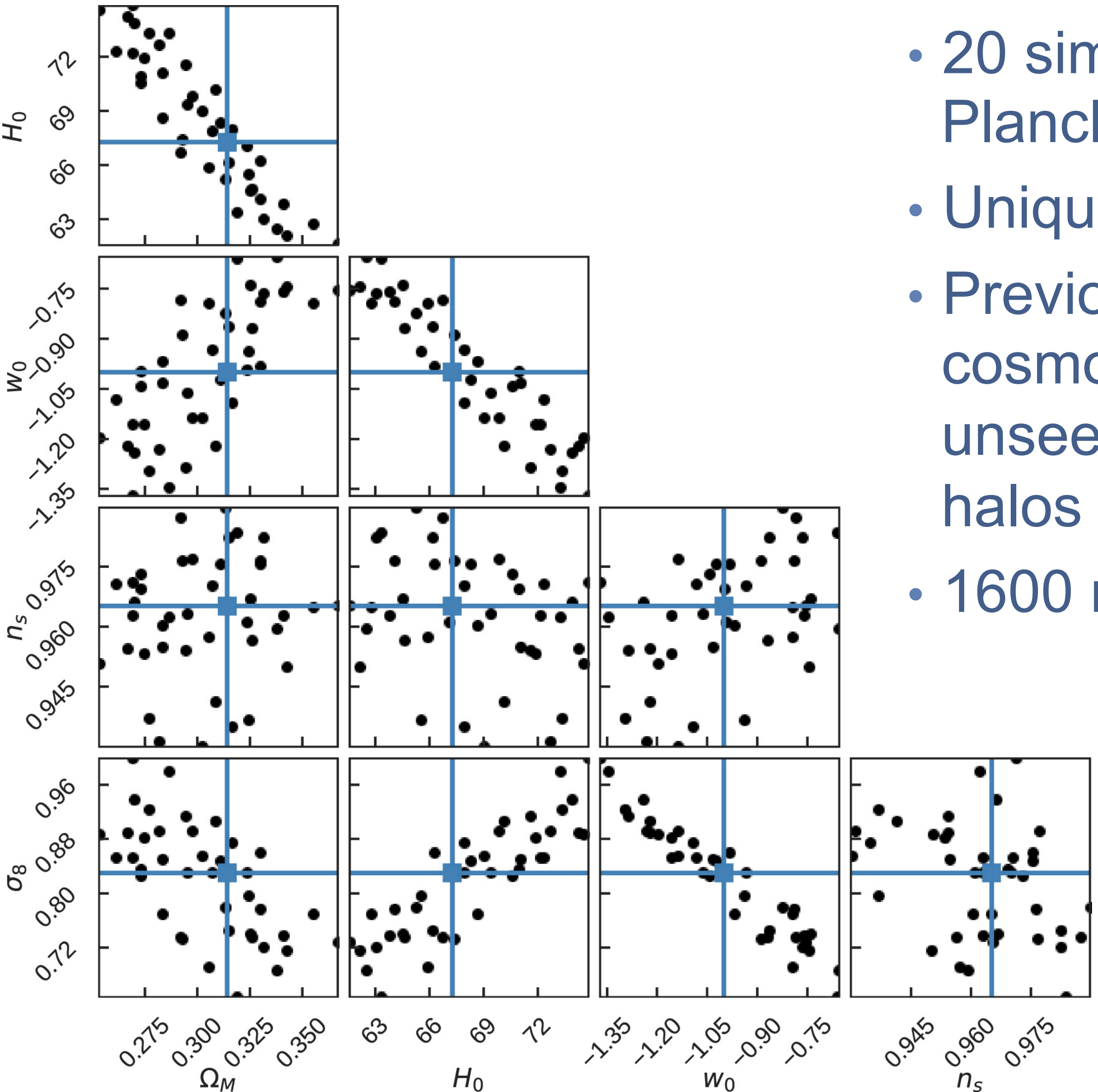


Ntampaka, Eisenstein, Yuan, & Garrison 2019 (1909.10527)

Proceed with Caution!

1. The network was trained to recognize *these* cosmologies – will it interpolate to cosmologies its never seen before?
2. The network was trained on matched-phase simulations – is it memorizing structures that correlate across simulations?

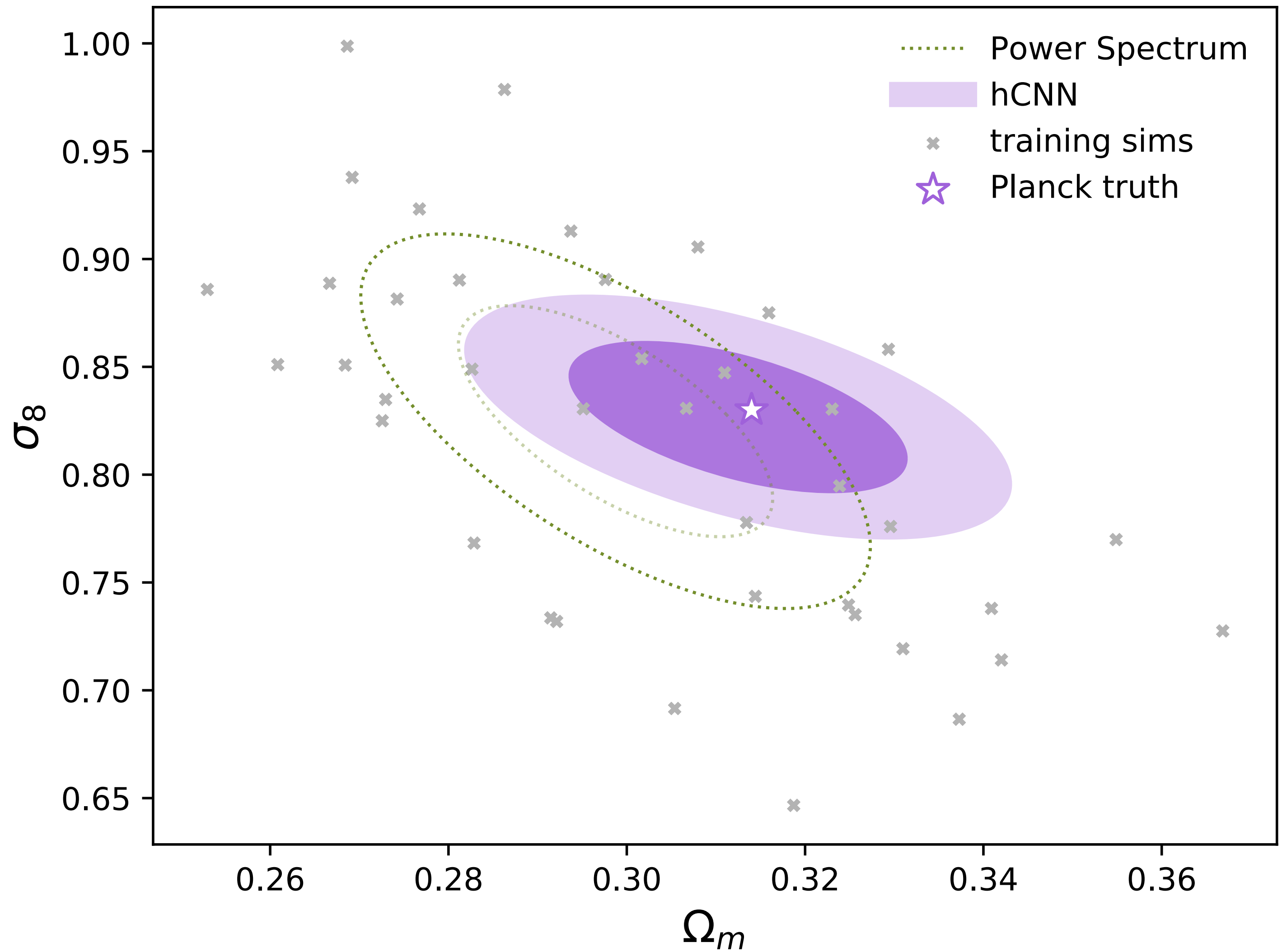
Planck Simulations Testing Catalog



- 20 simulation suite at the Planck cosmology.
- Unique initial conditions.
- Previously unseen cosmology and previously unseen model for populating halos with galaxies.
- 1600 mock observations.

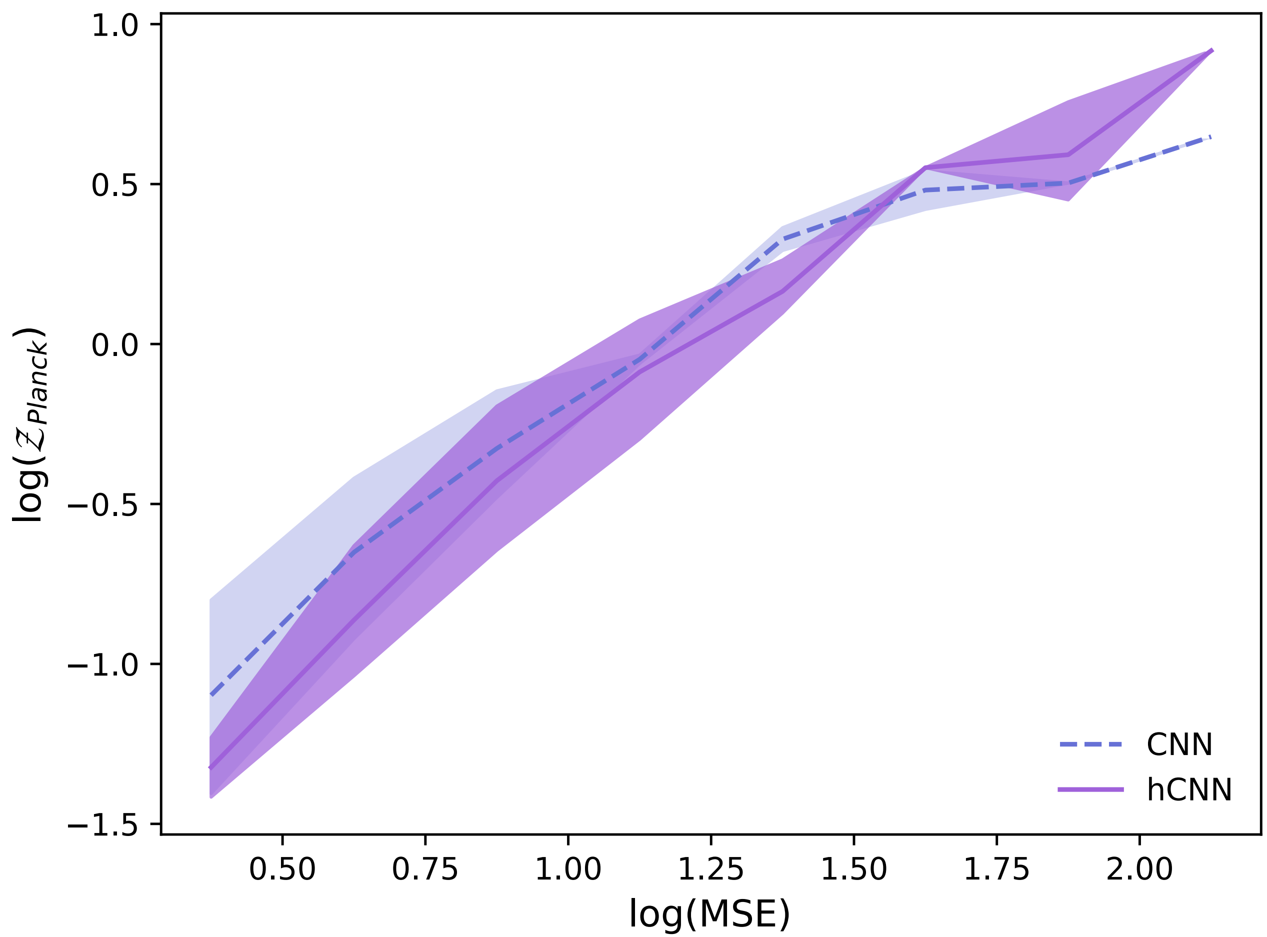
Garrison+ 2018

Planck Test Set Constraints

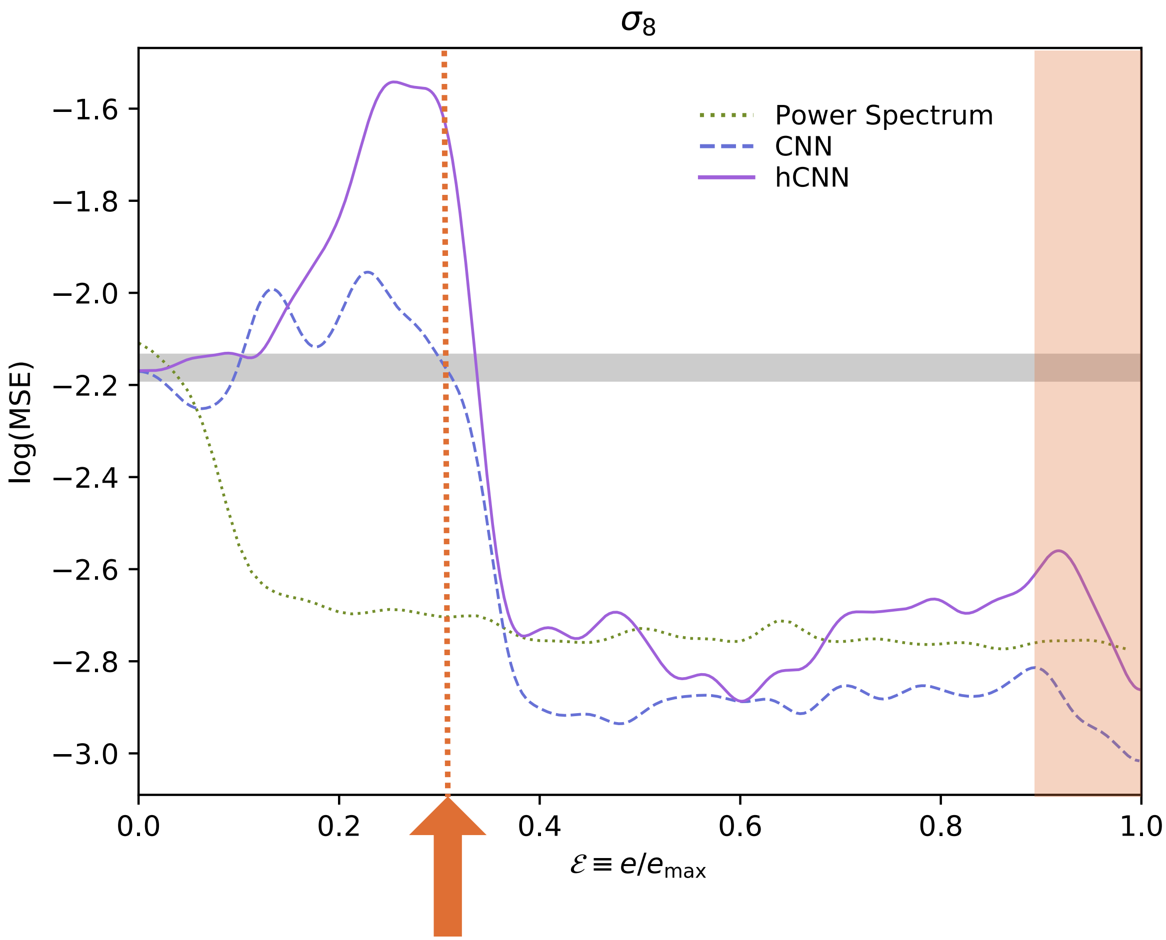


Ntampaka, Eisenstein, Yuan, & Garrison 2019 (1909.10527)

Lessons From Training #1: The Validation Set Helps!



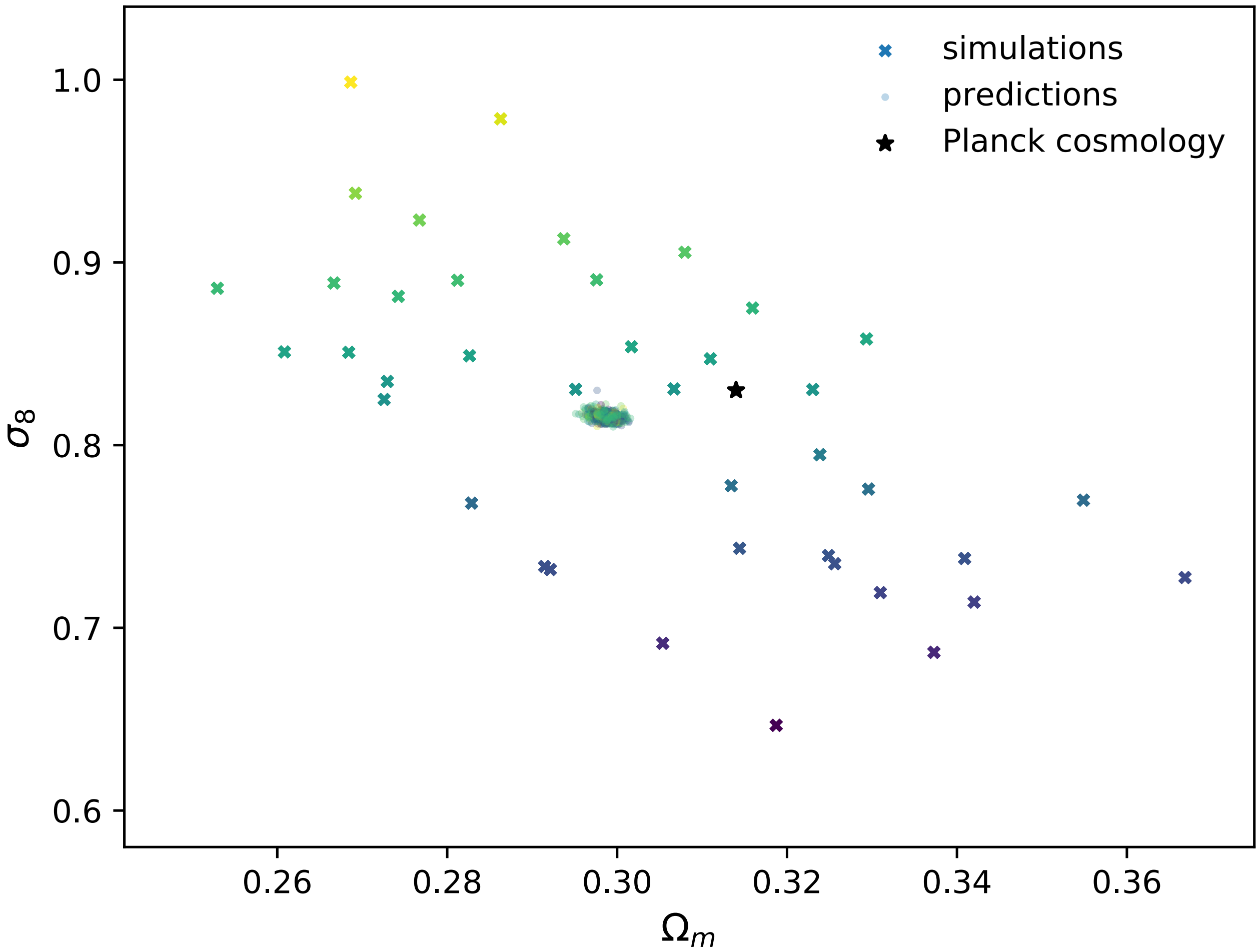
Lessons From Training #2: Summary Stats Can Lie



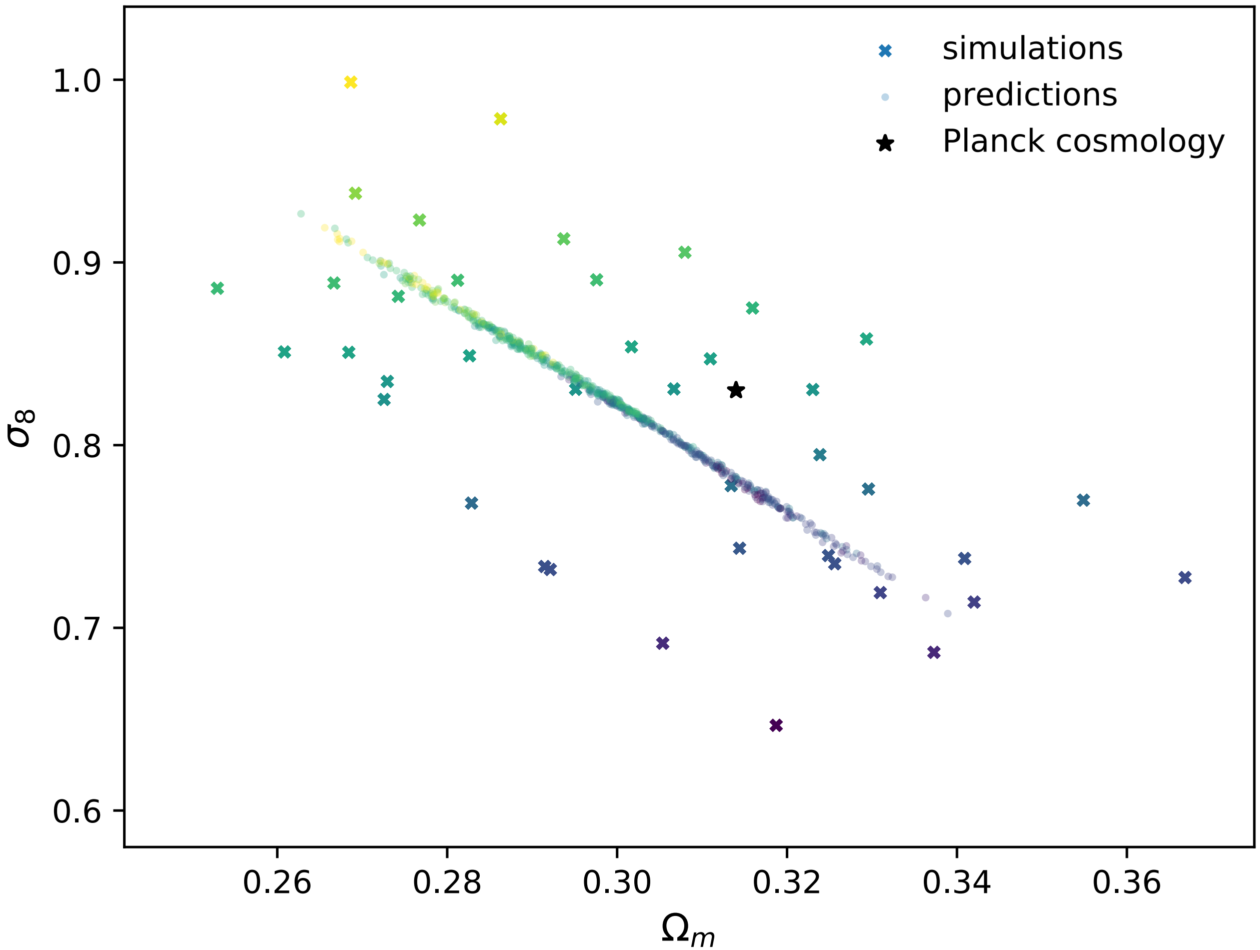
Assess best fit

Transition to lower learning rate

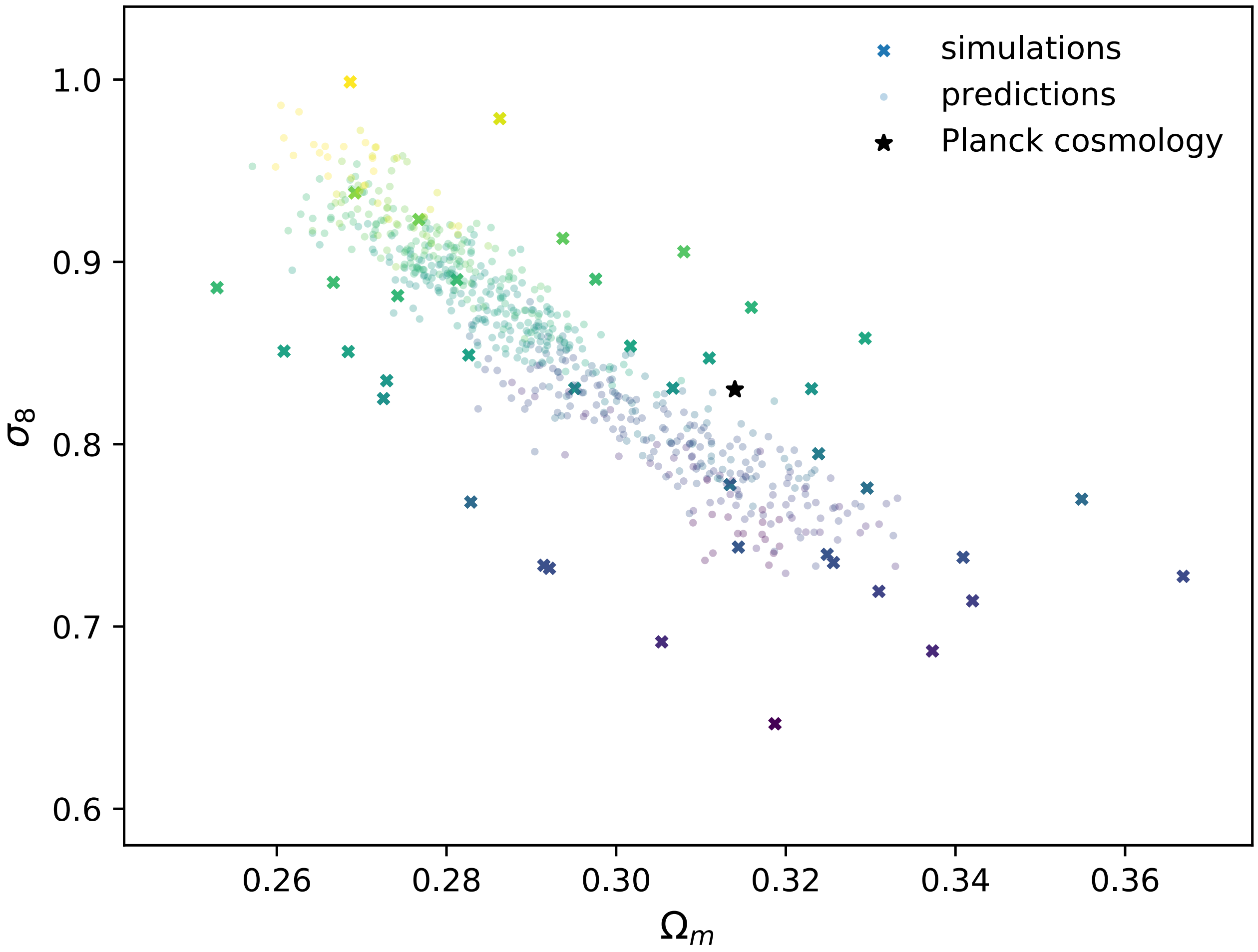
Lessons From Training #3: Diversity is Important



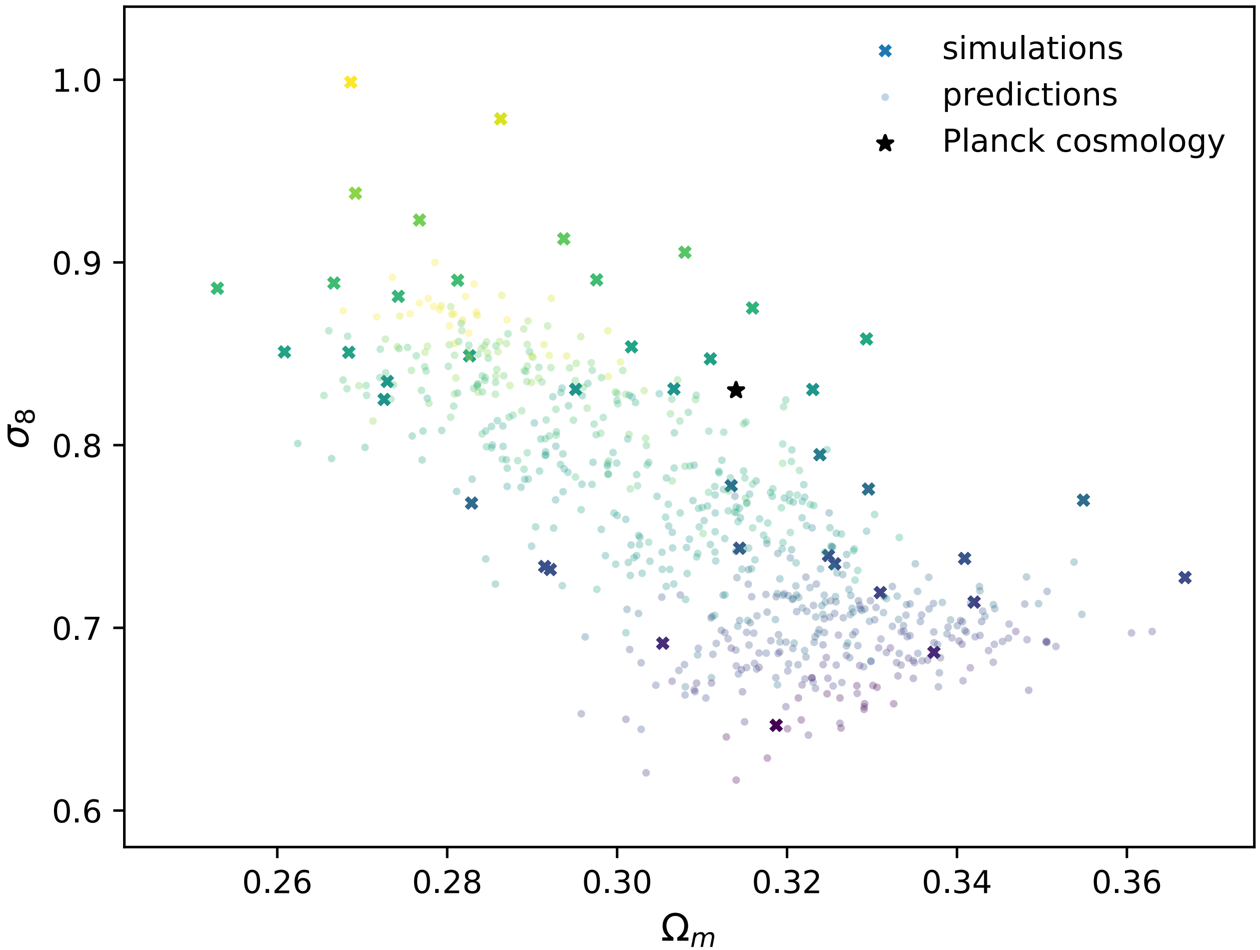
Lessons From Training #3: Diversity is Important



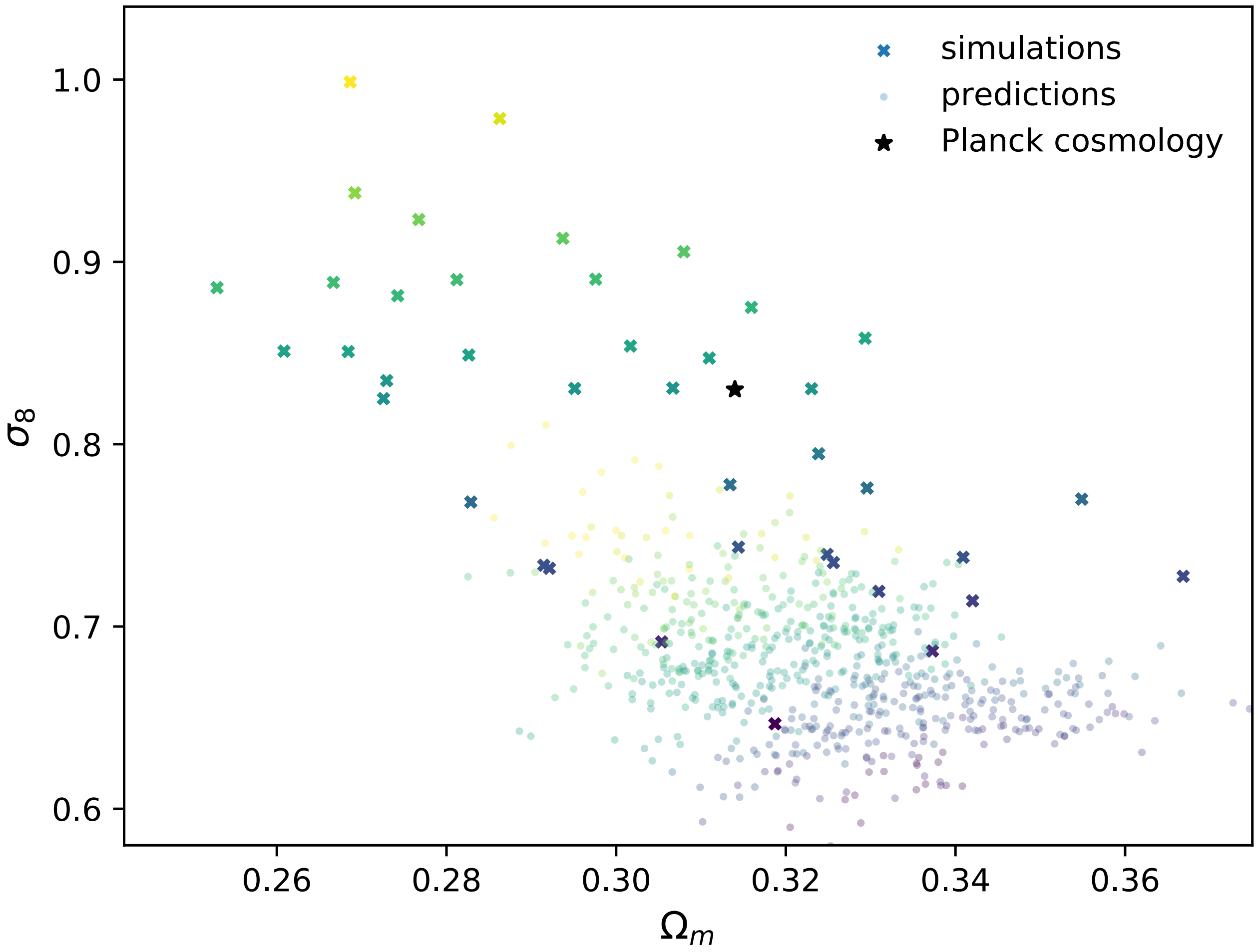
Lessons From Training #3: Diversity is Important



Lessons From Training #3: Diversity is Important



Lessons From Training #3: Diversity is Important



Using 3D CNNs to Constrain Cosmological Parameters

- Can put ~3% constraints on σ_8 and ~4% constraints on Ω_m (compare to Planck constraints: ~1% and ~2%, respectively)
- Small volume: 0.07 h^{-3} Gpc^3
  (the SDSS DR11 BOSS observation is ~60x larger!)
- The hCNN extracted useful patterns *in spite of complicating factors* such as small observation volume, varying cosmological parameters, and uncertainties in galaxy formation models
- Ntampaka, Eisenstein, Yuan, & Garrison 1909.10527

- **Deep neural network interpretability: Ntampaka+ 2019 1810.07703**

michelle.ntampaka@cfa.harvard.edu