Estimating Cosmological Constraints from Galaxy Cluster Abundance using Simulation-Based Inference

Moonzarin Reza [*1,2] **Yuanyuan Zhang** [*1,2] **Brian Nord** [3,4,5] **Jason Poh** [3,4] **Aleksandra Ciprijanovic** [5] **Louis Strigari** [1,2]

Abstract

Inferring the values and uncertainties of cosmological parameters in a cosmology model is of paramount importance for modern cosmic observations. In this paper, we use the simulation-based inference (SBI) approach to estimate cosmological constraints from a simplified galaxy cluster observation analysis. Using data generated from the Quijote simulation suite and analytical models, we train a machine learning algorithm to learn the probability function between cosmological parameters and the possible galaxy cluster observables. The posterior distribution of the cosmological parameters at a given observation is then obtained by sampling the predictions from the trained algorithm. Our results show that the SBI method can successfully recover the truth values of the cosmological parameters within the 2σ limit for this simplified galaxy cluster analysis, and acquires similar posterior constraints obtained with a likelihood-based Markov Chain Monte Carlo method, the current state-of the-art method used in similar cosmological studies.

1. Introduction

Studying cosmic structure formation and matter clustering statistics (e.g., Frieman et al., 2008; Weinberg et al., 2013; Abbott et al., 2022) in the late Universe (after redshift 2.0) has provided a rich ground for constraining cosmology models. These studies include analyses of galaxy clusters (see reviews in Allen et al., 2011; Kravtsov & Borgani, 2012), the largest gravitationally-bound structures in the Universe. The abundances and weak lensing mass measurements of galaxy clusters have been used to provide competitive constraints for ΛCDM models in ongoing cosmic surveys like the Dark Energy Survey (DES) (Abbott et al., 2020; To et al., 2021) , and is also projected to be powerful for studying wCDM models in furture experiment like the Legacy Survey of Space and Time (LSST) (The LSST Dark Energy Science Collaboration et al., 2018) at the Vera C. Rubin Observatory.

Moving forward, one potential challenge with those cosmic structure studies is about efficiently acquiring Bayesian posterior constraints of cosmological parameters, with the increasingly complicated cosmological and astrophysical models, and the larger parameter space generated by those models. To date, many of the cosmic structure formation studies rely on a Markov Chain Monte Carlo (MCMC) method to sample parameter posterior distributions, which further requires calculating a likelihood at running time based on theoretical models for the observables, and takes an increasingly long computing time which is no longer convenient (Lemos et al., 2022).

A potential improvement to those analyses is to adopt Machine Learning through a so-called "Simulation-Based Inference" (see a review in Cranmer et al., 2020), sometimes also known as "likelihood-free" approach (e.g., see Tam et al., 2022, for a closely related application). In this approach, we may precompute mock observables based on the theoretical models, known as "simulations", and then use those "simulations" to train a machine learning based inferer to map out the probabilistic function (or the "likelihood" function in alternative set-ups) between the model parameters and their possible observables. This probabilistic function can then be used to derive the posterior probability distribution of the model parameters at a given observable.

In this note, we demonstrate the potential of this SBI approach by applying it to a simplified galaxy cluster abundance analysis described in Section 2. In Section 3, we describe the SBI method, the Quijote simultaions and the analytical models. We present our results and conclusions

[*]Equal contribution [1] Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA [2] Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA [3]Department of Astronomy and Astrophyiscs, University of Chicago, Chicago, IL 60637, USA [4]Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA [5]Fermi National Accelerator Laboratory, Batavia, IL 60510, USA. Correspondence to: Moonzarin Reza <moonzarin@tamu.edu>.

ICML 2022 Workshop on Machine Learning for Astrophysics, Baltimore, Maryland, USA, 2022.

in Sections 4 and 5 respectively.

2. The Galaxy Cluster Analysis

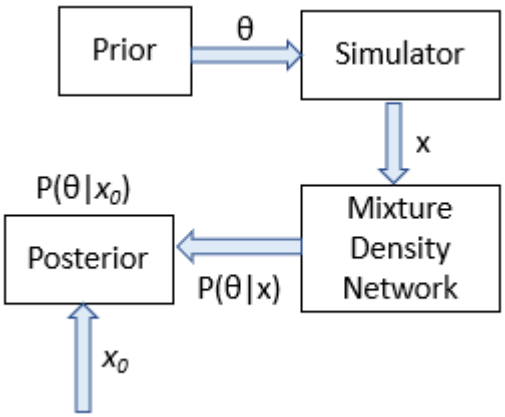

Figure 1. Block diagram illustrating the simulation-based inference method applied in this analysis.

Galaxy clusters correspond to the most massive (typical mass larger than $10^{14}\mathrm{M}_{\odot}/h$), gravitationally-bound structures in the universe, also known as dark matter halos in cosmic simulations. The abundance and mass distribution of massive dark matter halos are sensitive to cosmology, as indicated by the mass function of dark matter halos (e.g., Press & Schechter, 1974; Tinker et al., 2008). Observationally, galaxy cluster cosmology studies often rely on the number counts of galaxy clusters and their average masses in an observational range as the observational data vectors, which can be considered as summary statistics of their observations.

To resemble a typical galaxy cluster cosmology analysis, in this note, we analyze the dark matter halo counts and their masses in the Quijote simulations (Villaescusa-Navarro et al., 2020) with a fiducial Planck cosmology model. These simulations use a box volume of $(1Gpc/h)^3$, and follow the evolution of 256^3 or more dark matter particles. Our cosmological observables consist of the mean masses and the number counts of galaxy cluster-sized dark matter halos in four mass bins $[10^{14.0}, 10^{14.2}]$, $[10^{14.2}, 10^{14.4}]$, $[10^{14.4}, 10^{14.6}]$ and $[10^{14.6}, +\infty)\mathrm{M}_{\odot}/h$ at two different values of redshifts (0 and 0.5).

We use those observables to constrain five cosmological parameters: matter density (Ω_m), baryonic density (Ω_b), Hubble's constant (h), power law index of density perturbation (n_s), and the amplitude fluctuation of matter power spectrum (σ_8). The fiducial cosmology parameters are set at Ω_m = 0.3175, Ω_b = 0.049, h = 0.6711, n_s = 0.9624 and σ_8 = 0.834 (Hahn et al., 2020). We use the dark matter halos in this fixed-cosmology simulation suite as our observational test sample.

3. SBI method, Quijote simulations and analytical models

3.1. Simulation-based inference (SBI) method

Simulation-based inference (SBI) (Cranmer et al., 2020) methods can incorporate complex physical processes and observational effects in forward simulations (Tam et al., 2022). SBI is a machine-learning approach to learn the sampling distribution of data as a function of the model parameters. The main goal of simulation-based inference is to identify parameter sets which are both compatible with prior knowledge and match empirical observations. The outputs of SBI are not point estimates; rather all probabilistic values of parameter space consistent with the inputs are identified (Lueckmann et al., 2021). We use `SBI` (Tejero-Cantero et al., 2020), a neural network-based PyTorch package (Ketkar & Moolayil, 2021). The process is explained by the block diagram shown in Figure 1. The prior is first sampled to obtain an initial set of parameters, which are used to create synthetic data using forward simulations. Data generated by forward simulations are then passed into a gaussian mixture density network. The output of this network is the probability distribution of the cosmological variables as a function of the observables (number counts and mean masses of dark matter halos). The posterior (probability distribution function) is then sampled to generate the parameter space of the cosmological variables for specific observables. One of the major advantages of SBI over traditional likelihood-based methods is that it can be implemented in stages – simulations, training, inference. It provides us with an opportunity to more efficiently perform complicated analysis or analysis of large volume of data.

3.2. Quijote latin-hypercube simulations

The Quijote simulation suite contains subsets of simulations which are run for different sets of cosmological parameters (Villaescusa-Navarro et al., 2020). One such example is the latin-hypercube simulation set which contains 2000 simulations and their cosmological parameters are varied from the standard fiducial cosmology model with Ω_m in the range of $[0.1 - 0.5]$, $\Omega_b = [0.03 - 0.07]$, $h = [0.5 - 0.9]$, $n_s = [0.8 - 1.2]$ and $\sigma_8 = [0.6 - 1.0]$. We use the same summary statistics (described in Section 2) for the galaxy cluster obervables from these simulations to train the machine learning model.

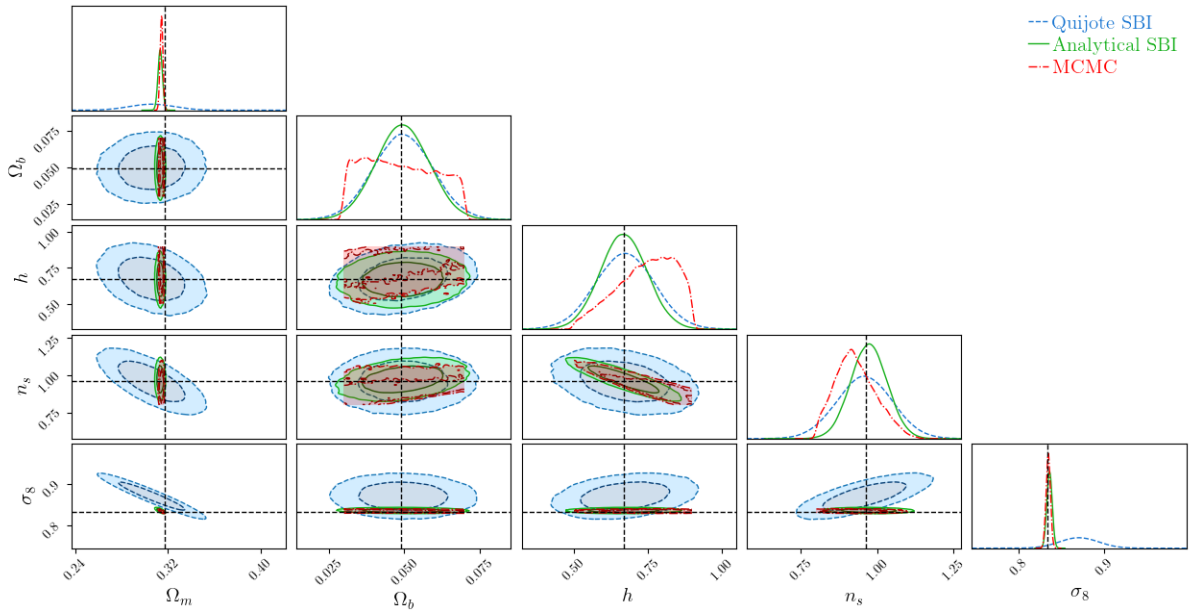

Figure 2. Posterior distributions of the cosmological parameters for analytical models (green solid), Quijote simulations (blue dashed) and MCMC (red dashdot)

3.3. Analytical models

The galaxy cluster observables can also be calculated using analytical formulas as:

$$
\begin{aligned}
N(\Delta_M|z,\Theta) &= V(z)\int_{\Delta_M} n(M|z,\Theta) + \delta_N,\\
NM(\Delta_M,z|\Theta) &= V(z)\int_{\Delta_M} M\times n(M|z,\Theta),\\
M(\Delta_M,z|\Theta) &= NM(\Delta_M|z,\Theta)/N(\Delta_M|z,\Theta) + \delta_M.
\end{aligned}
\tag{1}
$$

In these equations, $N(\Delta_M|z,\Theta)$ and $M(\Delta_M,z|\Theta)$ are the mock observables calculated from analytical models. $V(z)$ is the cosmic volume of the Quijote simulations that the problem is based upon. $n(M,z|\Theta)$ is the theoretical halo mass function that describes the number density of dark matter halos at redshift z, which has an analytical form depending on cosmology, indicated by Θ. In this analysis Θ refers to the five cosmological parameters, Ω_m, Ω_b, h, n_s and σ_8. We make use of the Bhattacharya et al. (2011) halo mass function for the FOF halo mass definitions implemented in Colossus, with a linear correction (Costanzi et al., 2019) to account for differences with the Quijote simulations. Furthermore, δ_N and δ_M represent noises added to the analytical models, caused by cosmic variance. We use the Quijote fixed cosmology simulations to estimate the cosmic variances for $N(\Delta_M|z,\Theta)$ and $M(\Delta_M|z,\Theta)$, and then randomly draw a gaussian uncertainty according to those variances, δ_N and δ_M, for each set of mock observables. In the end, we generate those analytical model simulations for over 10,000 sets of cosmological parameters. Specifically, we use a total of 11378 simulations to train the model. It is to be noted that the fast analytical simulations give valid insights and generating these simulations is feasible for this particular study.

Later in this analysis, the cosmic variances used in this fast analytical methods are used as the covariance matrices in a multi-dimensional Gaussian likelihood implemented in a Markov Chain Monte Carlo method as a comparison analysis.

4. Results

4.1. SBI method

We derive the posterior distributions of the five cosmological parameters with SBI method using the Quijote simulations (blue dashed lines), and the analytical models (green solid lines) as the training samples. The black dashed lines represent the truth values of those parameters. For comparison, we show the posterior results from Markov Chain Monte Carlo (MCMC) process, a state-of-the-art default method for sampling the posterior parameter distributions in such an analysis. In this comparison, the MCMC results are shown

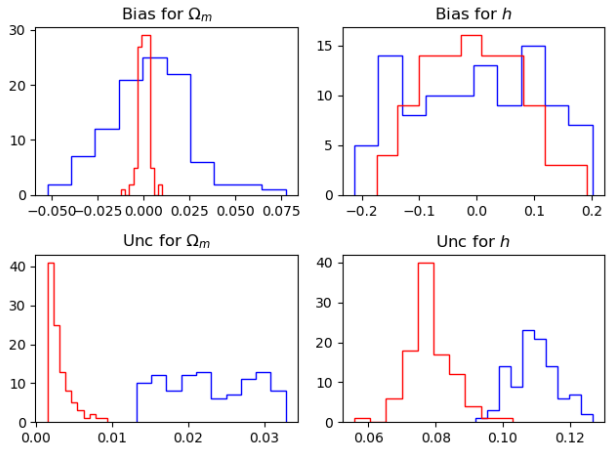

Figure 3. Distribution of bias (upper panel) and uncertainty (lower panel) for 100 test samples for analytical models (red) and Quijote simulations (blue)

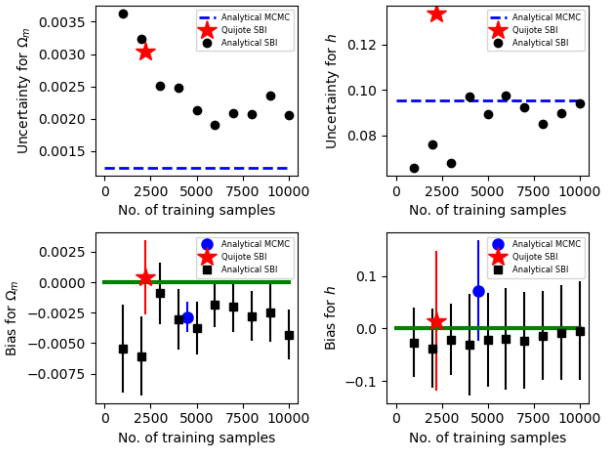

Figure 4. Uncertainty (upper panel) and bias (lower panel) for different number of training samples for analytical models, Quijote simulations and MCMC

by red dashdot lines.

As shown in Figure 2 and listed in Table 1, the truth values of the parameters are reasonably recovered by the MCMC method, and the SBI method applied to both the Quijote simulations and the analytical-model based simulations. For Ω_m, the truth value is recovered at 0.1σ, 1.6σ and 2.2σ level by the Quijote simulations, analytical models and MCMC method respectively. For the other four parameters, the truth values are recovered within the 1σ range by all three methods. For h and n_s, the SBI bias (both Quijote and analytical simulations) is much smaller than the corresponding MCMC bias.

In cosmological analysis, it is also important to accurately estimate the posterior uncertainties of the cosmological parameters, so as to correctly evaluate consistency and tensions between different cosmological models. While the MCMC method (state-of-the-art model) and the analytical-model based SBI method yield similar levels of uncertainties, the Quijote simulations based SBI method results in larger uncertainties than the other two methods for all parameters except Ω_b. The histograms in Figure 3 show the bias and uncertainty distributions for 100 test samples for the analytical simulations (red) and the Quijote simulations (blue) for Ω_m and h. Both methods result in symmetric bias distributions. For the analytical simulations, the bias tends to be distributed over a smaller range of values, and uncertainties are smaller (same result as Figure 2) than the Quijote simulations. We suspect that this might be due to the training sample size being too small in the latin-hypercube (Quijote) based results, and explore further in the next subsection.

Table 1. Estimates of cosmological parameters obtained using SBI and MCMC along with the truth values

Parameters	Truth	Analytical	Quijote	MCMC
Ω_m	0.317	$0.314^{+0.002}_{-0.002}$	$0.317^{+0.003}_{-0.003}$	$0.315^{+0.001}_{-0.001}$
Ω_b	0.049	$0.051^{+0.009}_{-0.009}$	$0.050^{+0.011}_{-0.011}$	$0.049^{+0.011}_{-0.011}$
h	0.671	$0.672^{+0.091}_{-0.091}$	$0.679^{+0.119}_{-0.119}$	$0.744^{+0.095}_{-0.095}$
n_s	0.962	$0.982^{+0.067}_{-0.067}$	$0.966^{+0.087}_{-0.087}$	$0.930^{+0.066}_{-0.066}$
σ_8	0.834	$0.836^{+0.003}_{-0.003}$	$0.831^{+0.004}_{-0.004}$	$0.835^{+0.003}_{-0.003}$

4.2. Bias and uncertainty variation with training sample size

We further evaluate how the bias and uncertainty level of the posterior constraints vary with different training sample sizes used in analytical-model based SBI method. We define bias and uncertainty (σ) as:

$$bias = \bar{\theta}_{Posterior} - \theta_{truth}$$

$$\sigma = \sqrt{\frac{\sum_{n=1}^{N}(\theta_{Posterior,i} - \bar{\theta}_{Posterior})^2}{N}}$$

In Figure 4, we plot the uncertainty (black circles) and bias (black squares) for different sizes of training sets using data generated from the analytical models for Ω_m and h . We start with 1000 samples and increase the training sample size in steps of 1000 until we reach 10000 samples. Uncertainty and bias corresponding to the MCMC (blue dashed lines and blue circles) and latin-hypercube simulations (red stars) are also plotted on the same axes. The uncertainty plots (upper panel) show that apart from some random fluc-

tuations, uncertainty reduces with an increase in the number of training samples for Ω_m and stablizes when the sample size reaches $\sim$5000, for which the MCMC bias is about 33% lower than the analytical-simulations SBI bias. A similar trend is also observed for σ_8 (not shown). There is no observed correlation between the training sample size and the uncertainty of h, Ω_b (not shown), and n_s(not shown), indicating that the contraints are not affected by the training sample size we tested here. The convergent value of analytical SBI uncertainty for these parameters is comparable to that of the MCMC method.

According to the results shown in the lower panel of Figure 4, on average, the SBI bias based on analytical simulations diminishes as the size of training sample increases for Ω_m. For h, the bias is nearly constant and independent of training sample size. Like the uncertainty, the bias for σ_8 shows a trend similar to Ω_m, and the bias for Ω_b and n_s follow the trend similar to h. MCMC bias is comparable to the SBI analytical bias for Ω_m, and is higher than the analytical bias for h.

5. Conclusions

We have applied a simulation-based inference method to a galaxy cluster cosmological analysis, using data generated from the Quijote latin-hypercube simulations and analytical models. Our results show that SBI method can recover the truth cosmological parameters for this galaxy cluster analysis within the 2σ limit. On average SBI method results in smaller bias than MCMC. We have also evaluated the dependence of bias and uncertainty on the training sample size for the analytical simulations and conclude that the uncertainty converges for a sample size $\sim$5000. The success of this attempt demonstrates that SBI is a promising method to be employed in future galaxy cluster cosmological analyses to shed light on the long-standing cosmological mysteries.

References

Abbott, T. M. C., Aguena, M., Alarcon, A., Allam, S., and et al., D. Dark Energy Survey Year 1 Results: Cosmological constraints from cluster abundances and weak lensing. , 102(2):023509, July 2020. doi: 10.1103/PhysRevD.102.023509.

Abbott, T. M. C., Aguena, M., Alarcon, A., Allam, S., and et al., D. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. , 105(2):023520, January 2022. doi: 10.1103/PhysRevD.105.023520.

Allen, S. W., Evrard, A. E., and Mantz, A. B. Cosmological Parameters from Observations of Galaxy Clusters. , 49(1):409–470, September 2011. doi: 10.1146/annurev-astro-081710-102514.

Bhattacharya, S., Heitmann, K., White, M., Lukić, Z., Wagner, C., and Habib, S. Mass Function Predictions Beyond ΛCDM. , 732(2):122, May 2011. doi: 10.1088/0004-637X/732/2/122.

Costanzi, M., Rozo, E., Simet, M., Zhang, Y., and et al. Methods for cluster cosmology and application to the SDSS in preparation for DES Year 1 release. , 488(4):4779–4800, October 2019. doi: 10.1093/mnras/stz1949.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Frieman, J. A., Turner, M. S., and Huterer, D. Dark energy and the accelerating universe. , 46:385–432, September 2008. doi: 10.1146/annurev.astro.46.060407.145243.

Hahn, C., Villaescusa-Navarro, F., Castorina, E., and Scoccimarro, R. Constraining M_ν with the bispectrum. Part I. Breaking parameter degeneracies. , 2020(3):040, March 2020. doi: 10.1088/1475-7516/2020/03/040.

Ketkar, N. and Moolayil, J. Introduction to pytorch. In *Deep learning with python*, pp. 27–91. Springer, 2021.

Kravtsov, A. V. and Borgani, S. Formation of Galaxy Clusters. , 50:353–409, September 2012. doi: 10.1146/annurev-astro-081811-125502.

Lemos, P., Weaverdyck, N., Rollins, R. P., Muir, J., and et al. Robust sampling for weak lensing and clustering analyses with the Dark Energy Survey. *arXiv e-prints*, art. arXiv:2202.08233, February 2022.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 343–351. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/lueckmann21a.html.

Press, W. H. and Schechter, P. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. , 187:425–438, February 1974. doi: 10.1086/152650.

Tam, S.-I., Umetsu, K., and Amara, A. Likelihood-free Forward Modeling for Cluster Weak Lensing and Cosmology. , 925(2):145, February 2022. doi: 10.3847/1538-4357/ac3d33.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020.

The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., Hložek, R., Collett, T., Gawiser, E., Scolnic, D., Alonso, D., Awan, H., Biswas, R., Blazek, J., Burchat, P., Chisari, N. E., Dell'Antonio, I., Digel, S., Frieman, J., Goldstein, D. A., Hook, I., Ivezić, Ž., Kahn, S. M., Kamath, S., Kirkby, D., Kitching, T., Krause, E., Leget, P.-F., Marshall, P. J., Meyers, J., Miyatake, H., Newman, J. A., Nichol, R., Rykoff, E., Sanchez, F. J., Slosar, A., Sullivan, M., and Troxel, M. A. The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document. *arXiv e-prints*, art. arXiv:1809.01669, September 2018.

Tinker, J., Kravtsov, A. V., Klypin, A., Abazajian, K., Warren, M., Yepes, G., Gottlöber, S., and Holz, D. E. Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality. , 688(2):709–728, December 2008. doi: 10.1086/591439.

To, C., Krause, E., Rozo, E., Wu, H., and et al., D. Dark Energy Survey Year 1 Results: Cosmological Constraints from Cluster Abundances, Weak Lensing, and Galaxy Correlations. , 126(14):141301, April 2021. doi: 10.1103/PhysRevLett.126.141301.

Villaescusa-Navarro, F., Hahn, C., Massara, E., Banerjee, A., Delgado, A. M., Ramanah, D. K., Charnock, T., Giusarma, E., Li, Y., Allys, E., et al. The quijote simulations. *The Astrophysical Journal Supplement Series*, 250 (1):2, 2020.

Weinberg, D. H., Mortonson, M. J., Eisenstein, D. J., Hirata, C., Riess, A. G., and Rozo, E. Observational probes of cosmic acceleration. , 530(2):87–255, September 2013. doi: 10.1016/j.physrep.2013.05.001.