Draft version August 22, 2024
Typeset using LaTeX twocolumn style in AASTeX631

Inferring Cosmological Parameters on SDSS via Domain-Generalized Neural Networks and Lightcone Simulations

Jun-Young Lee [1,2] Ji-hoon Kim [1,2,3] Minyong Jung [2] Boon Kiat Oh [4] Yongseok Jo [5,6] Songyoun Park,[2]
Jaehyun Lee [7] Yuan-Sen Ting [8,9,10,11] and Ho Seong Hwang [3,12]

[1] Institute for Data Innovation in Science, Seoul National University, Seoul 08826, Korea
[2] Center for Theoretical Physics, Department of Physics and Astronomy, Seoul National University, Seoul 08826, Korea
[3] Seoul National University Astronomy Research Center, Seoul 08826, Korea
[4] Department of Physics, University of Connecticut, Storrs, CT 06269, USA
[5] Columbia Astrophysics Laboratory, Columbia University, New York, NY 10027, USA
[6] Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA
[7] Korea Astronomy and Space Science Institute, Daejeon 34055, Republic of Korea
[8] Research School of Astronomy & Astrophysics, Australian National University, Canberra, ACT 2611, Australia
[9] School of Computing, Australian National University, Acton ACT 2601, Australia
[10] Department of Astronomy, The Ohio State University, Columbus, OH 43210, USA
[11] Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH 43210, USA
[12] Astronomy Program, Department of Physics and Astronomy, Seoul National University, Seoul 08826, Korea

(Received July XX, 2024; Revised July XX, 2024; Accepted August XX, 2024)

ABSTRACT

We present a proof-of-concept simulation-based inference on Ω_m and σ_8 from the SDSS BOSS LOWZ NGC catalog using neural networks and domain generalization techniques without the need of summary statistics. Using rapid lightcone simulations, L-PICOLA, mock galaxy catalogs are produced that fully incorporate the observational effects. The collection of galaxies is fed as input to a point cloud-based network, `Minkowski-PointNet`. We also add relatively more accurate Gadget mocks to obtain robust and generalizable neural networks. By explicitly learning the representations which reduces the discrepancies between the two different datasets via the semantic alignment loss term, we show that the latent space configuration aligns into a single plane in which the two cosmological parameters form clear axes. Consequently, during inference, the SDSS BOSS LOWZ NGC catalog maps onto the plane, demonstrating effective generalization and improving prediction accuracy compared to non-generalized models. Results from the ensemble of 25 independently trained machines find Ω_m=0.339±0.056 and σ_8=0.801±0.061, inferred only from the distribution of galaxies in the lightcone slices without relying on any indirect summary statistics. A single machine that best adapts to the Gadget mocks yields a tighter prediction of Ω_m=0.282±0.014 and σ_8=0.786±0.036. We emphasize that adaptation across multiple domains can enhance the robustness of the neural networks in observational data.

Keywords: *N*-body simulations (1083), Cosmological parameters from large-scale structure (340), Redshift surveys (1378), Neural networks (1933)

1. INTRODUCTION

Following its success in explaining the clustering of matter over a wide range of scales, the ΛCDM model has now ushered in the era of precision cosmology. The small perturbations imprinted in the cosmic microwave background grow as cold dark matter falls into and deepens potential wells. Small structures gravitationally evolve to create the characteristic cosmic webs and voids referred to as the large scale structures (LSS; Peebles 1981; Davis et al. 1985; Bond et al. 1996), which are observable in galaxy surveys (de Lapparent et al. 1986; Geller & Huchra 1989). The LSS serves as a widely used probe for constraining the cosmological parameters constituting the ΛCDM model, as it maps the distribution and motion of matter throughout the universe over time. Over the past few decades, a series of galaxy redshift surveys

Corresponding author: Ji-hoon Kim
mornkr@snu.ac.kr

have been conducted extensively to trace the distribution of galaxies and the growth history of LSS across a large spatial extent and depth (Huchra et al. 1983; York et al. 2000; Colless et al. 2001; Sohn et al. 2023).

Considering the galaxy distribution as a (biased) proxy for the total matter content of the universe, power spectrum multipoles and n-point correlation functions (n-pCF) can be derived to express matter clustering at different scales. These summary statistics serve as essential components in the development of mock catalogs and in the inference of cosmological parameters. The construction of survey-specific mocks, which mimic similar summary statistics and the geometry of the survey, imposes constraints on certain cosmological parameters (Kitaura et al. 2016; White et al. 2014a; Saito et al. 2016). Through high-resolution simulations in large volumes and by assigning adequate band magnitudes and spectroscopic information, generic catalogs applicable to various observational surveys can also be generated (Fosalba et al. 2015a; Crocce et al. 2015; Fosalba et al. 2015b; Dong-Páez et al. 2022). Other than producing the mocks that best match the observational catalog, derived summary statistics from realizations simulated with varying cosmology can be compared with the observational counterpart to make inference on the cosmological parameters, an approach referred to as simulation-based inference (Villaescusa-Navarro et al. 2020; Hahn & Villaescusa-Navarro 2021). While these cited works rely on predefined summary statistics, the simulation-based inference framework allows for the potential use of raw inputs together with the neural networks' flexible featurizations, which permits the exploration beyond summary statistics.

With the advent of artificial intelligence and machine learning, simulation-based inference of cosmological parameters has been accelerated. This involves inferring cosmological parameters from simulations by matching summary statistics or features, with neural networks serving as an option alongside more traditional measures of statistical inference such as Markov chain Monte Carlo (MCMC; Alsing et al. 2019; Jeffrey & Wandelt 2020). In particular, classic summary statistics such as the n-pCF and power spectra, which convey limited information about the matter distribution of the universe, can be replaced with features extracted by neural networks that capture much more complex information engraved inside (Shao et al. 2023). Attributed to this capability of extracting rich information not hinted at in the summary statistics, simulation-based inference with neural networks has shown the possibility of producing tight predictions on the cosmological parameters (Lemos et al. 2023). Therefore, the importance of simulation-based inference is being recognized as it can serve as an alternative for verifying and possibly resolving tensions in the cosmological parameters predicted from CMB observations and galaxy surveys, especially concerning H_0 and $S_8 \equiv \sigma_8 \sqrt{\Omega_\mathrm{m}/0.3}$ (Anchordoqui et al. 2021).

In this context, AI-driven projects have been launched to perform diverse tasks, including parameter estimation (Villaescusa-Navarro et al. 2022; Ni et al. 2023; Kreisch et al. 2022). Especially in the estimation of cosmological parameters, 21-cm tomography light cones (Neutsch et al. 2022), weak lensing (WL) convergence and shear maps (Fluri et al. 2018; Fluri et al. 2019; Fluri et al. 2022; Kacprzak & Fluri 2022; Lu et al. 2023), dark matter density fields (Pan et al. 2020; Lazanu 2021; Giri et al. 2023; Hortúa et al. 2023), and halo catalogs (Ravanbakhsh et al. 2016; Mathuriya et al. 2019; Ntampaka et al. 2020; Hwang et al. 2023; Shao et al. 2023) were utilized as inputs for various neural network architectures, typically in a traditional supervised learning setup. In contrast to the direct input of mocks, derived summary statistics such as the n-pCF, count-in-cell, void probability function, star formation rate density, and stellar mass functions (SMF) were also used as inputs (Boruah et al. 2023; Veronesi et al. 2023; Hahn et al. 2023a; Perez et al. 2022; Jo et al. 2023). In addition, individual galaxy properties (Villaescusa-Navarro et al. 2022), galaxy cluster properties (Qiu et al. 2023), or snapshots of galaxy catalogs (de Santi et al. 2023) were shown to be useful as inputs for neural networks.

Among the listed works, most tested their pipeline on simulated data sets, and only a few successfully generalized their neural networks to the actual observational data. Hahn et al. (2023a) and Hahn et al. (2023b) created a mask autoregressive flow using the power spectrum and bispectrum as summary statistics to provide constraints on cosmological parameters, based on the SDSS BOSS CMASS catalog (Reid et al. 2016). In contrast, Veronesi et al. (2023) leveraged 2-pCF from lognormal mocks as input to fully connected layers (FCL). Jo et al. (2023) used FCL emulators to perform implicit likelihood inference on observed SMF (Leja et al. 2020) and SFRD (Leja et al. 2022). Parameter inferences using WL convergence maps as probes, including the Kilo Degree Survey (Hildebrandt et al. 2017; Asgari et al. 2021) and Subaru Hyper Suprime-Cam first-year surveys (HSC-Y1; Hikage et al. 2019) were also performed with Convolutional Neural Networks (CNNs) or Graph CNNs (GCNNs; Fluri et al. 2019; Fluri et al. 2022; Lu et al. 2023). Notably, recent studies regard neural networks' outputs of predicted parameters as summary statistics due to their centrally biased nature (Gupta et al. 2018; Ribli et al. 2019; Fluri et al. 2019; Lemos et al. 2023), and perform additional Bayesian inferences.

In line with efforts to use deep learning for constraining cosmological parameters, this paper aims to perform a proof-of-concept test of conducting cosmological inference using the galaxy redshift survey, *without* relying on any indirect

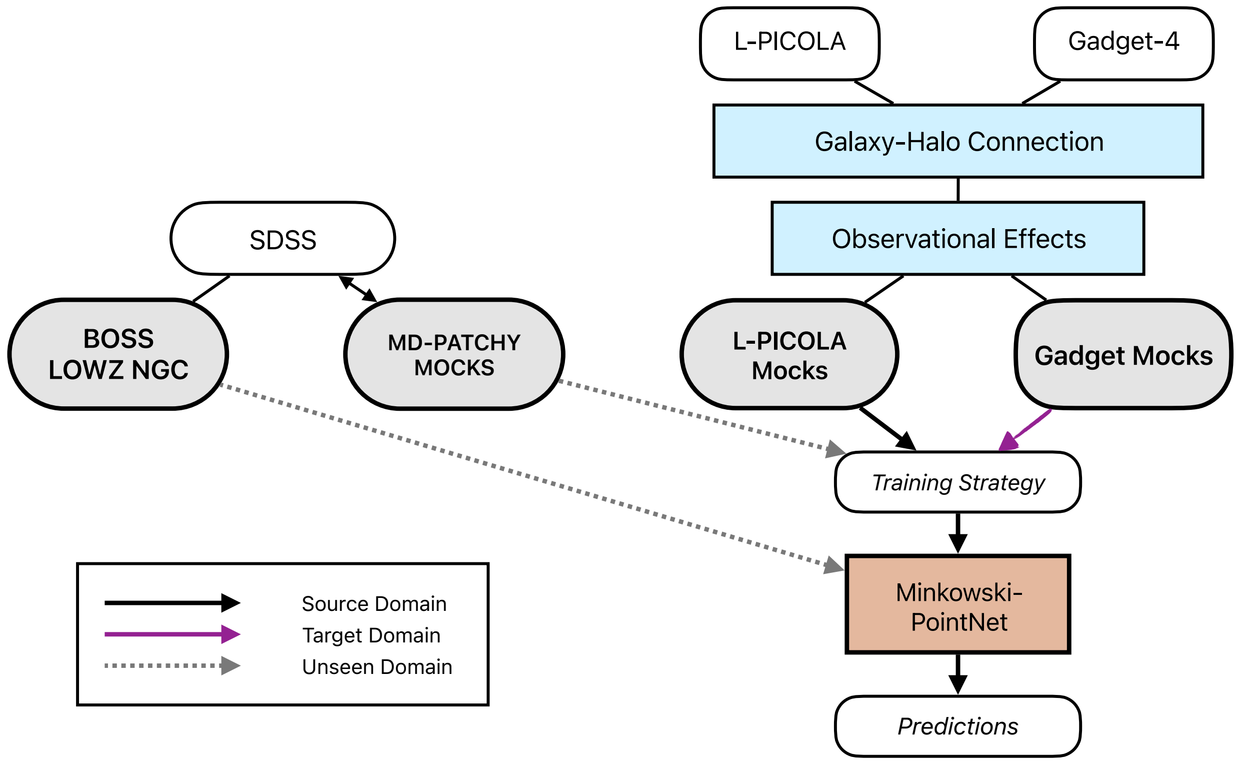

Figure 1. Diagram that exhibits the overall structure of this study with simulation and deep learning pipelines. We aim to infer cosmological parameters from the observation-driven catalog, SDSS BOSS LOWZ NGC. We produce two lightcone mock suites, L-PICOLA and GADGET mocks, combining *N*-body simulations (sections 2.2 and 2.3) and galaxy-halo connection models (Section 2.5) while fully accounting for observational effects (Section 2.6). We then utilize a point-cloud-based network, `Minkowski-PointNet`, which takes individual galaxies as inputs to predict $\Omega_{\rm m}$ and σ_8, and their errors (Section 3). The L-PICOLA mocks (source domain) are trained together with the GADGET mocks (target domain) using the training strategy for the domain adaptation and generalization techniques (see Section 4.3). In this process, we use training strategies to align the representation of each mock (Section 4). The adapted machines are then applied to unseen domains including the fine-tuned MD-PATCHY mocks and the SDSS BOSS LOWZ NGC sample. The main results including the predictions for the actual observation are shown in Section 5.

summary statistics, but rather utilizing the total raw distribution of galaxies as input to the neural network. For this test, we focus mainly on $\Omega_{\rm m}$ and σ_8, which are directly related to the $S_8 \equiv \sigma_8\sqrt{\Omega_{\rm m}/0.3}$ tension as mentioned above. As mentioned in Hahn et al. (2023a), this choice is due to the fact that $\Omega_{\rm m}$ and σ_8 are the parameters that are sensitive to the cosmological information of the clustering galaxies, while others are less constrained. In order to reduce any artificial priors arising from survey-specific observational biases, we rapidly generate a large mock suite that fully includes observational effects such as redshift space distortion, survey footprint, stellar mass incompleteness, radial selection, and fiber collision in the SDSS BOSS LOWZ Northern Galactic Cap (NGC) catalog. Then, using the position and mass information of individual and neighboring galaxies, we make inference on $\Omega_{\rm m}$ and σ_8, again *without* relying on any indirect summary statistics.

The biggest difficulty in using the whole galaxy catalog as input instead of the summary statistics is that the selection of codes begets overall differences in the resultant realizations. The differences are easily discernible and distinguishable by complex neural networks. Consequently, naively merging the different sets of mocks or domains limits the machines to merely learning fragmented domain-specific knowledge. Recent studies have tried to address such issues, as machines failing to attain robustness exhibit poor performances and lack predictability on unseen domains (Ni et al. 2023; Shao et al. 2023; Roncoli et al. 2023). Moreover, as simulated catalogs do not perfectly portray the actual universe, such discrepancies may significantly aggravate the performance of machines onto unseen observed data. Especially, the rapid generation of mocks trades off with the inaccuracies compared to the relatively time-consuming simulations, leading to a clear deviation. In order to make effective inferences on different types of simulation or domains, the neural network must achieve generalizability. This study focuses mainly on extracting and learning unified representations originating from distinct domains and exploiting generalized and integrated knowledge on the observational data.

This paper is organized as follows. In Section 2, we illustrate the creation of our mock data that thoroughly integrate the observational effects. We produce two suites of mocks using two distinct simulations, L-PICOLA and GADGET. The footprint and lightcone slices are shown together with the observational target, the SDSS BOSS LOWZ NGC catalog, and its specific set of mock catalogs, MD-PATCHY, for comparison. In Section 3, input features and the neural network architecture are introduced together with the training strategies in Section 4, to align the latent space representations of different mocks and achieve domain generalization or robustness. In Section 5, implicit likelihood estimates in Ω_m and σ_8 using the SDSS BOSS LOWZ NGC catalog are shown. We also discuss the impact of fine-tuned MD-PATCHY mocks on the predictability and generalizability of the machine. Finally, the results and the following conclusions are summarized in Section 7. The overall approach taken by the paper is schematically shown in Figure 1.

2. GALAXY CATALOG: OBSERVATION AND SIMULATION

2.1. *The Reference SDSS Catalog*

In this study, we utilize the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), part of SDSS-III (Eisenstein et al. 2011), which extends the previously studied distribution of luminous red galaxies (LRGs; Eisenstein et al. 2001) from SDSS I/II, adding fainter galaxies and thus larger number densities, for the purpose of measuring baryon acoustic oscillations. The survey consists of the LOWZ (Tojeiro et al. 2014) and CMASS (Reid et al. 2016) catalogs, which have different color and magnitude cuts. The LOWZ catalog targets galaxies at a low redshift of $z \lesssim 0.4$, while CMASS targets a higher redshift range of $0.4 \lesssim z \lesssim 0.7$. The LOWZ samples are roughly considered as volume-limited, whereas the CMASS samples, representing 'constant mass', are considered volume-limited within the mass and redshift ranges of $M_\star > 10^{11.3}\mathrm{M}_\odot$ and $z \lesssim 0.6$ (Reid et al. 2016; Maraston et al. 2013). Using the MKSAMPLE code, the LSS catalogs for both LOWZ and CMASS were created for BOSS DR12, fully equipped with survey masks and random samples. These samples include completeness and weights calculated for the analysis of large-scale structure (Reid et al. 2016).

To account for the stellar mass incompleteness of the survey and to incorporate cosmological information from the stellar masses of galaxies later on, we obtain stellar mass data from the value-added Portsmouth SED-fits catalog (Maraston et al. 2013), assuming a passive evolution model with the Kroupa IMF (Kroupa 2001). Since the Portsmouth SED-fits catalog includes both BOSS and LEGACY targets, we need to select those that are included in the LSS catalog. Following Rodríguez-Torres et al. (2016), we match galaxies using the unique combination of tags MJD, PLATEID, and FIBERID, and then assign the stellar masses from the matched galaxies in the Portsmouth catalog to the corresponding entries in the LSS catalog.

In this work, we use the Northern Galactic Cap (NGC) of the LOWZ samples with RA=150°–240° and DEC>0°. The selection of the LOWZ samples and the cropped regions is due to the limited volume of the lightcone simulations that will be used to generate mocks. Using this catalog as a benchmark, we generate mocks that incorporate the same observational effects: redshift space distortions, survey footprint geometry, stellar mass incompleteness, radial selection matching, and fiber collision (see Section 2.6 for more information).

2.2. *Rapidly Generated Lightcone Mocks,* L-PICOLA

L-PICOLA is a rapid dark matter simulation that employs the COmoving Lagrangian Acceleration method (COLA; Tassev et al. 2013) and supports on-the-fly generation of lightcones. At the expense of minute errors—2% in the power spectrum and 5% in the bispectrum—the code allows for the rapid generation of dark matter distributions in large box sizes (Howlett et al. 2015a). Numerous studies have leveraged on this computational efficiency to produce a vast amount of mock catalogs aimed for diverse observations (Howlett et al. 2015b; Howlett et al. 2022; Ishikawa et al. 2023).

In a box volume of $(1.2h^{-1}\mathrm{Gpc})^3$ we simulate the evolution of 1200^3 dark matter particles on 1200^3 meshes. Each particle has a mass of approximately $M_p \approx 8.3 \times 10^{10}\left(\frac{\Omega\mathrm{m}}{0.3}\right)h^{-1}\mathrm{M}_\odot$. The simulation starts with a 2LPT initial condition generated with 2LPTIC (Scoccimarro et al. 2012) at z_initial=9 and progresses in 10 steps to z=0.45, as Howlett et al. (2015a) suggests for sufficient precision in the resolution adopted here, with 10 lightcone slices generated from z=0.45 to z=0. A total of 1500 simulations are produced, incorporating cosmic variance across varying Ω_m and σ_8. Each of the two parameters is randomly sampled from a uniform distribution of $\Omega_\mathrm{m}\in[0.1, 0.5]$ and $\sigma_8\in[0.6, 1.0]$. We assume H_0=100h km s^{-1} Mpc^{-1} with h=0.674, n_s=0.96 following the results from Planck Collaboration et al. (2020). We select a realization from one pair of cosmological parameters most similar to the fiducial cosmology of MD-PATCHY with $\Omega_\mathrm{m} = 0.3067, \sigma_8 = 0.8238$ and name it `L-PICOLA fiducial`. We obtain the halos using the ROCKSTAR halo finder (Behroozi et al. 2013b) in lightcone mode, considering a minimum number of 10 particles as a seed halo (most detailed layer of subgroup hierarchy determined by the friends-of-friends algorithm). Thus, we impose a cut in the halo mass of $\log(M_\mathrm{h}/h^{-1}\mathrm{M}_\odot)$=11.45. Subsequently, the 1500 catalogs are rotated and reflected in six directions following Ravanbakhsh et al. (2016), generating a total of 9000 realizations referred to as L-PICOLA mocks. These mocks will be further cropped and masked separately according to the observational effects. From this we establish a one-to-one correspondence between the subhalos and galaxies.

2.3. *Adaptation: Gravitational N-body Simulation Mocks,* GADGET

The L-PICOLA mocks described in Section 2.2 lack accuracy in the clustering statistics on small scales compared to full *N*-body simulations (see Section 4.1). Therefore, we similarly generate mocks using GADGET-4 (Springel et al. 2021) in lightcone mode, which we refer to as GADGET mocks. Although they require more computational time and resources to generate than L-PICOLA mocks, GADGET mocks, are generally considered to offer higher fidelity at smaller scales (see Howlett et al. 2015a). Consequently, we use GADGET mocks as adaptation standards of the neural networks, to refine the code-specific knowledge from L-PICOLA mocks, implementing a training strategy that aligns the neural networks' extracted representations. For additional details, refer to Section 4.

The simulation resolution is the same as that of mock suites generated with L-PICOLA: a box volume of $(1.2h^{-1}\mathrm{Gpc})^3$ and 1200^3 dark matter particles with a softening length of $10h^{-1}$kpc. The simulation initiates with a 2LPT initial condition generated with N-GenIC (Springel 2015) at $z_{initial}$=10, similar to L-PICOLA, and ends at z=0.[1] The cosmological parameters of the fiducial run, `GADGET fiducial`, are set to be identical to those of MD-PATCHY mocks in Section 2.4: Ω_m=0.307115, σ_8=0.8288, and h=0.6777, with other parameters fixed to the previously stated values. Furthermore, in order to test the machine's predictability for non-fiducial mocks, we produce `GADGET low` with Ω_m=0.2, σ_8=0.7 and `GADGET high` with Ω_m=0.4, σ_8=0.8. We generate 6 samples each by rotating and reflecting the three GADGET simulations, totalling 18 samples.

2.4. *Adaptation: Fine-Tuned Mocks,* MD-PATCHY

MULTIDARK PATCHY Mocks (hereafter MD-PATCHY) are mock galaxy catalogs designed to match the SDSS-III BOSS survey (Kitaura et al. 2016; Rodríguez-Torres et al. 2016). They referenced the BIGMULTIDARK simulation (Klypin et al. 2016), a *N*-body simulation run on GADGET-2 (Springel 2005). The halos from the BIGMULTIDARK are populated using the stochastic halo abundance matching technique and the observational effects including redshift space distortion, survey footprint, stellar mass incompleteness, radial selection, and fiber collision are considered using the SUGAR code (Rodríguez-Torres et al. 2016). The reference catalog is used to calibrate PATCHY (Kitaura et al. 2013), which employs augmented Lagrangian Perturbation Theory (ALPT; Kitaura & Heß 2013) to generate dark matter fields. These fields are biased and the halo masses are identified using the HADRON code (Zhao et al. 2015), which takes the halos' environmental information into account. The halo catalog is further processed into galaxy mocks using the halo abundance matching procedure in the SUGAR code. Specifically, the clustering statistics are fitted by fine-tuning a single parameter–the scatter in the HAM procedure ($\sigma_\mathrm{HAM}(V_\mathrm{peak}|M_\star)$), where $M_\star$ represents the stellar mass and V_peak the peak velocity observed throughout the history of the halo. In total, 10240 MD-PATCHY mocks that mimic the clustering statistics, stellar mass functions, and observational effects are produced. The cosmological parameters used are Ω_m=0.307115, σ_8=0.8288, and h=0.6777. In this work, we focus on the 2048 mocks of the Northern Galactic Cap (NGC) of the LOWZ samples. Similarly to the GADGET mocks in Section 2.3, the MD-PATCHY mocks are used as reference mocks for adaptation of the neural networks during the training phase (see Section 4 for more information).

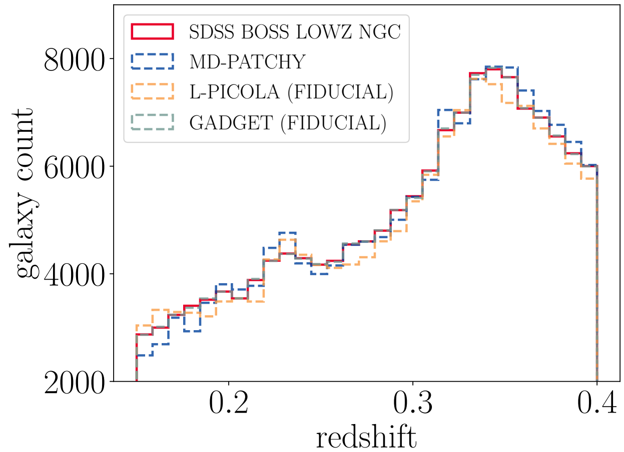

Figure 2. Count of galaxies per radial bins for SDSS BOSS LOWZ NGC (*red solid line*), `L-PICOLA fiducial` (*orange dashed line*), `GADGET fiducial` (*green dashed line*), and averaged count for all 2048 MD-PATCHY (*blue dashed line*). The radial bins from redshift 0.15 to 0.4 are defined to evenly divide the redshift space volume. All lightcone mocks with fiducial cosmological parameters exhibit a consistent number of galaxies across different radial bins compared to the SDSS BOSS LOWZ NGC catalog. See Section 2.6 for more information.

2.5. *Galaxy-Halo Connection*

[1] We acknowledge that starting a full *N*-body simulation, GADGET, at low redshifts may lead to inaccuracies, unlike L-PICOLA, despite the reduction of computational resources. The choice of the initial redshift was based on the comparative analyses presented in Howlett et al. (2015a). We leave such improvements to be addressed in our future work.

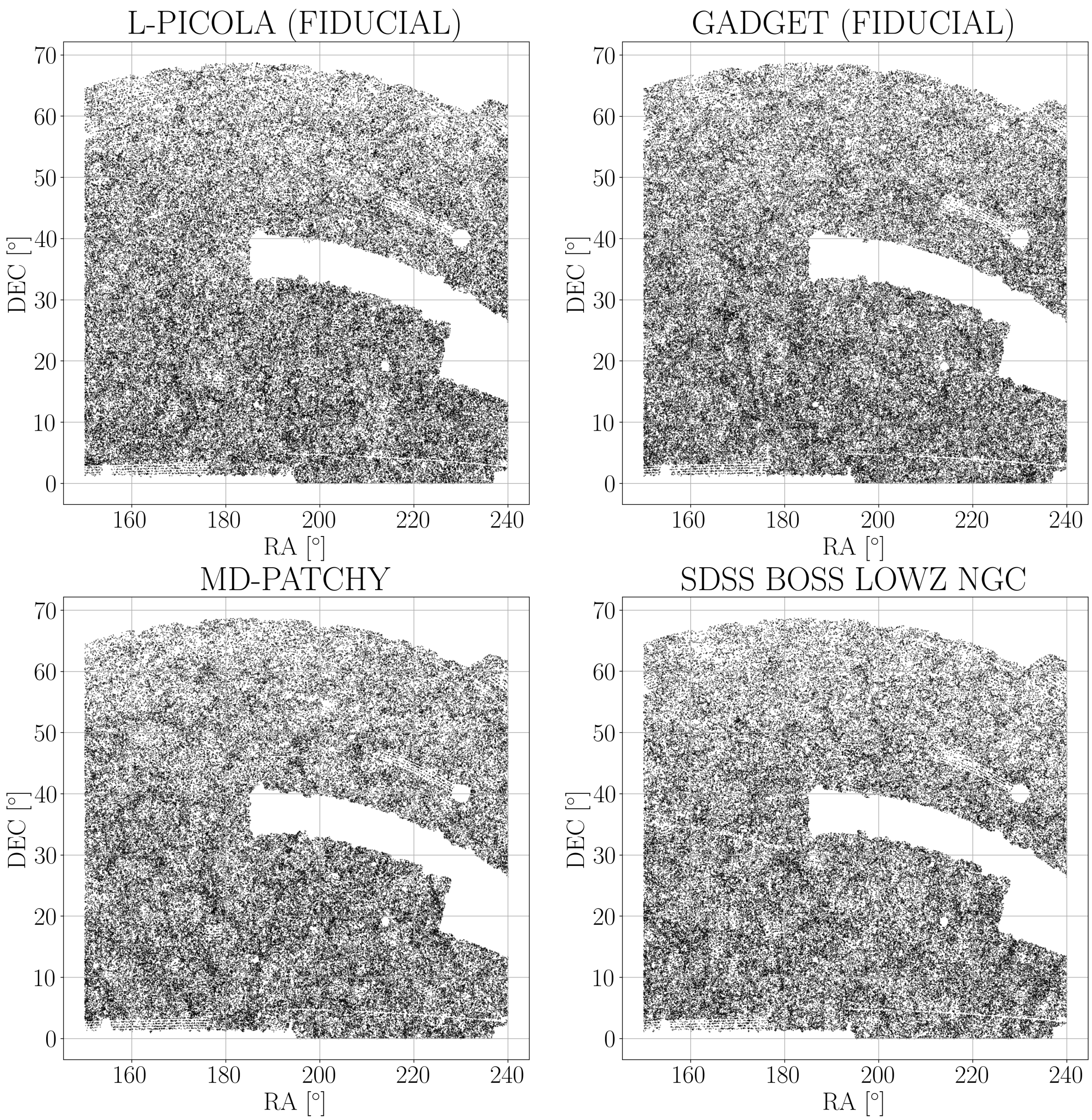

Figure 3. The footprints of a single realization from `L-PICOLA fiducial` (*top left*), `GADGET fiducial` (*top right*), MD-PATCHY (*bottom left*), and SDSS BOSS LOWZ NGC catalog (*bottom right*). The same acceptance and veto masks are employed to reproduce the overall topology. We further cut the region into RA=150°–240° and DEC>0°. See Section 2.6 for more information.

The galaxy-halo connection is a crucial statistical relation that summarizes the interplay between gravitational evolution and baryonic physics in galaxies and halos, widely studied in the fields of galaxy formation and cosmology (see Wechsler & Tinker 2018, for review). Numerous approaches in modeling are available, including the halo occupation distribution (HOD; Peacock & Smith 2000; Berlind et al. 2003), subhalo abundance matching (SHAM; Kravtsov et al. 2004; Conroy et al. 2006), and also the combined models such as subhalo clustering and abundance matching (SCAM; Guo et al. 2016; Ronconi et al. 2020). In the following, we introduce the two galaxy-halo connection methods: the fixed stellar-to-halo-mass relation and the SHAM.[2]

2.5.1. *Fixed Stellar-to-Halo-Mass Relation*

Here, we adopt the minimal model that connects N-body simulations to galaxy catalogs. Assuming a one-to-one

[2] The galaxy-halo connection models introduced here are indeed simplistic. To account for the detailed connection relation, it may be necessary to track the halo assembly history or apply varying population models by introducing few additional parameters. Here, we focus on proof-of-concept objective, rather than investigating deeply into this complex relation. Such limitations are left for future work.

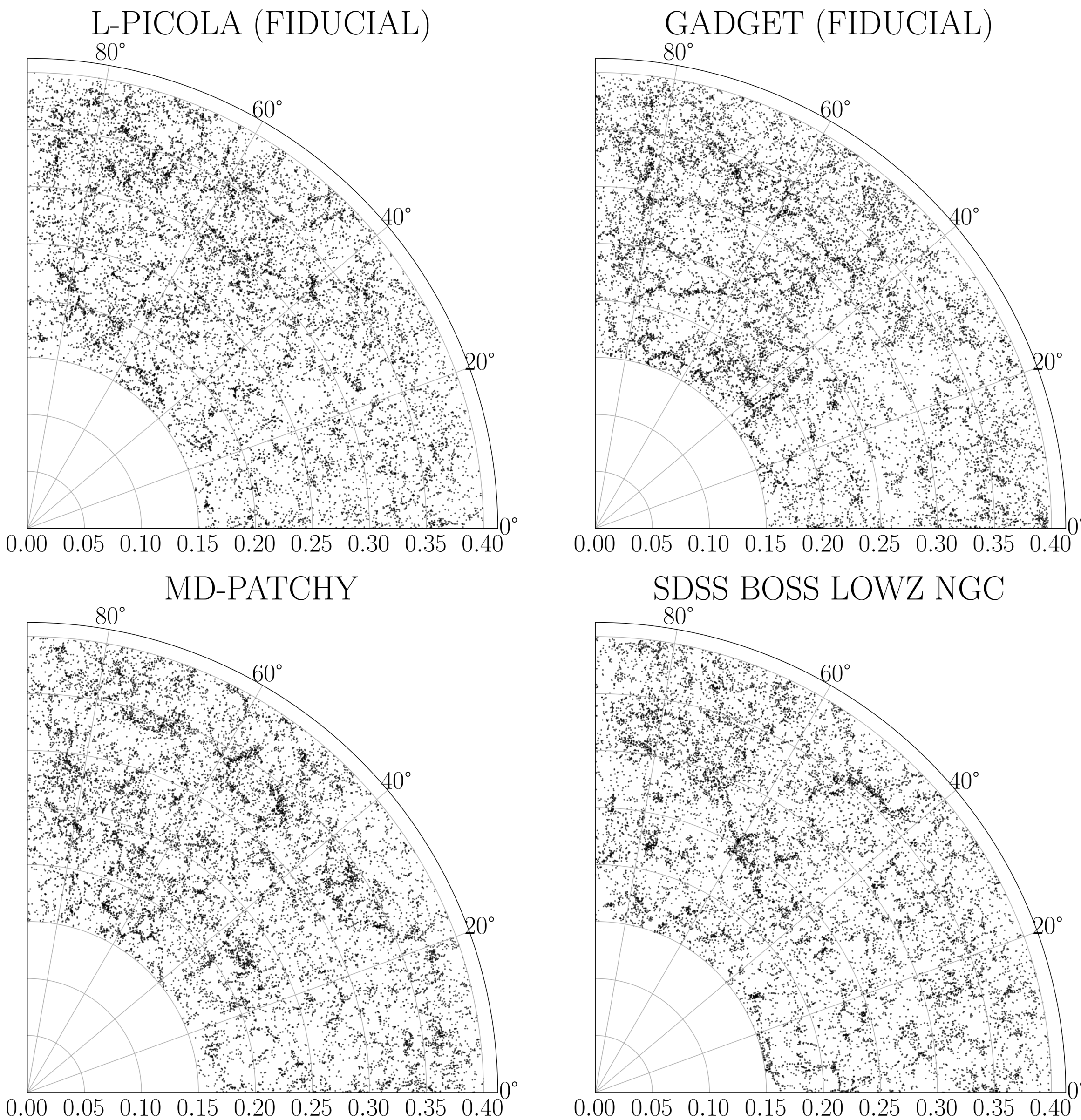

Figure 4. Lightcone slices of Figure 3 from $0^\circ < \mathrm{DEC} < 6^\circ$. See Section 2.6 for more information.

galaxy-subhalo correspondence as employed in the previous works (e.g., Kim et al. 2008; Hwang et al. 2016), we impose a fixed stellar-to-halo-mass relation (SHMR) across different realizations. In other words, we assume that the star formation efficiency of galaxies in halos is equivalent across different cosmologies, within the redshift range of this study.[3]

We use the SHMR obtained by Girelli et al. (2020), which compares the DUSTGRAIN-*pathfinder* simulation (Giocoli et al. 2018) with the SMF determined in (Ilbert et al. 2013)

[3] This is a strong assumption made to derive the stellar masses of each subhalo identified from a dark matter-only simulation, and where cosmology-dependent information intervenes. This is due to the impracticality of performing full hydrodynamic simulations of such a spatial and temporal extent across varying cosmologies. Despite introducing weak dependency, we emphasize that this assumption is made for a proof-of-concept test. For a model free of cosmological priors, refer to the `SHAM` model in Section 2.5.2.

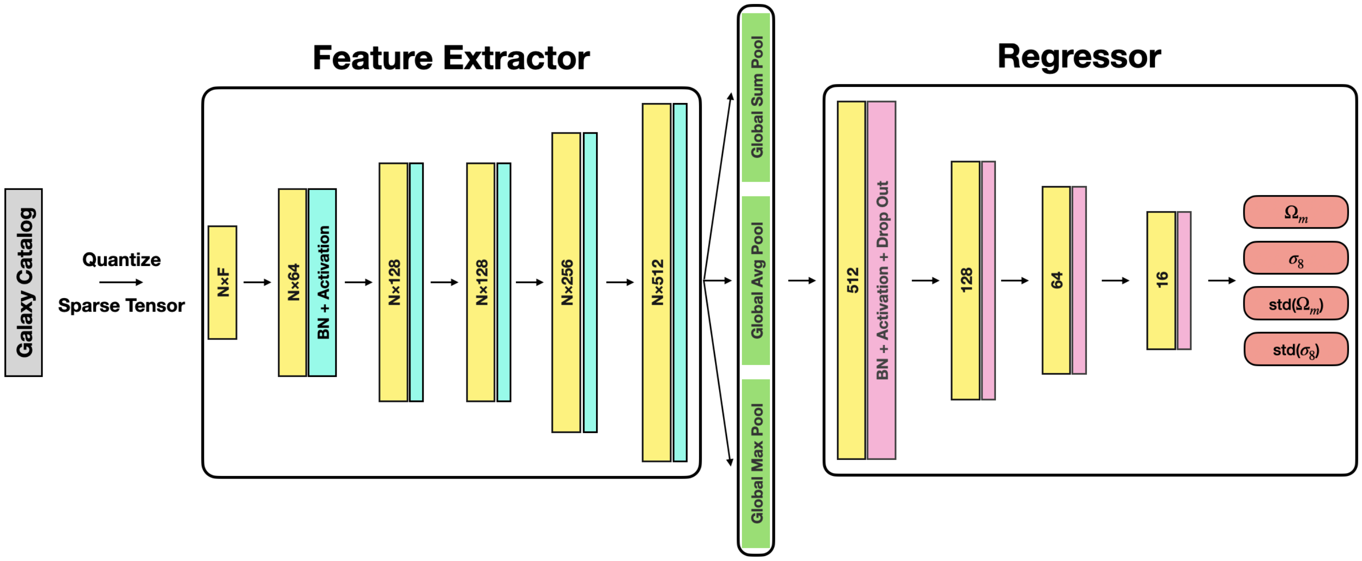

Figure 5. Architecture of the `Minkowski-PointNet` in this work. Each input is transformed into a sparse tensor with N galaxies, each with features F, and passed through five consecutive linear layers for feature extraction. Global sum, average and max pooling are done to extract a 1536-dimensional feature vector. Then it passes through the regressor consisting of four linear and dropout layers to predict Ω_m, σ_8 and their standard deviations. See Section 3.1 for more information.

from the Cosmological Evolution Survey (COSMOS) (Scoville et al. 2007). The SHMR is analyzed per different redshift bins to account for the temporal variability of the efficiency, parameterized as

$$\frac{M_\star}{M_h}(z) = 2A(z)\left[\left(\frac{M_\star}{M_A(z)}\right)^{-\beta(z)} + \left(\frac{M_h}{M_A(z)}\right)^{-\gamma(z)}\right]^{-1} \quad (1)$$

where M_h is the halo mass and $A(z)$ is the normalization factor at M_A, at which the double power-law breaks. Since our mock galaxies are selected within $0.15<z<0.40$, we utilize the SHMR parameters estimated for $0.2<z<0.5$. The best-fit parameters are $A(z)=0.0429$, $M_A=11.87$, $\beta=0.99$, and $\gamma=0.669$ when scatter of $\sigma_r=0.2$ dex is introduced. We will use these parameters, including the 0.2 dex scatter, for this work.

2.5.2. *Subhalo Abundance Matching*

In Section 2.5.1, the `Fixed SHMR` establishes cosmological priors as it selects the specific relation of connecting the halo mass properties to the baryonic physics. To tackle this issue, we alternatively utilize the non-parametric version of `SHAM`—a well-known basic galaxy-halo connection, as previously discussed, which is also used for constraining cosmological parameters (Simha & Cole 2013). The halo catalogs are painted with stellar masses using a monotonic relation between the simulated halo masses and the stellar masses identified from the observed SDSS BOSS LOWZ NGC catalog. Therefore, the difference between mocks with different cosmologies arises from the clustering of the galaxies instead of stellar mass itself as compared to the `Fixed SHMR` model.

We acknowledge that the prescription in our `SHAM` model is simplistic and may not fully describe the galaxy-halo connection. Numerous studies on SHAM have employed the historical peak mass or circular velocity of the halo (Reddick et al. 2013; Behroozi et al. 2013a). However, the nature of the on-the-fly generation of lightcones precludes the possibility of utilizing historical information. In order to bypass such limitations, Ishikawa et al. (2023) use snapshots instead of generating lightcones on-the-fly and employs post-processing to generate lightcones. However, since our focus here is on the proof-of-concept test of inferring cosmological parameters without summary statistics and using neural networks, we accept the inherent crudeness in the galaxy-halo connection model.

2.6. *Observational Effects*

We include the following observational effects of the SDSS BOSS LOWZ NGC catalog into the L-PICOLA simulations: redshift space distortions, survey footprint geometry, stellar mass incompleteness, radial selection matching, and fiber collision. By fully accounting for these observational effects, we can assess how observables from realizations endowed with different sets of cosmological parameters would have deviated from the actual observation.

Firstly, the positions of the model galaxies are shifted using their peculiar velocities to account for the redshift space distortion (RSD; Kaiser 1987). In order to match the footprint geometry of our mocks to that of the SDSS BOSS LOWZ NGC, we apply acceptance and veto masks. Galaxies are filtered out by applying the MANGLE masks (Swanson et al.

2008) using the MAKE_SURVEY code (White et al. 2014b). Next, for both `Fixed SHMR` and `SHAM` models, we restrict the area of interest to RA=150°–240° and DEC>0°.[4]

For the `Fixed SHMR` model, we further apply the incompleteness in the galaxy stellar mass function of the SDSS BOSS LOWZ NGC catalog, a statistical bias due to the observational constraints of the survey. Here, we apply the incompleteness of the LOWZ NGC sample, which is modeled by Leauthaud et al. (2016), using the Stripe 82 Massive Galaxy Catalog to measure the SMF. The incompleteness function is shown in Equation 2, where f, σ, and M_1 are free parameters for fitting. We calculate the interpolated incompleteness using the stellar mass and redshift of the galaxies, and decide whether to use or discard a galaxy based on the result.

$$c = \frac{f}{2}\left[1+\mathrm{erf}\left(\frac{\log M_\star/M_1}{\sigma}\right)\right] \quad (2)$$

After identifying the galaxies that are not observable due to stellar mass incompleteness and survey geometry, we randomly downsample the galaxies to match the radial selection. This is achieved by finely dividing the redshift range into 260 radial bins with equal redshift space volume spacing.

For the `SHAM` model we perform massive downsampling. Unlike the `Fixed SHMR` model, the SHAM model inherently includes stellar mass incompleteness because we use the observed galaxy catalog, which already has inherent incompleteness, as our reference. Also, we perform massive sampling instead of random sampling in order to match the monotonicity of the `SHAM` process. Similarly to the `Fixed SHMR` model, the sampled galaxies are filtered once more through the fiber collision algorithm, and then finally assigned with the appropriate stellar masses.

Furthermore, we mimic the fiber collision in the SDSS BOSS LOWZ NGC catalog. The SDSS galaxy spectra were obtained from fibers inserted into perforated plates. Since the fibers have a finite size with a collision radius of 62″, a portion of fiber-collided galaxies has not been assigned with any fibers. Using NBODYKIT (Hand et al. 2018), we classify the galaxies into two populations: decollided galaxies (D_1) and potentially collided galaxies (D_2) (Guo et al. 2012) using the angular friends-of-friends algorithm as in Rodríguez-Torres et al. (2016). The actual abundance matching of the `SHAM` model is performed after accounting for the fiber collisions in order to fully preserve the number of galaxies. However, for the `Fixed SHMR` model, the stellar mass incompleteness already includes the incompleteness due to fiber collisions. Nevertheless, this reduction should be applied since fiber collisions are an important systematic biases in the small-scale geometry of the survey. We consider the potential double-counting of fiber collisions within the stellar mass incompleteness to have a negligible impact on our final results.

[4] The trimming of the footprint was necessary to accommodate the generation of lightcones in octants of the sphere. This adjustment results in a slight deviation in the data used compared to earlier studies, such as those of Ivanov et al. (2020) and Hahn et al. (2023a). Nonetheless, we expect these differences to be minimal, given the modest nature of the change.

Figure 2 compares the galaxy count per radial bins for SDSS BOSS LOWZ NGC, MD-PATCHY, `L-PICOLA fiducial`, and `GADGET fiducial` mocks generated with the `Fixed SHMR` model. The similarity in the distributions verify the consistency across all three mocks and the observational catalog. In realizations with low Ω_m and σ_8 generated with the `Fixed SHMR` model, the absolute number of galaxies is relatively small, and thus the total number of galaxies may be less than that of the fiducial cosmology. Such a deficit can provide critical information to inform the neural network that the real universe is unlikely to have such cosmological parameters. However, the mocks produced by the `SHAM` model does not have difference in the total number of galaxies, as it directly matches the observed galaxy mass to the halo catalog.

Finally, for both `Fixed SHMR` and `SHAM` models, we restrict the area of interest to RA=150°–240° and DEC>0°. The four panels of Figure 3 show the footprint of the `L-PICOLA fiducial`, `GADGET fiducial`, MD-PATCHY, and the SDSS BOSS LOWZ NGC catalog. Notice that the masks are equally applied, showing the same apparent streaks and holes. Figure 4 shows the lightcone slices from 0°<DEC<6° for each of the four mocks, with the observational effects fully taken into account.

3. NEURAL NETWORK ARCHITECTURE

3.1. *Backbone:* `Minkowski-PointNet`

A large portion of the universe is empty, as galaxies are predominantly clustered along the filaments of the LSS. Therefore, depositing galaxies into uniform voxels can be highly inefficient, resulting in many voxels with few or even no galaxies assigned. To mitigate this problem, galaxies are represented as point clouds, with each galaxy depicted as a single point characterized by distinct positions and properties. This representation is then processed through a deep neural network called `Minkowski-PointNet`, which is a PointNet (Qi et al. 2016) implementation in the Minkowski Engine (Choy et al. 2019).

PointNet is a neural network architecture that captures the structure of point clouds, a simplified graph with no edges. PointNet is an architecture that can be generalized as DeepSets (Zaheer et al. 2017), which captures the permutation invariance and equivariance of point clouds (Bronstein et al. 2021). Such geometric priors are captured from the 1D convolution layers and the global pooling layers. Despite PointNet's use of rotation and translation invariance to handle point clouds, such procedures are omitted in our approach because of the redshift dependence of features and clustering, as well as the (RA,DEC) dependence of masking. Moreover, to explicitly introduce local properties, we apply the k-nearest-neighbor (kNN) algorithm to survey the characteristics of neighboring galaxies and explicitly add them to the feature vector. Such a step is inevitable since we are not able to perform message-passing between the nodes or the points, as the computational costs involving calculation on the edges are extremely demanding for the mocks comprising more than 150,000 galaxies. Therefore, we add the local information to the feature vector, to enrich the information fed to the machine.[5]

The Minkowski Engine is a library that efficiently handles sparse tensors, including operations such as auto-differentiation and convolution. Galaxies are grouped and quantized into sparse tensors based on their (RA, DEC, z) positions using the engine, where z denotes the redshift. The main advantage of this implementation lies in its ability to handle a variable number of points as inputs to the machine, whereas the original implementation of PointNet operates on fixed sizes. Additionally, it efficiently utilizes memory by grouping galaxies into sparse tensors. This approach results in approximately 25% of the quantized cells containing more than one galaxy, and around 5% containing more than two galaxies. This strategy effectively preserves the local structure while ensuring better memory consumption and performance.

The specific network layout is illustrated in Figure 5. The `Minkowski-PointNet` is capable of receiving point clouds of arbitrary size. The input catalog is transformed into a sparse tensor and passes through a total of five linear layers. Each linear layer is followed by a batch normalization layer (Ioffe & Szegedy 2015) and a leaky ReLU activation function. The tensor is then passed through the global sum, average, and max-pooling layers and concatenated to a 1536-dimensional vector. Global aggregators are crucial to reflecting the permutation invariance of the neural network. Unlike the original implementation of PointNet, solely using the global max-pooling as the aggregator, we add other aggregators to better capture the embedded information as suggested in Corso et al. (2020). After four consecutive linear layers, the machine predicts the Ω_m, σ_8, and their standard deviations, which will be used for implicit likelihood inference.

During the training process, we use the ADAM optimizer (Kingma & Ba 2014) with a learning rate of 10^{-7} and a ReduceLRonPlateau scheduler, which reduces the learning rate when the validation loss is not decreased, for a total of 20 epochs. We make use of 80% of the samples as a training data set and 10% each as validation and test data sets. We adopt the loss function for implicit likelihood inference as described in Jeffrey & Wandelt (2020), which is the sum of the following two loss functions, where y is the label and σ^2 the variance.

$$L_1 = \ln\left[\sum_{i\in batch}(y_{i,pred} - y_{i,true})^2\right] \tag{3}$$

$$L_2 = \ln\left[\sum_{i\in batch}((y_{i,pred} - y_{i,true})^2 - \sigma_i^2)^2\right] \tag{4}$$

By minimizing the combined loss function $L_\mathrm{vanilla}=L_1+L_2$, we optimize both prediction accuracy and enable the representation of the second moment, which corresponds to the standard deviation. Such approaches have recently been utilized in many machine learning projects to estimate the model's error in the absence of likelihoods (Villaescusa-Navarro et al. 2022; Villanueva-Domingo et al. 2022).

3.2. *Input Features*

The input features of galaxies should align with those derivable from observational data. Thus, we utilize the position and stellar mass of each galaxy, as well as information from its neighbors, to extract details about the local environment, following the methodology presented in Jo & Kim (2019). Moreover, it is important to note that we do not provide the machine with physical or comoving distances since they already imply a certain cosmology when converted from observed redshifts. Instead, we introduce a transformed position of each galaxy by $(X,Y,Z) = (z\sin(DEC)\cos(RA), z\sin(DEC)\sin(RA), z\cos(DEC))$. The redshift will be re-introduced as one feature, allowing the machine to infer the redshift dependence of features.

Additionally, we explicitly incorporate information from neighboring galaxies. This addresses the limitations of `Minkowski-PointNet`, which does not support message-passing between edges due to computational constraints arising from the large number of inputs. By introducing neighboring information, we expect these features to serve as proxies for relational local information. From the nine nearest neighbors, four local features are selected: mean distance, maximum distance, mean stellar mass, and maximum stellar mass. Again, since we apply a metric in the redshift space, the distances become unitless. The redshift and stellar mass of each galaxy are used as point-specific features. In total, the six features are aggregated per galaxy, combining both local and point-specific characteristics. Figure 6 displays a pair plot of features with contours for 1000 randomly sampled galaxies for the mocks generated with the `Fixed SHMR` model. The distribution exhibits fair consistency across the

[5] In contrast to PointNet++ (Qi et al. 2017), which uses kNN for grouping and non-uniform sampling of points, we do not adopt such set abstraction layers since the absolute number of galaxies comprising each realization needs to be informed to the machine.

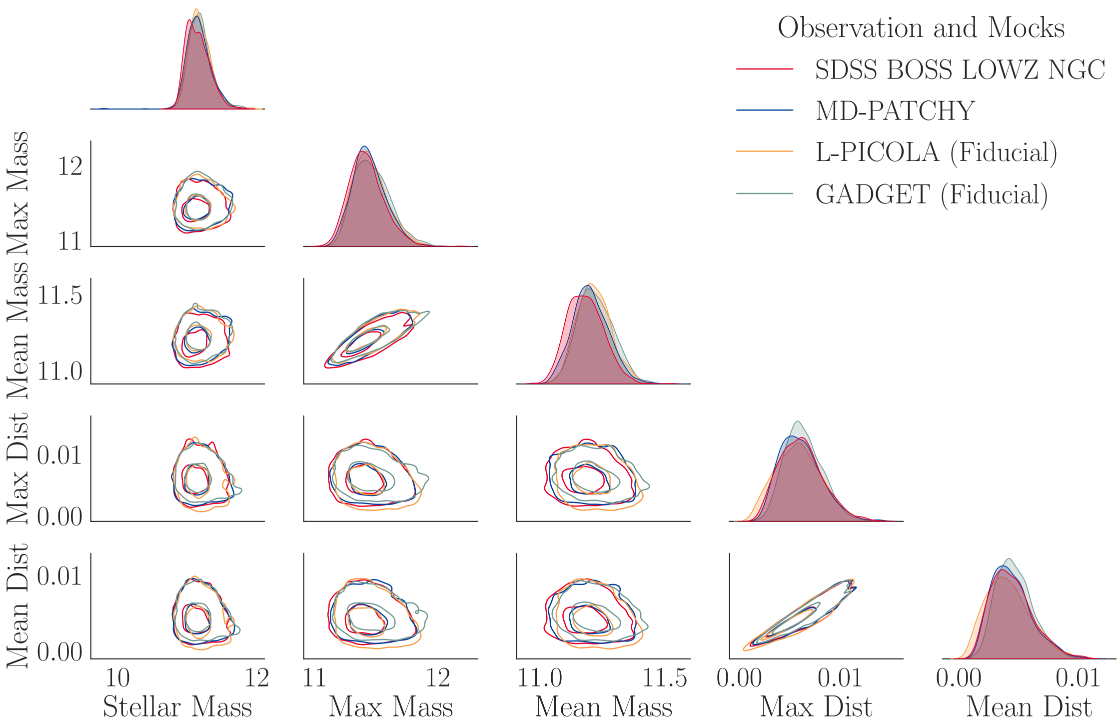

Figure 6. Pair plot of five features of a galaxy randomly sampled from each mock generated with the `Fixed SHMR` model: stellar mass of itself, maximum and mean neighbor masses, and maximum and mean neighbor distances, for a single realization of SDSS BOSS LOWZ NGC (*red*), MD-PATCHY (*blue*), `L-PICOLA fiducial` (*orange*), and `GADGET fiducial` (*green*). The plot shows for 1000 randomly sampled galaxies for each mocks. Masses are in units of $\log(M_\star/h^{-1}\mathrm{M}_\odot)$ and distances are expressed in terms of the newly assumed metric in redshift space. The distribution exhibits fair consistency across the three mocks and the SDSS BOSS LOWZ NGC catalog. See sections 3.1 and 3.2 for more information.

three mocks and the SDSS BOSS LOWZ NGC catalog. Another comparison between different cosmologies is available in Appendix A. Although not displayed for brevity, the `SHAM` models exhibit similar levels of consistency in the mocks.

4. TRAINING STRATEGIES

4.1. *Why is Domain Shift Critical?*

The small-scale clustering statistic and the low mass end of a halo mass function may have distortion because of its approximate nature in the L-PICOLA code. This is due to the dispersive behavior of dark matter particles that leads to an imprecise subhalo determination (Howlett et al. 2015b). Moreover, the on-the-fly lightcone simulation restricts us from exploiting the historical information of individual halos. The evolution of individual subhalos can be tracked using merger trees derived from simulation snapshots. From this, accurate modeling of the galaxy-halo connection through SHAM is feasible using $V_{\rm peak}$ or $V_{\rm max}$, even for dark matter fields generated with COLA simulations as opposed to the lightcone simulation (Ding et al. 2023). In an attempt to mitigate the intrinsic limitation of the rapid lightcone simulation, L-PICOLA, Howlett et al. (2022) introduce two free parameters to represent the subhalo number and mass ratio. These values are tuned by fitting the power spectrum monopole of the observational catalog. However, since we aim at performing inference rather than fine-tuning simulations to match observational data, such an adaptation step is inapplicable. We can enhance the flexibility of the models by incorporating extra free parameters and marginalizing over them during inference, particularly with the HOD framework. However, this approach restricts the use of stellar mass information used in modeling the stellar mass incompleteness and as features in the neural networks. We plan to address such issues in future work.

`Minkowski-PointNet` demonstrates strengths in its lack of specific limits on clustering scale, allowing for analysis across a wide range of scales, unlike most studies that impose an upper bound $k_{\rm max}$ (Hahn et al. 2023a,b; Ivanov et al. 2020; Philcox & Ivanov 2022). Even CNNs inherently impose an effective clustering scale through voxelization (Lemos et al. 2023). However, our approach is sensitive to small scales, offering rich clustering information while also being susceptible to small-scale distortions specific to each domain's codes. Therefore, it is critical to regularize the training of neural networks to acquire domain-agnostic knowledge.

Addressing the domain shift is crucial to ensuring the robustness of machines and their applicability to real-world observations. We adapt the machines using prepared suites of mocks: 9000 L-PICOLA mocks as the source, along with either 18 GADGET mocks or 2048 MD-PATCHY mocks as targets. By training them with specific strategies aimed at achieving domain adaptation and generalization, we expect the machines to learn domain-agnostic information. Consequently, they will be capable of extracting representations that can be generalized to multiple domains, particularly observational data.

4.2. *Training Objective: Domain Generalization*

The primary goal of this research is to conduct simulation-based inference on actual observational data using machines robust across different codes for generating mocks. A critical question arises: Can we establish a unified approach to forward modeling our universe and making fair inferences on the cosmological parameters? Unfortunately, current neural networks show apparent discrepancies when applied to other domains (Ni et al. 2023; Shao et al. 2023). However, recent trials in generating domain-adaptive graph neural networks to incorporate various sources have shown the possibility of achieving a more robust inference (Roncoli et al. 2023).

In the context of transfer learning, which involves the transfer of knowledge from a set of task to relevant tasks, each of the mock suites can be viewed as n mocks sampled from individual domains $\mathscr{D}_i$, or $S_i = \left\{(x^i_j, y^i_j)\right\}_{j=1}^{n} \sim (\mathscr{D}_i)^n$, where $x \in \mathscr{X}, y \in \mathscr{Y}$. $\mathscr{X}$ is the feature space and $\mathscr{Y}$ is the space for labels (cosmological parameters), while $\mathscr{D}_i \subset \mathscr{P}_{\mathscr{X}\mathscr{Y}}$ is a joint distribution on $\mathscr{X}$ and $\mathscr{Y}$ (Ganin et al. 2016; Wang et al. 2023). Our aim is to develop a machine that generalizes across multiple domains, even those unseen during the training phase, particularly the observational catalog. Attempts to test the generalizability of a machine trained on a single domain have been initiated by various projects in astronomy using machine learning and deep learning, referred to as “robustness tests” (Ni et al. 2023; Shao et al. 2023). In the language of transfer learning, testing on uninvolved domains in the training phase can be viewed as domain generalization (DG; Wang et al. 2023).

To achieve effective domain generalization, it is crucial that the distributions of the target (unseen domains) and source domains (domains involved in the training phase) are similar, which can be achieved through accurate modeling of mocks and training strategies to extract common features. Due to limitations in the accuracy of L-PICOLA mocks, non-negligible discrepancies exist compared to GADGET or MD-PATCHY mocks. Such domain shift (expressed by $\mathscr{H}$-divergence, $d_{\mathscr{H}}(\cdot,\cdot)$) is crucial in setting the upper bound on the empirical risk of any hypothesis (Ben-David et al. 2006, 2010; Albuquerque et al. 2019). Thus, achieving single-domain generalization solely through training on L-PICOLA mocks can be challenging.

To enhance the machine’s generalization capabilities, we utilize GADGET or MD-PATCHY mocks, which enable the machine to acquire common knowledge. Unlike domain generalization, GADGET or MD-PATCHY mocks are incorporated during the training phase, hence, this approach is termed domain adaptation. By employing a training strategy to learn from the relatively accurate mocks, the neural networks learn consistent semantics from the two domains, and finally generalize on the observational data, unseen at training phase.

A method includes utilizing the domain-adversarial neural network (DANN; Ganin et al. 2016), which seeks to derive domain-invariant features through the use of a domain classifier as a regularizer. This technique has recently been adopted for performing classification tasks in the field of astronomy (Huertas-Company et al. 2023; Ćiprijanović et al. 2020). However, multiple trials show that DANN still suffers from overfitting and there are discrepancies between domains (see Appendix C for more information). We find that such issues can be effectively mitigated by an alternative training strategy, which will be explained in **[Section 4.3]**.

4.3. *Training Strategy: Semantic Alignment*

Our strategy explicitly aligns representations from different domains with similar labels. In other words, given that the samples have similar cosmological parameters, regardless of the selection of simulations, the neural networks extract features that are similar to each other. Aligning the representations can explicitly bring about consistency in terms of their semantics across domains and be effective in domain generalization (Motiian et al. 2017). We adapt the semantic alignment loss in Motiian et al. (2017) to a regression task setup by adding the following loss term:

$$L_{\mathrm{SA}} = \sum_{i \in B_S} \sum_{j \in B_T} \frac{1}{\left\| \mathbf{y}^S_i - \mathbf{y}^T_j \right\|} \left\| g(\mathbf{x}^S_i) - g(\mathbf{x}^T_j) \right\| \tag{5}$$

Here, B_S and B_T represent batches from domains S (source) and T (target), respectively, with $g(\cdot)$ denoting the function that maps input to the representation vector. We apply the semantic alignment loss to the 16-dimensional representation, which can be obtained just before the terminal layer of the

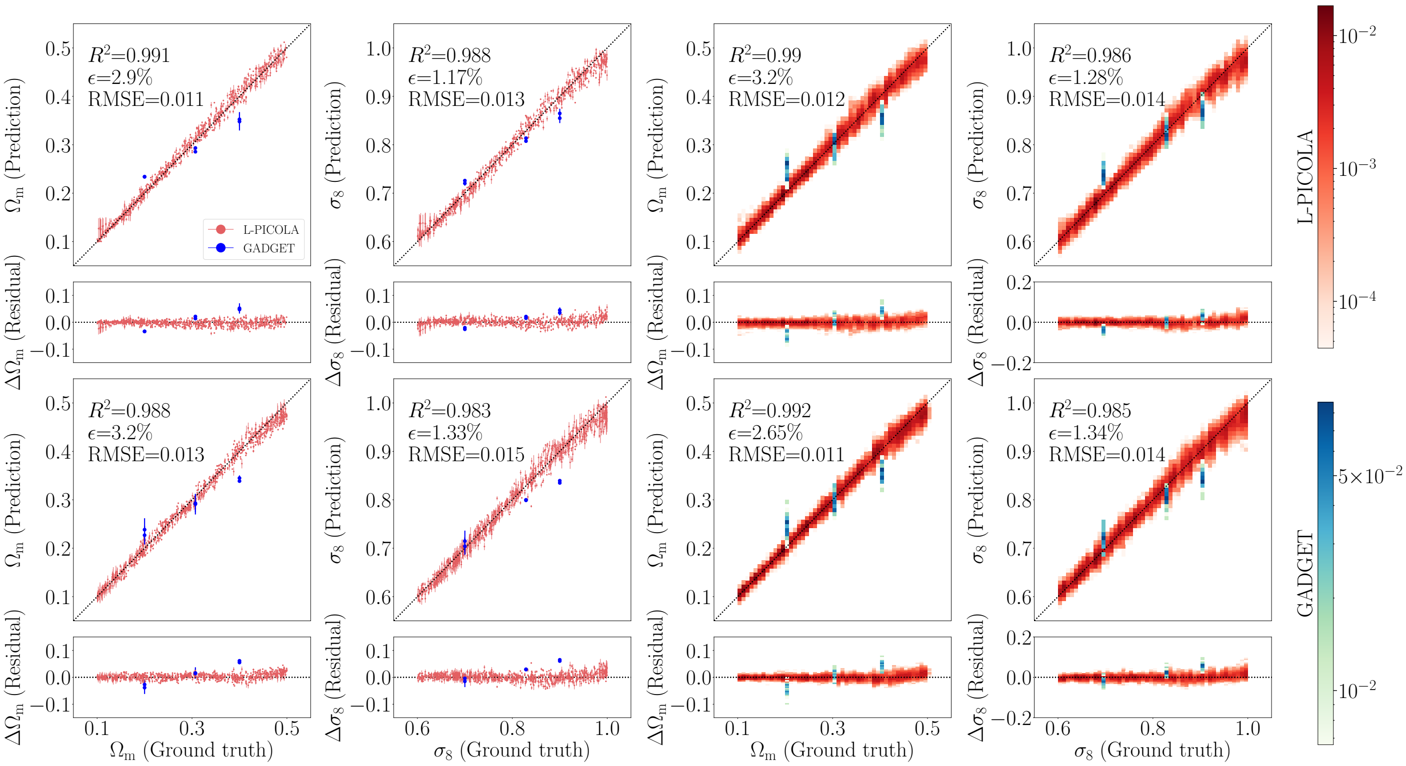

Figure 7. Comparison of the ground truth and the predicted values of Ω_m and σ_8 on the test set. Predictions are made by `Minkowski-PointNet` machines with L-PICOLA (*red*) and GADGET (*blue*) mocks, trained with the semantic alignment strategy. The *top two panels* display the results from the `SHAM` model and the *bottom two panels* display the results from the `Fixed SHMR` model. The *left two columns* show the results of a single arbitrarily selected machine, while *the right two* show the results for the 25 independently trained machines, with the normalized count expressed in logarithmic color bars. R^2, relative error (ε) and root-mean-squared-error (RMSE) metrics calculated altogether are shown. Residuals $\Delta y = y_{\rm true} - y_{\rm pred}$ are depicted in the *bottom panels*. The machine is trained, validated and tested on the two suites of mocks: L-PICOLA mocks and GADGET mocks. Error bars of the *two left columns* indicate the 1σ values derived from the implicit likelihood inference. *Black dotted lines* depicts the complete match with null residual. The results from the ensemble of 25 machines for Ω_m and σ_8 show a relative error of 3.20% and 1.28% for the `SHAM` model. The `Fixed SHMR` model yields 2.65% and 1.34%. See Section 5.1 for more information.

neural network, as depicted in Figure 5.[6] The generalization strength can be modified by adjusting the weight α_p in $L_{\rm total} = L_{\rm vanilla} + \alpha_p L_{\rm SA}$. Here, we slightly modify the adaptation parameter setup proposed by Ganin et al. (2016),

$$\alpha_p = \alpha_0 \left[\frac{2}{1+\exp(-\gamma p)} - 1 \right] \tag{6}$$

where p linearly increases from 0 to 1 as training epochs increase, with γ=5 and α_0=5. This gradual increase in the strength of the adaptation term allows the machine to first gain predictability on the labels before aligning the representations' semantics. Hyperparameters are chosen based on multiple trials to balance the trade-off between prediction accuracy and the strength of domain adaptation. To observe the effectiveness of the alignment process, or domain adaptation, we do not include the samples from the target domain in calculating the vanilla loss (see Equations 3 and 4). Therefore, the labels of targets are only implied to the machine through the semantic alignment loss. When incorporating GADGET mocks, we reserve 2/3 of the mocks for training and 1/3 for testing. For MD-PATCHY mocks, we use 80% as a training data set and 10% each for validation and test data sets, same as L-PICOLA mocks.

[6] In this study, we opted for a reduced representation of 16 dimensions instead of the comprehensive 1536-dimensional representation due to challenges in balancing accuracy and adaptability within our machine learning model. The use of the penultimate layer of linear networks as the representation vector was also used in Lin et al. (2022). Modifying the architecture of the neural network and performing detailed fine-tuning of hyperparameters are strategies that could enhance adaptability, which we aim to explore in future research.

5. PREDICTION OF `Minkowski-PointNet`

In this section, we conduct a series of performance tests of `Minkowski-PointNet` and make predictions on the cosmological parameters of the observational catalog. Given the stochastic nature of the training outcome arising from the existing trade-off between domain adaptability and the accuracy of individual predictions, we train 25 different machines, whose model parameters are randomly initialized.

Table 1. Summary of Predictions on the Cosmological Parameters

Models	Training Strategy	Ω_m	σ_8	$\varepsilon_{\Omega_m}(\%)$	$\varepsilon_{\sigma_8}(\%)$
`SHAM`	Semantic Alignment	**0.339±0.056**	**0.801±0.061**	**7.6**	**1.2**
`SHAM`	Vanilla	0.357±0.044	0.858±0.045	13.3	5.8
`Fixed SHMR`	Semantic Alignment	0.227±0.035	0.743±0.039	27.9	8.4
`Fixed SHMR`	Vanilla	0.196±0.021	0.705±0.019	37.8	13.1
Planck Collaboration et al. (2020)	-	0.315±0.007	0.811±0.006	-	-
Ivanov et al. (2020)	-	0.295±0.010	0.721±0.043	-	-

NOTE—Summary of cosmological parameter predictions from different models trained with L-PICOLA as *source* and GADGET mocks as *target*. For this work, we refer to the galaxy-halo connection model as the model names, together with the two training strategies: Semantic Alignment (*with domain adaptation*) and Vanilla (*without domain adaptation*). The predicted values for each models are given with their respective uncertainties, which include both the uncertainty of individual machine and all 25 independently trained machines combined. Together with our main results, we also display the results from the CMB measurements (Planck Collaboration et al. 2020) and the full-shape power spectrum analyses of BOSS (Ivanov et al. 2020) for reference. Relative differences ε_{Ω_m} and ε_{σ_8}, calculated with respect to the results of Planck Collaboration et al. (2020), are displayed for the models studied in this work. See Section 5.2 for more information on the results, and Section 6.1 for the discussion on the comparison between the two training strategies.

Before predicting on the actual SDSS BOSS LOWZ NGC data, we perform the same feature sampling by identifying their neighbors, as explained in Section 3.2. The designated local and global features are then fed to the trained machines. We compare and discuss the results from a set of machines adapted to different domains, as summarized in Table 1.

5.1. *Performance Tests of* `Minkowski-PointNet`

Following the training procedures discussed in the previous sections 3 and 4, machines are trained to predict Ω_m, σ_8 and their standard deviations. Figure 7 displays test results of machines trained with the semantic alignment strategy on the L-PICOLA and GADGET mocks. We present results for an arbitrarily selected single machine and for all 25 individually trained machines. The top two panels show the results of the `SHAM` model, and the bottom two panels show the results of the `Fixed SHMR` model. In each case, the upper panels show the comparison between the true and predicted values, while the bottom shows the residual. The test results are promising for both Ω_m and σ_8, regardless of the galaxy-halo connection model. The results from the ensemble of 25 machines for Ω_m and σ_8 show a relative error of 3.20% and 1.28% for the `SHAM` model and ε=2.65% and 1.34% for the `Fixed SHMR` model, respectively. A single machine shows a relative error of 2.90% and 1.17% for the `SHAM` model, and ε=3.20% and 1.33% for the `Fixed SHMR` model. The difficulty in trying to accurately predict σ_8 seen in recent studies (Villaescusa-Navarro et al. 2022; Villanueva-Domingo & Villaescusa-Navarro 2022; de Santi et al. 2023) is not apparent.

The blue markers and bins in Figure 7 show the domain adaptation results in GADGET mocks. Due to semantic loss, we are able to marginalize the selection of domains, which leads to the degradation of accuracy in each simulation set (for more information on the error analysis, see Section 6.1). Since the machine only implicitly infers the cosmological parameters of the GADGET mocks through semantic alignment loss during the training phase, a noticeable bias is observed in the predictions when comparing GADGET mocks to L-PICOLA mocks. However, the fact that the machine can make predictions solely by aligning the semantics of the source and target domains is encouraging.

Moreover, considering that the parameter space of input labels is constrained within a range of $\Omega_m \in [0.1, 0.5]$ and $\sigma_8 \in [0.6, 1.0]$, samples like `GADGET low` and `high` may encounter asymmetry when calculating the semantic alignment loss. In an extreme scenario, if a sample is characterized by the cosmological parameters $\Omega_m = 0.6$ and $\sigma_8 = 1.0$, it may suffer from bias due to the lack of samples with larger values of the cosmological parameters. This could lead to center-biased predictions as their representations may experience excessive center-ward pull. Overall, the adaptation results remain quite promising, indicating effective alignment of representations from the two domains by the machine.

5.2. *Predictions on the SDSS BOSS LOWZ NGC Catalog*

In this section, we present predictions on the SDSS BOSS LOWZ NGC Catalog made by the `Minkowski-PointNet` machines trained with different galaxy-halo connection models and training strategies. Table 1 summarizes the results of the machines trained with L-PICOLA and GADGET mocks. Figure 8 illustrates the aggregated outcomes of 25 distinct machines, each trained using semantic alignment with L-PICOLA and GADGET mocks, alongside benchmark values from Planck

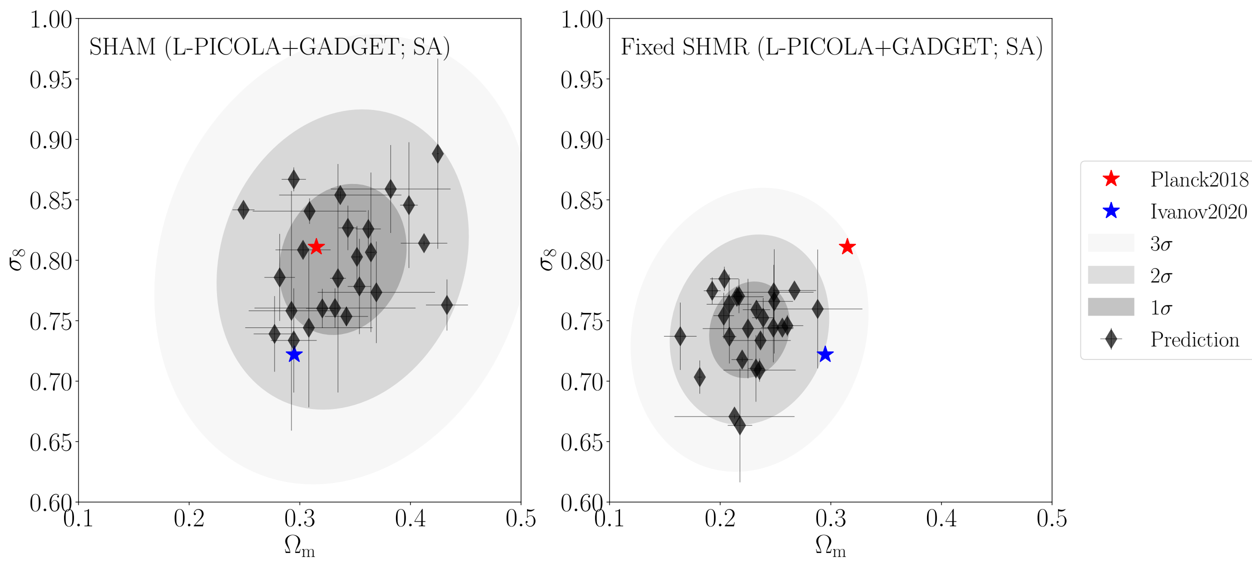

Figure 8. Prediction on the actual SDSS BOSS LOWZ NGC catalog from the ensemble of 25 independently trained `Minkowski-PointNet` machines. The *left* figure displays our results when using the `SHAM` model, and the *right* figure displays the results when using the `Fixed SHMR` model. The machines are trained with L-PICOLA and GADGET mocks with the semantic alignment (SA) strategy, a domain adaptation and generalization technique that enables the machines to extract consistent features regardless of their simulation domains (see Section 4.3). Predictions are shown with error bars. A *red star* shows the result from the Planck 2018 (Planck Collaboration et al. 2020) measurements and a *blue star* from Ivanov et al. (2020). *Elliptic contours* show the bounds of 1σ, 2σ, and 3σ bounds, calculated from the Gaussian Mixture Model (GMM) to incorporate the individual errors. Our results yield Ω_m=0.339±0.056, σ_8=0.801±0.061 (*left, SHAM*), and Ω_m=0.227±0.035, σ_8=0.743±0.039 (*right, Fixed SHMR*). See Section 5.2 for more information.

Collaboration et al. (2020) and Ivanov et al. (2020).[7] Even within a single training scheme, the predicted results vary significantly between machines, illustrating the stochastic nature of the training process. This suggests that there is degeneracy in the final state of the machine, with multiple configurations exhibiting similar, suboptimal performance. In other words, although different machines demonstrate consistent accuracy and precision on the test set, their predictions on the observational catalog unseen during training phase shows notable variability. This justifies our approach of training multiple machines instead of selecting only those with the best performance.

Next, we compare how the machine predicts on the observational data when trained *with* the domain-adaptive training strategy (semantic alignment) and when trained *without* it (vanilla). For the `Fixed SHMR` model, the prediction of the ensemble of 25 machines yield Ω_m=0.196±0.021 and σ_8=0.705±0.019 in vanilla scheme, while after applying the semantic alignment loss, Ω_m=0.227±0.035 and σ_8=0.743±0.039. The `SHAM` model yields Ω_m=0.357±0.044 and σ_8=0.858±0.045 in the vanilla scheme, and Ω_m=0.339±0.056 and σ_8=0.801±0.061 with semantic alignment. The semantic alignment worsens the precision compared to when not applied, despite increasing the accuracy of prediction, assuming Planck 2018 cosmology as the ground truth. Thus, although the same data sets are being used, the differences in how they are employed to train the machines severely affect the accuracy and precision of prediction on unseen domains.

The predictions vary significantly depending on the galaxy-halo connection model used to generate the mock catalogs. Especially, `Fixed SHMR` models exhibit considerable divergence from the Planck 2018 cosmology (Ω_m=0.315±0.007 and σ_8=0.811±0.006), while `SHAM` models are largely in agreement, within the 1σ error. Moreover, the Ω_m predicted by the `SHAM` models show consistent values with the most recent dark energy survey, DES Collaboration et al. (2024), which yields Ω_m=0.352±0.017 for the flat ΛCDM model, a higher value than the Planck 2018 cosmology. Although `SHAM` is the most favorable in terms of both accuracy and the absence of any cosmological priors involved in the forward modeling processes, `Fixed SHMR` exhibits better precision. This discrepancy likely stems from the additional cosmological priors incorporated via stellar

[7] The main result from Ivanov et al. (2020), which we cite in Table 1 and figures 8, 11, 13, and 14, combines the likelihoods from the Northern Galactic Cap (NGC) and the Southern Galactic Cap (SGC) across two redshift ranges: *low-z* ($z_\mathrm{eff}=0.38$) and *high-z* ($z_\mathrm{eff}=0.61$). Although our LOWZ NGC mocks differ from the *low-z* definition, having a lower effective redshift of $z_\mathrm{eff}=0.29$, the results from the *low-z* NGC used in Ivanov et al. (2020) yield $\Omega_\mathrm{m}=0.290\pm0.017$ and $\sigma_8=0.808\pm0.073$ (see sections 5.2 and 6.2 for more information).

masses in `Fixed SHMR` models, as opposed to `SHAM` models, which rely solely on clustering information.

This discrepancy can be due to several factors, although the precise cause of this bias in the `Fixed SHMR` model remains unclear. One potential reason is that, for the `Fixed SHMR` model, regardless of cosmology, any halo with a similar mass will be assigned a similar stellar mass following the SHMR. As discussed in Section 2.5.1, the SHMR from Girelli et al. (2020) was obtained from a different survey, COSMOS, which could also explain the variations. Additionally, the stellar masses of the galaxies in the observational catalog are determined on the basis of the Kroupa IMF (Kroupa 2001) with passive evolution from Maraston et al. (2013), whereas the SHMR we utilized is based on the SMF adjusted for the Chabrier IMF (Chabrier 2003) and the stellar population synthesis models from Bruzual & Charlot (2003), which can result in such differences. The exact cause of this discrepancy still being unclear, we stress the limitations of our naive assumption in the `Fixed SHMR` model, and that results may vary depending on the galaxy-halo-connection models. Here, we aim to demonstrate the feasibility of inferring without using summary statistics and leave further investigation into the impact of galaxy-halo connection models for future studies.

As mentioned above, when calculating the uncertainty of the inferred parameters, we adopt the most conservative approach. We consider both the error of individual predictions and the 25 independently trained machines, without cherry-picking. However, selecting a single machine that best adapts to and predicts on GADGET mocks, characterized by the smallest distance measured by $\sqrt{\Delta\Omega_{\rm m}^2+\Delta\sigma_8^2}$, yields results of $\Omega_{\rm m}$=0.267±0.020 and σ_8=0.775±0.0003 for `Fixed SHMR` and $\Omega_{\rm m}$=0.282±0.014 and σ_8=0.786±0.036 for `SHAM`. This suggests further potential for performing more precise inference on the cosmological parameters, achieved through the convergence of individual machines and enhanced robustness (see Section 6.3 for a discussion).

6. DISCUSSION

6.1. *Effect of Aligning Representations*

The improvement in generalizability can be attributed to the distribution of different domains aligned in the feature space. To compare the extracted features from machines trained by the vanilla scheme and the semantic alignment strategy, we visually inspect the distributions of their representations in a lower dimension (Jo et al. in prep). Figure 9 exhibits the latent space configuration of the targeted 16-dimensional vector reduced to two dimensions, deduced by the t-distributed Stochastic Neighbor Embedding algorithm (t-SNE; van der Maaten & Hinton 2008). In the semantic alignment strategy, the samples are evenly distributed in the reduced dimensions and the parameters gradually change along one direction, while being almost independent in the other direction[8]. This behavior naturally suggests that the machine is extracting features and representing them effectively in a way that removes degeneracy and gains predictability in the two parameters.

The vanilla scheme fails to achieve an adaptation of the GADGET mocks to the L-PICOLA mocks, resulting in a clear separation between the distributions. The proximity of the observation target to the GADGET mocks in comparison to the L-PICOLA mocks demonstrates that the GADGET mocks provide a more precise representation of our real universe for the `SHAM` model. On the other hand, when the semantic alignment strategy is employed, the two distinct domains blend into a single distribution. Consequently, this supports the claim that the machine is extracting common features from the two domains and less weighting on the domain-specific information, which improves prediction accuracy on the observational data.

However, there exists a clear trade-off as the semantic alignment loss degrades precision although showing better accuracy. To analyze the effect of semantic alignment on precision, we can first decompose the error into two sources: the *aleatoric* (statistical) error and the *epistemic* (model or systematic) error. The two distinct sources of errors can easily be seen in Figure 8—the aleatoric error estimated from the individual error bars of the machines and the epistemic error from the variance in the prediction from the ensemble of machines.

Figure 10 shows the two sources of error for the test sets of GADGET and L-PICOLA mocks, which are the domains seen during training phase, and MD-PATCHY and SDSS BOSS LOWZ NGC samples, unseen during training phase, for the `SHAM` model. As we have applied the trained machines to the SDSS BOSS LOWZ NGC catalog, we make inferences on the MD-PATCHY mocks for further analysis (See Section 6.3 for more information on the results). The epistemic errors are calculated by the standard deviation of the predictions on a single input data from the ensemble of 25 machines. On the other hand, the aleatoric errors are calculated by the root mean square of the predicted errors (see Equation 4) of the individual machines. Largely, the aleatoric and epistemic errors have comparable values for both the L-PICOLA test set and the SDSS BOSS LOWZ NGC catalog. However, the errors for the GADGET test set show a larger epistemic error compared to

[8] The gaps in the latent space can arise for several reasons. Firstly, the randomness in sampling the parameter space disrupts the dataset's uniformity. Secondly, the dimension reduction technique relies on the distribution's local structure and is inherently nonlinear. Furthermore, because of the discriminative nature of our neural networks, the distribution is not required to be uniform. Generative models such as normalizing flows and variational autoencoders are better suited for accurately modeling the distributions within specific probability distribution functions.

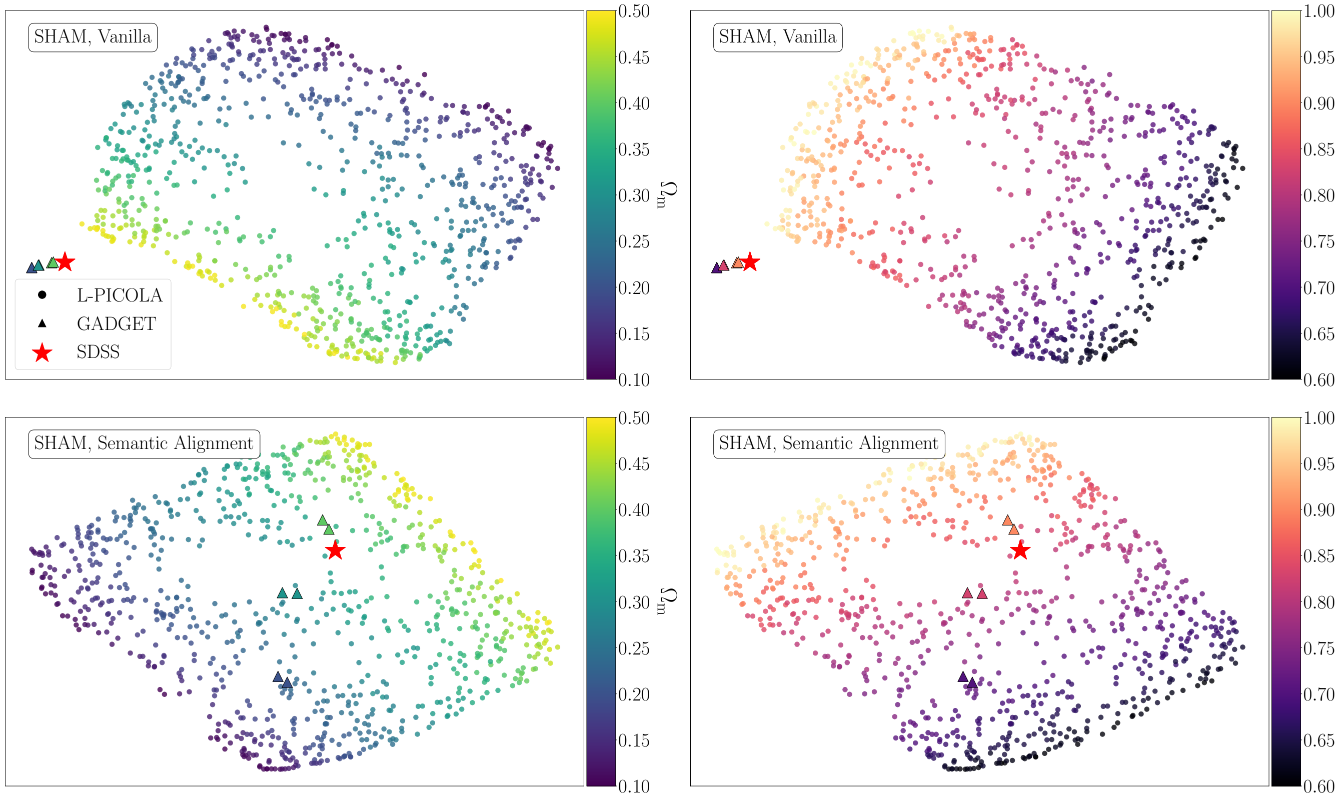

Figure 9. A visualization of the latent space configuration from a typical neural network trained with SHAM L-PICOLA and GADGET mocks with the vanilla scheme (*upper panels*) and the semantic alignment strategy (*lower panels*). The 16-dimensional vectors are reduced to two dimensions using the t-SNE algorithm (van der Maaten & Hinton 2008). L-PICOLA (*circle*) and GADGET (*triangle*) samples are colored according to their cosmological parameters: Ω_m (*left*) and σ_8 (*right*), alongside with SDSS BOSS LOWZ NGC (*red star*). In the *lower panels*, where the semantic alignment strategy is applied, two distinct axes are evident. Along one axis, the parameters gradually change in one direction, while remaining almost independent along the other. This pattern indicates that the two cosmological parameters are effectively represented. Moreover, the GADGET samples are effectively integrated and generalized in these two panels, in stark contrast to the *upper panels* of the vanilla scheme which show apparent distinction in the distribution. See Section 6.1 for more information.

the aleatoric error for the semantic alignment training strategy.

The alignment scheme has a positive effect in reducing errors when predicting in L-PICOLA samples. In particular, the epistemic and aleatoric errors in Ω_m show improvements by 23% and 4% each, respectively, and 17% and 33% for σ_8. Conversely for GADGET samples, epistemic and aleatoric errors on Ω_m show degradation by 92% and 38% each respectively, and 86% and 34% for σ_8. Thus, we can interpret that the domain-adapted machines exhibit weaker constraints, mostly due to the model-wise uncertainty on the target domain. In other words, the alignment scheme is unstable and can lead to significant variability in the machine's end-of-training state. This considerable variability in model performance on the target domain after domain adaptation can be attributed to the implicit provision of cosmological parameters to the models via the semantic alignment loss, in contrast to the vanilla models. However, the prediction on the unseen observational target shows no significant inclination towards either of the two sources of error. Specifically, the ratio of epistemic to aleatoric error increases by 21% for Ω_m and decreases by 23% for σ_8 after adaptation. Likewise, for the MD-PATCHY mocks, which are also unseen during the training phase, both epistemic and aleatoric errors arise, but the focus is on the aleatoric error, thus reducing the ratio of epistemic to aleatoric error.

Seen from the analyses above, domain adaptation with semantic alignment improves the overall generalizability on the unseen domains and precision in the source domain while sacrificing precision in target and unseen domains. Although its detailed impact on the precisions are indeed complex, the improvement on generalizability can be mathematically modeled by the *domain generalization error bound* (Albuquerque et al. 2019; Wang et al. 2023). The upper bound of the domain generalization error can also be decomposed into a few sources. Firstly, the machines have to perform well in each of the source domains individually and jointly. Moreover, the source domains should well depict the un-

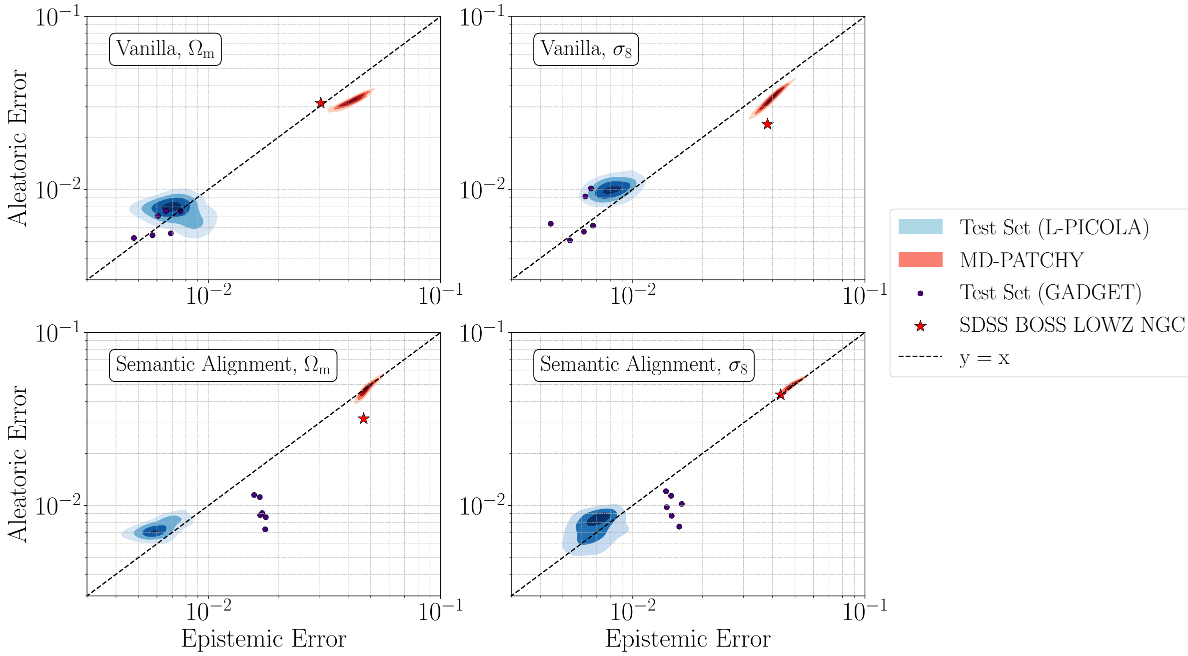

Figure 10. Comparison of *epistemic* and *aleatoric* error in logarithmic scale, from the ensemble prediction of 25 machines, trained with the vanilla scheme (*upper panels*) and the semantic alignment strategy (*lower panels*). The machines are trained in the two mock suites, L-PICOLA +GADGET, with the `SHAM` model. The *left two columns* are the results for $\Omega_{\rm m}$ and the *the right two* are the results for σ_8. *Blue contours* represent the test set samples of L-PICOLA, *red contours* the MD-PATCHY mocks, *dark purple circles* the test set samples of GADGET, and *red stars* the SDSS BOSS LOWZ NGC catalog. *Black dotted lines* depict the complete match between the two types of errors. The epistemic errors are calculated by the standard deviation of the predictions on a single input data from the ensemble of 25 machines. The aleatoric errors are calculated by the root mean square of the predicted errors from individual machines. See Section 6.1 for more information.

seen domain while reducing the discrepancy between the source domains. The discrepancy between the source domains can be explicitly reduced by the semantic alignment as seen from Figure 9, while the discrepancy between the source and the unseen domain can be reduced with the addition of accurate mocks.[9] The vanilla scheme has increased performance on the target sources by distinguishing between the domains, while semantic alignment aligns the distribution at the expense of degraded performance on the target domains. Thus, while domain adaptation shows a significant advantage in that it enables generalization through the alignment of domains, it still suffers from other trade-offs resulting in variability in the machines' end-of-training state, leading to weaker constraints on the cosmological parameters.

[9] Precisely, given multiple sources $\mathscr{D}_S^i$, we define a convex hull $\Lambda_S = \{\bar{\mathscr{D}}|\bar{\mathscr{D}}=\sum_{i=1}^N \pi_i \mathscr{D}_S^i,\ \pi \in \Delta_{N-1}\}$ with Δ_{N-1} being a $N-1$ dimensional simplex. We can then find an optimal distribution $\mathscr{D}^*=\sum_{i=1}^N \pi_i^* \mathscr{D}_S^i$ where π^* minimizes the distance between the optimal distribution $\mathscr{D}^*$ and the target unseen distribution $\mathscr{D}_U$. Therefore, the domain discrepancy between the optimal distribution $\mathscr{D}^*$ and the unseen domain $\mathscr{D}_U$ measured by the $\mathscr{H}$-divergence term ($d_{\mathscr{H}}(\mathscr{D}^*, \mathscr{D}_U)$), and the discrepancy between the two domains inside the convex hull ($\sup_{\mathscr{D}',\mathscr{D}''\in\Lambda_S} d_{\mathscr{H}}(\mathscr{D}', \mathscr{D}'')$) are the two major sources of error. Refer to Albuquerque et al. (2019) and Wang et al. (2023) for more information.

6.2. *Comparison with Previous Studies Using the SDSS BOSS Catalog*

Our simulation-based inference with neural networks, which replaces the use of summary statistics, yields results that can be compared with several notable studies utilizing the SDSS BOSS catalog. This comparison provides a broader context for evaluating the constraints on cosmological parameters. In the following, we compare our results with previous studies that used summary statistics from the full-shape power spectrum and bispectrum, as well as neural network-based approaches.

Compared to the full-shape power spectrum analyses that yield $\Omega_{\rm m} = 0.295 \pm 0.010$ and $\sigma_8 = 0.721 \pm 0.043$ (Ivanov et al. 2020) and the bispectrum analyses yielding $\Omega_{\rm m} = 0.338^{+0.016}_{-0.017}$ and $\sigma_8 = 0.692^{+0.035}_{-0.041}$ (Philcox & Ivanov 2022), our main results from `SHAM` show weaker constraints of $\Omega_{\rm m}$=0.339±0.056 and σ_8=0.801±0.061. However a direct comparison is not possible as our analyses are limited to the BOSS LOWZ NGC sample. In contrast, Ivanov et al. (2020) utilizes the likelihoods combining from the Northern Galactic Cap (NGC) and the Southern Galactic Cap (SGC) across two redshift ranges: *low-z* ($z_{\rm eff} = 0.38$) and *high-z* ($z_{\rm eff} = 0.61$), and Philcox & Ivanov (2022) both NGC and SGC samples from CMASSLOWZTOT, which combine the LOWZ, LOWZE2, LOWZE3 and CMASS catalogs. Al-

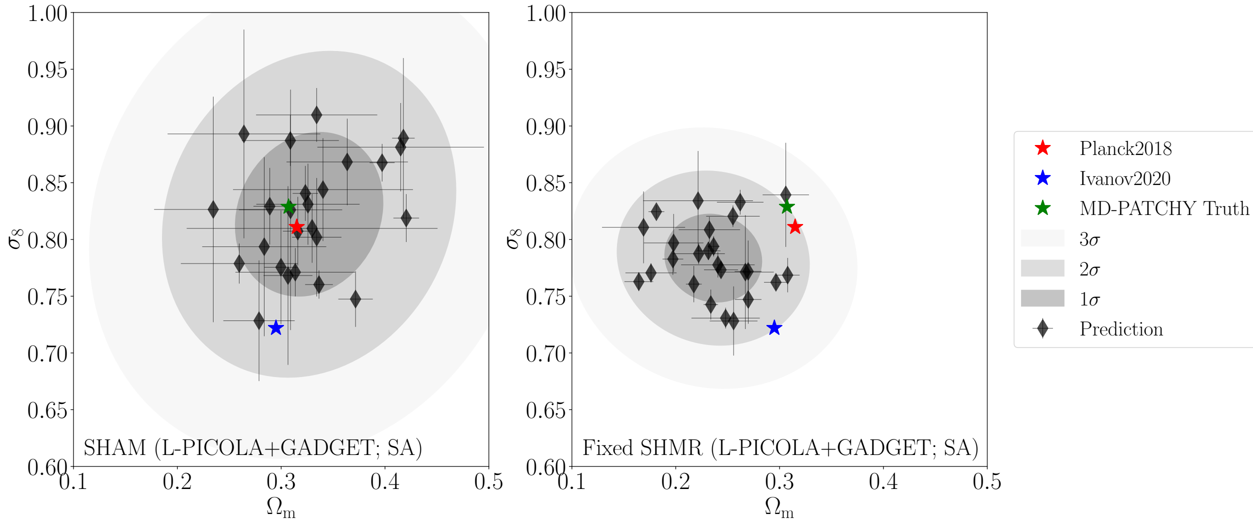

Figure 11. Prediction on the 2048 MD-PATCHY mocks from 25 independently trained `Minkowski-PointNet` machines. The *left* figure displays our results when using the `SHAM` model, and the *right* figure displays the results when using the `Fixed SHMR` model. The machines are trained with L-PICOLA and GADGET mocks with the semantic alignment strategy, a domain adaptation and generalization technique that enables the machines to extract consistent features regardless of their simulation domains (see Section 4.3). Predictions are shown with error bars. A *red star* shows the result from the Planck 2018 (Planck Collaboration et al. 2020) measurements, a *blue star* from Ivanov et al. (2020), and a *green star* the ground truth values of MD-PATCHY mocks. Individual error bars include the statistical error attributed to the cosmic variance of the 2048 MD-PATCHY mocks, which are calculated from the Gaussian Mixture Model (GMM). Individual errors are once again combined by the GMM for the *elliptic contours*, showing the 1σ, 2σ, and 3σ bounds. Our results yield $\Omega_{\rm m}$=0.327±0.070, σ_8=0.822±0.071 (*left, `SHAM`*), and $\Omega_{\rm m}$=0.236±0.046, σ_8=0.784±0.038 (*right, `Fixed SHMR`*). See Section 6.3 for more information.

though our LOWZ NGC mocks differ from the *low-z* definition, having a lower effective redshift of $z_{\rm eff} = 0.29$, the results from the *low-z* NGC used in Ivanov et al. (2020) yield $\Omega_{\rm m} = 0.290 \pm 0.017$ and $\sigma_8 = 0.808 \pm 0.073$.

Next, we compare our results with the recently developed simulation-based inference framework, SIMBIG, which uses BOSS CMASS samples (Hahn et al. 2023a,b; Lemos et al. 2023). Hahn et al. (2023a) used the power spectrum information up to $k_{\rm max}$=0.5h/Mpc together with normalizing flows, resulting in $\Omega_{\rm m}$=$0.292^{+0.055}_{-0.040}$ and σ_8=$0.812^{+0.067}_{-0.068}$. Compared to these results, we obtain a slightly better constraint on σ_8. On the other hand, Hahn et al. (2023b) analyzed the bispectrum monopole up to $k_{\rm max} = 0.5h$/Mpc conducted by using normalizing flows, yielding $\Omega_{\rm m}$=$0.293^{+0.027}_{-0.027}$ and σ_8=$0.783^{+0.040}_{-0.038}$. **[Therefore, (Hahn et al. 2023a) and (Hahn et al. 2023b) explicitly input the cosmological information derived from the clustering statistics at various scales into the machine. In contrast, Lemos et al. (2023) employ a 3D CNN applied to voxelized galaxy positions in real space, effectively capturing clustering characteristics up to $k_{\rm max} = 0.28h$/Mpc. The trained CNN's predictions serve as an intermediate summary statistic, which are then used to generate the final predictions through a flow-based neural network, yielding $\Omega_{\rm m}$=$0.267^{+0.033}_{-0.029}$ and σ_8=$0.762^{+0.036}_{-0.035}$. Our results show both weaker constraints compared to the results.]**

However, our study implements a more direct form of simulation-based inference using the embedding extracted by `Minkowski-PointNet`. As Lemos et al. (2023) point out, such direct inference from neural network embeddings shows weaker constraints. Thus, recent studies consider the predictions of neural networks as summary statistics and perform additional Bayesian inferences (Gupta et al. 2018; Ribli et al. 2019; Fluri et al. 2019; Lemos et al. 2023). Moreover, the major difference in our approach is that we adopt the most conservative form of setting constraints, presenting the ensemble results of 25 individually trained machines instead of a single machine. This highlights the degeneracy of the machines, which show similar performances on known datasets but produce varying predictions on unseen datasets. As mentioned above, using a single machine that is best adapted to the target GADGET samples, we obtain comparably tight constraints of $\Omega_{\rm m}$=0.282±0.014 and σ_8=0.786±0.036.

6.3. *Towards Improved Robustness*

The ultimate goal of replacing summary statistics with raw input from the mock catalogs for the inference of cosmological parameters would be to give tight and accurate constraints. However, since the neural networks capture the complexities engraved in the input data regardless of the physical importance, such methodology involves advantages and disadvantages at the same time. To maximize the advantage, one must consider building machines robust against the choice of domains.

An example of robustness is shown in Figure 11, where our domain-adapted machines are applied to the 2048 MD-PATCHY samples. The results show Ω_m=0.327±0.070 and σ_8=0.822±0.071 for the `SHAM` model, and Ω_m=0.236±0.046 and σ_8=0.784±0.038 for the `Fixed SHMR` model. The uncertainties are increased compared to the prediction results on SDSS BOSS LOWZ NGC catalog, partly due to the cosmic variance of the samples. In particular, the predicted values show differences from the SDSS BOSS LOWZ NGC catalog, despite the high degree of similarity of the MD-PATCHY mocks in the summary statistics. Especially, for the `SHAM` model, the machines correctly predict the lower value of Ω_m and the higher value of σ_8 for MD-PATCHY compared to the observational counterpart assuming Planck 2018 as the ground truth. In contrast to the domain-adapted machines, the vanilla machines yield Ω_m=0.365±0.055 and σ_8=0.875±0.054 for the `SHAM` model, and Ω_m=0.199±0.024 and σ_8=0.715±0.016 for the `Fixed SHMR` model. Again, as we have seen from the prediction results on the SDSS BOSS LOWZ NGC catalog, domain adaptation effectively boosts generalizability at the expense of precision.

To enhance the robustness of neural networks across diverse simulation and observation domains with varying cosmological parameters, we need more samples from the target domains. Currently, insufficient target domain data affects our ability to adapt and generalize effectively, resulting in increased epistemic or model uncertainties, as discussed in Section 6.1. This in turn leads to degraded precision in the final predictions, as shown in figures 7 and 10. Moreover, biases may arise from the discriminative nature of our current neural network model as seen for the GADGET samples in Figure 7. Generative models such as normalizing flows and its variants can be helpful in mitigating such biases and better approximate posterior distributions (Tang & Ting 2022). Addressing these biases is crucial to making reliable inferences in data-driven approaches, as emphasized by Lin et al. (2022).

To accommodate a broader range of cosmological parameters while retaining robustness, not only do we require more sophisticated neural network architectures, but also a focus on the accuracy and correctness of input data. In such data-driven approaches using highly sophisticated neural networks, unreliable input data will distort the extracted domain-agnostic representation. Furthermore, as demonstrated in Section 5.2, achieving both precision and accuracy in individual predictions is critical. By improving domain adaptation strategies and utilizing augmented target data, we can potentially enhance the precise inference of cosmological parameters, especially by focusing on reducing the model uncertainties. We plan to explore this potential further in future work.

6.4. *Limatations and Considerations*

We have demonstrated a proof-of-concept test of inferring cosmological parameters *without* relying on summary statistics, yet there are several limitations and considerations that merit discussion. L-PICOLA mocks, which are our main source domain, show inaccuracies, especially in modeling the halo mass function and small-scale clustering. These inaccuracies are worsened by the simplified assumptions in our galaxy-halo connection models, `SHAM` and `Fixed SHMR`. Our machine learning models, particularly `Minkowski-PointNet`, do not enforce explicit cut-offs, making them sensitive to such inaccuracies. Although we introduced GADGET mocks and performed domain adaptation to address these issues and improved the models' generalizability, this method involves trade-offs in precision.

To tackle these challenges, we suggest several strategies. To begin with, enhancing the flexibility of our galaxy-halo connection models by incorporating additional modeling parameters may improve both accuracy and robustness. Secondly, our target domain samples currently lack diversity in the domains and cosmologies, which might limit the generalizability of our models. Addressing this issue involves considering the inclusion of more mock samples from diverse codes, despite the higher computational costs. Additionally, exploring alternative techniques for domain adaptation and generalization could foster improvements in model performance across various datasets.

[Additionally, it is essential to explore the application of our new methodology to a range of galaxy redshift surveys, which vary in observational effects such as color magnitude cuts, survey depths, completeness, and footprints. Given that our mocks are explicitly modeled to include observational effects unique to the SDSS BOSS LOWZ NGC catalog, our present neural network cannot be applied to other observational surveys. In order to enhance the neural network's robustness against varying observations, we could augment our mock dataset with random cuts and masks, along with modifying radial selection functions. We plan to explore these strategies in our upcoming research.]

7. SUMMARY & CONCLUSION

We propose a novel approach to rapidly model vast quantities of galaxy catalogs through lightcone simulations, while fully incorporating the observational effects of the SDSS BOSS LOWZ NGC catalog and inferring Ω_m and σ_8 from the actual observations using trained neural networks. This addresses the question of whether performing simulation-based inference on observed galaxy redshift surveys using neural networks is feasible in the absence of summary statistics, but only with the position and mass information of individual galaxies. Our method extends previous works that perform "robust field-level inference" on different codes without adopting summary statistics (Shao et al. 2023; de Santi et al. 2023), and works that use summary statistics to infer values from the actual galaxy redshift surveys (Hahn et al. 2023a).

Using lightcone simulation L-PICOLA, we generate 9000 galaxy catalogs with varying cosmological parameters in a volume of $(1.2h^{-1}\mathrm{Gpc})^3$. Subhalos are identified using ROCKSTAR, with each subhalo assumed to host a single galaxy. We propose two models of galaxy-halo connection, `Fixed SHMR` and `SHAM`. The `Fixed SHMR` model assumes a constant star formation efficiency is assumed within a certain halo mass range across different cosmologies, allowing us to identify stellar masses with varying values across different redshift bins. However, the `Fixed SHMR` model suffers from the inclusion of cosmological priors, since they are determined from simulations assuming fiducial cosmology. Therefore, we introduce the `SHAM` model, free of cosmological priors, which paints the halo catalog by assuming a monotonic relation with the observed catalog. The catalogs undergo further processing to mimic the observational effects of the SDSS BOSS LOWZ NGC catalog, including RSD, survey footprint using the MANGLE masks, stellar mass incompleteness (for `Fixed SHMR`), radial selection, and fiber collision (Section 2).

The results and key takeaways are summarized below. *Without* employing summary statistics and using galaxies as point cloud inputs (Section 3), we perform implicit likelihood inference (Jeffrey & Wandelt 2020) and derive constraints on Ω_m and σ_8 from the SDSS BOSS LOWZ NGC sample. Rapidly generated L-PICOLA mock representations can be aligned with the more accurate GADGET mocks to achieve effective domain generalization using the semantic alignment loss (Section 4). Machines trained and adapted independently with L-PICOLA and GADGET mocks infer values of Ω_m=0.227±0.035 and σ_8=0.743±0.039 for the `Fixed SHMR` model and Ω_m=0.339±0.056 and σ_8=0.801±0.061 for the `SHAM` model, when applied to the SDSS BOSS LOWZ NGC catalog. Despite the divergence in the prediction results from the `Fixed SHMR` model, the `SHAM` model, which is free of cosmological priors, agrees with the Planck Collaboration et al. (2020) results within 1σ (Section 5.2 and Figure 8).

Although the constraints highlighted in Section 6.4 exist, we have demonstrated advancements in performing simulation-based inference on observations *without* the use of any summary statistics. This was primarily achieved by adapting across two different code domains, to extract a unified knowledge applicable to real-world observations. Moving forward, we aim to incorporate precise data from various fields and utilize more advanced models to enhance the robustness of our models. This could potentially establish the new method as a competitive approach in precisely constraining cosmological parameters.

Jun-Young Lee would like to thank Aleksandra Ćiprijanović, Francisco Villaescusa-Navarro, Yong-uk Cho, Cullan Howlett, Hyeonyong Kim, Seungjae Lee, Jubee Sohn, Jun Yong Park, and Eun-jin Shin for insightful discussions. He would also like to thank Francisco-Shu Kitaura and Cheng Zhao for providing the MD-PATCHY mocks. Jun-Young Lee's work was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korean Government (Ministry of Education) (P0025681-G02P22450002201-10054408, Semiconductor–Specialized University). Ji-hoon Kim's work was supported by the Global-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301976). His work was also supported by the NRF grant funded by the Korea government (MSIT) (No. 2022M3K3A1093827 and No. 2023R1A2C1003244). His work was also supported by the National Institute of Supercomputing and Network/Korea Institute of Science and Technology Information with supercomputing resources including technical support, grants KSC-2020-CRE-0219, KSC-2021-CRE-0442 and KSC-2022-CRE-0355. Jaehyun Lee is supported by the National Research Foundation of Korea (NRF-2021R1C1C2011626). HSH acknowledges the support by Samsung Electronic Co., Ltd. (Project Number IO220811-01945-01) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT), NRF-2021R1A2C1094577.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

REFERENCES

Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., & Mitliagkas, I. 2019, arXiv e-prints, arXiv:1911.00804, doi: 10.48550/arXiv.1911.00804

Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440, doi: 10.1093/mnras/stz1960

Anchordoqui, L. A., Di Valentino, E., Pan, S., & Yang, W. 2021, Journal of High Energy Astrophysics, 32, 28, doi: 10.1016/j.jheap.2021.08.001

Asgari, M., Lin, C.-A., Joachimi, B., et al. 2021, A&A, 645, A104, doi: 10.1051/0004-6361/202039070

Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013a, ApJ, 770, 57, doi: 10.1088/0004-637X/770/1/57

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013b, ApJ, 762, 109, doi: 10.1088/0004-637X/762/2/109

Ben-David, S., Blitzer, J., Crammer, K., et al. 2010, Machine Learning, 79, 151, doi: 10.1007/s10994-009-5152-4

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. 2006, in Advances in Neural Information Processing Systems, ed. B. Schölkopf, J. Platt, & T. Hoffman, Vol. 19 (MIT Press). https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf

Berlind, A. A., Weinberg, D. H., Benson, A. J., et al. 2003, ApJ, 593, 1, doi: 10.1086/376517

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603, doi: 10.1038/380603a0

Boruah, S. S., Eifler, T., Miranda, V., & Krishanth, P. M. S. 2023, MNRAS, 518, 4818, doi: 10.1093/mnras/stac3417

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. 2021, arXiv e-prints, arXiv:2104.13478, doi: 10.48550/arXiv.2104.13478

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000, doi: 10.1046/j.1365-8711.2003.06897.x

Chabrier, G. 2003, PASP, 115, 763, doi: 10.1086/376392

Choy, C., Gwak, J., & Savarese, S. 2019, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3075–3084

Ćiprijanović, A., Kafkes, D., Jenkins, S., et al. 2020, in 34th Conference on Neural Information Processing Systems. https://arxiv.org/abs/2011.03591

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039, doi: 10.1046/j.1365-8711.2001.04902.x

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, ApJ, 647, 201, doi: 10.1086/503602

Corso, G., Cavalleri, L., Beaini, D., Liò, P., & Veličković, P. 2020, arXiv e-prints, arXiv:2004.05718, doi: 10.48550/arXiv.2004.05718

Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, MNRAS, 453, 1513, doi: 10.1093/mnras/stv1708

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371, doi: 10.1086/163168

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10, doi: 10.1088/0004-6256/145/1/10

de Lapparent, V., Geller, M. J., & Huchra, J. P. 1986, ApJL, 302, L1, doi: 10.1086/184625

de Santi, N. S. M., Shao, H., Villaescusa-Navarro, F., et al. 2023, ApJ, 952, 69, doi: 10.3847/1538-4357/acd1e2

DES Collaboration, Abbott, T. M. C., Acevedo, M., et al. 2024, arXiv e-prints, arXiv:2401.02929, doi: 10.48550/arXiv.2401.02929

Ding, J., Li, S., Zheng, Y., et al. 2023, arXiv e-prints, arXiv:2311.00981, doi: 10.48550/arXiv.2311.00981

Dong-Páez, C. A., Smith, A., Szewciw, A. O., et al. 2022, arXiv e-prints, arXiv:2208.00540, doi: 10.48550/arXiv.2208.00540

Eisenstein, D. J., Annis, J., Gunn, J. E., et al. 2001, AJ, 122, 2267, doi: 10.1086/323717

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72, doi: 10.1088/0004-6256/142/3/72

Fluri, J., Kacprzak, T., Lucchi, A., et al. 2019, PhRvD, 100, 063514, doi: 10.1103/PhysRevD.100.063514

Fluri, J., Kacprzak, T., Lucchi, A., et al. 2022, Phys. Rev. D, 105, 083518, doi: 10.1103/PhysRevD.105.083518

Fluri, J., Kacprzak, T., Refregier, A., et al. 2018, Phys. Rev. D, 98, 123518, doi: 10.1103/PhysRevD.98.123518

Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015a, MNRAS, 448, 2987, doi: 10.1093/mnras/stv138

Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2015b, MNRAS, 447, 1319, doi: 10.1093/mnras/stu2464

Ganin, Y., Ustinova, E., Ajakan, H., et al. 2016, Journal of Machine Learning Research, 17, 1. http://jmlr.org/papers/v17/15-239.html

Geller, M. J., & Huchra, J. P. 1989, Science, 246, 897, doi: 10.1126/science.246.4932.897

Giocoli, C., Baldi, M., & Moscardini, L. 2018, Monthly Notices of the Royal Astronomical Society, 481, 2813, doi: 10.1093/mnras/sty2465

Girelli, G., Pozzetti, L., Bolzonella, M., et al. 2020, A&A, 634, A135, doi: 10.1051/0004-6361/201936329

Giri, U., Münchmeyer, M., & Smith, K. M. 2023, Phys. Rev. D, 107, L061301, doi: 10.1103/PhysRevD.107.L061301

Guo, H., Zehavi, I., & Zheng, Z. 2012, ApJ, 756, 127, doi: 10.1088/0004-637X/756/2/127

Guo, H., Zheng, Z., Behroozi, P. S., et al. 2016, Monthly Notices of the Royal Astronomical Society, 459, 3040, doi: 10.1093/mnras/stw845

Gupta, A., Matilla, J. M. Z., Hsu, D., & Haiman, Z. 2018, Phys. Rev. D, 97, 103515, doi: 10.1103/PhysRevD.97.103515

Hahn, C., & Villaescusa-Navarro, F. 2021, JCAP, 2021, 029, doi: 10.1088/1475-7516/2021/04/029

Hahn, C., Eickenberg, M., Ho, S., et al. 2023a, JCAP, 2023, 010, doi: 10.1088/1475-7516/2023/04/010

Hahn, C., Lemos, P., Parker, L., et al. 2023b, arXiv e-prints, arXiv:2310.15246, doi: 10.48550/arXiv.2310.15246

Hand, N., Feng, Y., Beutler, F., et al. 2018, AJ, 156, 160, doi: 10.3847/1538-3881/aadae0

Hikage, C., Oguri, M., Hamana, T., et al. 2019, Publications of the Astronomical Society of Japan, 71, 43, doi: 10.1093/pasj/psz010

Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454, doi: 10.1093/mnras/stw2805

Hortúa, H. J., García, L. Á., & Castañeda C., L. 2023, Frontiers in Astronomy and Space Sciences, 10, 1139120, doi: 10.3389/fspas.2023.1139120

Howlett, C., Manera, M., & Percival, W. J. 2015a, Astronomy and Computing, 12, 109, doi: 10.1016/j.ascom.2015.07.003

Howlett, C., Ross, A. J., Samushia, L., Percival, W. J., & Manera, M. 2015b, MNRAS, 449, 848, doi: 10.1093/mnras/stu2693

Howlett, C., Said, K., Lucey, J. R., et al. 2022, Monthly Notices of the Royal Astronomical Society, 515, 953, doi: 10.1093/mnras/stac1681

Huchra, J., Davis, M., Latham, D., & Tonry, J. 1983, ApJS, 52, 89, doi: 10.1086/190860

Huertas-Company, M., Iyer, K. G., Angeloudi, E., et al. 2023, arXiv e-prints, arXiv:2305.02478, doi: 10.48550/arXiv.2305.02478

Hwang, H. S., Geller, M. J., Park, C., et al. 2016, ApJ, 818, 173, doi: 10.3847/0004-637X/818/2/173

Hwang, S. Y., Sabiu, C. G., Park, I., & Hong, S. E. 2023, The Universe is worth 64^3 pixels: Convolution Neural Network and Vision Transformers for Cosmology. https://arxiv.org/abs/2304.08192

Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55, doi: 10.1051/0004-6361/201321100

Ioffe, S., & Szegedy, C. 2015, in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15 (JMLR.org), 448–456

Ishikawa, S., Okumura, T., & Nishimichi, T. 2023, arXiv e-prints, arXiv:2308.03871, doi: 10.48550/arXiv.2308.03871

Ivanov, M. M., Simonović, M., & Zaldarriaga, M. 2020, Journal of Cosmology and Astroparticle Physics, 2020, 042, doi: 10.1088/1475-7516/2020/05/042

Jeffrey, N., & Wandelt, B. D. 2020, in 34th Conference on Neural Information Processing Systems. https://arxiv.org/abs/2011.05991

Jo, Y., & Kim, J.-h. 2019, Monthly Notices of the Royal Astronomical Society, 489, 3565, doi: 10.1093/mnras/stz2304

Jo, Y., Genel, S., Wandelt, B., et al. 2023, ApJ, 944, 67, doi: 10.3847/1538-4357/aca8fe

Kacprzak, T., & Fluri, J. 2022, Physical Review X, 12, 031029, doi: 10.1103/PhysRevX.12.031029

Kaiser, N. 1987, MNRAS, 227, 1, doi: 10.1093/mnras/227.1.1

Kim, J., Park, C., & Choi, Y.-Y. 2008, ApJ, 683, 123, doi: 10.1086/589566

Kingma, D., & Ba, J. 2014, International Conference on Learning Representations

Kitaura, F.-S., & Heß, S. 2013, Monthly Notices of the Royal Astronomical Society: Letters, 435, L78, doi: 10.1093/mnrasl/slt101

Kitaura, F.-S., Yepes, G., & Prada, F. 2013, Monthly Notices of the Royal Astronomical Society: Letters, 439, L21, doi: 10.1093/mnrasl/slt172

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, MNRAS, 456, 4156, doi: 10.1093/mnras/stv2826

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, MNRAS, 457, 4340, doi: 10.1093/mnras/stw248

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, ApJ, 609, 35, doi: 10.1086/420959

Kreisch, C. D., Pisani, A., Villaescusa-Navarro, F., et al. 2022, ApJ, 935, 100, doi: 10.3847/1538-4357/ac7d4b

Kroupa, P. 2001, MNRAS, 322, 231, doi: 10.1046/j.1365-8711.2001.04022.x

Lazanu, A. 2021, Journal of Cosmology and Astroparticle Physics, 09, 039, doi: 10.1088/1475-7516/2021/09/039

Leauthaud, A., Bundy, K., Saito, S., et al. 2016, Monthly Notices of the Royal Astronomical Society, 457, 4021, doi: 10.1093/mnras/stw117

Leja, J., Speagle, J. S., Johnson, B. D., et al. 2020, ApJ, 893, 111, doi: 10.3847/1538-4357/ab7e27

Leja, J., Speagle, J. S., Ting, Y.-S., et al. 2022, ApJ, 936, 165, doi: 10.3847/1538-4357/ac887d

Lemos, P., Parker, L. H., Hahn, C., et al. 2023, in Machine Learning for Astrophysics, 18, doi: 10.48550/arXiv.2310.15256

Lin, Q., Fouchez, D., Pasquet, J., et al. 2022, A&A, 662, A36, doi: 10.1051/0004-6361/202142751

Lu, T., Haiman, Z., & Li, X. 2023, Monthly Notices of the Royal Astronomical Society, 521, 2050, doi: 10.1093/mnras/stad686

Maraston, C., Pforr, J., Henriques, B. M., et al. 2013, MNRAS, 435, 2764, doi: 10.1093/mnras/stt1424

Mathuriya, A., Bard, D., Mendygral, P., et al. 2019, in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC '18 (IEEE Press), doi: 10.1109/SC.2018.00068

Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. 2017, in Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Neutsch, S., Heneka, C., & Brüggen, M. 2022, MNRAS, 511, 3446, doi: 10.1093/mnras/stac218

Ni, Y., Genel, S., Anglés-Alcázar, D., et al. 2023, arXiv e-prints, arXiv:2304.02096, doi: 10.48550/arXiv.2304.02096

Ntampaka, M., Eisenstein, D. J., Yuan, S., & Garrison, L. H. 2020, ApJ, 889, 151, doi: 10.3847/1538-4357/ab5f5e

Pan, S., Liu, M., Forero-Romero, J., et al. 2020, Science China Physics, Mechanics & Astronomy, 63, doi: 10.1007/s11433-020-1586-3

Peacock, J. A., & Smith, R. E. 2000, MNRAS, 318, 1144, doi: 10.1046/j.1365-8711.2000.03779.x

Peebles, P. J. E. 1981, The Large-Scale Structure of the Universe (Princeton: Princeton University Press), doi: doi:10.1515/9780691206714

Perez, L. A., Genel, S., Villaescusa-Navarro, F., et al. 2022, arXiv e-prints, arXiv:2204.02408, doi: 10.48550/arXiv.2204.02408

Philcox, O. H. E., & Ivanov, M. M. 2022, Phys. Rev. D, 105, 043517, doi: 10.1103/PhysRevD.105.043517

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6, doi: 10.1051/0004-6361/201833910

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. 2016, arXiv preprint arXiv:1612.00593

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. 2017, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 5105–5114

Qiu, L., Napolitano, N. R., Borgani, S., et al. 2023, Cosmology with Galaxy Cluster Properties using Machine Learning. https://arxiv.org/abs/2304.09142

Ravanbakhsh, S., Oliva, J. B., Fromenteau, S., et al. 2016, in International Conference on Machine Learning. https://api.semanticscholar.org/CorpusID:2360303

Reddick, R. M., Wechsler, R. H., Tinker, J. L., & Behroozi, P. S. 2013, ApJ, 771, 30, doi: 10.1088/0004-637X/771/1/30

Reid, B., Ho, S., Padmanabhan, N., et al. 2016, MNRAS, 455, 1553, doi: 10.1093/mnras/stv2382

Ribli, D., Pataki, B. A., Zorrilla Matilla, J. M., et al. 2019, Monthly Notices of the Royal Astronomical Society, 490, 1843, doi: 10.1093/mnras/stz2610

Rodríguez-Torres, S. A., Chuang, C.-H., Prada, F., et al. 2016, MNRAS, 460, 1173, doi: 10.1093/mnras/stw1014

Roncoli, A., Ćiprijanović, A., Voetberg, M., Villaescusa-Navarro, F., & Nord, B. 2023, arXiv e-prints, arXiv:2311.01588, doi: 10.48550/arXiv.2311.01588

Ronconi, T., Lapi, A., Viel, M., & Sartori, A. 2020, Monthly Notices of the Royal Astronomical Society, 498, 2095, doi: 10.1093/mnras/staa2201

Saito, S., Leauthaud, A., Hearin, A. P., et al. 2016, Monthly Notices of the Royal Astronomical Society, 460, 1457, doi: 10.1093/mnras/stw1080

Scoccimarro, R., Hui, L., Manera, M., & Chan, K. C. 2012, PhRvD, 85, 083002, doi: 10.1103/PhysRevD.85.083002

Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1, doi: 10.1086/516585

Shao, H., Villaescusa-Navarro, F., Villanueva-Domingo, P., et al. 2023, ApJ, 944, 27, doi: 10.3847/1538-4357/acac7a

Simha, V., & Cole, S. 2013, Mon. Not. Roy. Astron. Soc., 436, 1142, doi: 10.1093/mnras/stt1643

Sohn, J., Geller, M. J., Hwang, H. S., et al. 2023, ApJ, 945, 94, doi: 10.3847/1538-4357/acb925

Springel, V. 2005, MNRAS, 364, 1105, doi: 10.1111/j.1365-2966.2005.09655.x

—. 2015, N-GenIC: Cosmological structure initial conditions, Astrophysics Source Code Library, record ascl:1502.003

Springel, V., Pakmor, R., Zier, O., & Reinecke, M. 2021, Monthly Notices of the Royal Astronomical Society, 506, 2871, doi: 10.1093/mnras/stab1855

Swanson, M. E. C., Tegmark, M., Hamilton, A. J. S., & Hill, J. C. 2008, Monthly Notices of the Royal Astronomical Society, 387, 1391, doi: 10.1111/j.1365-2966.2008.13296.x

Tang, K. S., & Ting, Y.-S. 2022, in Machine Learning for Astrophysics, 13, doi: 10.48550/arXiv.2207.02786

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, JCAP, 2013, 036, doi: 10.1088/1475-7516/2013/06/036

Tojeiro, R., Ross, A. J., Burden, A., et al. 2014, MNRAS, 440, 2222, doi: 10.1093/mnras/stu371

van der Maaten, L., & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579. http://jmlr.org/papers/v9/vandermaaten08a.html

Veronesi, N., Marulli, F., Veropalumbo, A., & Moscardini, L. 2023, Astronomy and Computing, 42, 100692, doi: 10.1016/j.ascom.2023.100692

Villaescusa-Navarro, F., Hahn, C., Massara, E., et al. 2020, ApJS, 250, 2, doi: 10.3847/1538-4365/ab9d82

Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., et al. 2022, arXiv e-prints, arXiv:2201.01300. https://arxiv.org/abs/2201.01300

Villaescusa-Navarro, F., Ding, J., Genel, S., et al. 2022, The Astrophysical Journal, 929, 132, doi: 10.3847/1538-4357/ac5d3f

Villanueva-Domingo, P., & Villaescusa-Navarro, F. 2022, ApJ, 937, 115, doi: 10.3847/1538-4357/ac8930

Villanueva-Domingo, P., Villaescusa-Navarro, F., Anglés-Alcázar, D., et al. 2022, The Astrophysical Journal, 935, 30, doi: 10.3847/1538-4357/ac7aa3

Wang, J., Lan, C., Liu, C., et al. 2023, IEEE Transactions on Knowledge & Data Engineering, 35, 8052, doi: 10.1109/TKDE.2022.3178128

Wechsler, R. H., & Tinker, J. L. 2018, Annual Review of Astronomy and Astrophysics, 56, 435, doi: 10.1146/annurev-astro-081817-051756

White, M., Tinker, J. L., & McBride, C. K. 2014a, MNRAS, 437, 2594, doi: 10.1093/mnras/stt2071

—. 2014b, MNRAS, 437, 2594, doi: 10.1093/mnras/stt2071

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579, doi: 10.1086/301513

Zaheer, M., Kottur, S., Ravanbakhsh, S., et al. 2017, in Advances in Neural Information Processing Systems, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, Vol. 30 (Curran Associates, Inc.). https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf

Zhao, C., Kitaura, F.-S., Chuang, C.-H., et al. 2015, Monthly Notices of the Royal Astronomical Society, 451, 4266, doi: 10.1093/mnras/stv1262

APPENDIX

A. FEATURES OF REALIZATIONS WITH DIFFERENT COSMOLOGICAL PARAMETERS

Features of individual and neighboring galaxies differ across realizations with varying cosmological parameters. In Figure 12, we provide the same pair plot as Figure 6 but for the `Fixed SHMR` model of the L-PICOLA mock suite with different cosmologies: `high` ($\Omega_{\rm m}$=0.4772, σ_8=0.9639), `low` ($\Omega_{\rm m}$=0.1185, σ_8=0.6163), and `fiducial` ($\Omega_{\rm m}$=0.3067, σ_8=0.8238). Notice that `low` deviates the most from `fiducial`, while `high` shows a better agreement in all features. This tendency becomes most extreme for distances to neighboring galaxies. This is due to the deficit of the total number of galaxies for `low`, which severely affects the separation between the galaxies. Although not displayed for brevity, the `SHAM` models exhibit consistency despite differences in the cosmological parameters. Such behavior arises from the fact that, in contrast to the `Fixed SHMR` model, the `SHAM` model matches the total galaxy count of the mocks to the SDSS BOSS LOWZ NGC catalog.

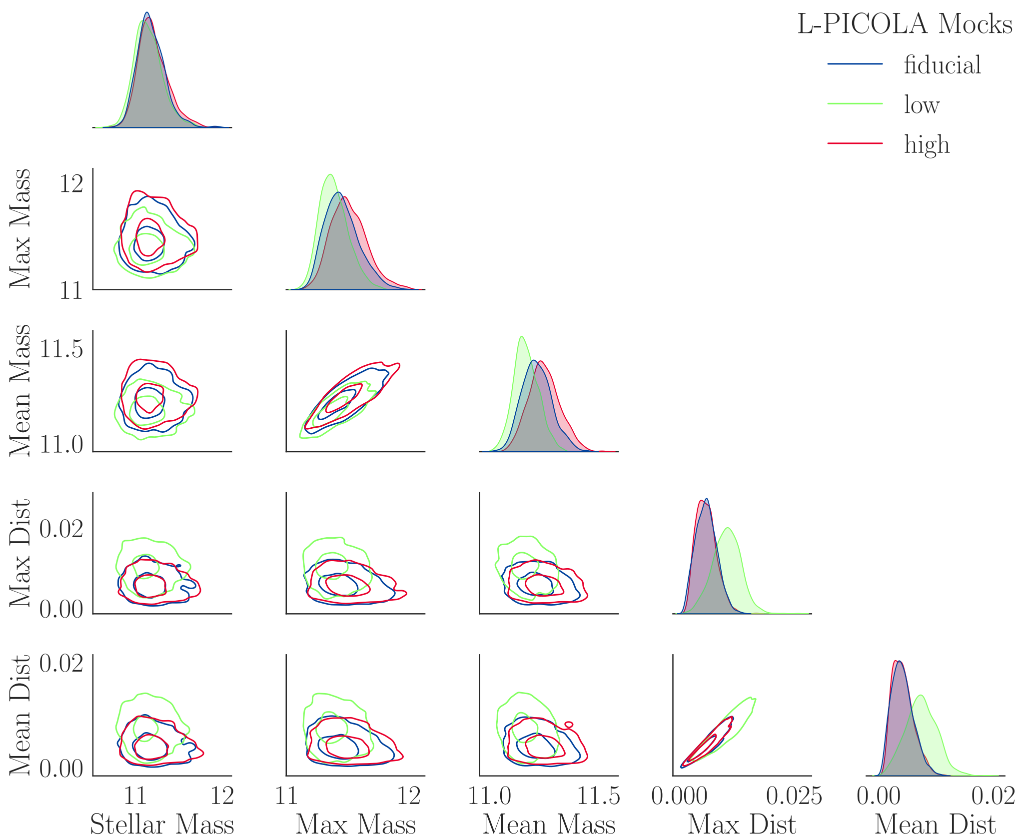

Figure 12. Pair plot of five features of a galaxy randomly sampled from each mock generated with the `Fixed SHMR` model, similar to Figure 6, but this time for three realizations of different cosmologies named `high` (*red*; $\Omega_{\rm m}$=0.4772, σ_8=0.9639), `low` (*green*; $\Omega_{\rm m}$=0.1185, σ_8=0.6163) and `fiducial` (*blue*; $\Omega_{\rm m}$=0.3067, σ_8=0.8238). The plot shows for 1000 randomly sampled galaxies for each mocks. Masses are in units of $\log(M_\star/h^{-1}\mathrm{M}_\odot)$ and distances are expressed in terms of the newly assumed metric in redshift space $(X,Y,Z) = (z\sin(DEC)\cos(RA), z\sin(DEC)\sin(RA), z\cos(DEC))$. See Appendix A for more information.

B. EFFECT OF FINE-TUNED MOCKS MD-PATCHY

We further investigate the possibility of increasing the accuracy and precision via the incorporation of fine-tuned MD-PATCHY mock samples. Similarly to machines trained with L-PICOLA and GADGET mocks, we train 25 different machines using L-PICOLA and MD-PATCHY with the semantic alignment loss applied. As shown in Figure 13, the results yield $\Omega_{\rm m}$=0.307$\pm$0.035 and σ_8=0.767$\pm$0.035 for the `Fixed SHMR` model, and $\Omega_{\rm m}$=0.343$\pm$0.053 and σ_8=0.796$\pm$0.051 for the `SHAM` model. Compared to

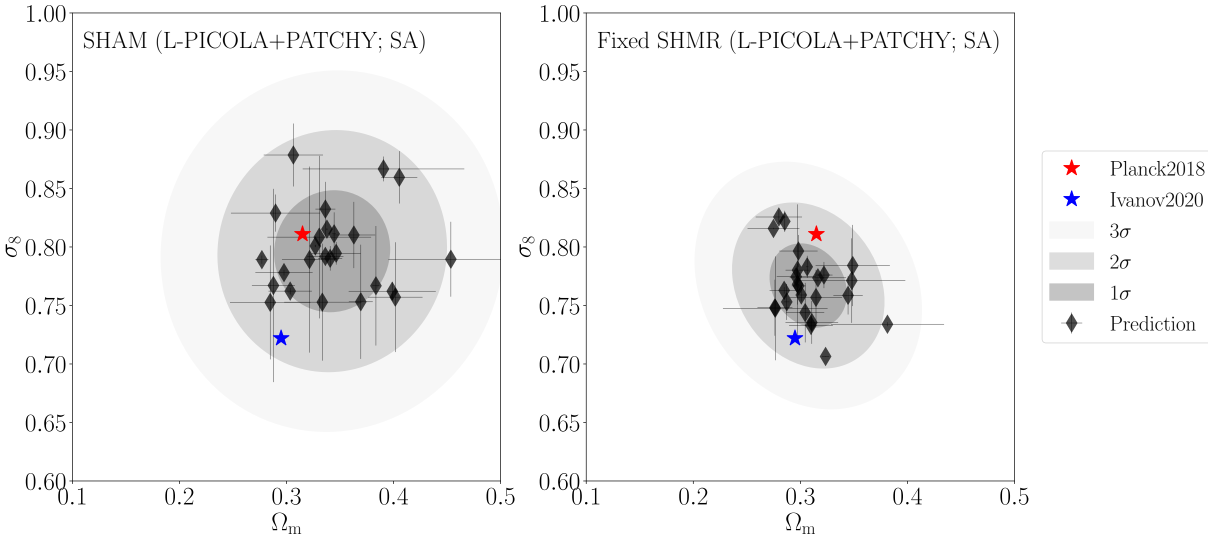

Figure 13. Prediction on the actual SDSS BOSS LOWZ NGC catalog from the ensemble of 25 independently trained `Minkowski-PointNet` machines. The *left* figure displays our results when using the `SHAM` model, and the *right* figure displays the results when using the `Fixed SHMR` model. The machines are trained with L-PICOLA and MD-PATCHY mocks without domain adaptation strategy (Vanilla), a domain adaptation and generalization technique that enables the machines to extract consistent features regardless of their simulation domains (see Section 4.3). Predictions are shown with error bars. A *red star* shows the result from the Planck 2018 (Planck Collaboration et al. 2020) measurements and a *blue star* from Ivanov et al. (2020). *Elliptic contours* show the 1σ, 2σ, and 3σ bounds, calculated from the Gaussian Mixture Model (GMM) to incorporate the individual errors. Our results yield Ω_m=0.343±0.053, σ_8=0.796±0.051 (*left, SHAM*), and Ω_m=0.307±0.035, σ_8=0.767±0.035 (*left, Fixed SHMR*). See Section 6.1 for more information.

when applying the GADGET mocks, better precision is achieved for both galaxy-halo connection models. Moreover, especially for the `Fixed SHMR` model, the accuracy drastically increases. Indeed, such behavior is well expected, as the machine can learn from the fine-tuned mocks, which better depict the observational sample.

Semantic alignment loss plays an explicit role in reducing the divergence of representations originating from different domains. For example, aligning the representations of $\mathscr{D}_{\mathrm{L-PICOLA}}$ and $\mathscr{D}_{\mathrm{MD-PATCHY}}$ to be close enough, adding MD-PATCHY mock samples will have a small impact on increasing the diameter of the convex hull of the domains. Moreover, assuming that the marginal distribution of MD-PATCHY is relatively similar to the SDSS BOSS LOWZ NGC catalog, the optimal domain, $\mathscr{D}^*$, will be weighted towards $\mathscr{D}_{\mathrm{MD-PATCHY}}$ and will effectively reduce the generalization risk. Therefore, this confirms not only the importance of aligning the representations from different domains but also the inclusion of accurate mocks involved in the training phase. This effect is maximized for the `Fixed SHMR` model, where the initially biased prediction, when trained with the L-PICOLA and GADGET domains, significantly alters to produce more accurate results assuming the Planck 2018 cosmology as the ground truth.

However, since MD-PATCHY mocks are based on a single cosmology, generalization is only effective locally. To train the machines to be globally robust, it is necessary that a multitude of high-fidelity mocks with diverse cosmologies are included as in Section 6.1. Such inclusion must be made across varying cosmological parameters, unlike the fine-tuned mocks with a single targeted value, as generalization is only performed locally in this case. We leave these aspects of improvement for future work.

C. ALTERNATIVE TRAINING STRATEGY: DOMAIN ADVERSARIAL TRAINING

An alternative training strategy for domain adaptation and generalization is to extract domain-invariant features through adversarial training. The essence of such a training strategy is to prevent the machine from learning domain-specific information. Here, we employ domain adversarial neural networks (DANN; Ganin et al. 2016), which adds a domain classifier to the backbone of the machine illustrated in Figure 5. The domain classifier is trained to classify whether the input originates from L-PICOLA mocks or MD-PATCHY mocks. Moreover, the preceding gradient reversal layer (GRL) enables forward propagation of the domain loss to the feature extractor. Consequently, the feature extractor weights are updated to produce domain-invariant features sufficient to deceive the domain classifier.

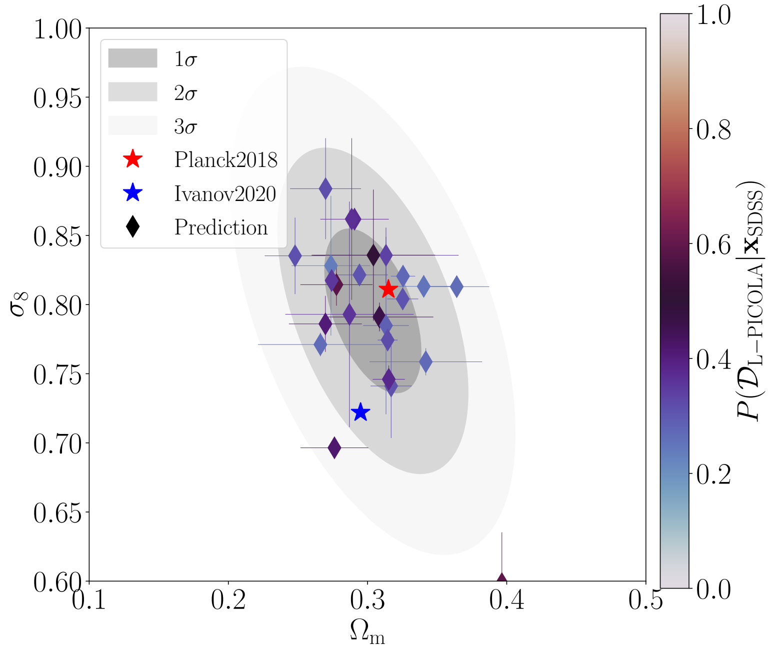

Figure 14. Prediction on the actual SDSS BOSS LOWZ NGC catalog from 25 independently trained `Minkowski-PointNet` machines, similar to Figure 8, but this time with the domain adversarial neural networks (DANN; Ganin et al. 2016) instead of the semantic alignment strategy. Predictions with error bars are shown and in different colors indicating the probability that the domain classifier classifies as L-PICOLA, $P(\mathscr{D}_{\text{L-PICOLA}}|\mathbf{x}_{\text{SDSS}})$. A *red star* shows the result from the Planck 2018 (Planck Collaboration et al. 2020) measurements and a *blue star* from Ivanov et al. (2020). *Elliptic contours* show the bounds of 1σ, 2σ, and 3σ bounds. The results yield Ω_{m}=0.304±0.033 and σ_8=0.795±0.057. See Appendix C for more information.

In this approach, we leverage the DANN strategy to perform regression tasks in a supervised domain adaptation setup using L-PICOLA and MD-PATCHY mocks. The loss function of the supervised DANN setup can be mathematically expressed as follows:

$$\begin{aligned} L(\theta_f, \theta_r, \theta_d; \mathbf{x}) = & L_{\text{vanilla}}(G_r(\theta_r; (G_f(\theta_f; \mathbf{x}))), \mathbf{y}) \\ & + \alpha L_{\text{domain}}(G_d(\theta_d; \mathscr{R}((G_f(\theta_f; \mathbf{x})))), d) \end{aligned} \tag{C1}$$

where θ_f, θ_r, θ_d denote the parameters and $G_f(\theta_f, \cdot)$, $G_r(\theta_r, \cdot)$, $G_d(\theta_d, \cdot)$ represent the function of the feature extractor, regressor, and domain classifier. Here, $\mathbf{x}$ represents the input, $\mathbf{y}$ represents the cosmological parameters, and d represents the domain. The GRL $\mathscr{R}(\mathbf{x})$ is a pseudo-function with properties $\mathscr{R}(\mathbf{x}) = \mathbf{x}$ and $\mathscr{R}'(\mathbf{x}) = -\mathbf{I}$. Introducing GRL reduces the DANN setup to a single minimization problem.

The terminal layer of the domain classifier passes through a sigmoid activation function, classifying input as L-PICOLA ("1") or MD-PATCHY ("0") based on a threshold of 0.5. The domain confusion loss L_{domain} is calculated using the binary cross-entropy loss with logits, accounting for the imbalance in the size of the data set between each domain. After training, we further train new domain classifiers, each with two trainable layers, for every machine while keeping the weights of the feature extractor frozen. This process allows us to evaluate the classifiability of the extracted features.

Figure 14 displays the results of the 25 independently trained DANN machines. Individual predictions are colored based on their probabilities as classified by the domain classifier, indicating whether they originated from L-PICOLA, denoted as $P(\mathscr{D}_{\text{L-PICOLA}}|\mathbf{x}_{\text{SDSS}})$. The results show Ω_{m}=0.304±0.033 and σ_8=0.795±0.057. However, we observe that compared to the semantic alignment strategy, the distribution between the two domains is not effectively reduced, making it susceptible to overfitting. Consequently, the adequacy of the training scheme can vary depending on the characteristics of the sources and targets and must be used judiciously.