Draft version July 20, 2023
Typeset using LaTeX **preprint** style in AASTeX631

Robust field-level likelihood-free inference with galaxies

Natalí S. M. de Santi [1,2] Helen Shao [3] Francisco Villaescusa-Navarro [1,3]
L. Raul Abramo [2] Romain Teyssier [3] Pablo Villanueva-Domingo [4] Yueying Ni [5,6]
Daniel Anglés-Alcázar [7,1] Shy Genel [1,8] Elena Hernandez-Martinez [9]
Ulrich P. Steinwandel [1] Christopher C. Lovell [10,11] Klaus Dolag,[9,12]
Tiago Castro [13,14,15] and Mark Vogelsberger[16,17]

[1] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY, 10010, USA*
[2] *Instituto de Física, Universidade de São Paulo, R. do Matão 1371, 05508-900, São Paulo, Brasil*
[3] *Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544 USA*
[4] *Computer Vision Center - Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Barcelona, Spain*
[5] *Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, US*
[6] *McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, US*
[7] *Department of Physics, University of Connecticut, 196 Auditorium Road, U-3046, Storrs, CT, 06269, USA*
[8] *Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, New York, NY 10027, USA*
[9] *Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 München, Germany*
[10] *Institute of Cosmology and Gravitation, University of Portsmouth, Burnaby Road, Portsmouth, PO1 3FX, UK*
[11] *Centre for Astrophysics Research, School of Physics, Engineering & Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK*
[12] *Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85741 Garching, Germany*
[13] *INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[14] *INFN, Sezione di Trieste, Via Valerio 2, I-34127 Trieste TS, Italy*
[15] *IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy*
[16] *Kavli Institute for Astrophysics and Space Research, Department of Physics, MIT, Cambridge, MA 02139, USA*
[17] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Massachusetts Institute of Technology, Cambridge MA 02139, USA*

ABSTRACT

We train graph neural networks to perform field-level likelihood-free inference using galaxy catalogs from state-of-the-art hydrodynamic simulations of the CAMELS project. Our models are rotational, translational, and permutation invariant and do not impose any cut on scale. From galaxy catalogs that only contain 3D positions and radial velocities of $\sim 1,000$ galaxies in tiny $(25\ h^{-1}\mathrm{Mpc})^3$ volumes our models can infer the value of Ω_m with approximately 12% precision. More importantly, by testing the models on galaxy catalogs from thousands of hydrodynamic simulations, each having a different efficiency of supernova and AGN feedback, run with five different codes and subgrid models – IllustrisTNG, SIMBA, Astrid, Magneticum, SWIFT-EAGLE –, we find that our models are robust to changes in astrophysics, subgrid physics, and sub-

Corresponding author: Natalí S. M. de Santi
natalidesanti@gmail.com

halo/galaxy finder. Furthermore, we test our models on 1,024 simulations that cover a vast region in parameter space – variations in 5 cosmological and 23 astrophysical parameters – finding that the model extrapolates really well. Our results indicate that the key to building a robust model is the use of both galaxy positions and velocities, suggesting that the network have likely learned an underlying physical relation that does not depend on galaxy formation and is valid on scales larger than $\sim 10\ h^{-1}$kpc.

Keywords: magnetohydrodynamics (MHD) – galaxies: statistics — cosmology: cosmological parameters — methods: statistics

1. INTRODUCTION

The standard model of Cosmology describes a Universe filled with dark matter (DM), baryonic matter and some form of dark energy (DE). Despite many observational constraints, such as the temperature and polarization fluctuations in the cosmic microwave background (Bennett et al. 2013; Planck Collaboration et al. 2020), many mysteries still remain, in particular, the fundamental natures of DE and DM. In order to solve the remaining puzzles and consolidate this physical description, cosmologists aim at constraining, with the highest precision and accuracy possible, the parameters of the model.

Since the distribution of matter and galaxies in the Universe depends on the cosmological parameters, the clustering of these objects can be used to infer the values of those parameters. In order to collect as much data as diversely as possible large international efforts are currently underway to survey the cosmos at different wavelengths: DESI (DESI Collaboration et al. 2016), Euclid (Laureijs et al. 2011; Amendola et al. 2013; Racca et al. 2016; Euclid Collaboration: Castro et al. 2022), Prime Focus Spectrograph (PFS) (Takada et al. 2014), J-PAS (Benitez et al. 2014), Square Kilometer Array (SKA) (Taylor & Braun 1999), Roman (Spergel et al. 2015), JWST (Pontoppidan et al. 2022), and others. The data from these missions will encompass larger volumes at different redshifts, using a variety of different types of galaxies, observed at many wavelengths. Extracting the maximum amount of relevant information from these data sets is of key importance in order to improve our understanding of fundamental physics.

To achieve that goal theoretical predictions and methods to extract that information are needed. On the one hand, we have traditional methods for extracting information from cosmological observations. In the case of Bayesian inference schemes for cosmological parameters, nearly all analyses make use of summary statistics, like the power spectrum. However, this approach is sub-optimal as we do not know what summary statistics contain all (or the majority) of the cosmological information (Hahn et al. 2020; Uhlemann et al. 2020; Gualdi et al. 2021; Banerjee & Abel 2021). Furthermore, usual methods normally require expensive simulations to either estimate covariance matrices or to forward-model the observations (Efron 1982; Taylor et al. 2013; Heavens et al. 2017; Chartier & Wandelt 2022; de Santi & Abramo 2022).

On the other hand, machine learning (ML) techniques have been shown to outperform traditional methods in a large variety of contexts and areas, including cosmology and astrophysics. In fact, the power of these new methods resides precisely in their ability to deal with large and complex data sets, providing nonlinear relations in high-dimensional feature spaces that allow us to solve regression and classification tasks (Ivezić et al. 2014). Using different summary statistics as input data, Perez

et al. (2022) are able to derive cosmological parameters without the need for additional input from theoretical models, thus providing a powerful generalization of the usual Monte Carlo-based methods. In particular, likelihood-free inference methods work by taking data directly from the simulations (without the need for summary statistics and, thus, model comparison), and many papers have shown competitive results compared with the usual statistical inference methods (Ravanbakhsh et al. 2017; Ntampaka et al. 2020; Mangena et al. 2020; Hassan et al. 2020; Villaescusa-Navarro et al. 2021a; Cole et al. 2022; Villanueva-Domingo & Villaescusa-Navarro 2022; Makinen et al. 2022; Shao et al. 2022a). At a level closer to the observations and simulations, many papers exploring the halo–galaxy connection are able to make predictions that are comparable to the output of numerical/analytical methods (Jo & Kim 2019; Yip et al. 2019; Zhang et al. 2019; Kamdar et al. 2016; Wadekar et al. 2020; Kasmanoff et al. 2020; Moster et al. 2021; McGibbon & Khochfar 2022; Shao et al. 2022b; von Marttens et al. 2022; Villanueva-Domingo et al. 2021; Delgado et al. 2022; de Santi et al. 2022; Jespersen et al. 2022; Villanueva-Domingo et al. 2022; Lovell et al. 2022; Rodrigues et al. 2023). Furthermore, a clear advantage of ML models is that, once they are trained, they typically make predictions much faster than traditional methods (Jespersen et al. 2022) and a disadvantage arises when these models fail to extrapolate their predictions across different data sets, from the ones with which they were trained with (Villaescusa-Navarro et al. 2021a; Villanueva-Domingo & Villaescusa-Navarro 2022; Villaescusa-Navarro et al. 2022a).

Machine learning algorithms can also work with sparse and irregular data; e.g. through graph neural networks (GNNs) (Gilmer et al. 2017; Battaglia et al. 2018; Bronstein et al. 2021). GNNs exhibit multiple advantages over convolutional neural networks (CNNs). For instance, in the context of cosmology, they do not impose any cut on the considered physical scales, and different physical symmetries (e.g. translational and rotational invariance) can be easily implemented in the models (see Villanueva-Domingo & Villaescusa-Navarro 2022). GNNs have been used for a variety of tasks, such as parameter inference (Villanueva-Domingo & Villaescusa-Navarro 2022; Shao et al. 2022a; Makinen et al. 2022; Anagnostidis et al. 2022), inferring halo masses (Villanueva-Domingo et al. 2021), speeding up semi-analytic models (Jespersen et al. 2022), and rediscovering Newton's law (Cranmer et al. 2020).

In particular, Villanueva-Domingo & Villaescusa-Navarro (2022) showed that GNNs were able to infer $\Omega_{\rm m}$ with $\sim 10\%$ accuracy just based on galaxy properties (e.g., positions, stellar mass, radius, and metallicity), without making use of any summary statistics and performing likelihood-free inference. However, their model was not robust. The lack of robustness in this field-level inference task with galaxies through GNNs can be attributed to many reasons: from intrinsic differences in the subgrid models of the different simulations to the models learning unique, numerical, artifacts. Developing robust models, that extrapolate properly even with real data, is one most important tasks needed to replace standard data analysis techniques (e.g. perturbation theory). Shao et al. (2022a) showed instead that the positions and velocities of DM halos were robust to numerics in N-body codes as well as to variations in astrophysical parameters when inferring $\Omega_{\rm m}$ using a field-level approach. Even so, this work still deals with non-observables, such as DM halos, and some of their properties. In our companion paper, Shao et al. (2023) we are providing analytical equations to predict $\Omega_{\rm m}$ from the positions and velocity modulus fields of DM halos. This model is robust across different DM N-body simulations and, by changing the normalization of the input velocity modulus for each

Table 1. Characteristics of the hydrodynamical simulations used in this work.

Model	Usage	Number of simulations used	Mean number of galaxies per catalog	Reference
Astrid	Train, validate & test	1000(LH) + 27(CV)	1114	Bird et al. (2022)
SIMBA	Train, validate & test	1000(LH) + 27(CV)	1093	Davé et al. (2019)
IllustrisTNG	Train, validate & test	1000(LH) + 27(CV) + 1024(SB)	737	Pillepich et al. (2018a)
IllustrisTNG300	Test	1(LH)	799	Nelson et al. (2019)
Magneticum	Test	50(LH) + 27(CV)	3655	Hirschmann et al. (2014a)
SWIFT-EAGLE	Test	64(LH)	1255	Schaye et al. (2015)

hydrodynamic simulation, it is able to perform predictions for galaxy catalogs too. These equations bring a big step towards a physical interpretation of the model.

In the present work, we extend all these previous efforts using GNNs to show that we can build models that perform field-level likelihood-free inference using galaxy catalogs that are robust to changes in numerics, astrophysics, subgrid physics, and the method to identify galaxies. We train GNNs using thousands of galaxy catalogs from state-of-the-art hydrodynamic simulations of the CAMELS project (Villaescusa-Navarro et al. 2021b). We also investigate which galaxy properties are robust and how they contribute to the network predictions, showing that we only need the phase-space information of the galaxies to achieve the best results.

The manuscript is organized as follows: in Section 2 we present the data set, describing the different simulations used and their different setups; in Section 3 we describe the methodology, where we explain the data pre-processing, the translation of the galaxy catalogs into graphs, and the general architecture employed; in Section 4, we present the results related to the best model, our efforts to improve it, and investigate which is the most important source of information for the GNNs to extract their inferences; and, in Section 5, we present the discussion and conclusions, analyzing the differences among the different simulations, and provide ideas for future work.

2. DATA

In this section, we describe the data we use to train, validate, and test our models. We emphasize that all the galaxy properties considered in this work are direct from the simulations. In this way, we are not performing any changes in order to consider realistic effects, such as taking into account errors in the peculiar velocities. These considerations will be addressed in future work.

2.1. *Simulations*

The galaxy catalogs we use to train, validate, and test our models come from thousands of hydrodynamic simulations of the Cosmology and Astrophysics with MachinE Learning Simulations – CAMELS project (Villaescusa-Navarro et al. 2021b, 2022b). The hydrodynamic simulations have been run with different codes that solve the hydrodynamic equations differently and implement different subgrid models: IllustrisTNG (Weinberger et al. 2017a; Pillepich et al. 2018a), SIMBA (Davé et al. 2019), Astrid (Bird et al. 2022), Magneticum (Hirschmann et al. 2014a), and SWIFT-EAGLE (Schaye et al. 2015; Schaller et al. 2016). All the simulations follow the evolution of 256^3 DM particles and are initialized with 256^3 fluid elements from $z = 127$ down to $z = 0$ in periodic boxes of

25 h^{-1}Mpc on a side. The catalogs used in this work correspond to $z = 0$. The fiducial values of the cosmological parameters are: $\Omega_{\rm m} = 0.3$, $\Omega_{\rm b} = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $\sigma_8 = 0.8$, $w = -1$, $M_\nu = 0$ eV.

The CAMELS simulations can be classified into different sets and suites depending on how their parameters are arranged and which code was used to run them. We start by classifying the catalogs into different sets:

- **Latin Hypercube (LH).** The simulations in this category have their cosmological and astrophysical parameter variations arranged in a Latin hypercube that spans: $\Omega_{\rm m} \in [0.1, 0.5]$ and $\sigma_8 \in [0.6, 1.0]$, $A_{\rm SN1} \in [0.25, 4.0]$, $A_{\rm SN2} \in [0.5, 2.0]$, $A_{\rm AGN1} \in [0.25, 4.0]$, and $A_{\rm AGN2} \in [0.5, 2.0]$. $A_{\rm SN}$ and $A_{\rm AGN}$ are astrophysical parameters that control the efficiency of supernova (SN) and active galactic nuclei (AGN) feedback (see Villaescusa-Navarro et al. (2021b); Ni et al. (2023) for a detailed description of the meaning of the astrophysical simulations in every simulation suite). Each of the simulations in the Latin hypercube has been run with a different initial random seed for the generation of the initial conditions. We used these simulations for training, validating, and testing.

- **Cosmic Variance (CV).** These simulations have been run with the fiducial value of the cosmological and astrophysical parameters. The initial conditions for each simulation in this set have been generated with a different initial random seed. These simulations are only used for testing the models.

- **Sobol Sequence (SB).** The simulations in this set have their cosmological and astrophysical parameters arranged in a Sobol sequence (Sobol' 1967). A total of 28 parameters are varied: 5 cosmological ($\Omega_{\rm m}$, $\Omega_{\rm b}$, h, n_s, σ_8) and 23 astrophysical. The astrophysical parameters varied include the usual ones ($A_{\rm SN1}, A_{\rm SN2}, A_{\rm AGN1}, A_{\rm AGN2}$) and incorporate many others such as star formation, galactic winds, black hole (BH) growth and quasar parameters. All of them vary in ranges around the fiducial values used in the IllustrisTNG set. Their range of variation is large enough to enable a broad sampling of the considered parameter (Ni et al. 2023). We note that this set covers the largest region in parameter space within CAMELS although at a much lower density given the high dimensionality of the considered space. We use these simulations only for testing and to investigate how well our models generalize.

The CAMELS simulations can also be classified into different model suites according to the code used to run them:

- **IllustrisTNG.** These simulations were run using Arepo (Springel 2010; Weinberger et al. 2020) applying the same subgrid physics as the IllustrisTNG simulations (Weinberger et al. 2017a; Pillepich et al. 2018a). This suite contains 1000 LH, 27 CV, and 1024 SB simulations.

- **SIMBA.** These simulations were run with the Gizmo code (Hopkins 2015) and employ the same subgrid physics as the SIMBA simulation (Davé et al. 2019). This suite contains 1000 LH and 27 CV simulations.

- **Astrid.** These simulations were run using MP-Gadget (Feng et al. 2018) applying some modifications to the subgrid model employed in the Astrid simulation (Ni et al. 2022; Bird et al. 2022; Ni et al. 2023). This suite contains 1000 LH and 27 CV simulations.

- **Magneticum.** These simulations were run with the parallel cosmological Tree-PM code P-Gadget3 (Springel 2005). The code uses an entropy-conserving formulation of Smoothed Particle Hydrodynamics (SPH) (Springel & Hernquist 2002), with SPH modifications according to Dolag et al. (2004, 2005, 2006). It includes also prescriptions for multiphase interstellar medium based on the model by Springel & Hernquist (2003) as well as Tornatore et al. (2007) for the metal enrichment prescription. The model follows the growth and evolution of BHs and their associated AGN feedback based on the model presented by Springel et al. (2005) and Di Matteo et al. (2005), but includes modifications based on Fabjan et al. (2011), Hirschmann et al. (2014b) and Steinborn et al. (2016). The set contains 50 LH and 27 CV simulations. The following subgrid parameters were varied in order to control the stellar and AGN feedback (with parameter ranges given in square brackets) on the Latin-hypercube:

  - A_{SN1}, energy per unit of SFR [0.25, 4.0] $\times 10^{51}$.
  - A_{SN2}, wind speed [250,1000].
  - A_{AGN1}, coupling efficiency of the BH feedback [0.25, 4.0].
  - A_{AGN2}, boost of the AGN mode feedback [0.5, 2.0].

- **SWIFT-EAGLE.** These simulations have been run with the SWIFT code (Schaller et al. 2016, 2018) using a new subgrid physics model based on the original Gadget-EAGLE simulations (Schaye et al. 2015; Crain et al. 2015), with some parameter changes (Borrow et al. 2022). The full model will be described in Borrow & et. al. (2023). This suite contains 64 LH simulations varying the following subgrid parameters controlling the stellar and AGN feedback (with parameter ranges given in square brackets) on the Latin hypercube:

  - $f_{\rm E,min}$, the minimal stellar feedback fraction, [0.18, 0.6].
  - $f_{\rm E,max}$, the maximal stellar feedback fraction, [5, 10].
  - $N_{\rm H,0}$, pivot point in density that the feedback energy fraction plane rotates around, [$10^{-0.6}$, $10^{-0.15}$].
  - $\sigma_{\rm n}$ and σZ, energy fraction sigmoid width, controlling the density and metallicity dependence, [0.1, 0.65].
  - $\varepsilon_{\rm f}$, coupling coefficient of radiative efficiency of AGN feedback, [10^{-2}, 10^{-1}].
  - $\Delta T_{\rm AGN}$, AGN heating temperature, [$10^{8.3}$, $10^{9.0}$].
  - α, BH accretion suppression/enhancement factor, [0.2, 1.1].

Finally, to quantify the robustness of our model to super-sample covariance effects, we made use of the IllustrisTNG300-1 simulation (Nelson et al. 2019), which covers a larger volume of $(205\ h^{-1}{\rm Mpc})^3$ with slightly higher resolution than our fiducial CAMELS simulations and has a slightly different cosmology: $\Omega_{\rm m} = 0.3089$, $\Omega_{\rm b} = 0.0486$, $\Omega_\Lambda = 0.6911$, $h = 0.6774$, $\sigma_8 = 0.8159$, and $n_s = 0.9667$. This simulation was run with AREPO and made use of the IllustrisTNG subgrid physics model (Weinberger et al. 2017b; Pillepich et al. 2018a,b; Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018; Springel et al. 2018).

We emphasize that although the name of the parameters $A_{\rm SN1}, A_{\rm SN2}, A_{\rm AGN1}, A_{\rm AGN2}$ is common among different simulations, their actual implementation and effect on galaxy properties and clustering can be very distinct. Therefore, it is important to keep in mind that those parameters are not meant to share physical effects, only their names.

2.2. *Galaxy catalogs*

Halos and subhalos are identified in the simulations for every snapshot using two different halo and subhalo finders: SubFind (Springel et al. 2001; Dolag et al. 2009) and VELOCIraptor (Elahi et al. 2019; Cañas et al. 2019). All galaxy catalogs are from SubFind with the exception of SWIFT-EAGLE, which only contains VELOCIraptor catalogs. The reason for using two different codes is to check the robustness of our results to the subhalo finding procedure, which can cause some differences in the number of galaxies as shown in Gómez et al. (2022).

Galaxies are defined in all cases as subhalos that contain at least one star particle. In this work, we only consider galaxies with stellar masses above $1.3 \times 10^8\ M_\odot/h$. A galaxy catalog is constructed by taking all galaxies whose stellar mass is higher than a given threshold. For every simulation, we produce several galaxy catalogs by varying the stellar mass threshold.

A summary of the simulation characteristics can be found in Table 1, where we present their usage, the number of catalogs, the mean number of galaxies per catalog and the reference for each of the original galaxy formation models.

3. METHODOLOGY

In this section, we describe: the method we use to construct graphs from galaxy catalogs (Section 3.1); the architecture of our GNN (Section 3.2); the method to carry out likelihood-free inference (Section 3.3); the training procedure and optimization choices (Section 3.4); and the evaluation of the methodology, where we present the scores for the metrics we analyzed (Section 3.5).

3.1. *Galaxy graphs: construction*

The input for our GNNs are graphs: mathematical structures characterized by nodes, edges, and global properties. Every element of the graph can be described by a set of properties: $\mathbf{n}_i$ represents the properties of node i, $\mathbf{e}_{ij}$ represents the features of the edge between node i and j, and $\mathbf{g}$ contains global properties of the graph (Gilmer et al. 2017; Zhou et al. 2018; Battaglia et al. 2018). We construct graphs from catalogs that contain the galaxy positions and their peculiar velocities (only the z component); in some models, we also include the stellar mass of the galaxies.

In this work, we follow the method presented in Villanueva-Domingo & Villaescusa-Navarro (2022) (and used in Shao et al. (2022a) and Makinen et al. (2022) for halos) where galaxies represent the graph nodes and two galaxies are connected by an edge if their distance is smaller than a given linking radius r_{link}. Additionally, we use as a global property of the graph the logarithm of the number of galaxies in the graph: $\log_{10}(N_g)$[1].

[1] We have checked that including the number of galaxies as global feature yields slightly better results. For that reason, we keep that property.

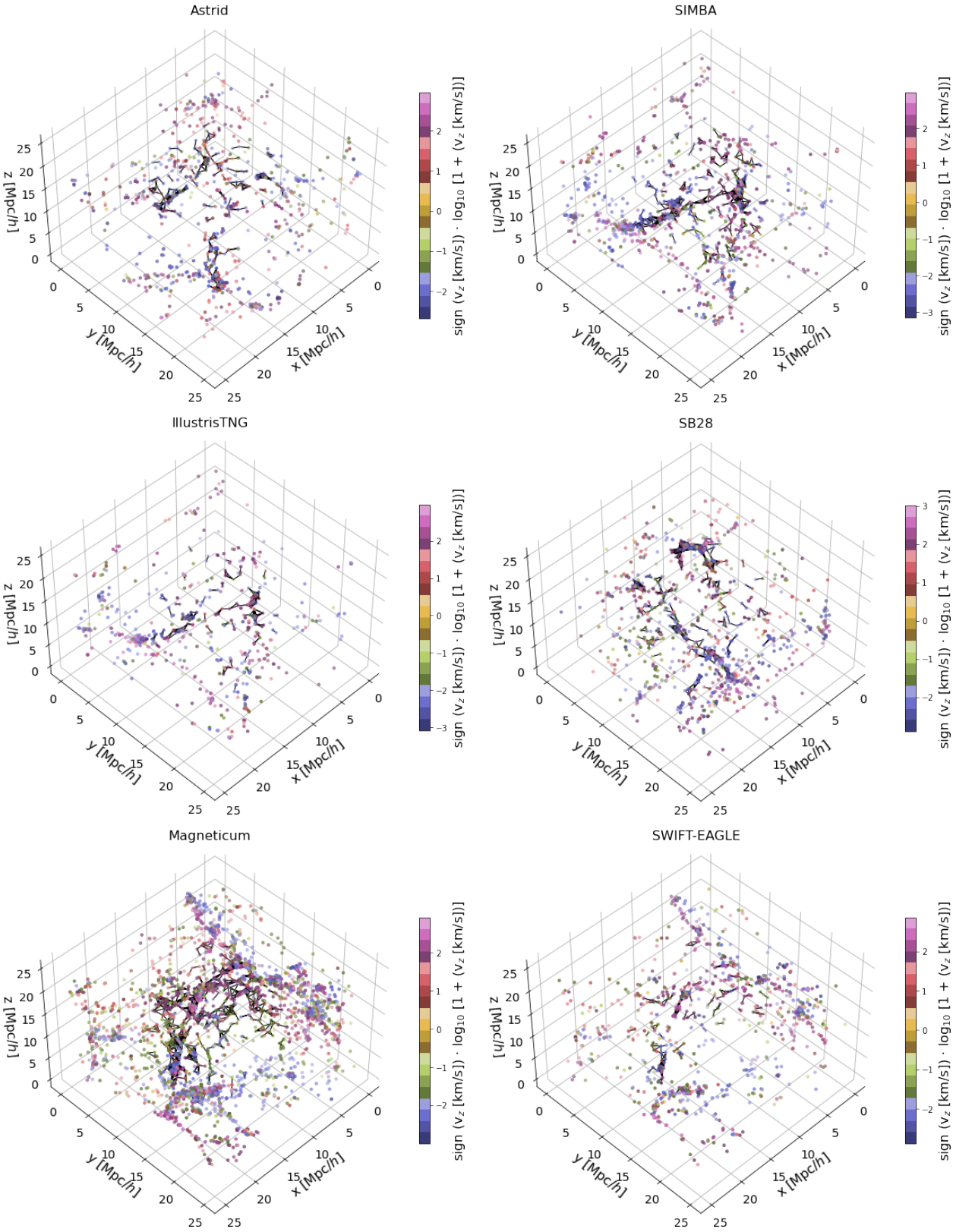

Figure 1. Examples of graphs constructed from galaxy catalogs from different CAMELS simulations: Astrid, SIMBA, IllustrisTNG, Magneticum, SB28, and SWIFT-EAGLE. The nodes represent the galaxies and their colors correspond to the normalization (Equation 1) of the z component of their peculiar velocity. Galaxies are connected by edges (shown as black lines) if their distance is smaller than $r_{\rm link} \sim 1.25\ h^{-1}$Mpc. We stress that we are nogalaxies which are linked due to periodic boundary conditions in these plots.

We investigate the contribution of the z component of the galaxy's peculiar velocities v_z and the stellar mass $M_\star$ as node attributes. We transform these features according to:

$$v_z \to \mathrm{sign}(v_z) \cdot \log_{10}\left[1 + \mathrm{abs}(v_z)\right], \tag{1}$$

$$M_\star \to \log_{10}(1 + M_\star)\ . \tag{2}$$

We chose to work with only one component for the galaxy velocity. This is because we want to be as close as possible to observational data, where we have access only to the radial peculiar velocity, i.e., the velocity measured along the line of sight.

The edge features contain information about the spatial distribution of galaxies (their positions), and those properties are designed to make the graph invariant under rotations and translations. We follow Villanueva-Domingo & Villaescusa-Navarro (2022) and set the edge features as:

$$\mathbf{e}_{ij} = \left[\frac{|\mathbf{d}_{ij}|}{r_{link}}, \alpha_{ij}, \beta_{ij}\right], \tag{3}$$

where:

$$\mathbf{d}_{ij} = [\mathbf{r}_i - \mathbf{r}_j] \tag{4}$$

$$\boldsymbol{\delta}_i = \mathbf{r}_i - \mathbf{c} \tag{5}$$

$$\alpha_{ij} = \frac{\boldsymbol{\delta}_i}{|\boldsymbol{\delta}_i|} \cdot \frac{\boldsymbol{\delta}_j}{|\boldsymbol{\delta}_j|} \tag{6}$$

$$\beta_{ij} = \frac{\boldsymbol{\delta}_i}{|\boldsymbol{\delta}_i|} \cdot \frac{\mathbf{d}_{ij}}{|\mathbf{d}_{ij}|}, \tag{7}$$

with $\mathbf{r}_i$ representing the position of a galaxy i and $\mathbf{c} = \sum_i^N \mathbf{r}_i/N$ being the centroid. Here, the *distance* $\mathbf{d}_{ij}$ is the difference of two galaxy (i and j) positions, the *difference vector* $\boldsymbol{\delta}_i$ denotes the position of a galaxy i with respect to the centroid, α_{ij} is the (cosine of) the angle between the difference vectors of two galaxies, while β_{ij} represents the angle between the difference vector of a galaxy i and its distance to another galaxy j. We account for periodic boundary conditions when computing both distances and angles. Moreover, we consider reverse edges – a copy of the graphs, with the same nodes and edges but with all of the edges reversed while compared to the orientation of the corresponding edges in the original graph; we do not consider self-loops (an edge that connects a node to itself). Note that, by construction, the model is rotational and translation invariant, as those operations will not change the edge features of the graph. In other words, they will remain the same while performing the usual rotation and translational matrix transformations to the galaxy positions (Villanueva-Domingo & Villaescusa-Navarro 2022).

In Figure 1 we show graphs constructed from galaxy catalogs of the different simulations: Astrid, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE. All these catalogs contain galaxies with minimum stellar mass: $M_\star = 1.95 \times 10^8\ M_\odot/h$. In all the graphs galaxies are colored according to their v_z (transformed according to Equation 1), and two galaxies are connected by a black line if their distance is within $r_{\rm link} \simeq 1.25\ h^{-1}$Mpc (this value was found with Optuna, as it will be described in Section 3.4). Notice that we are not connecting galaxies which are linked due to the periodic boundary conditions in this representation, i.e., a galaxy near the border of the box is not showing to be connected to some other galaxy in the other box extreme, even when they are linked

due to these conditions. This simple visual comparison shows that the spatial distribution of galaxies and their peculiar velocities are similar among all simulations. We note that the graph constructed from the Magneticum simulation exhibits a significantly larger number of galaxies than the others; this happens due to the employed AGN model used in Magneticum.

Every graph is characterized by a set of labels that we aim at inferring (e.g. $\Omega_{\rm m}$). We normalize these labels as θ_i, using

$$\theta_i \to \frac{(\theta_i - \theta_{\rm min})}{(\theta_{\rm max} - \theta_{\rm min})}, \tag{8}$$

where $\theta_{\rm min}$ and $\theta_{\rm max}$ represent the minimum and the maximum values of the corresponding parameter.

More details about the construction of the graphs, as well as an analysis of the different graphs (for the different simulations) are presented in Appendix A.

3.2. *GNN architecture*

The architecture we employ in this work follows the one presented in CosmoGraphNet[2] (Villanueva-Domingo & Villaescusa-Navarro 2022). Basically, the GNN is trained to infer the value of some cosmological parameter ($\Omega_{\rm m}$) from an input graph. Because GNNs are designed to deal with irregular and sparse data, the main idea behind their work was to perform a transformation of their components information (nodes $\mathbf{n}_i$, edges $\mathbf{e}_{ij}$, and global $\mathbf{g}$ attributes are updated), while the graph structure is preserved. In the end, the information is compressed, being converted by a usual multi-layer perceptron (MLP), to deliver the final property of the graph. By construction, GNNs preserve the graph symmetries (permutational invariance in the nodes, edge, and global attributes (Gilmer et al. 2017; Battaglia et al. 2018; Bronstein et al. 2021)). Besides, as done in Villanueva-Domingo & Villaescusa-Navarro (2022), the edge attributes consider translational and rotational symmetries (and here account for periodic boundary conditions too).

We have used a *message passing scheme* where each message passing layer updates the node and edge features[3], taking as input the graph and delivering as output its updated version. The node and edge features at layer $\ell+1$ are found from the node and edge features at layer ℓ as:

- **Edge model:**

$$\mathbf{e}_{ij}^{(\ell+1)} = \mathcal{E}^{(\ell+1)}\left(\left[\mathbf{n}_i^{(\ell)}, \mathbf{n}_j^{(\ell)}, \mathbf{e}_{ij}^{(\ell)}\right]\right), \tag{9}$$

  where $\mathcal{E}^{(\ell+1)}$ represents a MLP;

- **Node model:**

$$\mathbf{n}_i^{(\ell+1)} = \mathcal{N}^{(\ell+1)}\left(\left[\mathbf{n}_i^{(\ell)}, \bigoplus_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}, \mathbf{g}\right]\right), \tag{10}$$

  where $\mathfrak{N}_i$ represents all neighbors of node i, $\mathcal{N}^{(\ell+1)}$ is a MLP, and $\oplus$ is a multi-pooling operation responsible to concatenate several permutation invariant operations:

$$\bigoplus_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)} = \left[\max_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}, \sum_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}, \frac{\sum_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}}{\sum_{j\in\mathfrak{N}_i}}\right]. \tag{11}$$

[2] Available on Github repository https://github.com/PabloVD/CosmoGraphNet, DOI: 10.5281/zenodo.6485804.

[3] Note that we do not employ a model to update the global attribute. This was used just in order to update the node information (see Equation 10).

The use of the multi-pooling operation in the equation above was made because it has been argued that several aggregators can enhance the expressiveness of GNNs (Corso et al. 2020). Additionally, the number of layers to perform this update is a hyperparameter to be chosen in the optimization scheme. We also we made use of residual layers in the intermediate layers. The use of residuals means adding the input of the layer to its respective output, i.e., adding node/edge attributes to node/edge models. A discussion about this use can be found in Li et al. (2017) and Villanueva-Domingo & Villaescusa-Navarro (2022).

Once the graph has been updated using the N message passing layers, we collapse it into a 1-dimensional feature vector using

$$\mathbf{y} = \mathcal{F}\left(\left[\bigoplus_{i\in\mathfrak{F}} \mathbf{n}_i^N, \mathbf{g}\right]\right), \tag{12}$$

where $\mathcal{F}$ is the last MLP, $\oplus_{i\in\mathfrak{F}}$ the last multi-pooling operation (done exactly according to Equation 11, but operating over all nodes in the graph $\mathfrak{F}$), and $\mathbf{y}$ represents the target of the GNN (e.g. $\Omega_{\rm m}$).

All the MLP are constructed by a series of fully connected layers with ReLU activation function (except for the last layer, which does not employ an activation function). The number of layers, the number of neurons per layer, the weight decay, and the learning rate were considered as hyperparameters. The implementation of all the architectures presented in this work was done using PyTorch Geometric (Fey & Lenssen 2019).

3.2.1. *Variations of the architecture*

In Section 4.3 we investigate whether the information of our model is due to clustering, the distribution of velocities, or both. For that test, we made use of slightly different architectures to the one outlined above. Their main differences are:

- **Galaxy positions.**

  This model is used to quantify how much information is coming from the clustering of galaxies, i.e., it only uses galaxy positions. For that reason, the graphs only contain edge features (in the same way outlined above) and no node features. Because of this, the first layer of the model (Equation 9) operates in a slightly different way:

$$\mathbf{e}_{ij}^{(1)} = \mathcal{E}^{(1)}\left(\mathbf{e}_{ij}^{(0)}\right), \tag{13}$$

$$\mathbf{n}_i^{(1)} = \mathcal{N}^{(1)}\left(\left[\bigoplus_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(1)}, \mathbf{g}\right]\right). \tag{14}$$

  Note that other layers operate in exactly the same way as described in Equations 9-10.

- **Galaxy velocities.**

  This model is used to quantify how much information is coming from the distribution of galaxy velocities. Therefore, the graphs do not contain any spatial information and we can use *deep sets*[4] (Zaheer et al. 2017) architectures. In this case, we only have a node model that implements:

$$\mathbf{n}_i^{(\ell+1)} = \mathcal{N}^{(\ell+1)}\left(\mathbf{n}_i^{(\ell)}\right). \tag{15}$$

[4] It is important to note that, different from a usual neural network (NN), this implementation is invariant to permutations (the main property of GNNs) and is made to deliver global information of a structured data.

The target quantity is computed using Equation 12.

3.3. *Likelihood-free inference and the loss function*

Our models are trained to infer the value of a given parameter (θ_i, e.g. $\Omega_{\rm m}$) by predicting the marginal posterior mean μ_i and standard deviation σ_i without making any assumption about the form of the posterior, i.e.

$$\mathbf{y}_i(\mathcal{G}) = [\mu_i(\mathcal{G}), \sigma_i(\mathcal{G})], \tag{16}$$

where

$$\mu_i(\mathcal{G}) = \int_{\theta_i} d\theta_i \; \theta_i \; p(\theta_i|\mathcal{G}) \tag{17}$$

$$\sigma_i^2(\mathcal{G}) = \int_{\theta_i} d\theta_i \; (\theta_i - \mu_i)^2 \; p(\theta_i|\mathcal{G}) \, . \tag{18}$$

$\mathcal{G}$ represents the input graph and $p(\theta_i|\mathcal{G})$ is the marginal posterior, taken according to

$$p(\theta_i|\mathcal{G}) = \int_{\theta_i} d\theta_1 d\theta_2 \ldots d\theta_n \; p(\theta_1, \theta_2, \ldots, \theta_n|\mathcal{G}). \tag{19}$$

In order to achieve this, we made use of a specific loss function following Jeffrey & Wandelt (2020):

$$\mathcal{L} = \log\left[\sum_{j\in\text{batch}} \left(\theta_{i,j} - \mu_{i,j}\right)^2\right] + \log\left\{\sum_{j\in\text{batch}} \left[\left(\theta_{i,j} - \mu_{i,j}\right)^2 - \sigma_{i,j}^2\right]^2\right\}, \tag{20}$$

where j represents the samples in a given batch and i represents the index of the considered parameter (e.g. $i = 1$ for $\Omega_{\rm m}$). We refer the reader to Villaescusa-Navarro et al. (2022c) for the justification of the usage of the logarithms in the above expression.

We note that throughout the paper we will be referring to the error of the model as the quantity described above σ_i. This error only represents the aleatoric error, and therefore does not include the epistemic one, i.e., the error intrinsically related to the ML model. We have quantified the magnitude of the epistemic errors by training 10 different models with the same value of the hyperparameters (the best ones for the considered setup) and calculating the variance between the predictions of the models. We find that error to be $10\times$ smaller than the aleatoric one. Therefore, from now on, we will only report aleatoric errors since they dominate the total error budget.

3.4. *Training procedure and optimization*

We train our models on graphs constructed from galaxy catalogs of the LH sets of a given suite (e.g. the LH set of the Astrid simulations). We initially split the 1000 LH simulations into training (850 simulations), validation (100 simulations), and testing (50 simulations). For each simulation, we generate 10 galaxy catalogs constructed by taking all galaxies with stellar masses larger than $1.3R \times 10^8 \; M_\odot/h$, where R is a random number uniformly distributed between 1 and 2. This strategy is made in order to marginalize over different minimum threshold values for stellar masses, as well to increase the number of catalogs used to train the models[5]. For each catalog, we produce a graph as outlined in Section 3.1.

[5] A similar trick was used in Shao et al. (2022a), where the authors employed a similar augmentation in the halo catalogs, choosing them according to a minimum number of dark matter particles as a threshold. Shao et al. (2023) also made use of this method.

We then train the models using the above architecture for 300 epochs making use of ADAM optimizer (Kingma & Ba 2014) to perform the gradient descent, and a batch size of 25 samples. The hyperparameter optimization (where we have used the learning rate, the weight decay, the linking radius, the number of message passing layers, and the number of hidden channels per layer of the MLPs) was carried out using the OPTUNA package (Akiba et al. 2019) to perform a Bayesian optimization with Tree Parzen Estimator (TPE) (Bergstra et al. 2011). We made use of at least 100 trials to perform this task and we directed OPTUNA to minimize the validation loss, computed using an early-stopping scheme, in order to save only the model with the minimum validation error. The selected model was used for test subsequently.

3.5. *Performance Metrics*

We quantify the accuracy and precision of our models using different metrics that we describe below. We consider the true value of the parameter in question for graph i as θ_i, while we denote as μ_i and σ_i the prediction of the network for the posterior mean and standard deviation, respectively.

- **Root Mean Squared Error (RMSE):**

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \mu_i)^2}. \tag{21}$$

  Low values of the RMSE indicate the model is precise.

- **Coefficient of determination:**

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\theta_i - \mu_i)^2}{\sum_{i=1}^{N}\left(\theta_i - \bar{\theta}_i\right)^2}, \tag{22}$$

  where $\bar{\theta}_i = \frac{1}{N}\sum_{i=1}^{N}\theta_i$. Values close to 1 indicate the model is accurate.

- **Pearson Correlation Coefficient (PCC):**

$$\mathrm{PCC} = \frac{\mathrm{cov}\,(\theta, \mu)}{\sigma_\theta \sigma_\mu}. \tag{23}$$

  This statistic measures the positive/negative linear relationship between truth values and inferences: good values are close to ± 1, and worse closer to 0. It gives an idea of the accuracy of the model.

- **Bias:**

$$b = \frac{1}{N}\sum_{i=1}^{N}(\theta_i - \mu_i)\,. \tag{24}$$

  This statistic quantifies how much the inferences are "biased" with respect to the truth values; better values are close to 0.

- **Mean relative error:**

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}\frac{|\theta_i - \mu_i|}{\mu_i}. \tag{25}$$

  Low values of this statistic indicate the model is precise.

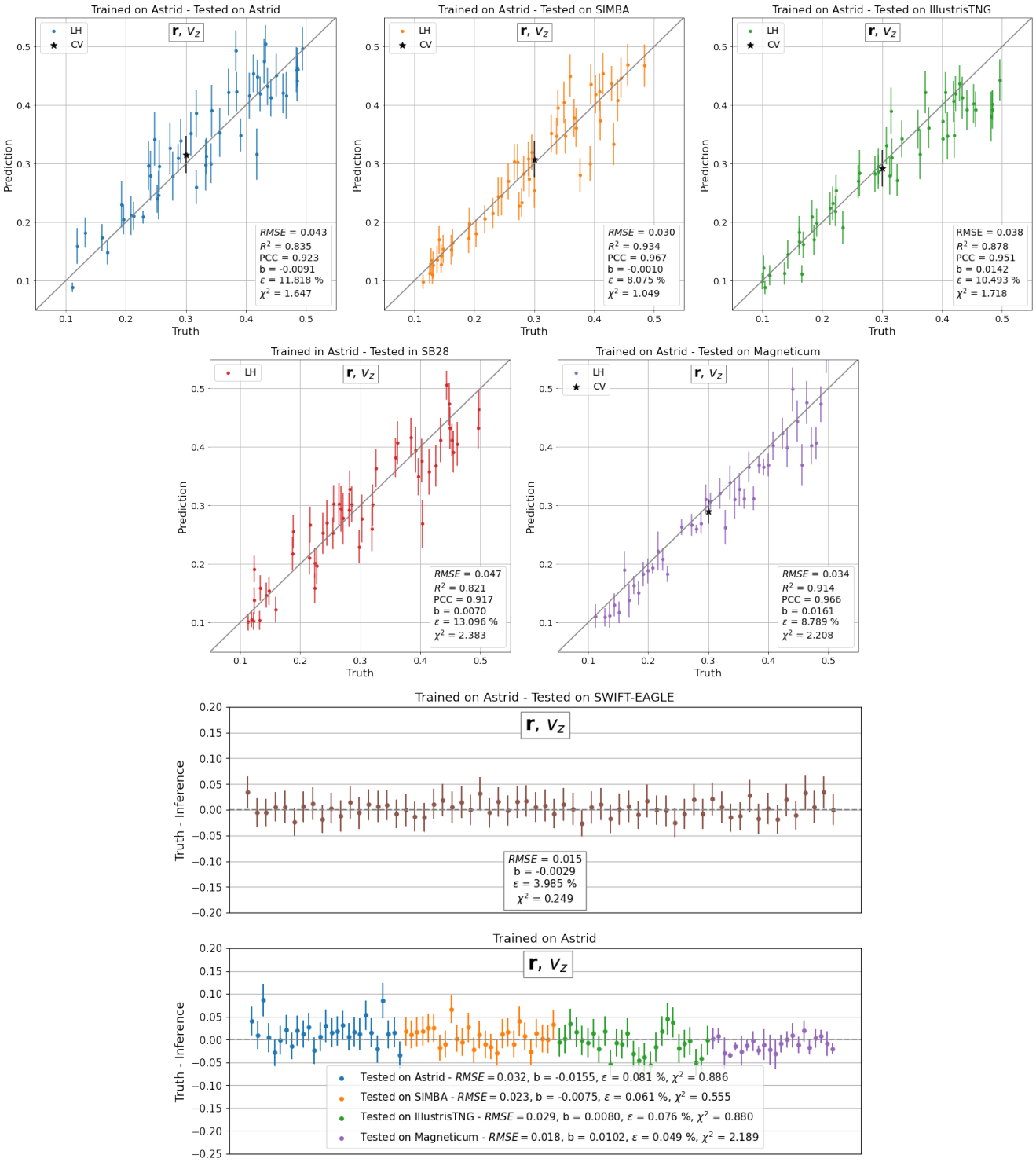

Figure 2. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions** and **velocities in the z direction**. We present the results for models *trained* on **Astrid** and *tested* on Astrid (top left), SIMBA (top middle), IllustrisTNG (top right), SB28 (second row left), Magneticum (second row right), and SWIFT-EAGLE (third row). The bottom panel shows the results of testing on CV sets of Astrid, SIMBA, IllustrisTNG, and Magneticum.

- **Reduced chi squared:**

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\theta_i - \mu_i}{\sigma_i}\right)^2. \tag{26}$$

  This statistic quantifies the accuracy of the estimated errors. Values of χ^2 close to 1 indicate the magnitude of the errors (posterior standard deviation in our case) is properly inferred, while values larger/smaller than 1 indicate the model is under/over predicting the errors.

We make use of these statistics to quantify the accuracy, precision, and bias of a given model in the test set. Note that in some cases we omit to report the value of some of these statistics for clarity, or when the statistics are not well defined (e.g. when tested on the CV set).

4. RESULTS

In this section we present the main results of testing our GNN models on galaxy catalogs with different cosmologies, astrophysical parameters, and subgrid physics models from the catalogs used for training. We start by showing the results of our best model, which only needs 3D galaxy positions and 1D velocity components, in Section 4.1. We then attempt to increase the precision of the model by adding more galaxy properties, particularly stellar mass, in Section 4.2. Next, we investigate the origin of the information extracted by our models in Section 4.3.

Note that we focus our analysis entirely on $\Omega_{\rm m}$. This is because our constraints on σ_8 are very weak. We provide further details on this in Appendix B. All results below are shown for catalogs built with galaxies with a minimum value of stellar mass as $M_\star = 1.95 \cdot 10^8\ M_\odot/h$, a value right in the middle of the threshold used in our training criteria[6].

4.1. *Positions & velocities*

We start by showing the results of training GNNs on catalogs that only contain the positions and velocities (only the z component)[7] of galaxies to infer the value of $\Omega_{\rm m}$. We have trained models using galaxy catalogs from the LH sets of the Astrid, IllustrisTNG, and SIMBA simulations. We then test these models on all other galaxy catalogs not included in their training set.

We found that the model trained on Astrid galaxy catalogs exhibits the best extrapolation properties, so we focus our analysis on it. The success of the model trained on Astrid can be associated with (1) the variety in the number of galaxies along the Astrid catalogs in LH sets, which vary from small to large numbers of galaxies ($N \in [30, 5,000]$ – see more details in the Appendix A) and (2) Astrid produces larger variations in some galaxy properties given the parameter variations in the LH set (Ni et al. 2023). We show the results of the models trained on IllustrisTNG and SIMBA catalogs in Appendix C. In addition, we trained a model on SB28 set, but even so, the model does not show good predictions when tested on the other simulations.

In Figure 2 we show the results of testing the model on galaxy catalogs from the LH sets of Astrid (top left), SIMBA (top middle), IllustrisTNG (top right), SB28 (second row left), Magneticum (second row right), and SWIFT-EAGLE (third row). In all these plots (apart from SB28 and SWIFT-EAGLE) we present the average (of their mean and standard deviation) across all of their CV boxes

[6] We have checked that our results are not very sensitive to the particular stellar mass cut we take, as long as we are not very close to the training boundaries.

[7] Due to homogeneity and isotropy, the results presented choosing the z component of the velocity are equivalent to choosing either x or y ones.

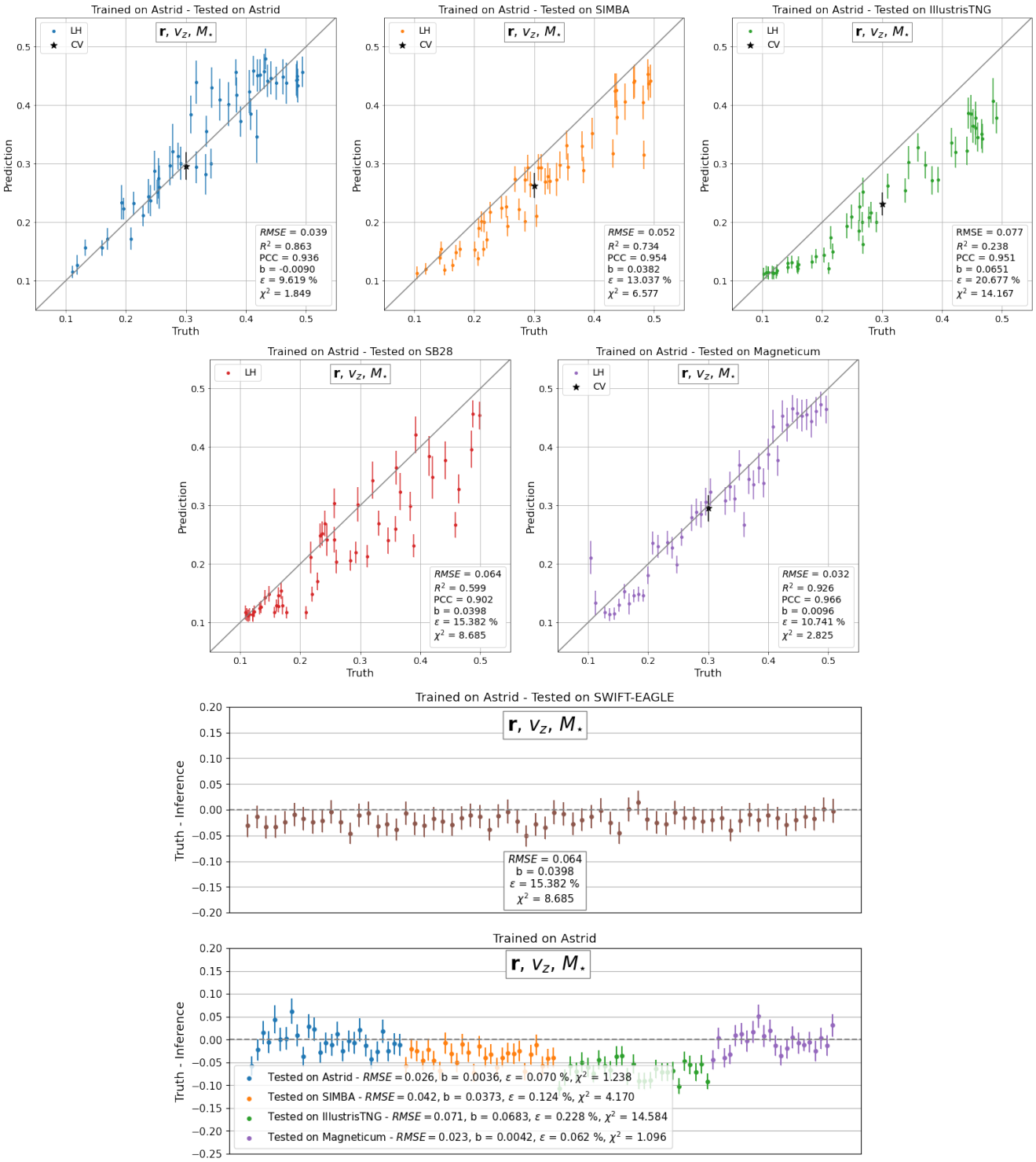

Figure 3. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions**, **velocities in the z direction**, and **stellar mass**. We present the results for a model *trained* on **Astrid** and *tested* on Astrid (top left), SIMBA (top middle), IllustrisTNG (top right), SB28 (second row left), Magneticum (second row right), and SWIFT-EAGLE (third row). The bottom panel shows the results of testing on CV sets of Astrid, SIMBA, IllustrisTNG, and Magneticum.

as a black point at $\Omega_{\rm m} = 0.3$. The results of testing the model on galaxies catalogs from the CV set of the different suites are shown in the bottom panel. Note that for clarity we only show 50 randomly selected samples of the predictions for all the LH results[8]. We stress that even if we only show the results for 50 random catalogs, the numbers reported for the different performance metrics (e.g. RMSE) are evaluated using all catalogs in the test set (e.g. 1000 catalogs for IllustrisTNG).

When using the model trained on Astrid and testing it on itself, we find that the GNN is able to infer $\Omega_{\rm m}$ with $RMSE = 0.043$, $R^2 = 0.835$, $PCC = 0.923$, $b = -0.0091$, $\epsilon = 11.8\%$, and $\chi^2 = 1.647$. These numbers indicate the model is accurate, precise, unbiased, and its errors are only slightly under predicted[9]. While testing that model on the other simulations the performance metrics are in the ranges: $RMSE \in [0.015, 0.047]$, $R^2 \in [0.821, 0.934]$, $PCC \in [0.917, 0.967]$, $b \in [-0.0010, 0.0161]$, $\epsilon \in [4.0, 13.1]\%$, and $\chi^2 \in [0.249, 2.383]$, showing that the model extrapolates very well, as can also be seen in Figure 2. Note that the model performs best on SIMBA and SWIFT-EAGLE, and worst on SB28. This indicates that, while the model is generally robust, even when tested on SB28, it becomes increasingly difficult to extrapolate predictions over distant regions in parameter space.

We have included a test using IllustrisTNG300 box in order to test the importance of super-sample covariance effects. Basically, the lack of power on scales larger than our boxes can affect both the abundance and clustering of galaxies (Hu & Kravtsov 2003; Hamilton et al. 2006; Takada & Bridle 2007; Li et al. 2014). We find that our method can partially account for these effects. We provide further details in Appendix D.

We now discuss the performance of the model on galaxy catalogs from the CV set. We find that our model works better when tested on the CV catalogs compared to the LH and SB sets. This could be due to the fact that the cosmology and astrophysics of those models lie exactly in the center of the training set. Those configurations are less prone to biased results, although it is interesting to observe that cosmic variance effects are not the main contribution to the error budget. Finally, all the different simulations end up with differences lower than 5% (apart from some boxes of Astrid or SIMBA, where we achieve differences {truth - inference} up to 10%) for the best model, once again being accurate, precise, and without any bias.

We conclude this part by emphasizing the overall good accuracy of our model, which accounts for cosmic variance, marginalizes over astrophysics, and is robust to changes in halo/subhalo finder and subgrid physics models. On top of this, the fact that the model works so well even in full extrapolation mode (e.g. when being tested on the SB28 simulations) indicates that the network may have learned physical relations (coming from the galaxies phase-space distribution) rather than a common feature among simulations.

4.2. *Positions, velocities, & stellar masses*

We now investigate whether we can make our model more precise, while keeping it robust, by considering an additional galaxy property: the stellar mass. For this, we construct graphs in the standard way (as described in Section 3) but taking as node features both velocity and stellar mass:

[8] In the case of Astrid we only have 50 samples in the test set since the majority of the LH set was used for training.

[9] To show the scores for the best model while testing it on Astrid and Magneticum, we removed respectively one and four predictions that correspond to a χ^2 larger than 14.0. They are points in the test set that achieved this bad inference and that we call "outliers". Outliers not only because of the bad scores but mainly because they correspond to particular realizations in the LH set with extreme values for the astrophysical parameters, which are realizations far away from the fiducial model. We do not follow this procedure in the other models (apart from the best model, trained on Astrid using only galaxy positions and z component of the velocity) because they end up with a huge number of "bad" predictions, not only in the matter of fact to this issue.

$[v_z, M_\star]$ (properties normalized as described in Section 3.1). We then train GNN models using catalogs from the Astrid LH set.

We present the results in Figure 3. When testing the model on galaxy catalogs from the Astrid LH set we find that the results improved for almost all the metrics: $RMSE = 0.039$, $R^2 = 0.863$, $PCC = 0.936$, $b = -0.0090$, $\epsilon = 9.62\%$, and $\chi^2 = 1.849$, which means that the GNN was able to extract more information from the catalogs. On the other hand, when testing the model on the galaxy catalogs from the other simulation suites the scores worsen: $RMSE \in [0.032, 0.077]$, $R^2 \in [0.238, 0.926]$, $PCC \in [0.902, 0.966]$, $b \in [0.0096, 0.0651]$, $\epsilon \in [10.7, 20.7]\%$, and $\chi^2 \in [2.825, 14.167]$. In other words, the model has become more precise when tested on itself, at the expense of becoming less accurate, when tested on other simulation sets. It is worth noting that some metrics actually improved when tested on galaxy catalogs from Magneticum, as seen in Figure 3. It is not clear to us what could be the explanation behind this: whether it is either a coincidence or due to the fact that galaxies in Astrid and Magneticum are more alike somehow while considering this specific galaxy property.

Our results are in agreement with those of Villanueva-Domingo & Villaescusa-Navarro (2022) who performed a similar analysis with galaxy catalogs whose node features were the maximum circular velocity, the stellar mass, the galaxy radius, and the star metallicity. While the model of those authors was more precise than ours (likely due to the use of additional galaxy properties), it was not robust. However, our models are slightly more robust; we believe this could be due to the fact that we use catalogs with different stellar mass thresholds to train the models, which overcomes the differences due to the fact that we are marginalizing over different stellar mass thresholds. This conclusion agrees with what Shao et al. (2022a) have found using the same idea of marginalization over an augmentation technique.

We reach similar conclusions when testing our models on galaxy catalogs from simulations of the CV sets (see the last panel of Figure 3), especially noticing that we have obtained a bias in the predictions for the different simulations. We emphasize the importance of testing the models on simulations as diverse as possible. Should we only have galaxy catalogs from Astrid and Magneticum simulations, we could reach the wrong conclusion that the model was both more precise and accurate than the one constructed using only positions and velocities.

4.3. *Where does the information come from?*

We now investigate where the information from our robust model (discussed in Section 4.1) comes from. Since in that model we only made use of galaxy positions and velocities, there are only three possibilities: 1) the information is coming from the positions of galaxies (clustering), 2) the information is coming from the distribution of galaxy velocities, and 3) the information is coming from both positions and velocities. Note that we are not considering attributing the importance to the level of information coming from the number of galaxies in the catalogs because: a) as mentioned in the Footnote 1, this global property only improved slightly the results, and b) we do not have a considerable number of catalogs with the same number, or even with the same range of the number, of galaxies (see Appendix A). This last reason should result in worse predictions due to the lack of data to train the machinery and would not allow to test it in all the different sub-grid physics models (which is the case of Magneticum, which only contains boxes with thousands of galaxies - see again Appendix A).

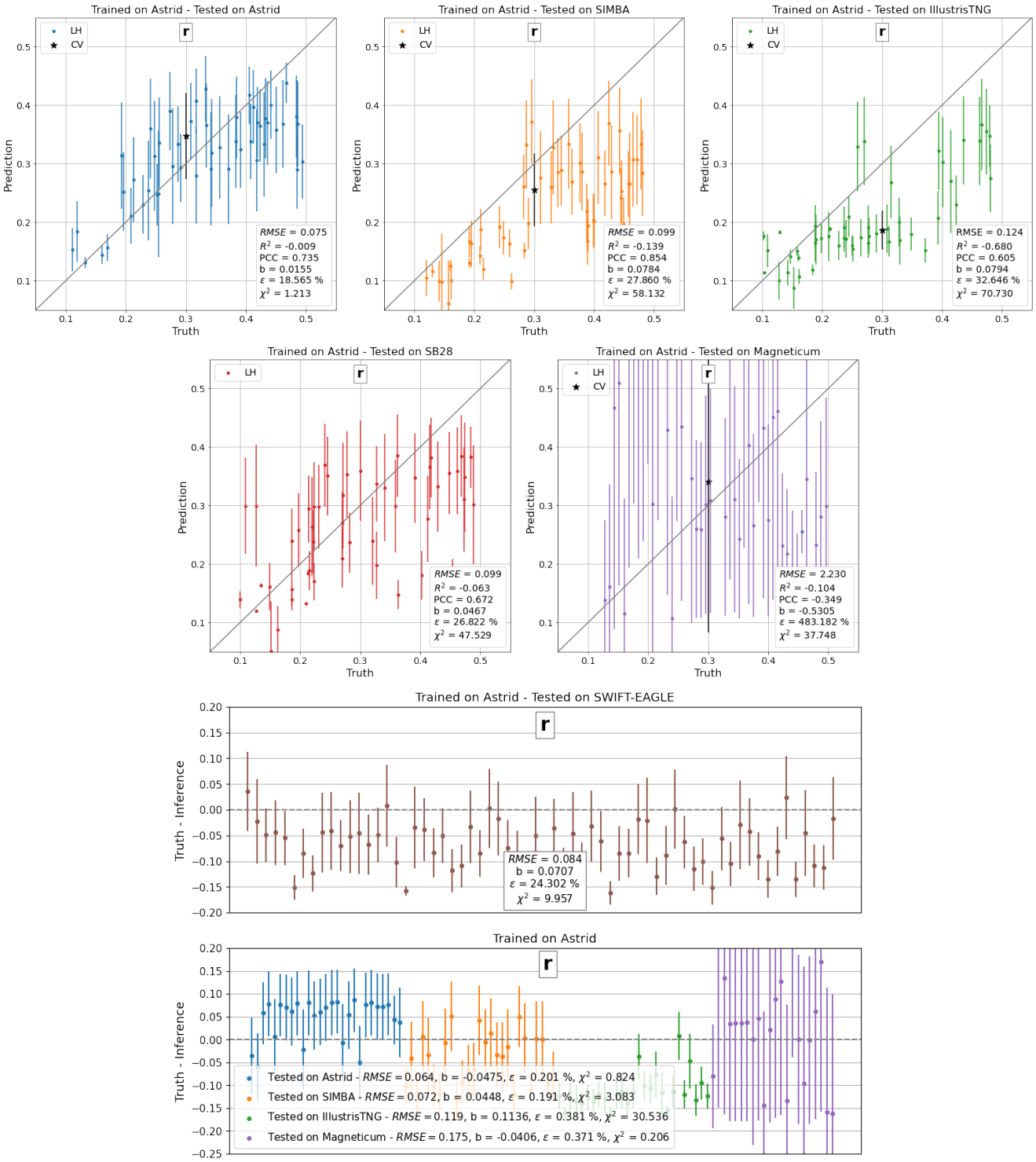

Figure 4. Likelihood-free inference of $\Omega_{\rm m}$ using **only** galaxy **positions**. We present the results for models *trained* on **Astrid** and *tested* on Astrid (top left), SIMBA (top middle), IllustrisTNG (top right), SB28 (second row left), Magneticum (second row right), and SWIFT-EAGLE (third row). The bottom panel shows the results of testing on CV sets of Astrid, SIMBA, IllustrisTNG, and Magneticum.

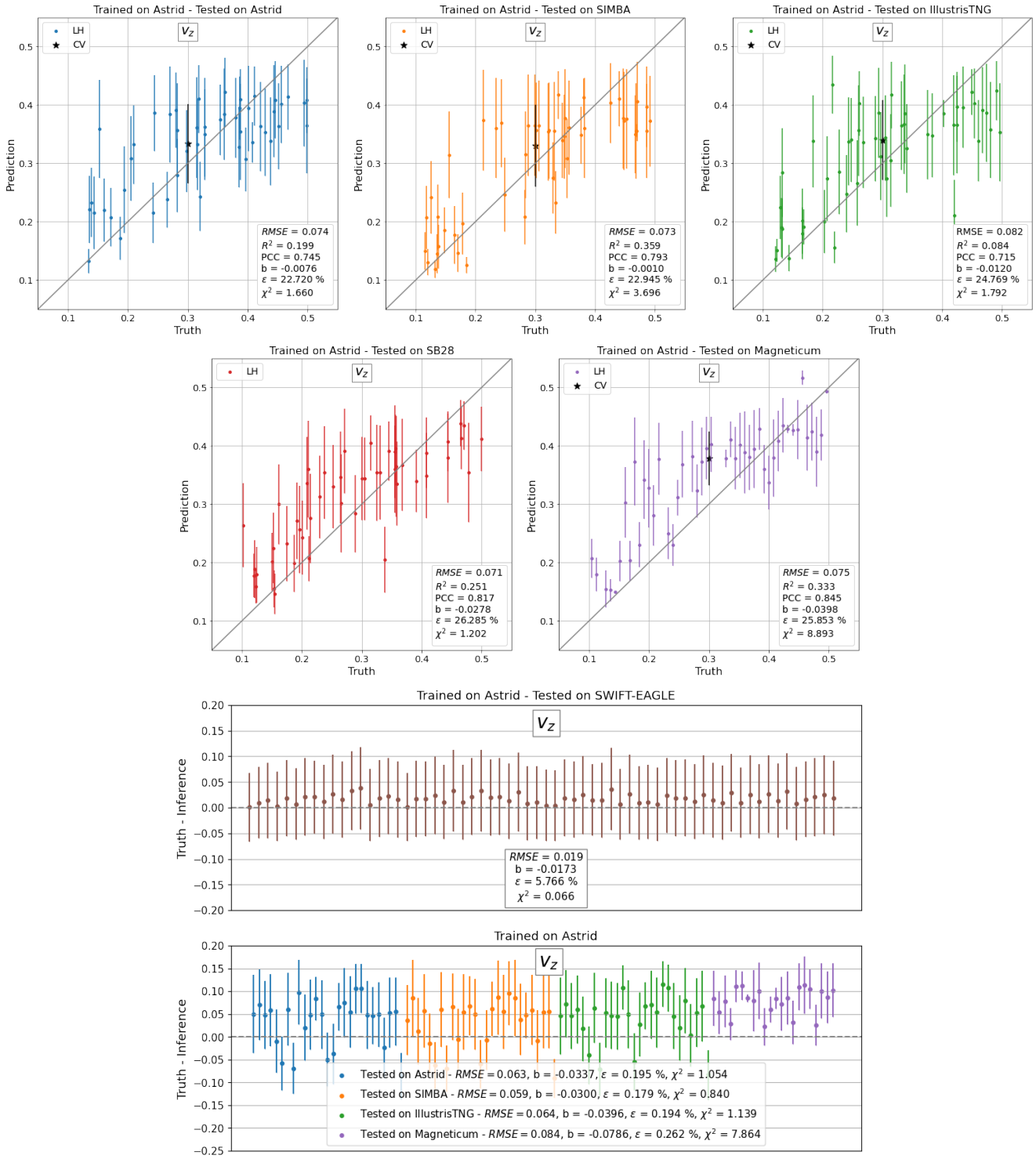

Figure 5. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **velocities in z direction**. We present the results for a model *trained* on **Astrid** and *tested* on Astrid (top left), SIMBA (top middle), IllustrisTNG (top right), SB28 (second row left), Magneticum (second row right), and SWIFT-EAGLE (third row). The bottom panel shows the results of testing on CV sets of Astrid, SIMBA, IllustrisTNG, and Magneticum.

In order to address the first possibility we have made use of graphs where the nodes do not contain any property. We train the model, in the Astrid suite, using the first slightly different GNN architecture described in Section 3.2.1: galaxy positions, i.e., using the prescription presented in Equations 13-14 for the first message passing layer. We then test the model on the different graphs from different simulation suites. The results are presented in Figure 4, following the same scheme as Figure 2. In all the tests the results are visibly worse (with large error bars) and significantly biased (when extrapolating to the other simulations). More specifically, we found: $RMSE \in [0.084, 2.230]$, $R^2 \in -[0.680, 0.063]$, $PCC \in [-0.349, 0.854]$, $b \in [-0.5305, 0.0467]$, $\epsilon \in [24.3, 483.2]\%$, and $\chi^2 \in [9.957, 70.730]$. While testing the model in the CV sets we found a low performance for all the metrics analyzed, with larger error bars. Our results are qualitatively in agreement with those of Villanueva-Domingo & Villaescusa-Navarro (2022), who performed a similar analysis but with galaxy catalogs with a fixed stellar mass threshold and did not use Astrid as the training set. From this test, we conclude that the network cannot be extracting the information just from galaxy clustering.

Next, we train a deep set model (see the second model presented in Section 3.2.1: galaxy velocities) on galaxy catalogs that only contain the z component of the galaxy velocities (i.e. there are no galaxy positions) and, then, we employed Equation 15. We used Astrid simulation to train the model. Figure 5 displays the results. Also in this case we find that the model performs poorly: $RMSE \in [0.019, 0.082]$, $R^2 \in [0.084, 0.359]$, $PCC \in [0.715, 0.845]$, $b \in -[0.0398, 0.0010]$, $\epsilon \in [5.8, 26.3]\%$, and $\chi^2 \in [0.066, 8.893]$. These results are distinct from what Villanueva-Domingo & Villaescusa-Navarro (2022) found (while using a deep set as well), whose scores were comparable to the ones from the GNN. Note that those authors used more galaxy properties and we only use the 1D velocity component. The results for catalogs of the CV sets have large error bars and poor values for all the metrics. We then conclude that galaxy velocities can not be alone the origin of the information extracted by the network.

The above tests indicate that the network is making use of both positions and velocities to infer the value of $\Omega_{\rm m}$. Another important point to highlight is that the models trained on galaxy positions alone and galaxy velocities alone, although not very precise, seem to also not be robust. This may indicate that the model that uses galaxy positions and velocities may be extracting robust information due to constraints in phase space (e.g. the necessity to fulfill the continuity equation), directly encoding effective information on $\Omega_{\rm m}$.

These findings may be related to some previous ideas that correlate with the matter content of the Universe to galaxy positions and peculiar velocities (Peebles 1980; Kaiser 1987; Cen et al. 1994; Strauss & Willick 1995), and that motivated a number of efforts towards peculiar velocity surveys (Howlett et al. 2017; Kourkchi et al. 2020; Howlett et al. 2022a). Besides, our results agree with the findings of Shao et al. (2022a) who used GNNs to predict $\Omega_{\rm m}$ based on positions and velocity modulus, but for DM halo catalogs. Moreover, this intricate relation motivates a deep analysis of the direct interpretation of the network predictions, which is being taken into account by us in Shao et al. (2023), combining symbolic regression with the GNNs.

5. DISCUSSION AND CONCLUSIONS

The quest to extract the maximum information from galaxy redshift surveys has motivated the development of many different approaches (Efron 1982; Feldman et al. 1994; Taylor et al. 2013; Abramo et al. 2016; Heavens et al. 2017; Hahn et al. 2020; Uhlemann et al. 2020; Gualdi et al. 2021;

Banerjee & Abel 2021; de Santi & Abramo 2022; Chartier & Wandelt 2022) and the upcoming data from the current and next generation of surveys (Taylor & Braun 1999; Laureijs et al. 2011; Takada et al. 2014; Amendola et al. 2013; Benitez et al. 2014; Spergel et al. 2015; DESI Collaboration et al. 2016; Euclid Collaboration: Castro et al. 2022; Pontoppidan et al. 2022), is pressing this field of research. While we do not have a final answer to this question ML techniques are appearing as a promising tool to tackle this problem (Ravanbakhsh et al. 2017; Ntampaka et al. 2020; Mangena et al. 2020; Hassan et al. 2020; Villaescusa-Navarro et al. 2021a; Cole et al. 2022; Perez et al. 2022). In particular, GNNs stand out as good machinery to extract cosmological information from galaxy and halo catalogs from simulations (Villanueva-Domingo & Villaescusa-Navarro 2022; Shao et al. 2022a; Makinen et al. 2022; Anagnostidis et al. 2022).

GNNs are ideal methods to analyze galaxy redshift surveys because: 1) they are designed to work with sparse and irregular data (Gilmer et al. 2017; Battaglia et al. 2018; Bronstein et al. 2021); 2) it is easy to construct models that fulfill physical symmetries (Villanueva-Domingo & Villaescusa-Navarro 2022); 3) they do not apply any cutoff on the scale to extract information. Perhaps the most challenging task associated with ML methods is their robustness (Hassani & Javanmard 2022), a hard question already explored using 2D maps with CNNs (Villaescusa-Navarro et al. 2021a), tabular data (Villaescusa-Navarro et al. 2022a), and galaxy catalogs (Villanueva-Domingo & Villaescusa-Navarro 2022). The reason behind the lack of robustness of the models is unclear and can be due to multiple factors: 1) data sets do not overlap; 2) models may be learning no physical effects (e.g. numerical artifacts); 3) data representation is different. We emphasize that precision is completely irrelevant without accuracy. The only way to deploy ML models to perform analysis with real data is to employ accurate models. Thus, robustness lies at the heart of this problem.

In this work, we have trained GNN models on thousands of galaxy catalogs from state-of-the-art hydrodynamic simulations of the CAMELS project to infer the value of $\Omega_{\rm m}$ at the field-level using a likelihood-free approach. More importantly, we have investigated the robustness of the models by testing them on galaxy catalogs from simulations run with completely different codes to the ones used for training. We now outline the main takeaways from this work:

- The model trained on Astrid catalogs that only contain galaxy positions and velocities (the z component) is able to infer the value of $\Omega_{\rm m}$ with $\sim 12\%$ precision and accuracy when tested on Astrid catalogs with different cosmologies and astrophysical parameters.

- The performance is similar when tested on galaxy catalogs from other galaxy formation simulations (each with different cosmology and astrophysics) run with four different hydrodynamic codes: IllustrisTNG, SIMBA, Magneticum, and SWIFT-EAGLE. This fact illustrates the robustness of the model under variations of the underlying subgrid physics.

- It also works well when tested on the SB28 set of the IllustrisTNG suite: a collection of 1024 simulations that varies 28 parameters (5 cosmological and 23 astrophysical) and therefore goes well beyond the diversity used to train the model (where only 6 parameters are varied).

- Our model is also robust to changes in the halo/subhalo finder: the galaxy catalogs of the SWIFT-EAGLE simulations were constructed employing VELOCIraptor, a different method than the one used for training (SubFind). When we tested our model on SWIFT-EAGLE catalogs we still obtained good predictions.

- The above constraints were obtained using a very small volume $(25\ h^{-1}\mathrm{Mpc})^3$ that only contains ~ 1000 galaxies with stellar masses above $\sim 2\times10^8\ M_\odot/h$ at $z=0$. We note that some galaxy catalogs contain a much larger ($\sim 5,000$, which is the case of Magneticum simulations) or smaller (~ 30, in some Astrid boxes) number of galaxies and the model still performs well on those.

- When training our models on galaxy catalogs that contain positions, velocities, and stellar masses we are able to build models that are more precise but less accurate. In fact, those models are no longer robust across different simulation codes and therefore could not be used with real data.

- We find that our models are extracting information from both galaxy positions and velocities. Furthermore, models trained using catalogs that only contain galaxy positions or galaxy velocities are not only less precise but also less accurate. We speculate that having both positions and velocities may improve the accuracy of the models as the phase-space distribution is constrained by physical arguments, such as the continuity equation, that need to be fulfilled independently of cosmology, astrophysics, and subgrid model employed.

Given the precision and accuracy of our model, it will be interesting applying it to peculiar velocity surveys such as the SLOAN catalog (Howlett et al. 2022b) or even the Cosmicflows-4 catalog (Kourkchi et al. 2020). We note that several steps need to be carried out before performing such a task:

- The method needs to be shown robust to changes in super-sample covariance. This is because in this analysis we did not account for such effect at the training stage. If the method is not robust to this effect, we should retrain our models on galaxy catalogs from larger volumes or catalogs that include the super-sample covariance effect. We note that preliminary work indicates that the models can deal with this effect, at least partially. We refer the reader to Appendix D) for further details.

- Through this work we are dealing with peculiar velocities from simulations. Therefore, we do not take into account any model or changes to consider observational errors in this quantity. The peculiar velocities of galaxies cannot be measured with infinite precision. A quantification of how the error on the peculiar velocities propagates into the constrain in Ω_m needs to be performed.

- An investigation on whether selection effects may affect the results is also needed, as some surveys rely on particular tracers (e.g. supernovae) that are not available on all galaxies above a certain stellar mass, as we consider here. Moreover, we plan to investigate the effect of increasing the number of galaxies and decreasing the number density in future work, competitive effects that will arise in real data and that are respectively low and high in the present analysis.

The possible application of this machinery to real data relies on one inherent limitation of the presented methodology, a question that is still related to robustness. This is because the GNN will be able to extrapolate their predictions only if applied to something compatible with the data set on which it has been trained. In other words, if the CAMELS suite of simulations will be able to capture

the main characteristics of our observable Universe. We plan to carry out these tasks in future work and show if the model can be robust, accurate, and precise to be able to infer a good estimate for $\Omega_{\rm m}$.

We now discuss the similarities and differences between this paper and previous works:

- Villanueva-Domingo & Villaescusa-Navarro (2022): our best model was achieved using a new CAMELS hydrodynamical set of simulations: the Astrid catalogs (different of IllustrisTNG and SIMBA suites, used in that work); only 3D galaxy positions and 1D velocity components carried all the information (differently from what these authors have considered, using stellar metallicity, galaxy radius, and maximum circular velocity too); we employed in the training stage a marginalization over different minimum values for stellar mass thresholds, instead of considering only one for all the ML stages; therefore, our results are robust over different subgrid physics, what does not happen in that work.
- Shao et al. (2022a): we have trained our models on galaxy, rather than of halo catalogs from N-body simulations; we make use of galaxy observables as input information (positions and velocity in only one direction), different from considering the modulus of galaxy peculiar velocities.
- Makinen et al. (2022): these authors used the Quijote halo suite (Villaescusa-Navarro et al. 2020), which does not consider subhalo properties. Besides, their analysis utilizes a different method to compute posteriors than the one employed here.
- Anagnostidis et al. (2022): in all our analyses we make use of hydrodynamical catalogs, considering astrophysics information. These authors consider lightcones from halo catalogs which is not taken into account here.

Overall, this paper presents a new method to study cosmology using the clustering and velocities of galaxies at the field-level, without imposing any cut on scale, that seems robust to changes in cosmology, astrophysics, subgrid physics, and galaxy identification algorithms.

ACKNOWLEDGMENTS

We would like to thank David Spergel, Ravi K. Sheth, Michael Strauss, Tamara Davis, Natália V. N. Rodrigues, Joop Schaye, Matthieu Schaller, Lucia A. Perez, and the CAMELS team for the enlightening discussions and valuable comments. We thank the São Paulo Research Foundation (FAPESP), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Simons Foundation for financial support. NSMS acknowledges financial support from FAPESP, grants 2019/13108-0 and 2022/03589-4. The CAMELS project is supported by the Simons Foundation and NSF grant AST 2108078. TC is supported by the INFN INDARK PD51 grant and the FARE MIUR grant "ClustersXEuclid" R165SBKTMA. EH acknowledges supported by the grant agreements ANR-21-CE31-0019 / 490702358 from the French Agence Nationale de la Recherche / DFG for the LOCALIZATION project. DAA acknowledges support by NSF grants AST-2009687 and AST-2108944, CXO grant TM2-23006X, and Simons Foundation award CCA-1018464. The research in this paper made use of the SWIFT open-source simulation code (http://www.swiftsim.com, Schaller et al. 2018) version 1.2.0. The training of the GNNs has been carried out using graphics processing units (GPUs) from Simons Foundation, Flatiron Institute, Center of Computational Astrophysics.

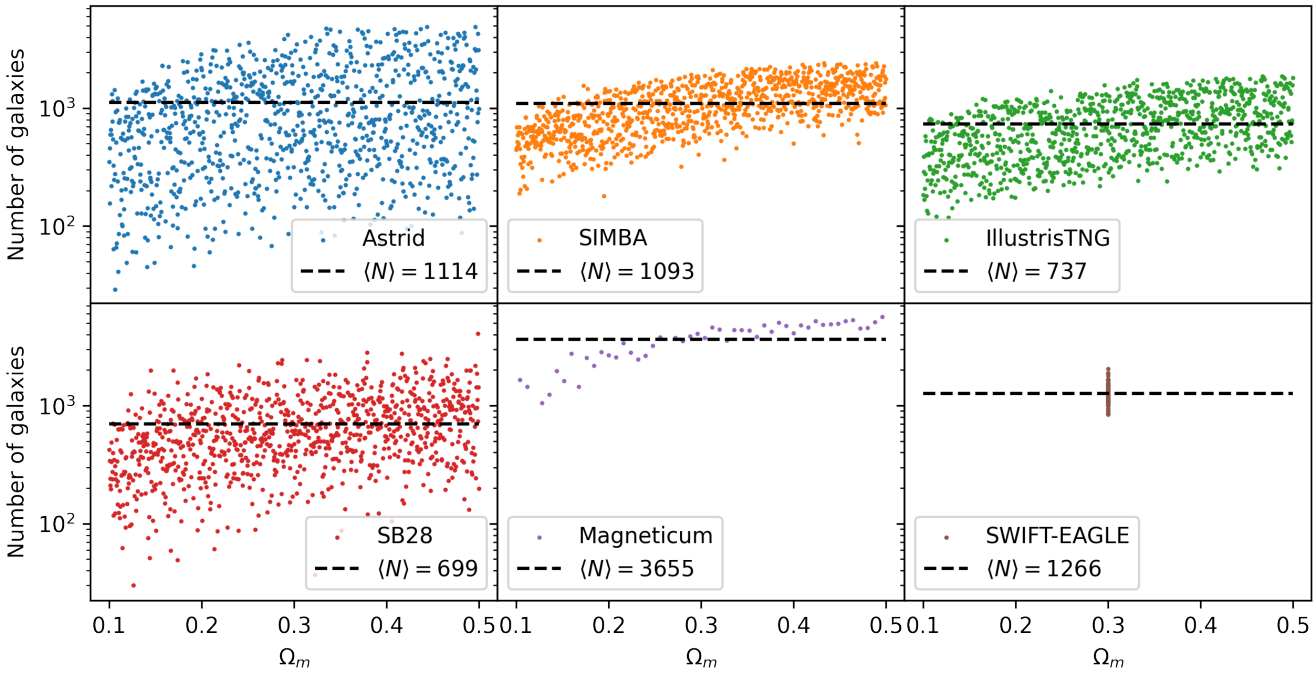

Figure 6. Comparison of the number of galaxies per LH catalog in CAMELS simulations for Astrid (top left), SIMBA (top middle), IllustrisTNG (top right), SB28 (bottom left), Magneticum (bottom middle), and SWIFT-EAGLE (bottom right). The horizontal lines correspond to the mean number of galaxies per simulation.

APPENDIX

A. GRAPH DETAILS

All the graphs built in this work follow the prescription presented in Section 3.1. The simple visual inspection of Figure 1 indicates some disparity among graphs from the different simulation suites. In this appendix, we explore some other aspects of the graphs and their characteristics according to the different simulations.

In Figure 6 we compare the number of galaxies in the LH catalogs for the different CAMELS simulations, considering a threshold in stellar mass as $M_\star = 1.95 \cdot 10^8\ M_\odot/h$. In almost all the cases the mean number of galaxies is ~ 1000, being a bit lower (~ 700) for IllustrisTNG and its variation SB28, and dramatically higher ($\sim 3,600$) for Magneticum. Besides, we can see that Astrid includes catalogs with a huge range of galaxy number ($N \in [30, 5,000]$), while the SIMBA and IllustrisTNG LH sets are much narrower (the same follows for SB28, with a higher dispersion of galaxy number, but not so broad as in Astrid). Finally, the range of the number of galaxies for Magneticum is $N \in [1000, 5500]$, including catalogs with such a large number of galaxies that do not have equivalent simulations in the SIMBA and IllustrisTNG data sets. As mentioned in Section 3.1 the large number of galaxies in Magneticum is related to the particular feedback model employed in those simulations.

The distances and the number of edges among the galaxies belonging to different catalogs were also investigated, as well the percentage of single galaxies per catalog. As expected, the distances among galaxies cover a range $d \in [10^{-2}, 21.65]\ h^{-1}$Mpc. All the catalogs have a similar shape in their spatial

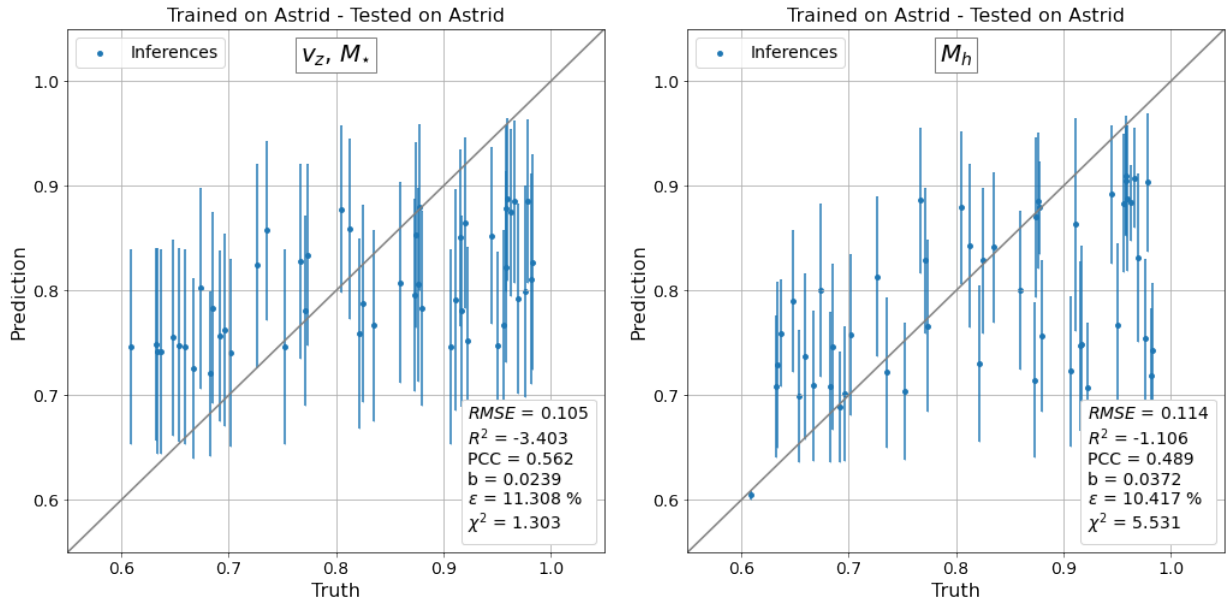

Figure 7. Likelihood-free inference of σ_8 using galaxy **velocities on the z direction** and **stellar mass** (on the left) and **halo mass** (on the right) as node attributes. We present the results for a model *trained* on **Astrid** and *tested* on Astrid.

distributions, with small differences on small scales. Single galaxies (the ones which are not connected to any other, and therefore only contribute to the propagation of their node information), on average, do not correspond to more than $\sim 20\%$ of the galaxies in the catalogs. This means that most of the information of the galaxies came from their connections (i.e. clustering properties). The number of edges per catalog is of order $\sim 10,000$, indicating that most galaxies have ~ 10 connections. Finally, the $r_{\rm link}$ found in all the models, for all different CAMELS sets in the hyperparameter training optimization, was around 1.25 h^{-1}Mpc.

B. INFERRING SIGMA 8

In this appendix, we present our efforts in trying to infer σ_8 using galaxy catalogs as graphs to feed GNN models. We made a sequence of tests of properties to include as node information in our graphs and none of them resulted in a robust model. Here we present two main results guided by: (1) Villanueva-Domingo & Villaescusa-Navarro (2022) while using galaxy velocities (in one direction) and including one more galaxy property, the stellar mass; and (2) Shao et al. (2022a) when using the host halo mass as node information for the graphs. The results are shown in Figure 7. In both models, we found poor performance: higher values for RMSE (> 0.1), negative values for R^2 ($-[3.4, 1.1]$) and low values for PCC ($[0.49, 0.56]$). In the case of the model which uses the halo mass, the χ^2 value is higher too (> 5.5). Furthermore, the predictions are around the fiducial/mean value, without covering the whole range of values and having higher error bars.

As already shown by Villanueva-Domingo & Villaescusa-Navarro (2022), it is a challenge to infer this cosmological parameter using galaxy information, which may need more galaxy properties (stellar mass, galaxy radius, metallicity, and maximum circular velocity) to achieve better performance. Then, because of relying on galaxy properties that differ substantially among the different simulations, it is hard to get a robust model. That is why our inference while using only galaxy velocity and stellar mass, is worse than these authors' results. On the other hand, because we are using all galaxies

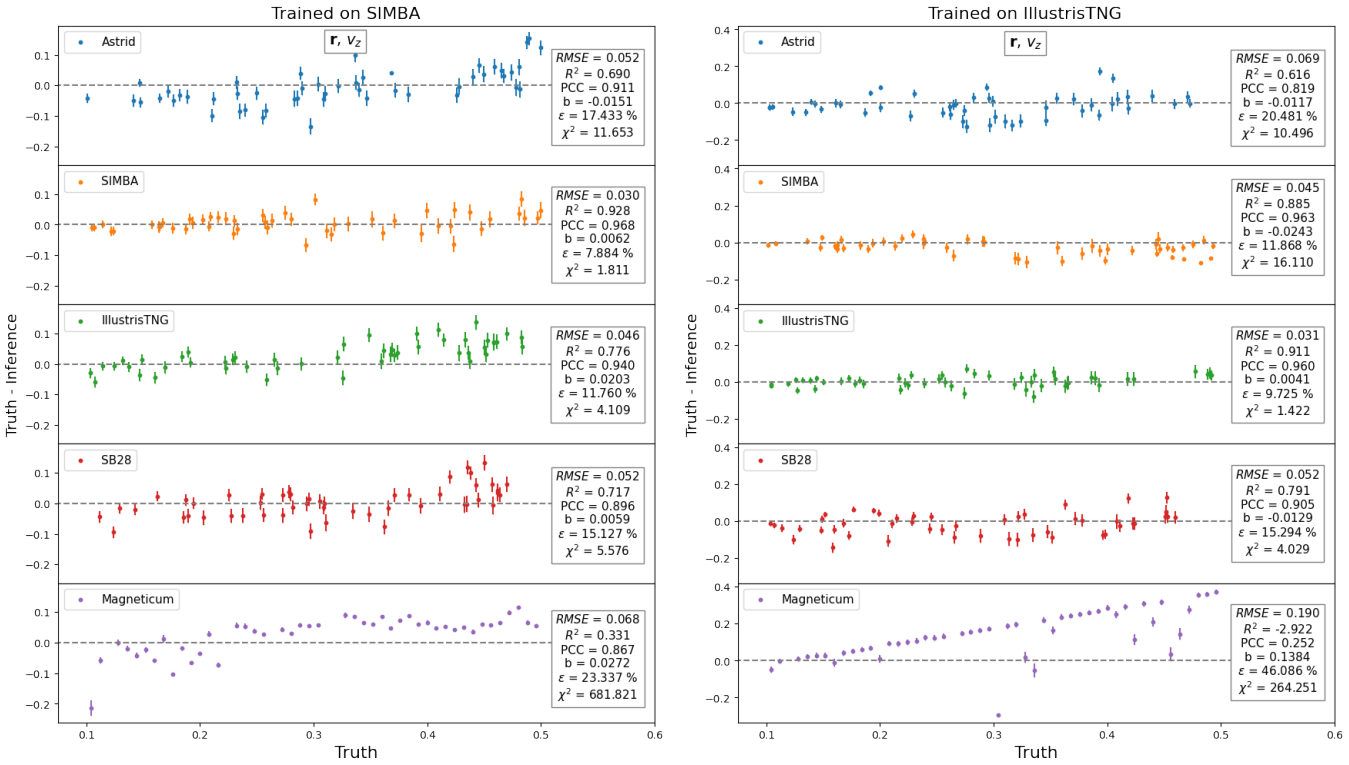

Figure 8. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions** and **velocities in the z direction**. We present the results for **LH set** tests of a model *trained* on **SIMBA** (on the left) and **IllustrisTNG** (on the right) and *tested* on Astrid, SIMBA, IllustrisTNG, SB28, and Magneticum respectively from the top to the bottom.

(centrals and satellites), our results are not directly comparable to the ones presented in Shao et al. (2022a), where only halo catalogs (without subhalos) are employed.

Therefore, we conclude, in agreement with Villanueva-Domingo & Villaescusa-Navarro (2022) and Villaescusa-Navarro et al. (2022a) that to constrain σ_8 precisely, we need larger volumes, as no ML technique was able to infer their value using only galaxy information. Besides, getting the correct value of this parameter can be challenging also for the standard approaches due to the small size of the boxes in the CAMELS suite. One possible solution can be found in Perez et al. (2022), where the authors obtained good constraints to predict σ_8 using machine learning methods to deal with the usual summary statistics, for larger boxes (100 h^{-1}Mpc). Another possible way to solve the puzzle related to σ_8 predictions should train a GNN on galaxy catalogs at higher redshifts and look for their impact on galaxy populations. This can be mostly related to the response of σ_8 in the abundance of more massive structures due to hierarchical structure formation, which does not happen at $z = 0$, where small galaxy populations dominate (Ni et al. 2023). This will be addressed in future work.

C. SIMBA AND ILLUSTRISTNG RESULTS

The present appendix follows the results of Section 4.1, for models trained using SIMBA and IllustrisTNG data sets. We stress that the GNN architecture follows the same structure as the one used in the best model (but with a different set of hyperparameters, found using Optuna).

All the results are presented in Figures 8, 9 and 10, where we plot the values for {truth - inference} in the y-axis, while the x-axis shows either the truth values of $\Omega_{\rm m}$ or an arbitrary order of the predictions

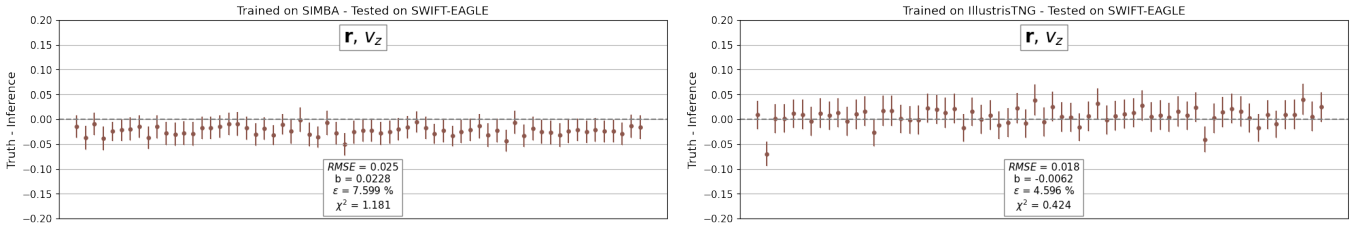

Figure 9. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions** and **velocities in the z direction**. We present the results for a model *trained* on **SIMBA** (on the left) and **IllustrisTNG** (on the right) and *tested* on SWIFT-EAGLE.

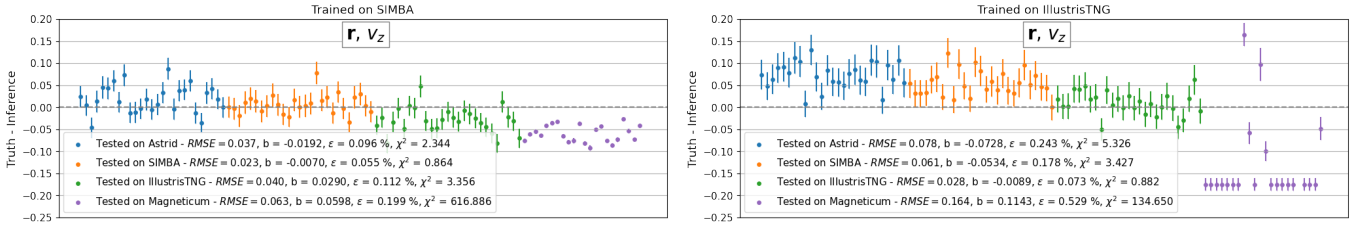

Figure 10. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions** and **velocities in the z direction**. We present the results for **CV set** tests of a model *trained* on **IllustrisTNG** and *tested* on Astrid, SIMBA, IllustrisTNG, SB28, and Magneticum.

by simulation suite. The metrics for the models trained on SIMBA/IllustrisTNG and tested on themselves are very good (even compared to the best model): $RMSE = [0.030, 0.031]$, $R^2 = [0.911, 0.928]$, $PCC = [0.960, 0.968]$, $b = [0.0041, 0.0062]$, $\epsilon = [7.9, 9.7]\%$, and $\chi^2 = [1.422, 1.811]$. However, all the tests on the other simulations are worse: $RMSE \in [0.018, 0.190]$, $R^2 \in [-2.922, 0.885]$, $PCC \in [0.252, 0.963]$, $b \in [0.0059, 0.1384]$, $\epsilon \in [4.6, 46.1]\%$, and $\chi^2 \in [0.424, 681.821]$. The worst predictions show up when the networks are tested on Magneticum (both, for the model trained on SIMBA and IllustrisTNG, but being worse for the latter). The tests on SWIFT-EAGLE and in the CV sets show that the scores are, in most cases, a bit worse compared to the best model (when we train the model using Astrid).

Our results suggest that the very poor predictions for Magneticum are due to the fact that the models trained on SIMBA and IllustrisTNG have never seen catalogs with such a high number of galaxies, which is the case for Magneticum catalogs (see Appendix A, specially Figure 6, which shows that Astrid covers a large range of number of galaxies when compared to SIMBA and IllustrisTNG). We have tested to increase the stellar mass cut in Magneticum catalogs and have obtained better predictions (comparable to the same models tested on the other catalogs apart themselves) while using the models trained on SIMBA/IllustrisTNG. This shows that reducing the number of galaxies in Magneticum catalogs improves their inferences significantly. Therefore, although the number of galaxies is not the most important property in the analysis, we can clearly see their effect on the model predictions while taking a look at these results.

Finally, in contrast to the robust model that was trained using Astrid, the inferences from the models trained using SIMBA and IllustrisTNG are, unfortunately, not robust across different simulations.

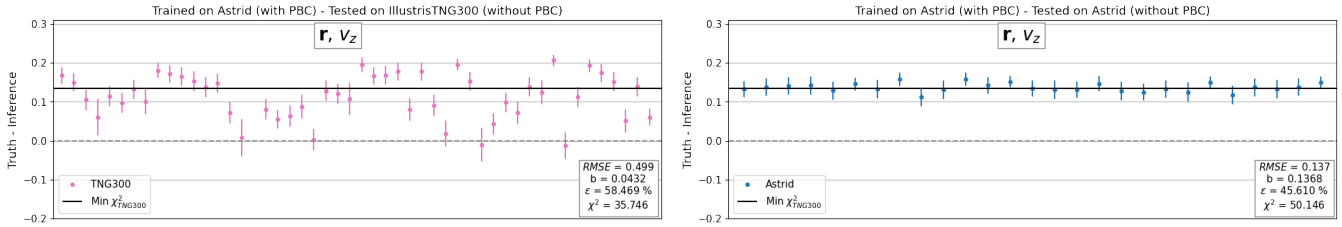

Figure 11. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions** and **velocities in the z direction**. We present the results for a model *trained* on **Astrid** and *tested* on: (1) 50 random $(25\ h^{-1}{\rm Mpc})^3$ sub-volumes of the **IllustrisTNG300** simulation (on the left) and (2) Astrid (on the right). In both cases the model was trained considering the periodic boundary conditions (PBC) and tested without this consideration.

D. SUPER-SAMPLE COVARIANCE ANALYSIS

We start noticing that our 25 h^{-1}Mpc boxes have a mean overdensity, $\langle\rho/\bar{\rho}\rangle = 1$. In the real Universe, $(25\ h^{-1}{\rm Mpc})^3$ patches will not satisfy that equality, and values larger or smaller will appear due to the presence of power on modes larger than the size of that region. Those modes are expected to affect both the clustering of galaxies and their internal properties. Here we investigate whether such effects will affect our models. To test this, we made use of the IllustrisTNG300-1 simulation, which covers a periodic volume of $(205\ h^{-1}{\rm Mpc})^3$ at a slightly higher resolution than the CAMELS simulations.

We have selected 50 random $(25\ h^{-1}{\rm Mpc})^3$ sub-volumes within the IllustrisTNG300 box, taking the galaxies in those sub-volumes and constructed graphs to input into our model. It is important to note that we have turned off the periodic boundary conditions (PBC) when constructing the graphs, due to the fact that the distribution of galaxies is not periodic within the sub-volumes. The results of testing our model with these galaxy catalogs are shown in Figure 11. We can see that the inferences for the IllustrisTNG300 catalogs have a positive bias of $b = 0.0432$ and the different estimations fluctuate around an offset that we indicate as Min $\chi^2_{\rm TNG300}$. This value represents the χ^2 minimization considering the IllustrisTNG300 inferences.

In order to test if this offset can be an effect of turning off the PBC we have tested the model on Astrid galaxy catalogs whose graphs have been constructed neglecting PBC. The results are presented in the right panel of Figure 11. We can see that we find almost the same offset for these new predictions.

Given the large effect that the PBC have on our results, we have retrained the GNN model on Astrid galaxy catalogs whose graphs are constructed without using PBC. We then test that model on galaxy catalogs from 100 random sub-volumes of the IllustrisTNG300 simulation. The results are presented in Figure 12. We can see that the inferences do not exhibit good scores: $RMSE = 0.089$, $b = -0.0073$, $\epsilon = 24.8\%$, and $\chi^2 = 34$. Even so, all the predictions fluctuate around the true values, indicating that we may have outliers. After removing predictions related to $\chi^2 > 14.0$ (35 points) we achieve better results that follows for: $RMSE = 0.059$, $b = -0.0118$, $\epsilon = 16.6\%$, and $\chi^2 = 4,0$.

From these results, we conclude that our method is not severely affected by super-sample covariance in the majority of the cases, although it does not work in all scenarios. We note that the fraction of outliers (i.e. cases where the model performs badly) is much higher in this test case than in, e.g., SB28 simulations. This indicates that further work is needed to either account for super-sample

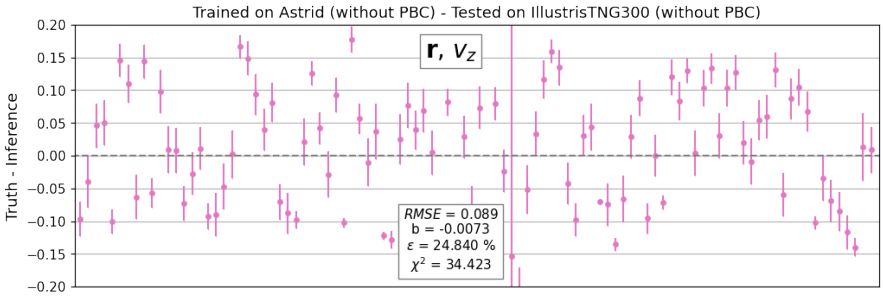

Figure 12. Likelihood-free inference of $\Omega_{\rm m}$ using galaxy **positions** and **velocities in the z direction**. We present the results for a model *trained* on **Astrid** and *tested* on 100 random $(25\ h^{-1}{\rm Mpc})^3$ sub-volumes within **IllustrisTNG300**. This specific model was trained without the periodic boundary conditions (PBC) and tested without this too.

covariance effects or to identify the range of validity of our models. We leave this task for future work.

REFERENCES

Abramo, L. R., Secco, L. F., & Loureiro, A. 2016, MNRAS, 455, 3871, doi: 10.1093/mnras/stv2588

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, arXiv e-prints, arXiv:1907.10902, doi: 10.48550/arXiv.1907.10902

Amendola, L., Appleby, S., Bacon, D., et al. 2013, Living Reviews in Relativity, 16, 6, doi: 10.12942/lrr-2013-6

Anagnostidis, S., Thomsen, A., Kacprzak, T., et al. 2022, arXiv e-prints, arXiv:2211.12346. https://arxiv.org/abs/2211.12346

Banerjee, A., & Abel, T. 2021, MNRAS, 500, 5479, doi: 10.1093/mnras/staa3604

Battaglia, P. W., Hamrick, J. B., Bapst, V., et al. 2018, arXiv e-prints, arXiv:1806.01261. https://arxiv.org/abs/1806.01261

Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv e-prints, arXiv:1403.5237. https://arxiv.org/abs/1403.5237

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, ApJS, 208, 20, doi: 10.1088/0067-0049/208/2/20

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. 2011, in Advances in Neural Information Processing Systems, ed. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger, Vol. 24 (Curran Associates, Inc.). https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

Bird, S., Ni, Y., Di Matteo, T., et al. 2022, MNRAS, 512, 3703, doi: 10.1093/mnras/stac648

Borrow, J., & et. al. 2023, In preparation.

Borrow, J., Schaller, M., Bahe, Y. M., et al. 2022, arXiv e-prints, arXiv:2211.08442, doi: 10.48550/arXiv.2211.08442

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. 2021, arXiv e-prints, arXiv:2104.13478. https://arxiv.org/abs/2104.13478

Cañas, R., Elahi, P. J., Welker, C., et al. 2019, MNRAS, 482, 2039, doi: 10.1093/mnras/sty2725

Cen, R., Bahcall, N. A., & Gramann, M. 1994, ApJL, 437, L51, doi: 10.1086/187680

Chartier, N., & Wandelt, B. D. 2022, MNRAS, 509, 2220, doi: 10.1093/mnras/stab3097

Cole, A., Miller, B. K., Witte, S. J., et al. 2022, JCAP, 2022, 004, doi: 10.1088/1475-7516/2022/09/004

Corso, G., Cavalleri, L., Beaini, D., Liò, P., & Veličković, P. 2020, arXiv e-prints, arXiv:2004.05718, doi: 10.48550/arXiv.2004.05718
Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, MNRAS, 450, 1937, doi: 10.1093/mnras/stv725
Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., et al. 2020, arXiv e-prints, arXiv:2006.11287. https://arxiv.org/abs/2006.11287
Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827, doi: 10.1093/mnras/stz937
de Santi, N. S. M., & Abramo, L. R. 2022, JCAP, 2022, 013, doi: 10.1088/1475-7516/2022/09/013
de Santi, N. S. M., Rodrigues, N. V. N., Montero-Dorta, A. D., et al. 2022, MNRAS, 514, 2463, doi: 10.1093/mnras/stac1469
Delgado, A. M., Wadekar, D., Hadzhiyska, B., et al. 2022, MNRAS, 515, 2733, doi: 10.1093/mnras/stac1951
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv e-prints, arXiv:1611.00036. https://arxiv.org/abs/1611.00036
Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604, doi: 10.1038/nature03335
Dolag, K., Borgani, S., Murante, G., & Springel, V. 2009, MNRAS, 399, 497, doi: 10.1111/j.1365-2966.2009.15034.x
Dolag, K., Jubelgas, M., Springel, V., Borgani, S., & Rasia, E. 2004, ApJL, 606, L97, doi: 10.1086/420966
Dolag, K., Meneghetti, M., Moscardini, L., Rasia, E., & Bonaldi, A. 2006, MNRAS, 370, 656, doi: 10.1111/j.1365-2966.2006.10511.x
Dolag, K., Vazza, F., Brunetti, G., & Tormen, G. 2005, MNRAS, 364, 753, doi: 10.1111/j.1365-2966.2005.09630.x
Efron, B. 1982, The Jackknife, the Bootstrap and other resampling plans (Society for Industrial and Applied Mathematics (SIAM))
Elahi, P. J., Cañas, R., Poulton, R. J. J., et al. 2019, PASA, 36, e021, doi: 10.1017/pasa.2019.12
Euclid Collaboration: Castro, T., Fumagalli, A., Angulo, R. E., et al. 2022, arXiv e-prints, arXiv:2208.02174, doi: 10.48550/arXiv.2208.02174
Fabjan, D., Borgani, S., Rasia, E., et al. 2011, MNRAS, 416, 801, doi: 10.1111/j.1365-2966.2011.18497.x
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23, doi: 10.1086/174036
Feng, Y., Bird, S., Anderson, L., Font-Ribera, A., & Pedersen, C. 2018, MP-Gadget/MP-Gadget: A tag for getting a DOI, FirstDOI, Zenodo, doi: 10.5281/zenodo.1451799
Fey, M., & Lenssen, J. E. 2019, arXiv e-prints, arXiv:1903.02428, doi: 10.48550/arXiv.1903.02428
Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. 2017, arXiv e-prints, arXiv:1704.01212. https://arxiv.org/abs/1704.01212
Gómez, J. S., Padilla, N. D., Helly, J. C., et al. 2022, MNRAS, 510, 5500, doi: 10.1093/mnras/stab3661
Gualdi, D., Gil-Marín, H., & Verde, L. 2021, JCAP, 2021, 008, doi: 10.1088/1475-7516/2021/07/008
Hahn, C., Villaescusa-Navarro, F., Castorina, E., & Scoccimarro, R. 2020, JCAP, 2020, 040, doi: 10.1088/1475-7516/2020/03/040
Hamilton, A. J. S., Rimes, C. D., & Scoccimarro, R. 2006, MNRAS, 371, 1188, doi: 10.1111/j.1365-2966.2006.10709.x
Hassan, S., Andrianomena, S., & Doughty, C. 2020, MNRAS, 494, 5761, doi: 10.1093/mnras/staa1151
Hassani, H., & Javanmard, A. 2022, arXiv e-prints, arXiv:2201.05149. https://arxiv.org/abs/2201.05149
Heavens, A. F., Sellentin, E., de Mijolla, D., & Vianello, A. 2017, MNRAS, 472, 4244, doi: 10.1093/mnras/stx2326
Hirschmann, M., Dolag, K., Saro, A., et al. 2014a, MNRAS, 442, 2304, doi: 10.1093/mnras/stu1023
—. 2014b, MNRAS, 442, 2304, doi: 10.1093/mnras/stu1023
Hopkins, P. F. 2015, Monthly Notices of the Royal Astronomical Society, 450, 53, doi: 10.1093/mnras/stv195
Howlett, C., Said, K., Lucey, J. R., et al. 2022a, MNRAS, 515, 953, doi: 10.1093/mnras/stac1681
—. 2022b, MNRAS, 515, 953, doi: 10.1093/mnras/stac1681
Howlett, C., Staveley-Smith, L., & Blake, C. 2017, MNRAS, 464, 2517, doi: 10.1093/mnras/stw2466
Hu, W., & Kravtsov, A. V. 2003, ApJ, 584, 702, doi: 10.1086/345846

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (Princeton University Press), doi: 10.1515/9781400848911

Jeffrey, N., & Wandelt, B. D. 2020, arXiv e-prints, arXiv:2011.05991. https://arxiv.org/abs/2011.05991

Jespersen, C. K., Cranmer, M., Melchior, P., et al. 2022, arXiv e-prints, arXiv:2210.13473. https://arxiv.org/abs/2210.13473

Jo, Y., & Kim, J.-h. 2019, MNRAS, 489, 3565, doi: 10.1093/mnras/stz2304

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

Kamdar, H. M., Turk, M. J., & Brunner, R. J. 2016, MNRAS, 457, 1162, doi: 10.1093/mnras/stv2981

Kasmanoff, N., Villaescusa-Navarro, F., Tinker, J., & Ho, S. 2020, arXiv e-prints, arXiv:2012.00186. https://arxiv.org/abs/2012.00186

Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980. https://arxiv.org/abs/1412.6980

Kourkchi, E., Tully, R. B., Eftekharzadeh, S., et al. 2020, ApJ, 902, 145, doi: 10.3847/1538-4357/abb66b

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193. https://arxiv.org/abs/1110.3193

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. 2017, arXiv e-prints, arXiv:1712.09913. https://arxiv.org/abs/1712.09913

Li, Y., Hu, W., & Takada, M. 2014, PhRvD, 89, 083519, doi: 10.1103/PhysRevD.89.083519

Lovell, C. C., Wilkins, S. M., Thomas, P. A., et al. 2022, MNRAS, 509, 5046, doi: 10.1093/mnras/stab3221

Makinen, T. L., Charnock, T., Lemos, P., et al. 2022, arXiv e-prints, arXiv:2207.05202. https://arxiv.org/abs/2207.05202

Mangena, T., Hassan, S., & Santos, M. G. 2020, MNRAS, 494, 600, doi: 10.1093/mnras/staa750

Marinacci, F., Vogelsberger, M., Pakmor, R., et al. 2018, MNRAS, 480, 5113, doi: 10.1093/mnras/sty2206

McGibbon, R. J., & Khochfar, S. 2022, MNRAS, 513, 5423, doi: 10.1093/mnras/stac1269

Moster, B. P., Naab, T., Lindström, M., & O'Leary, J. A. 2021, MNRAS, 507, 2115, doi: 10.1093/mnras/stab1449

Naiman, J. P., Pillepich, A., Springel, V., et al. 2018, MNRAS, 477, 1206, doi: 10.1093/mnras/sty618

Nelson, D., Pillepich, A., Springel, V., et al. 2018, MNRAS, 475, 624, doi: 10.1093/mnras/stx3040

Nelson, D., Springel, V., Pillepich, A., et al. 2019, Computational Astrophysics and Cosmology, 6, 2, doi: 10.1186/s40668-019-0028-x

Ni, Y., Di Matteo, T., Bird, S., et al. 2022, MNRAS, 513, 670, doi: 10.1093/mnras/stac351

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

Peebles, P. J. E. 1980, The large-scale structure of the universe (Princeton University Press)

Perez, L. A., Genel, S., Villaescusa-Navarro, F., et al. 2022, arXiv e-prints, arXiv:2204.02408. https://arxiv.org/abs/2204.02408

Pillepich, A., Nelson, D., Hernquist, L., et al. 2018a, MNRAS, 475, 648, doi: 10.1093/mnras/stx3112

Pillepich, A., Springel, V., Nelson, D., et al. 2018b, MNRAS, 473, 4077, doi: 10.1093/mnras/stx2656

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

Pontoppidan, K. M., Barrientes, J., Blome, C., et al. 2022, ApJL, 936, L14, doi: 10.3847/2041-8213/ac8a4e

Racca, G. D., Laureijs, R., Stagnaro, L., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9904, Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave, ed. H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, & E. C. Tong, 99040O, doi: 10.1117/12.2230762

Ravanbakhsh, S., Oliva, J., Fromenteau, S., et al. 2017, arXiv e-prints, arXiv:1711.02033. https://arxiv.org/abs/1711.02033

Rodrigues, N. V. N., de Santi, N. S. M., Montero-Dorta, A. D., & Abramo, L. R. 2023, arXiv e-prints, arXiv:2301.06398. https://arxiv.org/abs/2301.06398
Schaller, M., Gonnet, P., Chalk, A. B. G., & Draper, P. W. 2016, in Proceedings of the Platform for Advanced Scientific Computing Conference, 2, doi: 10.1145/2929908.2929916
Schaller, M., et al. 2018, SWIFT: SPH With Inter-dependent Fine-grained Tasking, Astrophysics Source Code Library. http://ascl.net/1805.020
Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521, doi: 10.1093/mnras/stu2058
Shao, H., Villaescusa-Navarro, F., Villanueva-Domingo, P., et al. 2022a, arXiv e-prints, arXiv:2209.06843. https://arxiv.org/abs/2209.06843
Shao, H., Villaescusa-Navarro, F., Genel, S., et al. 2022b, ApJ, 927, 85, doi: 10.3847/1538-4357/ac4d30
Shao, H., de Santi, N. S. M., Villaescusa-Navarro, F., et al. 2023, arXiv e-prints, arXiv:2302.14591, doi: 10.48550/arXiv.2302.14591
Sobol', I. 1967, USSR Computational Mathematics and Mathematical Physics, 7, 86, doi: https://doi.org/10.1016/0041-5553(67)90144-9
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757. https://arxiv.org/abs/1503.03757
Springel, V. 2005, MNRAS, 364, 1105, doi: 10.1111/j.1365-2966.2005.09655.x
—. 2010, MNRAS, 401, 791, doi: 10.1111/j.1365-2966.2009.15715.x
Springel, V., Di Matteo, T., & Hernquist, L. 2005, MNRAS, 361, 776, doi: 10.1111/j.1365-2966.2005.09238.x
Springel, V., & Hernquist, L. 2002, MNRAS, 333, 649, doi: 10.1046/j.1365-8711.2002.05445.x
—. 2003, MNRAS, 339, 289, doi: 10.1046/j.1365-8711.2003.06206.x
Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726, doi: 10.1046/j.1365-8711.2001.04912.x
Springel, V., Pakmor, R., Pillepich, A., et al. 2018, MNRAS, 475, 676, doi: 10.1093/mnras/stx3304
Steinborn, L. K., Dolag, K., Comerford, J. M., et al. 2016, MNRAS, 458, 1013, doi: 10.1093/mnras/stw316
Strauss, M. A., & Willick, J. A. 1995, PhR, 261, 271, doi: 10.1016/0370-1573(95)00013-7
Takada, M., & Bridle, S. 2007, New Journal of Physics, 9, 446, doi: 10.1088/1367-2630/9/12/446
Takada, M., Ellis, R. S., Chiba, M., et al. 2014, PASJ, 66, R1, doi: 10.1093/pasj/pst019
Taylor, A., & Braun, R., eds. 1999, Science with the Square Kilometer Array : a next generation world radio observatory
Taylor, A., Joachimi, B., & Kitching, T. 2013, MNRAS, 432, 1928, doi: 10.1093/mnras/stt270
Tornatore, L., Borgani, S., Dolag, K., & Matteucci, F. 2007, MNRAS, 382, 1050, doi: 10.1111/j.1365-2966.2007.12070.x
Uhlemann, C., Friedrich, O., Villaescusa-Navarro, F., Banerjee, A., & Codis, S. 2020, MNRAS, 495, 4006, doi: 10.1093/mnras/staa1155
Villaescusa-Navarro, F., Hahn, C., Massara, E., et al. 2020, ApJS, 250, 2, doi: 10.3847/1538-4365/ab9d82
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021a, arXiv e-prints, arXiv:2109.09747. https://arxiv.org/abs/2109.09747
—. 2021b, ApJ, 915, 71, doi: 10.3847/1538-4357/abf7ba
Villaescusa-Navarro, F., Ding, J., Genel, S., et al. 2022a, ApJ, 929, 132, doi: 10.3847/1538-4357/ac5d3f
Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., et al. 2022b, arXiv e-prints, arXiv:2201.01300. https://arxiv.org/abs/2201.01300
—. 2022c, ApJS, 259, 61, doi: 10.3847/1538-4365/ac5ab0
Villanueva-Domingo, P., & Villaescusa-Navarro, F. 2022, ApJ, 937, 115, doi: 10.3847/1538-4357/ac8930
Villanueva-Domingo, P., Villaescusa-Navarro, F., Genel, S., et al. 2021, arXiv e-prints, arXiv:2111.14874. https://arxiv.org/abs/2111.14874
Villanueva-Domingo, P., Villaescusa-Navarro, F., Anglés-Alcázar, D., et al. 2022, ApJ, 935, 30, doi: 10.3847/1538-4357/ac7aa3
von Marttens, R., Casarini, L., Napolitano, N. R., et al. 2022, MNRAS, 516, 3924, doi: 10.1093/mnras/stac2449

Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2020, arXiv e-prints, arXiv:2012.00111. https://arxiv.org/abs/2012.00111

Weinberger, R., Springel, V., & Pakmor, R. 2020, ApJS, 248, 32, doi: 10.3847/1538-4365/ab908c

Weinberger, R., Springel, V., Hernquist, L., et al. 2017a, MNRAS, 465, 3291, doi: 10.1093/mnras/stw2944

—. 2017b, MNRAS, 465, 3291, doi: 10.1093/mnras/stw2944

Yip, J. H. T., Zhang, X., Wang, Y., et al. 2019, arXiv e-prints, arXiv:1910.07813. https://arxiv.org/abs/1910.07813

Zaheer, M., Kottur, S., Ravanbakhsh, S., et al. 2017, arXiv e-prints, arXiv:1703.06114. https://arxiv.org/abs/1703.06114

Zhang, X., Wang, Y., Zhang, W., et al. 2019, arXiv e-prints, arXiv:1902.05965. https://arxiv.org/abs/1902.05965

Zhou, J., Cui, G., Hu, S., et al. 2018, arXiv e-prints, arXiv:1812.08434. https://arxiv.org/abs/1812.08434