Title
Galaxy Phase-space and Field-level Cosmology: The Strength of Semianalytic Models

Journal
The Astrophysical Journal, 1007(1)

ISSN
0004-637X

Authors
de Santi, Natalí SM
Villaescusa-Navarro, Francisco
Araya-Araya, Pablo
et al.

Publication Date
2026-08-10

DOI
10.3847/1538-4357/ae84b4

Draft version August 4, 2026
Typeset using LaTeX default style in AASTeX631

Galaxy Phase-Space and Field-Level Cosmology: The Strength of Semi-Analytic Models

Natalí S. M. de Santi [1,2,3] Francisco Villaescusa-Navarro [3,4] Pablo Araya-Araya [5,6] Gabriella De Lucia,[7,8] Fabio Fontanot [7,8] Lucia A. Perez [3,9] Manuel Arnés-Curto [10,11] Violeta Gonzalez-Perez [10,12] Ángel Chandro-Gómez [10,13,14] Rachel S. Somerville [3] and Tiago Castro [15,16,17,18]

[1] *Berkeley Center for Cosmological Physics, University of California, Berkeley, 341 Campbell Hall, Berkeley, CA 94720, USA*
[2] *Physics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[3] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY, 10010, USA*
[4] *Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544 USA*
[5] *Cosmic Dawn Center (DAWN), Copenhagen, Denmark*
[6] *DTU Space, Technical University of Denmark, Elektrovej 327, DK2800 Kgs. Lyngby, Denmark*
[7] *INAF - Astronomical Observatory of Trieste, via G.B. Tiepolo 11, I-34143 Trieste, Italy*
[8] *IFPU - Institute for Fundamental Physics of the Universe, via Beirut 2, 34151, Trieste, Italy*
[9] *Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA*
[10] *Departamento de Física Teórica, Universidad Autónoma de Madrid, 28049 Madrid, Spain*
[11] *Departamento de Física Fundamental and IUFFyM, Universidad de Salamanca, E-37008 Salamanca, Spain*
[12] *Centro de Investigación Avanzada en Física Fundamental (CIAFF), Facultad de Ciencias, Universidad Autónoma de Madrid, ES-28049 Madrid, Spain*
[13] *International Centre for Radio Astronomy Research, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia*
[14] *ARC Centre for All-Sky Astrophysics in 3 Dimensions (ASTRO 3D)*
[15] *INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[16] *INFN, Sezione di Trieste, Via Valerio 2, I-34127 Trieste TS, Italy*
[17] *IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy*
[18] *CSC - Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing, Via Magnanelli 2, Bologna, Italy*

ABSTRACT

Semi-analytic models are a widely used approach to simulate galaxy properties within a cosmological framework, relying on simplified yet physically motivated prescriptions. They have also proven to be an efficient alternative for generating accurate galaxy catalogs, offering a faster and less computationally expensive option compared to full hydrodynamical simulations. In this paper, we demonstrate that using only galaxy 3D positions and radial velocities, we can train a graph neural network coupled to a moment neural network to obtain a robust machine learning based model capable of estimating the matter density parameters, $\Omega_{\rm m}$, with a precision of approximately 10%. The network is trained on $(25h^{-1}{\rm Mpc})^3$ volumes of galaxy catalogs from L-Galaxies and can successfully extrapolate its predictions to other semi-analytic models (GAEA, SC-SAM, and Shark) and, more remarkably, to hydrodynamical simulations (`Astrid`, SIMBA, IllustrisTNG, and SWIFT-EAGLE). Our results show that the network is robust to variations in astrophysical and subgrid physics, cosmological and astrophysical parameters, and the different halo-profile treatments used across simulations. This suggests that the physical relationships encoded in the phase-space of semi-analytic models are largely independent of their specific physical prescriptions, reinforcing their potential as tools for the generation of realistic mock catalogs for cosmological parameter inference.

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

Keywords: Semi-analytical models — hydrodynamical simulations — machine learning — cosmological parameter inference

1. INTRODUCTION

It is well known that galaxies are not randomly distributed in space; rather, they are observables that trace the cosmic web, the large scale structure of the Universe, where dark matter (DM) plays a central role in structure formation. Galaxies not only trace the underlying DM distribution but also maintain an intricate relationship with the DM halos in which they form and evolve, what is commonly referred to as the halo-galaxy connection. Modeling galaxy evolution within a cosmological context therefore represents one of the greatest challenges in astrophysics, owing to the vast range of scales and the multitude of physical processes involved (Somerville & Davé 2015).

Over the past decades, various methods have been developed to model the physics underlying the halo-galaxy connection. These approaches generally fall into two main categories: physical and empirical models (Wechsler & Tinker 2018; De Lucia 2019). Physical models, including hydrodynamical simulations and semi-analytic models (SAMs), aim to capture the fundamental processes that govern galaxy formation within DM halos. In contrast, empirical models such as subhalo abundance matching (SHAM, e.g. Kravtsov et al. 2004; Yu et al. 2024) and halo occupation distribution (HOD, e.g. Zheng et al. 2005; Avila et al. 2020) frameworks are more data-driven, relying on statistical correlations between galaxy and halo properties inferred from simulations or observations.

In this work, we focus on the two principal physics based techniques that serve as our datasets: hydrodynamical simulations and SAMs. This choice is motivated by the fact both methods produce remarkably consistent predictions, that are in good agreement with observations (Somerville & Davé 2015; Wechsler & Tinker 2018). Furthermore, using a range of models allow us to account for the complex physics of the halo-galaxy connection and to mitigate the impact of differences between modeling approaches.

Hydrodynamical simulations solve the equations of gravity, hydrodynamics, and thermodynamics using numerical techniques applied to particles and/or grid cells representing DM and gas (Somerville & Davé 2015; Crain & van de Voort 2023). Their main limitations, however, stem from their high computational cost and from the still incomplete understanding of the small-scale physics governing baryonic processes, which must be incorporated through various subgrid prescriptions.

In this context, SAMs represent one of the most effective current alternatives for predicting observables in the Universe (Baugh 2006; Benson 2010; Somerville & Davé 2015; Gonzalez-Perez et al. 2020; Gómez et al. 2025). They are built on top of DM-only simulations, in which various physical processes are approximated through analytic prescriptions that can be tracked along the merger histories of DM halos (Somerville & Primack 1999; Hirschmann et al. 2016; Henriques et al. 2020). As a result, SAMs drastically reduce computational demands compared to fully numerical hydrodynamical simulations, enabling both the efficient exploration of a wide parameter space and the generation of catalogs over much larger volumes. Their main limitation, however, lies in the approximations themselves: they are less capable of capturing small-scale physical effects and can differ in their assumptions for the underlying halo profile and gas dynamic prescriptions.

Recent studies (Anagnostidis et al. 2022; Villanueva-Domingo & Villaescusa-Navarro 2022; Shao et al. 2022; Makinen et al. 2022; de Santi et al. 2023a,b; Shao et al. 2023; Roncoli et al. 2023; Chatterjee & Villaescusa-Navarro 2024; Cuesta-Lazaro & Mishra-Sharma 2024) have demonstrated that a powerful approach to constraining cosmological parameters from galaxy and/or halo catalogs involves representing them as point clouds or graphs, and performing field-level simulation-based inference (SBI), also referred to as likelihood-free inference (LFI) or implicit likelihood inference (ILI). These works have primarily relied on DM-only or hydrodynamical simulations, leveraging their exact numerical solutions, which provide precise positions and velocities of the simulated structures.

Building upon the use of these methods for cosmological parameter inference, a key challenge, especially when aiming for applications to real data, is the development of robust machine learning (ML) suites. These models are inherently data-driven, meaning they learn patterns based on the intrinsic properties of the data they are trained on and, consequently, often fail to extrapolate accurately to datasets that differ from their training domain (Hassani & Javanmard 2022). Since we do not yet know which physical model best describes our Universe, it is essential to design methods that exhibit consistent predictive power across different simulation frameworks, and then more importantly, are able to extract cosmological and astrophysical information once applied to real observations.

The lack of robustness in current ML models remains an open issue and can be attributed to several factors: (1) limited overlap between datasets, arising from differences in the chosen astrophysical parameter variations and in the resulting galaxy-property distributions; (2) models learning spurious or non-physical features (e.g., numerical artifacts); and (3) differences in data representation, such as comparing raw phase-space information with compressed latent or manifold-based representations.

Robustness has previously been explored across the CAMELS hydrodynamical simulations using a variety of data formats, including 2D projected maps with convolutional neural networks (Villaescusa-Navarro et al. 2021a), galaxy tabular data (Villaescusa-Navarro et al. 2022a), neutral hydrogen maps (Jo et al. 2025), and galaxy and/or halo catalogs represented as graphs or point clouds (Villanueva-Domingo & Villaescusa-Navarro 2022; Shao et al. 2022, 2023; de Santi et al. 2023a). Additionally, analyses using different N-body simulations were conducted by Bayer et al. (2025), who showed that statistical out-of-distribution behavior can arise from both resolution effects and the sensitivity of ML methods to small-scale fluctuations.

Regarding galaxy catalogs, several approaches have already been explored to mitigate this lack of robustness. One of them relies on domain adaptation (DA) techniques (Csurka 2017; Wang & Deng 2018; Farahani et al. 2020; Kumari & Singh 2023), which aim to create models that generalize across different data domains. Most of these methods focus on modifying the loss function to include a measure of the discrepancy between the latent-space probability distributions of the different datasets.

DA techniques, when combined with graph neural networks (GNNs), have already been applied to the estimation of $\Omega_{\rm m}$ across different hydrodynamical simulations, specifically IllustrisTNG and SIMBA, using the galaxies' phase-space information (Roncoli et al. 2023). The authors found that performance improved when training on SIMBA and testing on IllustrisTNG, although a bias persisted.

An alternative strategy is to train ML models on datasets that are sufficiently diverse to encompass the variability present across different simulations. This approach was adopted in de Santi et al. (2023a), where a GNN coupled with a moment neural network (MNN), trained on `Astrid` using galaxy phase-space information, was shown to successfully extrapolate its predictions of $\Omega_{\rm m}$ across five different subgrid physical models. In that case, the model's success was attributed to the greater diversity of the training dataset, both in the total number of galaxies per catalog and in the variation of their properties (Ni et al. 2023), combined with the chosen graph construction and GNN architecture.

Beyond assessing the robustness of different ML approaches, it is essential to build datasets using methods that are computationally fast and scalable. In de Santi et al. (2023a), the first robust field-level model across multiple hydrodynamical simulations was demonstrated. However, extending this idea by running thousands of large-volume hydrodynamical simulations, such as `Astrid`, is simply not feasible with current computational resources. This is precisely why SAMs are so promising: they offer a rapid yet physically motivated way to generate diverse galaxy catalogs, making them strong candidates to match the performance of hydrodynamical simulations for SBI and, ultimately, for the creation of realistic mocks needed for real-data applications.

In this paper we make use of galaxy 3D positions and radial velocities from L-Galaxies to train a GNN associated with a MNN to predict Ω_m, looking for a robust model that could overcome the main differences over different SAMs and different hydrodynamic simulations. The manuscript is structured as follows: in Section 2 we describe the dataset used to train and test the model (specifications related to the SAMs are provided in Appendix A); in Section 3 we describe the methodology, presenting an overview of the ML model used as well as the training procedure (more details can be found in the Appendix B); in Section 4 the results are shown and explained; Section 5 presents a discussion and conclusions. Additionally, a Normalizing Flows (NFs) were coupled with the GNN instead of the MNN, to perform the full posterior inferences. The results of this experiment are shown and discussed in Appendix C.

2. DATA

2.1. *Simulations*

The galaxy catalogs we use to train, validate, and test our models come from thousands of hydrodynamic and semi-analytic simulations of the Cosmology and Astrophysics with MachinE Learning Simulations – CAMELS project (Villaescusa-Navarro et al. 2021b, 2022b), that include periodic boxes of 25 h^{-1}Mpc on a side. All the simulations follow the evolution of 256^3 DM particles and are initialized with 256^3 fluid elements from $z = 127$ down to $z = 0$. The catalogs used in this work correspond to $z = 0$.

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:

Table 1. Characteristics of the SAMs and hydrodynamic simulations used in this work.

Model	Type	Usage	Number of simulations used	Mean number of galaxies per catalog	Reference
L-Galaxies	SAM	Train, validate & Test	1,000(LH) + 27(CV)	5,233	Henriques et al. (2020)
GAEA	SAM	Test	1,000(LH) + 27(CV)	1,868	De Lucia et al. (2024)
SC-SAM	SAM	Test	50(LH) + 23(CV)	1,337	Perez et al. (2023)
Shark	SAM	Test	1,000(LH) + 27(CV)	940	Lagos et al. (2024)
Astrid	Hydro	Test	1,000(LH) + 27(CV)	1,114	Bird et al. (2022)
SIMBA	Hydro	Test	1,000(LH) + 27(CV)	1,093	Davé et al. (2019)
IllustrisTNG	Hydro	Test	1,000(LH) + 27(CV) + 1,024(SB)	737	Pillepich et al. (2018)
Magneticum	Hydro	Test	50(LH) + 27(CV)	3,655	Hirschmann et al. (2014a)
SWIFT-EAGLE	Hydro	Test	1,000(LH) + 27(CV)	1,220	Schaye et al. (2015)

- **Latin Hypercube (LH).** The simulations in this category have their cosmological and astrophysical parameters sampled according to a LH design. For the cosmological parameters, their values span the following ranges: $\Omega_{\rm m} \in [0.1, 0.5]$ and $\sigma_8 \in [0.6, 1.0]$. In the hydrodynamical simulations, only four astrophysical parameters are varied: $A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, and $A_{\rm AGN2}$. These parameters control the efficiency of supernova (SN) and active galactic nuclei (AGN) feedback, although their specific definitions vary across the different hydrodynamic models. For explicit details, see Villaescusa-Navarro et al. (2021b); Ni et al. (2023). The parameter ranges are: $A_{\rm SN1} \in [0.25, 4.0]$, $A_{\rm SN2} \in [0.5, 2.0]$, $A_{\rm AGN1} \in [0.25, 4.0]$, and $A_{\rm AGN2} \in [0.5, 2.0]$. For the SAMs, a much larger number of parameters are varied, and they are different depending on the SAM in question (due to their different physical prescriptions); these parameters are detailed in Appendix A. Other cosmological parameters such as $\Omega_{\rm b}$, h, n_s, and w follow the same values of the Cosmic Variance set. Each of the simulations in the LH 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 the ML model.

- **Cosmic Variance (CV).** These simulations have been run with the fiducial value of the cosmological and astrophysical parameters, that follows for: $\Omega_m = 0.3$, $\sigma_8 = 0.8$, $\Omega_{\rm b} = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $w = -1$, $M_\nu = 0$ eV and $A_{SN1} = A_{SN2} = A_{AGN1} = A_{AGN2} = 1$. The fiducial values employed for each SAM can also be found in Appendix A. 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 ML models.

- **Sobol Sequence (SB).** The simulations in this set have their cosmological and astrophysical parameters arranged following a Sobol sequence (Sobol' 1967). A total of 28 parameters are varied: five cosmological ($\Omega_{\rm m}$, $\Omega_{\rm b}$, h, n_s, σ_8) and 23 astrophysical. The astrophysical parameters include the commonly used ones ($A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, $A_{\rm AGN2}$) as well as others related to star formation, galactic winds, black hole (BH) growth, and quasar activity. All parameters vary within ranges centered around their fiducial values. We use these simulations only for testing (and only under the IllustrisTNG prescription accordingly to Ni et al. 2023) to assess how well our ML models generalize to the new effects introduced by these parameter variations, which differ from those explored during training.

The SAMs have been run over DM halo merger trees, generated by the DM-only counterpart from the Astrid hydrodynamic simulation, ran using MP-Gadget (Feng et al. 2018). The merger trees have been obtained with SubLink (Nelson et al. 2015; Rodriguez-Gomez et al. 2015), using their extended format and halos found with SubFind (Springel et al. 2001; Dolag et al. 2009), for L-Galaxies and GAEA. SC-SAM and Shark were run with Rockstar subhalo catalogs (Behroozi et al. 2013a) in combination with ConsistentTrees (Behroozi et al. 2013b). All these merger trees follow the same particle and redshift evolution, for the same size boxes and cosmological parameters of the former simulations described.

The CAMELS-SAM simulations can also be classified into different model suites according to the code used to run them. See Appendix A for more details about the parameters that are varied:

- **L-Galaxies.** This SAM was run on the DM-only counterpart from the `Astrid` simulations, using merger trees obtained using SubLink and subhalos from SubFind and adopts the galaxy formation prescriptions from the Henriques et al. (2020) version of L-Galaxies[1]. The positions and velocities of satellite galaxies are normally tracked using the most bound DM particle, obtained from an input file optimized for the Millennium simulation (Springel et al. 2005b). However, because generating this input is not straightforward for the `Astrid` simulation, these quantities are instead extrapolated from the properties of the last subhalo before it becomes disrupted. The Henriques et al. (2020) model was calibrated using a Markov Chain Monte Carlo (MCMC) approach, constrained by the stellar mass function and the fraction of passive (quiescent) galaxies at $z = 0$ and $z = 2$, as well as by the neutral hydrogen mass function at $z = 0$. L-Galaxies includes 19 free parameters in total; following the calibration framework of Henriques et al. (2020), we varied 14 of them[2]. These parameters govern the equations describing key astrophysical processes, including secular star formation, merger-induced starbursts, supernova feedback, the growth of supermassive BHs and their associated AGN feedback, the reincorporation timescale of ejected gas, and tidal disruption. The 14 free parameters, their corresponding variation ranges, and fiducial values are listed in Table 3. They were sampled using the LH algorithm, generating $1,000$ LH catalogs, while the default Henriques et al. (2020) values were used to produce 27 control (CV) catalogs.

- **GAEA.** The GAEA[3] (GAlaxy Evolution and Assembly) SAM is based on De Lucia et al. (2014); Hirschmann et al. (2016); Xie et al. (2017); Fontanot et al. (2017, 2020); Xie et al. (2020); De Lucia et al. (2024); Xie et al. (2024) and their boxes were run over merger trees obtained with SubLink and subhalos with SubFind from the DM counterpart of `Astrid`. In GAEA, galaxy positions and velocities are treated as follows: central galaxies are placed at the center of the most bound subhalo and inherit its velocity. Satellite galaxies are associated with distinct DM subhalos identified with SubFind (Springel et al. 2001; Dolag et al. 2009), and their positions and velocities correspond to those of their host subhalo. When a satellite's subhalo is stripped below the resolution limit of the simulation, the galaxy becomes an orphan and its position and velocity are assigned to those of the most bound particle at the last snapshot in which the subhalo could still be identified (Xie et al. 2020). From the free parameters available, four of them were varied in a LH and used to run $1,000$ LH catalogs: two related to the SN efficiency and two related to AGN feedback. A more detailed explanation about these parameters can be found in Table 4, as well as their varied range and fiducial values. The model has been formally calibrated on the Millennium simulation (Springel et al. 2005b) and their parameters have been used to produce 27 CV boxes (De Lucia et al. 2024). Additionally, the same set of parameters has been proved to be robust at different Millennium resolutions (Fontanot et al. 2025).

- **SC-SAM.** These SAMs have been run over DM-only merger trees from `Astrid`, obtained from Rockstar catalogs in combination with ConsistentTrees, and based on the model presented in Somerville et al. (2008); Somerville et al. (2015). Three free parameters ($A_{\rm SN1}$, multiplicative pre-factor to $\epsilon_{\rm SN}$; $A_{\rm SN2}$, additive pre-factor to α_{rh}; and $A_{\rm AGN}$, multiplicative pre-factor to $\kappa_{\rm radio}$; Perez et al. 2023) have been varied to produce 50 LH catalogs (see Table 5 for more details, ranges and fiducial values). The orbits of satellite galaxies are tracked using an analytic dynamical friction model (Somerville et al. 2008), and satellite positions are re-assigned in post-processing to follow an NFW profile (Navarro et al. 1997). The model has been calibrated in order to follow observable relations such as stellar mass function, the stellar mass-halo mass relation, the cold gas fraction vs. stellar mass, stellar metallicity-stellar mass, and BH mass-bulge mass relationships accordingly to observations and IllustrisTNG300 (Gabrielpillai et al. 2022; Perez et al. 2023), producing 23 CV catalogs.

- **Shark.** Catalogs of model galaxies from Shark are based on the fiducial calibration presented in Lagos et al. (2018, 2024), with the exception of $v_{\rm hot}$, ν_{SF}, $\rm e_{sb}$ and $\epsilon_{\rm disc}$. These four parameters have been set to match different stellar mass function datasets (Baldry et al. 2012, Moustakas et al. 2013, Ilbert et al. 2013, Li & White 2009) when using halos from the CV boxes. Hence, these parameters are slightly different from the fiducial ones.

[1] https://lgalaxiespublicrelease.github.io/

[2] Henriques et al. (2020) also calibrated an additional parameter related to ram pressure stripping, which we fixed at their best-fit value.

[3] An introduction to GAEA, a list of recent work, as well as a data file containing published model predictions, can be found at https://sites.google.com/inaf.it/gaea/home.

The fiducial calibration was performed using merger trees from SURFS DM-only simulations (Elahi et al. 2018), with halos identified with the VELOCIraptor finder (Elahi et al. 2019a; Cañas et al. 2019) and links made with TreeFrog (Elahi et al. 2019b). Shark tracks satellite galaxies using subhalo information whenever the subhalo is resolved. Once the subhalo disappears, the resulting orphan galaxies are followed analytically using a NFW-based prescription for their positions and velocities (Navarro et al. 1997). These model has been run over Astrid's DM-only simulations also using Rockstar catalogs in combination with ConsistentTrees, although post-processed in this case by the code DHalo (Jiang et al. 2014). This model was used to generate 27 CV boxes, sampling 16 free parameters over 1,000 LH simulations. These model free parameters govern astrophysical processes such as stellar and AGN feedbacks, star formation, BH growth, gas reincorporation, tidal stripping and disc instabilities. A summary of the parameters and their fiducial values and corresponding ranges is provided in Table 6.

We stress that the free parameters varied across the different SAMs do not follow a common pattern. For instance, L-Galaxies and Shark vary 14 and 16 parameters, respectively, whereas GAEA and SAM-SC vary only four and three. Such differences naturally lead to variations in their resulting galaxy populations. We stress that the free parameters varied over the different SAMs do not follow a pattern.

Although the SAMs presented in this paper are all built on the same Astrid DM-only counterpart, they have been calibrated and run using different halo finders and merger tree builders (e.g., SubLink+SubFind, Rockstar+ConsistentTrees, and VELOCIraptor+TreeFrog). Such differences can introduce significant scatter in galaxy properties (Knebe et al. 2015). For example, Avila et al. (2014) showed that variations in halo-finder algorithms can substantially alter the mass-assembly histories from which merger trees are constructed. Similarly, even when the same SAM is applied to different merger trees, large discrepancies in galaxy predictions may arise (Lee et al. 2014), although other studies have found good consistency in some cases (Gómez et al. 2022). In addition, numerical artifacts inherent to each merger tree algorithm can further affect SAM predictions (Chandro-Gómez et al. 2025), reflecting both differences in merger tree construction (Srisawat et al. 2013) and the halo finders on which they rely (Avila et al. 2014). Finally, although beyond the scope of this work, different N-body codes can yield noticeably different DM fields that also require marginalization (Bayer et al. 2025), potentially compounding these effects.

The hydrodynamic simulations have been run with different codes that solve the hydrodynamic equations differently and implement different subgrid models: IllustrisTNG (Weinberger et al. 2017; Pillepich et al. 2018), SIMBA (Davé et al. 2019), Astrid (Bird et al. 2022), Magneticum (Hirschmann et al. 2014a), and SWIFT-EAGLE (Schaye et al. 2015; Crain et al. 2015; Schaller et al. 2016; Schaller et al. 2024).

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

- **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 1,000 LH and 27 CV 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 1,000 LH and 27 CV simulations.

- **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. 2017; Pillepich et al. 2018). This suite contains 1,000 LH, 27 CV, and 2048 SB 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 also includes 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. (2005a) 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.

- **SWIFT-EAGLE.** These simulations have been run with the SWIFT code (Schaller et al. 2016, 2018; Schaller et al. 2024) 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 Lovell & et. al. (2025). This suite contains 1,000 LH and 27 CV simulations.

Finally, the halos and subhalos are identified in the hydrodynamical simulations for every snapshot using the halo and subhalo finder: SUBFIND (Springel et al. 2001; Dolag et al. 2009).

2.2. *Galaxy catalogs*

In this work, we consider only galaxies with stellar masses above $1.3 \times 10^8\ M_\odot/h$, a threshold that lies well above the resolution and convergence limits of all simulations employed. It is important to note that stellar masses are defined differently in SAMs and hydrodynamical simulations. In SAMs, stellar masses arise from analytic prescriptions (e.g., SHARK separates bulge and disk components, while SAM-SC and GAEA track the total stellar content analytically). In contrast, hydrodynamical simulations compute stellar masses directly from the numerical evolution, by summing the masses of all star particles bound to each galaxy's subhalo.

Each galaxy catalog is then constructed by selecting all galaxies whose stellar masses exceed the chosen threshold, and for every simulation we generate multiple catalogs by varying this mass limit.

A summary of the simulation characteristics is provided in Table 1, which lists if they are SAMs or hydrodynamic simulations, their respective usages (train, validation and test) for the ML model, the number of catalogs for the different sets (LH, SB, and CV), the mean number of galaxies per catalog[4], 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 associated to the MNNs (Section 3.2) and the training procedure and optimization choices (Section 3.3). Any further detail related to this section can be found in Appendix B.

3.1. *Galaxy graphs*

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 galaxy catalogs that contain the galaxy positions and their peculiar velocities (only the z component), following the same prescription as used in de Santi et al. (2023a). 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}, a value which we consider as a hyperparameter for the network. We use as a global property of the graph the logarithm of the number of galaxies in the graph: $\log_{10}(N_g)$. 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. The exact node and edge attributes are described in B.1.

3.2. *Architecture*

The architecture we employ in this work also follows the one presented in de Santi et al. (2023a) and Villanueva-Domingo & Villaescusa-Navarro (2022)[5]. Basically, the GNN is trained to infer another global property of each input graph, the value of $\Omega_{\rm m}$ from each galaxy catalog. The main idea is to perform a transformation of the graph components' information (i.e. nodes $\mathbf{n}_i$, edges $\mathbf{e}_{ij}$, and global $\mathbf{g}$ attributes are updated), while the graph structure is preserved. This is known as message passing scheme (see the exact implementation in Section B.2). In the end, the compressed graph information is converted by a multi-layer perceptron (MLP) into the final predicted property of the graph ($\Omega_{\rm m}$). The loss function is designed so that the MNNs infer the first (μ, see Equation B13) and second moments (σ, see Equation B14) of the posterior distribution. 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 architecture presented in this work was done using PYTORCH GEOMETRIC (Fey & Lenssen 2019).

[4] See Section 4 for a discussion regarding the different number of galaxies in the different simulations.

[5] See COSMOGRAPHNET github repository: https://github.com/PabloVD/CosmoGraphNet.

Table 2. Hyperparameters values selected for the best model.

Hyperparameter	Best value
Hidden Features	27
Number of GNN layers	2
r_{link}	1.11Mpc/h
Learning Rate	$2.13 \cdot 10^{-7}$
Weight Decay	$9.5 \cdot 10^{-3}$

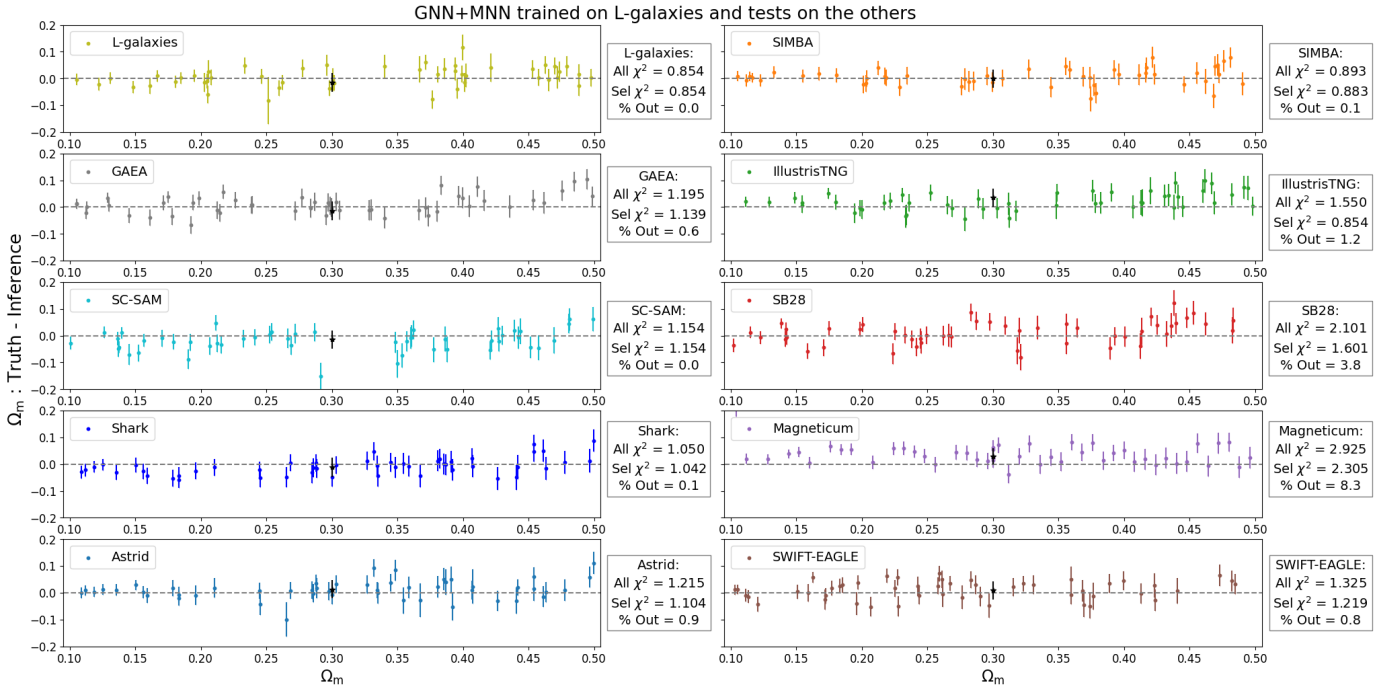

Figure 1. Truth - Inference values for Ω_m: GNN + MNN. The model trained on L-Galaxies is evaluated on L-Galaxies, GAEA, SC-SAM, `Astrid`, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE catalogs. We plot the deviation of our predicted Ω_m from the true values (truth - μ; see Equation B13), with error bars corresponding to σ (see Equation B14) predictions. Colored points represent the testing points (LH or SB sets), while the black points correspond to the average over the CV set inferences. We also present in the side legends the χ^2 score (see Equation B22) for all samples in the test sets and the selected samples, as well as the percentage of samples removed per test set. Overall, the model remains robust across all simulations considered, including both SAMs and hydrodynamical simulations.

3.3. *Training procedure and optimization choices*

We train our models on graphs constructed from galaxy catalogs of the LH sets from L-Galaxies. We initially split the 1,000 LH simulations into training (850 simulations), validation (100 simulations), and testing (50 simulations) sets. 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 model. The same trick was employed on de Santi et al. (2023a). For each catalog, we produce a graph as outlined in Section 3.1.

We then train the model using the above architecture for 300 epochs making use of the 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 the *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. The best set of parameters are in Table 2.

We also trained the network on GAEA, but the resulting model was not robust, and therefore we do not present those results in detail. Further discussion of this point can be found in Section 4.2.

4. RESULTS

In this section, we present the main results obtained with the GNN-MNN model trained on L-Galaxies and tested across the different SAM prescriptions and hydrodynamical simulations. Section 4.1 reports the model performance for both the cosmological and astrophysical parameter variations (LH and SB28 testing sets), as well as for the fiducial cosmology (CV set). In Section 4.2, we analyze the impact of the number of galaxies per catalog on the robustness of the model.

4.1. *Field-Level Inference and Performance Metrics*

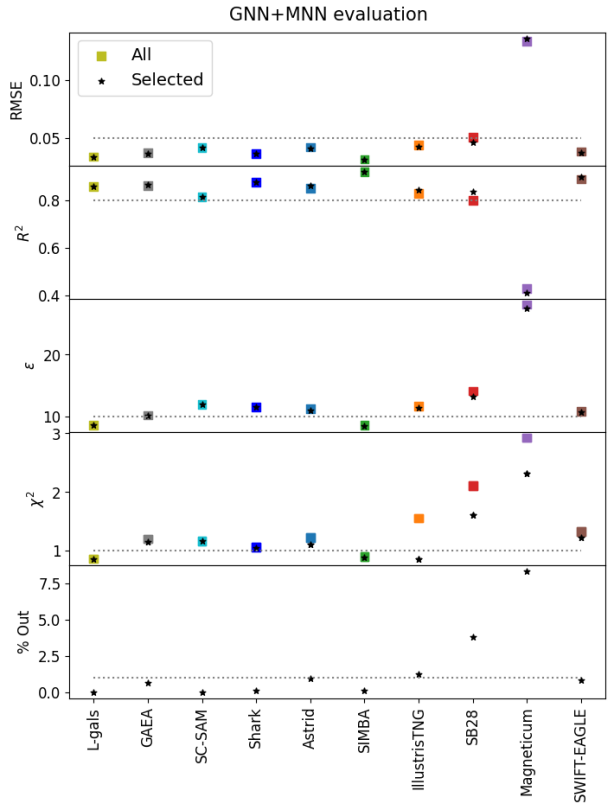

Figure 2. GNN+MNN: Model Evaluation. We present RMSE (root mean square error), R^2 (coefficient of determination), ϵ (mean relative error), χ^2 (reduced chi squared), and % Out (percentage of samples removed) for the test sets of L-Galaxies, GAEA, SC-SAM, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE indicated in the x-axis. Results are presented to all and selected samples according to χ^2 values.

In Figure 1, we present the difference between the inferred and true values of $\Omega_{\rm m}$ as a function of the true values. The true values are represented accordingly to the first moment μ (see Equation B13) as well as their error bars correspond to the second moment σ (see Equation B14), predicted by the MNN (see more details in Section B.3). The model was evaluated on the following galaxy formation models: L-Galaxies (testing set), GAEA, SC-SAM, Shark, Astrid, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE. Horizontal gray dashed lines indicate the ideal case of perfect predictions (i.e., zero residual).

For clarity and to avoid overcrowding the figure, we display only 50 randomly chosen samples from each dataset, along with the average inference over the 27 realizations of the CV set, represented by black star markers. Additionally, we report the values of χ^2 (see Equation B22) in the figure legends, comparing the results for the entire test set ("all") and for the subset obtained after excluding a fraction of catalogs with large residuals ("sel"). This excluded fraction, also indicated in the legend, corresponds to catalogs with $\chi^2 > 10$ (see the definition of χ^2 in Equation B22 as well as more information related to other evaluation metrics in the Appendix B.4). The same selection procedure was adopted in de Santi et al. (2023b), where predictions were filtered using the same χ^2 threshold. Even though this selection improves performance, the reason why the model performs poorly on these catalogs remains unclear and will require further investigation in future work.

The first test was performed using the model trained on L-Galaxies and evaluated on its corresponding L-Galaxies testing set. Ultimately, the model provides accurate predictions for both the mean (μ) and dispersion (σ) moments across the LH samples (covering the full $\Omega_{\rm m}$ range) and the CV realizations. No catalogs needed to be excluded from the testing set, confirming the good performance of the model.

The second analysis evaluates the model's performance on the GAEA, SC-SAM, and Shark galaxy catalogs. In general, the predictions remain accurate across the entire range of $\Omega_{\rm m}$ values, including for the CV realizations. A small fraction of catalogs (0.6% for GAEA and 0.1% for Shark) were excluded for exceeding the imposed χ^2 threshold. However, these represent a negligible portion of the datasets, and the corresponding χ^2 values for the selected predictions remain close to unity, confirming that the model extrapolates well across the different SAM prescriptions.

Finally, the third analysis evaluates the model on the hydrodynamical simulations. The predictions exhibit small error bars and values mostly centered around zero, although the corresponding χ^2 values are noticeably higher. This effect is particularly evident for SB28 (all $\chi^2 = 2.10$) and Magneticum (all $\chi^2 = 2.93$). Another key difference, when compared to the performance on the SAMs, lies in the fraction of catalogs removed due to the χ^2 selection cut. These fractions remain below or around 1% for `Astrid`, SIMBA, IllustrisTNG, and SWIFT-EAGLE, but increase to 3.8% and 8.3% for SB28 and Magneticum, respectively.

In Figure 2, we report several performance metrics: RMSE (root mean square error, measuring the typical prediction error; Equation B17), R^2 (coefficient of determination, quantifying how much variance is explained by the model; Equation B18), ϵ (mean relative error, expressing the average fractional deviation; Equation B21), χ^2 (reduced chi-squared, assessing consistency between predictions and uncertainties; Equation B22), and the percentage of catalogs excluded based on their χ^2 values (% Out). Full definitions of these evaluation metrics are provided in Appendix B.4. These metrics are shown for the model tested on L-Galaxies, GAEA, SC-SAM, Shark, `Astrid`, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE. For each case, we present results computed over both the entire testing set ("all") and the selected subset of catalogs ("selected").

Across all tests, the model exhibits consistently strong performance: RMSE values remain low (around 0.05), R^2 values are close to 0.8 (approaching unity), and the relative errors (ϵ) are typically near 10%. The fraction of catalogs excluded due to high χ^2 values is generally below 2.5% for most galaxy formation models. A moderate decline in performance is observed for SB28 and Magneticum, which show higher RMSE, ϵ, and χ^2 values, along with a larger fraction of excluded catalogs (exceeding 7.5% for Magneticum). Nevertheless, all metrics remain largely stable after applying the $\chi^2 < 10$ selection, indicating that outlier removal does not significantly alter the overall model performance.

The achieved precision of $\sim 10\%$ is consistent with expectations for field-level inference using galaxy phase-space information alone (e.g., Villanueva-Domingo & Villaescusa-Navarro (2022); de Santi et al. (2023a)). In particular, the model trained and tested on L-Galaxies reaches an accuracy of $\sim 8.6\%$, improving over the $\sim 12\%$ obtained when training and testing on `Astrid` in de Santi et al. (2023a), likely due to the broader diversity of SAM-based catalogs. However, when evaluated on other SAMs and hydrodynamical simulations, the precision degrades to $\sim 12\%$, reflecting the challenges of extrapolating across different galaxy formation models. This behavior highlights an intrinsic trade-off between accuracy and robustness, and can be attributed to the limited information content of the input features (positions and line-of-sight velocities), the finite simulation volume, and degeneracies with astrophysical processes. Further improvements are expected from larger volumes, more diverse training datasets, and the inclusion of additional observables that preserve robustness.

While the best extrapolation performance is achieved across the different SAMs rather than when applying the model on the hydrodynamical simulations, the model still extrapolates well regardless of the merger tree similarities and differences among the datasets. The model was trained on L-Galaxies, whose merger trees are based on the DM-only counterpart of `Astrid` (as is the case for GAEA, SAM-SC, and Shark), and yet it not only generalizes well to `Astrid`, but also to SIMBA, IllustrisTNG, Magneticum, and SWIFT-EAGLE.

4.2. *The influence of the number of galaxies per catalog*

One of the main findings in de Santi et al. (2023a) was that the success of our previous model could be attributed to the use of a diverse training dataset, particularly in terms of the number of galaxies per catalog across different $\Omega_{\rm m}$ values. A similar analysis is presented in this work, where we compare the number of galaxies per catalog as a function of $\Omega_{\rm m}$ in Figure 3. In these plots, each point represents an individual catalog from the LH or SB samples (colored according to the corresponding simulation) and from the CV set (shown as black points). Horizontal black dotted lines indicate the average number of galaxies for each galaxy formation model, with the corresponding values also reported in the legend.

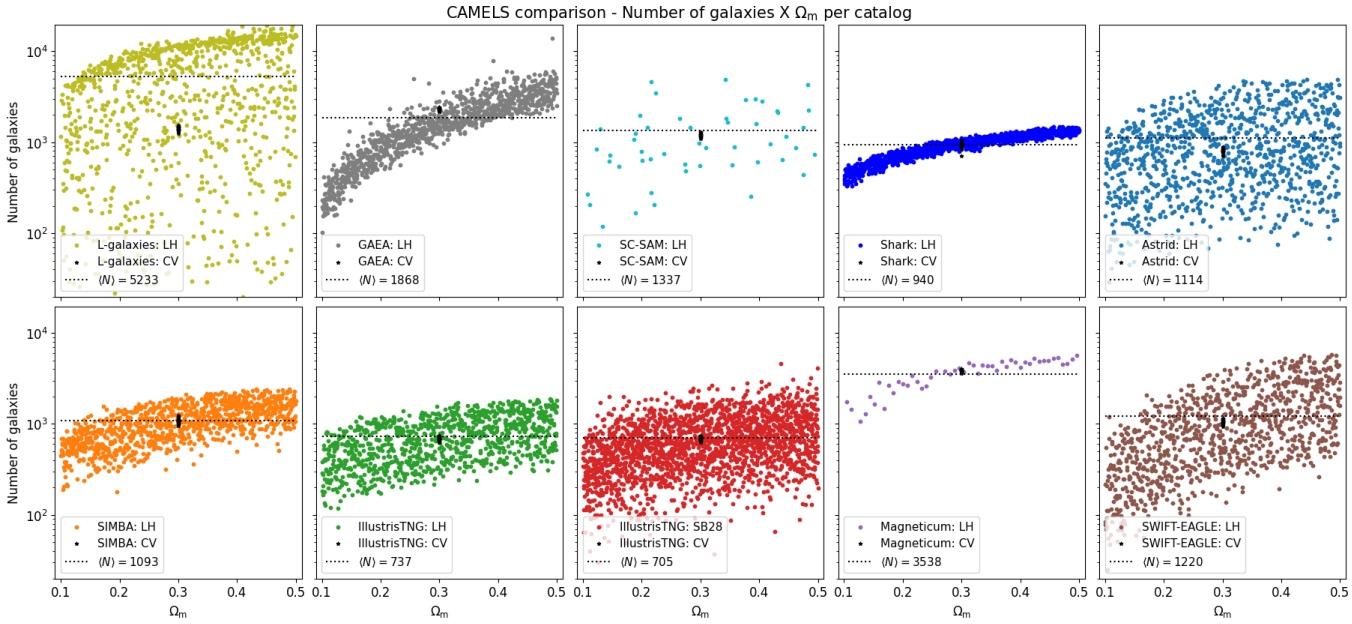

Figure 3. Comparison of the number of galaxies per Ω_m value. We present the number of galaxies per Ω_m value for L-Galaxies, GAEA, SC-SAM, Shark, `Astrid`, IllustrisTNG, SIMBA, SB28, Magneticum, and SWIFT-EAGLE catalogs. The horizontal dotted lines correspond to the mean number of galaxies per different model, while the black points represent the CV catalogs.

As expected due to the success of the model trained on L-Galaxies, their catalogs exhibit a wide variation in the number of galaxies per realization, ranging from just one up to 19,308. This spread is even broader than that found in `Astrid` and encompasses all other galaxy formation models. Moreover, the average number of galaxies in L-Galaxies is the highest among all models ($\langle N \rangle = 5,233$), while the corresponding value for the CV set follows the general trend of about 1,000 galaxies ($\langle N_{\rm CV} \rangle = 1,398$). This extreme variation arises from the parameter choices adopted in the L-Galaxies LH samples used to generate their galaxy catalogs.

The other SAMs (GAEA, Shark, and SC-SAM) do not exhibit as broad a variation in the number of galaxies, but they follow the same trend as the hydrodynamical simulations, with the number of galaxies increasing as Ω_m increases. Their average numbers of galaxies per catalog are $\langle N \rangle = 1,868$, $\langle N \rangle = 1,337$, and $\langle N \rangle = 940$, respectively for GAEA, Shark, and SC-SAM, which are consistent with the average values found in the hydrodynamical simulations. Nevertheless, a significant variation in the number of galaxies is observed for SC-SAM, suggesting that variations in its parameters may lead to a pattern similar to that seen in L-Galaxies. However, additional realizations of the model are required to confirm this claim.

Finally, the superior diversity of L-Galaxies becomes clear when training the model on other SAMs. Models trained on GAEA, for example, are significantly less robust than those trained on L-Galaxies. This can be attributed to the fact that GAEA does not produce a sufficiently wide range of galaxy counts, reducing the diversity seen by the model and limiting its ability to generalize. As a result, the model trained on GAEA struggles to predict $\Omega_{\rm m}$ for other simulations.

This behavior also provides insight into the impact of the training dataset choice. It mirrors the findings of de Santi et al. (2023a), where models trained on IllustrisTNG or SIMBA exhibited weaker generalization compared to those trained on the more diverse `Astrid` dataset. Together, these results suggest that training on datasets with broader galaxy population diversity, such as L-Galaxies, is a key requirement for achieving robust extrapolation across simulations.

Another contributing factor may be differences in the galaxy properties produced by L-Galaxies compared to those found in the other SAMs. For example, Ni et al. (2023) reported that `Astrid` exhibits larger variations in several galaxy properties across its LH parameter space-variations that may also be present in L-Galaxies but not in models such as GAEA. Nevertheless, a direct comparison is still needed to identify which specific galaxy properties contribute

to the success of the model trained on L-Galaxies, as well as how these properties differ across the other SAMs and the hydrodynamical simulations.

5. DISCUSSION AND CONCLUSIONS

Many different approaches have been, and continue to be, developed to construct comprehensive models of halo and galaxy formation and evolution, as well as to connect these models with observable properties (Henriques et al. 2020; De Lucia et al. 2014; Perez et al. 2023; Lagos et al. 2024; Bird et al. 2022; Davé et al. 2019; Pillepich et al. 2018; Hirschmann et al. 2014b; Schaye et al. 2015). From hydrodynamical simulations to semi-analytic models (SAMs), subhalo abundance matching (SHAM), and halo occupation distribution (HOD) frameworks, all of these methods aim to reproduce the Universe across its diverse scales and properties (Wechsler & Tinker 2018; Somerville & Davé 2015). However, it remains unclear which of these approaches best reproduces the relevant observables, and many of them, particularly hydrodynamical simulations, are prohibitively expensive to run at the scales needed for modern cosmological analyses.

At the same time, the pursuit of the most effective tools for generating mock catalogs and predicting cosmological parameters, with the highest possible accuracy and precision for current and upcoming observational surveys (Taylor & Braun 1999; Laureijs et al. 2011; Takada et al. 2014; Amendola et al. 2013; Benitez et al. 2014; Spergel et al. 2015; Abdul Karim et al. 2025; Euclid Collaboration: Castro et al. 2022; Pontoppidan et al. 2022), has never been more pressing. While numerous models exist to describe galaxy formation, machine learning (ML) techniques are emerging as powerful tools that can assist the community in making more data-driven and informed choices, even when integrated with traditional analysis pipelines (de Santi & Abramo 2022).

In this context, graph neural networks (GNNs) emerge as powerful tools for analyzing galaxy catalogs because: (1) they are inherently designed to handle sparse and irregular data structures (Gilmer et al. 2017; Battaglia et al. 2018; Bronstein et al. 2021); (2) it is straightforward to construct GNN architectures that respect physical symmetries (Villanueva-Domingo & Villaescusa-Navarro 2022); and (3) they can extract information from galaxy catalogs without imposing a predefined scale cutoff. Furthermore, GNNs have already been employed to develop robust models across different hydrodynamical simulations (de Santi et al. 2023a; Roncoli et al. 2023) and DM-only simulations (Shao et al. 2023, 2022), and have also proven effective for incorporating systematic effects (de Santi et al. 2023b).

In this work, we trained a GNN coupled with a moment neural network (MNN) on thousands of catalogs generated with the fast semi-analytic model L-Galaxies, in order to infer $\Omega_{\rm m}$ at the field level. More importantly, we assessed the robustness of the model by testing it on galaxy catalogs generated from simulations that employ completely different codes from those used for training. Our main findings can be summarized as follows:

- The model trained on L-Galaxies, using only galaxy positions and velocities, is capable of inferring $\Omega_{\rm m}$ with an accuracy and precision of approximately 8.6% when tested on L-Galaxies catalogs with varying cosmological and astrophysical parameters.

- The robustness tests performed across different SAM prescriptions, specifically, GAEA, SC-SAM, and Shark, show that the model's relative error remains below 12%, even without excluding catalogs based on the χ^2 threshold (see Section 4.1). Moreover, more than 99.4% of catalogs are retained after applying the χ^2 selection, indicating stable performance across different semi-analytic implementations.

- The GNN model demonstrates remarkable extrapolation power when applied to hydrodynamical simulations. Its accuracy and precision remain below 12% without excluding any catalogs for `Astrid`, SIMBA, IllustrisTNG, and SWIFT-EAGLE. Only SB28 and Magneticum exhibit higher relative errors (14% and 28%, respectively), reflecting differences in cosmological and astrophysical parameter variations compared to those used for training, as well as possible intrinsic discrepancies between the Magneticum and L-Galaxies models. In all cases, the fraction of catalogs exceeding the χ^2 threshold remains below 8.3%, underscoring the overall robustness of the approach.

Importantly, the GNN model presented here remains robust across different prescriptions for satellite galaxies. Although the model was trained on L-Galaxies which, like GAEA, models satellite galaxies based on subhalo merger trees, it still achieves strong predictions for SC-SAM and Shark, both of which model satellites using simpler NFW halo mass profiles, after the subhalos disappears from being tracked. Moreover, the model performs well on the hydrodynamical simulations, where satellite and central galaxies are identified directly through numerical solutions of

the underlying physics, rather than analytic prescriptions. This demonstrates that the model is not sensitive to the specific theoretical choices underlying satellite modeling, further reinforcing its robustness.

A meaningful point of comparison is between the GNN model presented here and our previous model in de Santi et al. (2023a). The model trained on L-Galaxies and tested on L-Galaxies achieves an accuracy of approximately 8.6%, representing a notable improvement over the model trained and tested on `Astrid` in de Santi et al. (2023a), which reached an accuracy of about 12%. For hydrodynamical simulations such as SIMBA and IllustrisTNG, the accuracies remain broadly comparable between the two works (around 10%). However, the previous model performed better on SB28 and Magneticum (below 13%), highlighting intrinsic differences between SAM- and hydro-based datasets and the challenges posed by their distinct galaxy populations. We note that the model in de Santi et al. (2023a) was designed specifically for hydrodynamical simulations and therefore was not intended to be applied to SAMs. The present work extends this line of investigation by evaluating robustness across both SAMs and hydrodynamical simulations.

Additionally, in Appendix C, we present a slight adaptation of the model, in which the GNN is coupled to a normalizing flows (NF) instead of the usual MNN. Although this adapted model also exhibits signs of robustness, we emphasize that these results are not among the main findings of this paper. The key conclusion from this exercise is that a larger number of simulations will be required to accurately predict full posterior distributions when employing generative approaches such as NFs. This insight will guide future work, where new simulation models will be incorporated to further explore and improve generative cosmological inference frameworks.

The potential application of this framework to real observational data depends on an inherent limitation of the methodology, one that remains closely tied to the question of robustness. A GNN model can reliably extrapolate its predictions only when applied to data that are compatible with the domain on which it was trained. In this work, we take an important step forward by demonstrating that these ML methods can be successfully applied to SAMs, which are far more computationally feasible to generate than hydrodynamical simulations. Moreover, our results show that SAMs exhibit a remarkable level of internal consistency, being largely independent of their specific physical prescriptions. This strongly reinforces their potential as realistic and efficient way to produce catalogs for field-level cosmological inference.

Nevertheless, applying these GNN models to real data will likely require larger simulation suites, whether CAMELS or others, that incorporate key observational effects such as light-cone geometries and survey selection functions. These ingredients are typically implemented through HOD-based mock constructions, which differ from the idealized setups used in this study. Integrating such observational features is therefore a crucial next step toward developing robust, SAM-based ML inference pipelines for real survey data.

In future work, we plan to address these challenges by extending our analysis to additional approximate methods and by incorporating redshift-space distortions and light-cone maps. Our goal is to assess whether the GNN model can remain robust, accurate, and precise enough to infer not only $\Omega_{\rm m}$ but also other cosmological parameters from more realistic galaxy catalogs.

ACKNOWLEDGMENTS

NSMS acknowledges partial support from the Simons Foundation, NSF CDSE grant AST-2408026 and the NASA TCAN grant 80NSSC24K0101. FF acknowledges support from the Next Generation European Union PRIN 2022 "20225E4SY5 - From ProtoClusters to Clusters in one Gyr". PA-A acknowledges support from the Independent Research Fund Denmark (DFF) under grant 4251-00086B. VGP has been supported by the Atracción de Talento Contract no. 2023-5A/TIC-28943 granted by the Comunidad de Madrid in Spain, and the national grants CNS2024-154242, PID2024-159420NB-C43, and PID2021-122603NB-C21.

The run and analysis for SHARK has been carried out in the computing cluster at UAM (TAURUS). L-Galaxies, GAEA, and SC-SAM have been run on the Simons Foundation, Flatiron Institute, Center of Computational Astrophysics computing cluster. The training of the GNNs has been carried out using graphics processing units (GPUs) from Simons Foundation, Flatiron Institute, Center of Computational Astrophysics. The Flatiron Institute is supported by the Simons Foundation. OpenAI's language model – ChatGPT was used for some language editing during the draft stage of the preparation of this manuscript.

Facilities: Flatiron Institute and UAM (TAURUS) computing clusters.

Software: Pytorch (Fey & Lenssen 2019), *Optuna* (Akiba et al. 2019), CosmoGraphNet (Villanueva-Domingo & Villaescusa-Navarro 2022), *Scikit-learn* (Pedregosa et al. 2011), Pyro (Bingham et al. 2019), TARP (Lemos et al. 2023).

APPENDIX

A. SAMS SPECIFICATIONS

In this appendix we present the specifications related to the SAMs utilized to train, validate and test the ML models, which are first presented in Section 2.1. The parameters being varied as well as their ranges and fiducial values are described in Tables 3, 4, 5, and 6, respectively for L-Galaxies, GAEA, SC-SAM, and Shark.

Table 3. Summary of the parameters sampled in L-Galaxies. The equations related to the free parameters are presented in the Supplementary Material of Henriques et al. (2020).

Parameter	Fiducial Value	Range	Reference
Star Formation			
$\alpha_{\rm H_2}$	0.06	$10^{-12} - 10^{7}$	Eq.(S16)
$\alpha_{\rm SF,burst}$	0.5	$10^{-4} - 1.0$	Eq.(S37)
$\beta_{\rm SF,burst}$	0.38	$10^{-4} - 1.0$	Eq.(S37)
AGN feedback & BH growth			
$k_{\rm AGN}$ $[M_\odot\ {\rm yr}^{-1}]$	$2.5 \cdot 10^{-3}$	$10^{-8} - 1$	Eq.(S28)
$f_{\rm BH}$	0.066	$0.01 - 1$	Eq. (S27)
$V_{\rm BH}$ [km s^{-1}]	700	$1 - 3 \cdot 10^{3}$	Eq.(S27)
Stellar Feedback			
$\epsilon_{\rm reheat}$	5.6	$10^{-4} - 10$	Eq.(S22)
$V_{\rm reheat}$ [km s^{-1}]	110	$1 - 10^{3}$	Eq.(S22)
$\beta_{\rm reheat}$	2.9	$10^{-4} - 5$	Eq.(S22)
$\eta_{\rm eject}$	5.5	$10^{-4} - 10$	Eq.(S20)
$V_{\rm eject}$ [km s^{-1}]	220	$1 - 10^{3}$	Eq.(S20)
$\beta_{\rm eject}$	2.0	$10^{-4} - 5$	Eq.(S20)
Reincorporation			
$\gamma_{\rm reinc}$ $[{\rm yr}^{-1}]$	$1.2 \cdot 10^{10}$	$10^{-7} - 10^{15}$	Eq.(S25)
Tidal Disruption			
$\alpha_{\rm friction}$	1.8	$0.01 - 10$	Eq.(S26)

Table 4. Summary of the parameters sampled in GAEA.

Parameter	Fiducial value	Range	Reference
Efficiency of supernovae			
$\epsilon_{\rm reheat}$	0.028	$0 - 1$	Tab.(1) in Hirschmann et al. (2016)
$\epsilon_{\rm eject}$	0.10	$0 - 1$	Tab.(1) in Hirschmann et al. (2016)
AGN feedback			
$k_{\rm radio}$	1.36	$0 - 2 \cdot 10^{-4}$	Eq.(2) in Fontanot et al. (2020)
$\epsilon_{\rm qw}$	4.86	$0.5 - 10^{3}$	Eq.(15) in Fontanot et al. (2020)

Table 5. Summary of the parameters sampled in SC-SAM. More details can be found in §2.3 Perez et al. (2023).

Parameter	Fiducial value	Range	Reference
Supernovae feedback			
$A_{\rm SN1}$	1	$[0.25, 4]$	Eq. (2) in Somerville et al. (2015)
$A_{\rm SN2}$	0	$[-2, 2]$	Eq. (2) in Somerville et al. (2015)
AGN feedback			
$A_{\rm AGN}$	1	$[0.25, 4]$	Eq. (20) in Somerville et al. (2008)

Table 6. Summary of the parameters sampled in SHARK.

Parameter	Fiducial value	Range	Reference
Stellar Feedback			
$v_{\rm hot}$	80 km.s^{-1}	$10^{-3} - 500$ km.s^{-1}	Eq. (25)-(28) in Lagos et al. (2018)
β	3.79	$5 \times 10^{-3} - 5$	Eq. (25)-(28) in Lagos et al. (2018)
$\beta_{\rm min}$	0.104	$0.01 - 1$	Appendix A2 in Lagos et al. (2024)
Star Formation			
$\nu_{\rm SF}$	1.7 Gyr^{-1}	$0.01 - 10$ Gyr^{-1}	Eq. (7) in Lagos et al. (2018)
Reincorporation			
$\tau_{\rm reinc}$	21.53 Gyr	$10^{-3} - 10^{3}$ Gyr	Eq. (30) in Lagos et al. (2018)
$M_{\rm norm}$	1.183×10^{11} M$_\odot$	$10^{9} - 10^{12}$ M$_\odot$	Eq. (30) in Lagos et al. (2018)
γ	-2.339	$-3 - 0$	Eq. (30) in Lagos et al. (2018)
AGN feedback & BH growth			
$f_{\rm smbh}$	0.008	$10^{-5} - 100$	Eq. (37) in Lagos et al. (2018)
$\rm e_{sb}$	20	$0.5 - 50$	Section 4.4.10 in Lagos et al. (2018)
η	4	$1 - 10$	Section 4.4.10 in Lagos et al. (2018)
κ	10.31	$10^{-5} - 100$	Eq. (40) in Lagos et al. (2018)
$\kappa_{\rm radio}$	0.023	$0.1 - 3$	Eq. (32) in Lagos et al. (2024)
$\epsilon_{\rm QSO}$	10	$0.1 - 100$	Eq. (37) in Lagos et al. (2024)
Tidal and Ram-pressure stripping			
$\alpha_{\rm RPS}$	1	$0.1 - 10$	Eq. (52) in Lagos et al. (2024)
minimum_halo_mass_fraction	0.01	$10^{-5} - 0.1$	Section 3.5 in Lagos et al. (2024)
Disc instabilities			
$\epsilon_{\rm disc}$	0	$0 - 50$	Section 4.4.8 in Lagos et al. (2018)

B. ML MODEL DETAILS

In this appendix, we elaborate on the details of the graphs (see Section B.1), of the employed architecture (Sections B.2 and B.3), and present the definitions of the metrics (Section B.4) used to evaluate the ML models presented in this work through all the Sections 3 and 4.

B.1. *Graph details*

As described in Section 3.1, the graphs are constructed following the methodology of de Santi et al. (2023a), where galaxies are represented as nodes carrying information about their peculiar velocities, v_z. This choice is primarily motivated by the analogy between v_z and the line-of-sight velocity measured in observations. The velocity information is incorporated as node attributes according to:

$$v_z \rightarrow \mathrm{sign}(v_z) \cdot \log_{10}\left[1 + \mathrm{abs}(v_z)\right]. \tag{B1}$$

We use the galaxy positions to establish the connections between galaxies, thereby defining the edges of the graphs. In addition, the positions are used to compute the edge features, which are defined as:

$$\mathbf{e}_{ij} = \left[\frac{|\mathbf{d}_{ij}|}{r_{link}}, \alpha_{ij}, \beta_{ij}\right], \tag{B2}$$

where:

$$\mathbf{d}_{ij} = [\mathbf{r}_i - \mathbf{r}_j] \tag{B3}$$

$$\boldsymbol{\delta}_i = \mathbf{r}_i - \mathbf{c} \tag{B4}$$

$$\alpha_{ij} = \frac{\boldsymbol{\delta}_i}{|\boldsymbol{\delta}_i|} \cdot \frac{\boldsymbol{\delta}_j}{|\boldsymbol{\delta}_j|} \tag{B5}$$

$$\beta_{ij} = \frac{\boldsymbol{\delta}_i}{|\boldsymbol{\delta}_i|} \cdot \frac{\mathbf{d}_{ij}}{|\mathbf{d}_{ij}|}, \tag{B6}$$

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.

Every graph is characterized by a label that we aim at inferring ($\Omega_{\rm m}$). We normalize it using

$$\Omega_m \to \frac{\left(\Omega_m - \Omega_m^{\rm min}\right)}{\left(\Omega_m^{\rm max} - \Omega_m^{\rm min}\right)}, \tag{B7}$$

where $\Omega_m^{\rm min}$ and $\Omega_m^{\rm max}$ represent the minimum and the maximum values.

B.2. *Architecture Details*

We have used a message passing scheme cited in Section 3.2, where each message passing layer updates the node and edge features, taking as input the graph and delivering as output its updated version. We do not employ a model to update the global attribute. The global attribute was used just in order to update the node information (see Equation B9). The node and edge features at layer $\ell + 1$ are found from the node and edge features at layer ℓ as:

- **Edge model:**

$$\mathbf{e}_{ij}^{(\ell+1)} = \mathcal{E}^{(\ell+1)}\left(\left[\mathbf{n}_i^{(\ell)}, \mathbf{n}_j^{(\ell)}, \mathbf{e}_{ij}^{(\ell)}\right]\right), \tag{B8}$$

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

- **Node model:**

$$\mathbf{n}_i^{(\ell+1)} = \mathcal{N}^{(\ell+1)}\left(\left[\mathbf{n}_i^{(\ell)}, \bigoplus_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}, \mathbf{g}\right]\right), \tag{B9}$$

  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:

$$\bigoplus_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)} = \left[\max_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}, \sum_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}, \frac{\sum_{j\in\mathfrak{N}_i} \mathbf{e}_{ij}^{(\ell+1)}}{\sum_{j\in\mathfrak{N}_i}}\right]. \tag{B10}$$

We also we made use of residual layers in the intermediate layers (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

$$\mathbf{y} = \mathcal{F}\left(\left[\bigoplus_{i\in\mathcal{G}} \mathbf{n}_i^N, \mathbf{g}\right]\right), \tag{B11}$$

where $\mathcal{F}$ is the last MLP, $\oplus_{i\in\mathfrak{F}}$ the last multi-pooling operation (done exactly according to Equation B10, but operating over all nodes in the graph $\mathcal{G}$), and $\mathbf{y}$ represents the target of the GNN ($\Omega_{\rm m}$).

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).

B.3. *Moment Neural Networks and the loss function*

The model is trained to infer the value of $\Omega_{\rm m}$ 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.

$$\mathbf{y}_i(\mathcal{G}) = [\mu_i(\mathcal{G}), \sigma_i(\mathcal{G})], \tag{B12}$$

where

$$\mu_i(\mathcal{G}) = \int_{\Omega_m^i} d\Omega_m^i \; \Omega_m^i \; p(\Omega_m^i|\mathcal{G}) \tag{B13}$$

$$\sigma_i^2(\mathcal{G}) = \int_{\Omega_m^i} d\Omega_m^i \; (\Omega_m^i - \mu_i)^2 \; p(\Omega_m^i|\mathcal{G})\,. \tag{B14}$$

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

$$p(\Omega_m^i|\mathcal{G}) = \int_{\Omega_m^i} d\Omega_m^1 d\Omega_m^2 \dots d\Omega_m^n \; p(\Omega_m^1, \Omega_m^2, \dots, \Omega_m^n|\mathcal{G}). \tag{B15}$$

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

$$\mathcal{L} = \log\left[\sum_{j\in\text{batch}} \left(\Omega_m^j - \mu_j\right)^2\right] + \log\left\{\sum_{j\in\text{batch}} \left[\left(\Omega_m^j - \mu_j\right)^2 - \sigma_j^2\right]^2\right\}, \tag{B16}$$

where j represents the samples in a given batch.

B.4. *Evaluation metrics*

We quantify the accuracy and precision of the model using different metrics that we describe below. These metrics have been used mainly in Section 4.1 and in the Appendix C. We consider the true value of the parameter in question for graph i as Ω_m^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):**

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\Omega_m^i - \mu_i)^2}. \tag{B17}$$

  Low values of the RMSE indicate the model is precise.

- **Coefficient of determination:**

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(\Omega_m^i - \mu_i\right)^2}{\sum_{i=1}^{N} \left(\Omega_m^i - \bar{\Omega}_m^i\right)^2}, \tag{B18}$$

  where $\bar{\Omega}_m^i = \frac{1}{N}\sum_{i=1}^{N} \Omega_m^i$. Values close to 1 indicate the model is accurate, while negative values and values close to zero do not.

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

$$\text{PCC} = \frac{\text{cov}\left(\Omega_{\rm m}, \mu\right)}{\sigma_{\Omega_{\rm m}}\sigma_\mu}. \tag{B19}$$

  This statistic measures the positive/negative linear relationship between truth values and inferences: good values are close to ± 1, and worse closer to 0.

- **Bias:**

$$b = \frac{1}{N}\sum_{i=1}^{N}\left(\Omega_m^i - \mu_i\right). \tag{B20}$$

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

- **Mean relative error:**

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}\frac{|\Omega_m^i - \mu_i|}{\mu_i}. \tag{B21}$$

Low values of this statistic indicate the model is precise.

- **Reduced chi squared:**

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\Omega_m^i - \mu_i}{\sigma_i}\right)^2. \tag{B22}$$

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. This was the statistic chosen for selecting good and bad predictions (see Section 4).

We make use of these statistics to quantify the accuracy, precision, and bias of the model on each different test sets (for L-Galaxies, GAEA, SC-SAM, Shark, Astrid, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE). We have utilized *Scikit-learn* (Pedregosa et al. 2011) for computing them.

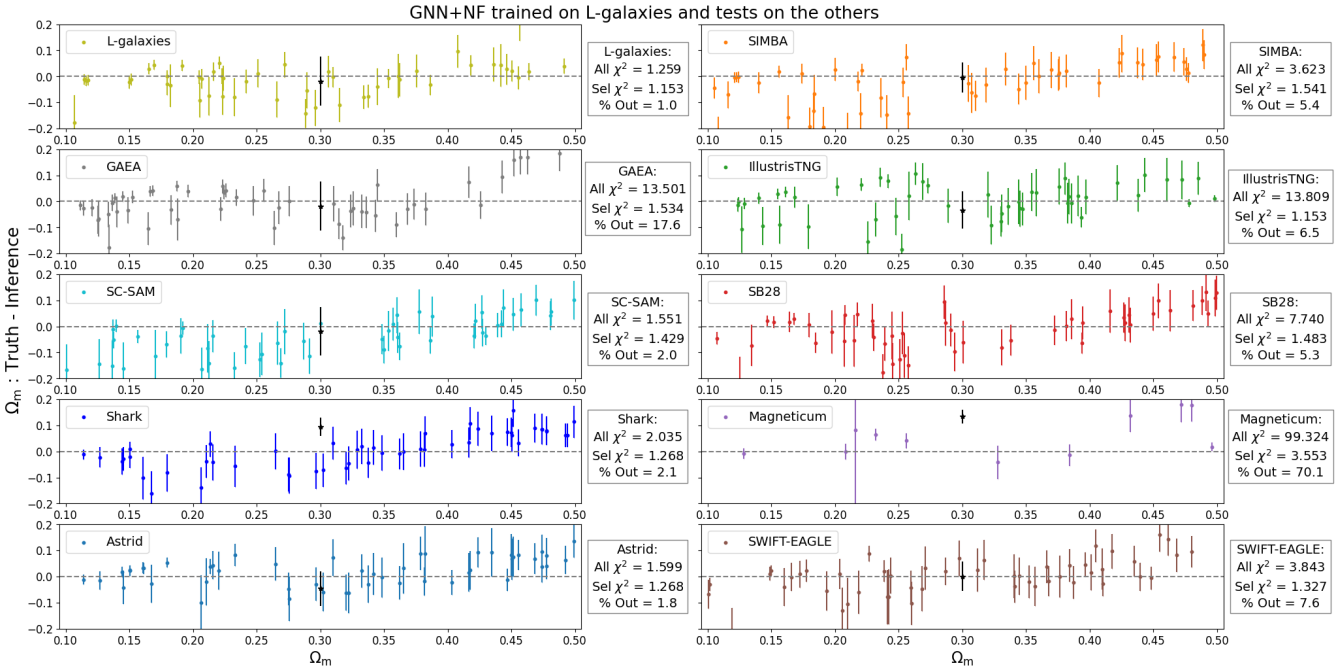

Figure 4. Truth - Inference values for Ω_m: GNN + NF. The model trained on L-Galaxies is evaluated on L-Galaxies, GAEA, SC-SAM, Astrid, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE catalogs. We plot the deviation of our predicted Ω_m from the true values (truth - μ; see Equation B13), with error bars corresponding to σ (see Equation B14) predictions. Colored points represent the testing points (LH or SB sets), while the black points correspond to the average over the CV set inferences. We also present in the side legends the χ^2 score (see Equation B22) for all samples in the test sets and the selected samples, as well as the percentage of samples removed per test set. By contrast, this model does not achieve the same level of performance as the primary GNN-MNN model, showing reduced accuracy and less robust extrapolation across the different simulations.

C. GNN+NF

Throughout this paper, we have presented predictions for $\Omega_{\rm m}$ using a GNN coupled with a MNN. In this framework, the model predicts values that effectively touch the posterior distribution by estimating its first μ and second σ moments (see Equations B13 and B14). In this appendix, we extend the approach by coupling the GNN with a generative model based on NFs, enabling the network to infer the full posterior distribution of $\Omega_{\rm m}$ directly from the data.

The general idea of this method is to, starting with a random variable $\mathbf{x}$ with simple distribution $p(\mathbf{x})$, perform a transformation $f(\mathbf{x})$ or a series of transformations $f_\phi(\mathbf{x}) = f_1 \circ \cdots \circ f_k(\mathbf{x})$ to learn a complex desired distribution for variable $\mathbf{z}$. Then, the flow represents this bijective map, which needs to be invertible and differentiable, and the two distributions are connected via

$$p(\mathbf{z}) = p(\mathbf{x}) \left| \det \left(\frac{\partial f^{-1}(\mathbf{z})}{\partial \mathbf{x}} \right) \right|. \tag{C23}$$

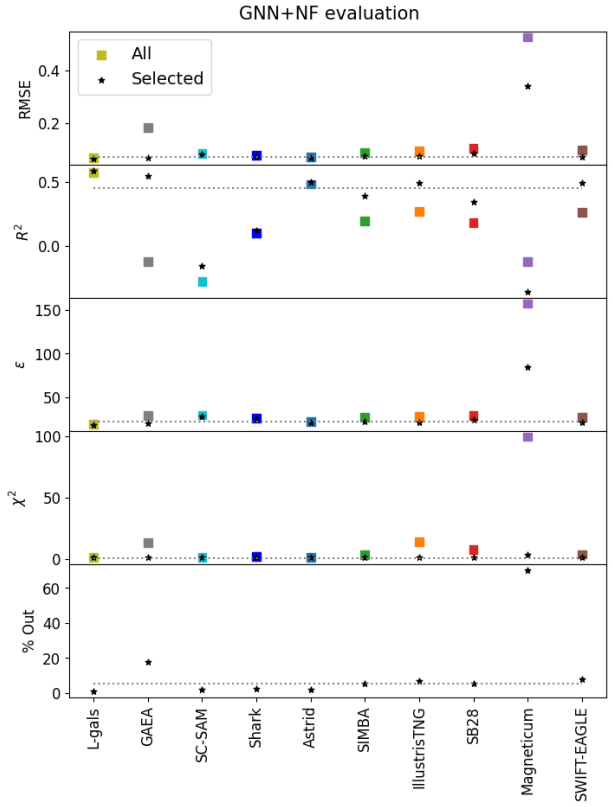

Figure 5. GNN+NF: Model Evaluation. We present RMSE (root mean square error), R^2 (coefficient of determination), ϵ (mean relative error), χ^2 (reduced chi squared), and % Out (percentage of samples removed) for the test sets of L-Galaxies, GAEA, SC-SAM, IllustrisTNG, SIMBA, SB28, Magneticum, and SWIFT-EAGLE indicated in the x-axis. Results are presented to all and selected samples according to χ^2 values.

In this work we have been using $p(\mathbf{x})$ as a Gaussian distribution and conditioning the latent space from the GNNs to an Autoregressive Spline (Durkan et al. 2019; Dolatabadi et al. 2020; Kingma et al. 2016) to infer Ω_m distributions. The implementation was performed using Pyro library (Bingham et al. 2019).

The graph construction, GNN architecture, and training procedure followed the same scheme described in the main part of the paper (see Section 3). The primary difference lies in the number of catalogs used for training, validation, and testing, which are 600, 200, and 200, respectively. In addition, the optimal set of hyperparameters was determined using the *Optuna* optimization package (Akiba et al. 2019), with the selected values listed in Table 7. The hyperparameters associated with the NF are bound and count bins, which respectively represent the number of spline segments and the size of the bounding box used in the transformation.

To facilitate a direct comparison between this model and the main one presented in the paper, we computed the mean and standard deviation of each posterior prediction for $\Omega_{\rm m}$. The results for the GNN coupled with the NF are shown in Figures 4 and 5, alongside the corresponding results for the baseline model in Figures 1 and 2.

The takeaway from Figure 4 is that, overall, the predictions exhibit a slight bias and larger uncertainties, while compared to the main model. These biases are particularly noticeable for the CV sets of Shark, Astrid, and Magneticum. It is also important to note that the χ^2 values increased across several models, reaching 13.5 for GAEA, approximately 2 for Shark, and even higher values for the hydrodynamical simulations (13.8 for IllustrisTNG and around 99 for Magneticum).

This increase in χ^2 propagated into the fraction of predictions excluded under the $\chi^2 > 10$ criterion. Among the SAMs, 17.6% of the GAEA catalogs were removed, while for the Magneticum hydrodynamical simulation this fraction increased to 70.1%. Nevertheless, after excluding these outlier catalogs, the remaining χ^2 values stabilized around unity for all tests, indicating good performance on the retained samples.

Table 7. Hyperparameters values selected for the best model.

Hyperparameter	Best value
Bound	3
Count Bins	9
Hidden Features	94
Number of GNN and NF layers	5
Number of NF transformations	4
r_{link}	0.25Mpc/h
Learning Rate	$8.76 \cdot 10^{-4}$
Weight Decay	$7.4 \cdot 10^{-4}$

The GNN+NF model evaluation summarized in Figure 5 also reveals that it performs slightly worse than the main GNN+MNN model. On average, the RMSE increased to about 0.07, the R^2 values stabilized near 0.45, the χ^2 remained close to 1, and the fraction of outliers increased to roughly 10%.

The final evaluation of the generative model is performed through a coverage test, a metric used to assess the accuracy of generative posterior estimators, $\hat{p}(\Omega_{\rm m}|G)$. Here, we employ the Tests of Accuracy with Random Points (TARP) (Lemos et al. 2023).

The basic idea of this test is to compute the expected coverage probability (ECP), denoted as ECP$(\hat{p}, \alpha, \mathcal{R})$, averaged over the data distribution G, where $\hat{p}$ is the posterior estimator, $\mathcal{R}$ is a credible region generator, and α represents the credibility level. The ECP is estimated using samples drawn from both the reference distribution $p(\Omega_{\rm m}|G)$, and the inferred posterior $\hat{p}(\Omega_{\rm m}|G)$.

A high-quality posterior estimator should yield a one-to-one correspondence between coverage and credibility level across the full range of $\Omega_{\rm m}$. Deviations from this relation indicate that the posterior estimator is either overconfident (undercovering) or underconfident (overcovering).

The results of the TARP test are shown in Figure 6, where the solid lines represent the mean values and the shaded regions correspond to the standard deviation across 100 realizations of the test. We present the results for the testing sets of L-Galaxies, GAEA, SC-SAM, Shark, Astrid, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE.

The model's quality can be most directly evaluated from the L-Galaxies test, which already shows a slight deviation from the one-to-one correspondence line, indicating a transition from mild overconfidence to underconfidence. However, this result should be interpreted with caution, as the test set includes only 200 samples of $\Omega_{\rm m}$, which may be insufficient for a reliable coverage assessment. Because the total number of available L-Galaxies realizations is $1,000$, the maximum number of independent test samples that could be used was necessarily limited to 200. A similar limitation may affect the results for SC-SAM and Magneticum, which include only 73 and 77 catalogs, respectively.

Apart from the Shark test, where the model appears slightly overconfident at credibility levels above 30%, the TARP results for GAEA, Astrid, SIMBA, IllustrisTNG, SB28, and SWIFT-EAGLE show an almost perfect correspondence between coverage and credibility, demonstrating the robustness of the generative model across these datasets.

REFERENCES

Abdul Karim, M., Aguilar, J., Ahlen, S., et al. 2025, Phys. Rev. D, 112, 083515, doi: 10.1103/tr6y-kpc6

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

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

Avila, S., Knebe, A., Pearce, F. R., et al. 2014, MNRAS, 441, 3488, doi: 10.1093/mnras/stu799

Avila, S., Gonzalez-Perez, V., Mohammad, F. G., et al. 2020, MNRAS, 499, 5486, doi: 10.1093/mnras/staa2951

Baldry, I. K., Driver, S. P., Loveday, J., et al. 2012, MNRAS, 421, 621, doi: 10.1111/j.1365-2966.2012.20340.x

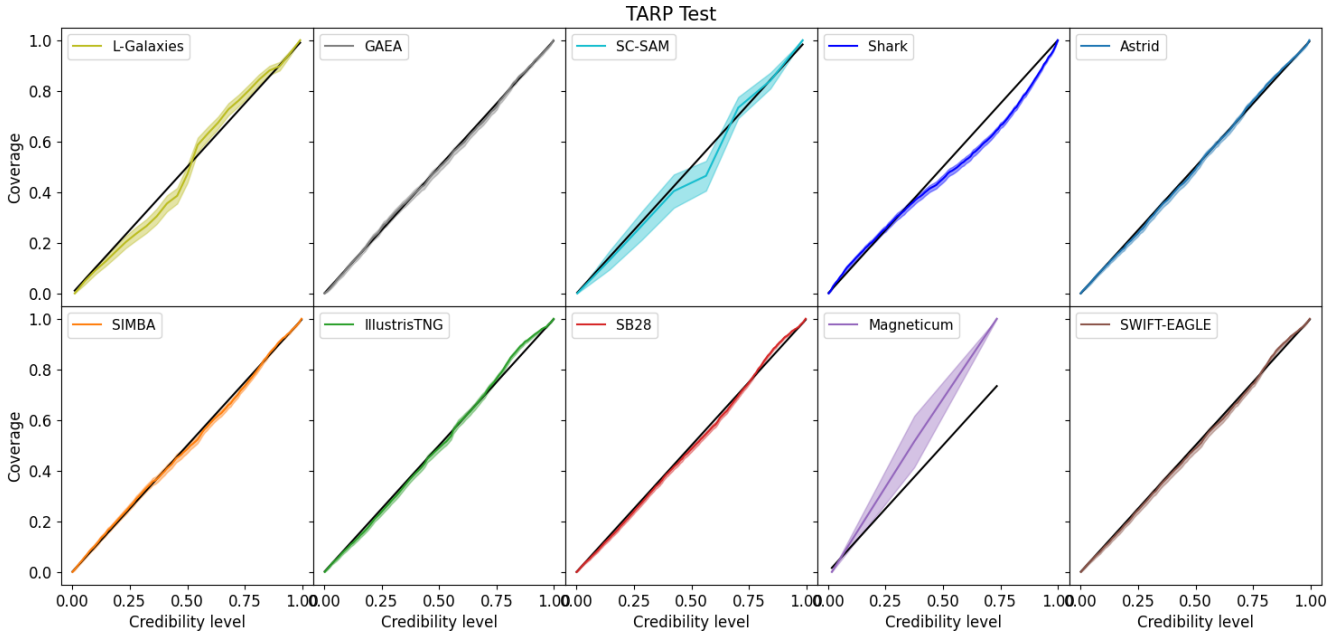

Figure 6. GNN+NF: TARP test. The TARP test was performed for the testing sets of L-Galaxies, GAEA, SC-SAM, Shark, Astrid, SIMBA, IllustrisTNG, SB28, Magneticum, and SWIFT-EAGLE catalogs. Here we present the coverage versus the credibility level for each dataset.

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

Baugh, C. M. 2006, Reports on Progress in Physics, 69, 3101, doi: 10.1088/0034-4885/69/12/R02

Bayer, A. E., Villaescusa-Navarro, F., Sharief, S., et al. 2025, ApJ, 989, 207, doi: 10.3847/1538-4357/adef4e

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

Behroozi, P. S., Wechsler, R. H., Wu, H.-Y., et al. 2013b, ApJ, 763, 18, doi: 10.1088/0004-637X/763/1/18

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

Benson, A. J. 2010, PhR, 495, 33, doi: 10.1016/j.physrep.2010.06.001

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

Bingham, E., Chen, J. P., Jankowiak, M., et al. 2019, J. Mach. Learn. Res., 20, 28:1. http://jmlr.org/papers/v20/18-403.html

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

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

Chandro-Gómez, Á., Lagos, C. d. P., Power, C., et al. 2025, MNRAS, 539, 776, doi: 10.1093/mnras/staf519

Chatterjee, A., & Villaescusa-Navarro, F. 2024, arXiv e-prints, arXiv:2405.13119, doi: 10.48550/arXiv.2405.13119

Crain, R. A., & van de Voort, F. 2023, Annual Review of Astronomy and Astrophysics, 61, 473, doi: https://doi.org/10.1146/annurev-astro-041923-043618

Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, MNRAS, 450, 1937, doi: 10.1093/mnras/stv725

Csurka, G. 2017, arXiv e-prints, arXiv:1702.05374, doi: 10.48550/arXiv.1702.05374

Cuesta-Lazaro, C., & Mishra-Sharma, S. 2024, PhRvD, 109, 123531, doi: 10.1103/PhysRevD.109.123531

Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827, doi: 10.1093/mnras/stz937

De Lucia, G. 2019, Galaxies, 7, 56, doi: 10.3390/galaxies7020056

De Lucia, G., Fontanot, F., Xie, L., & Hirschmann, M. 2024, A&A, 687, A68, doi: 10.1051/0004-6361/202349045

De Lucia, G., Tornatore, L., Frenk, C. S., et al. 2014, MNRAS, 445, 970, doi: 10.1093/mnras/stu1752
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., Shao, H., Villaescusa-Navarro, F., et al. 2023a, ApJ, 952, 69, doi: 10.3847/1538-4357/acd1e2
de Santi, N. S. M., Villaescusa-Navarro, F., Abramo, L. R., et al. 2023b, arXiv e-prints, arXiv:2310.15234, doi: 10.48550/arXiv.2310.15234
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
Dolatabadi, H. M., Erfani, S., & Leckie, C. 2020, arXiv e-prints, arXiv:2001.05168, doi: 10.48550/arXiv.2001.05168
Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. 2019, arXiv e-prints, arXiv:1906.04032, doi: 10.48550/arXiv.1906.04032
Elahi, P. J., Cañas, R., Poulton, R. J. J., et al. 2019a, PASA, 36, e021, doi: 10.1017/pasa.2019.12
Elahi, P. J., Poulton, R. J. J., Tobar, R. J., et al. 2019b, PASA, 36, e028, doi: 10.1017/pasa.2019.18
Elahi, P. J., Welker, C., Power, C., et al. 2018, MNRAS, 475, 5338, doi: 10.1093/mnras/sty061
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
Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. 2020, arXiv e-prints, arXiv:2010.03978, doi: 10.48550/arXiv.2010.03978
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
Fontanot, F., De Lucia, G., Hirschmann, M., et al. 2017, MNRAS, 464, 3812, doi: 10.1093/mnras/stw2612
Fontanot, F., De Lucia, G., Xie, L., et al. 2025, A&A, 699, A108, doi: 10.1051/0004-6361/202452029
Fontanot, F., De Lucia, G., Hirschmann, M., et al. 2020, MNRAS, 496, 3943, doi: 10.1093/mnras/staa1716
Gabrielpillai, A., Somerville, R. S., Genel, S., et al. 2022, MNRAS, 517, 6091, doi: 10.1093/mnras/stac2297
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., Hough, T., Jiménez Muñoz, A., et al. 2025, A&A, 697, A171, doi: 10.1051/0004-6361/202554064
Gómez, J. S., Padilla, N. D., Helly, J. C., et al. 2022, MNRAS, 510, 5500, doi: 10.1093/mnras/stab3661
Gonzalez-Perez, V., Cui, W., Contreras, S., et al. 2020, MNRAS, 498, 1852, doi: 10.1093/mnras/staa2504
Hassani, H., & Javanmard, A. 2022, arXiv e-prints, arXiv:2201.05149. https://arxiv.org/abs/2201.05149
Henriques, B. M. B., Yates, R. M., Fu, J., et al. 2020, MNRAS, 491, 5795, doi: 10.1093/mnras/stz3233
Hirschmann, M., De Lucia, G., & Fontanot, F. 2016, MNRAS, 461, 1760, doi: 10.1093/mnras/stw1318
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
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55, doi: 10.1051/0004-6361/201321100
Jeffrey, N., & Wandelt, B. D. 2020, arXiv e-prints, arXiv:2011.05991. https://arxiv.org/abs/2011.05991
Jiang, L., Helly, J. C., Cole, S., & Frenk, C. S. 2014, MNRAS, 440, 2115, doi: 10.1093/mnras/stu390
Jo, Y., Genel, S., Sengupta, A., et al. 2025, ApJ, 991, 120, doi: 10.3847/1538-4357/adec78
Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980. https://arxiv.org/abs/1412.6980
Kingma, D. P., Salimans, T., Jozefowicz, R., et al. 2016, arXiv e-prints, arXiv:1606.04934, doi: 10.48550/arXiv.1606.04934
Knebe, A., Pearce, F. R., Thomas, P. A., et al. 2015, MNRAS, 451, 4029, doi: 10.1093/mnras/stv1149
Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, ApJ, 609, 35, doi: 10.1086/420959
Kumari, S., & Singh, P. 2023, arXiv e-prints, arXiv:2308.01265, doi: 10.48550/arXiv.2308.01265
Lagos, C. d. P., Tobar, R. J., Robotham, A. S. G., et al. 2018, Monthly Notices of the Royal Astronomical Society, 481, 3573, doi: 10.1093/mnras/sty2440
Lagos, C. d. P., Bravo, M., Tobar, R., et al. 2024, Monthly Notices of the Royal Astronomical Society, 531, 3551, doi: 10.1093/mnras/stae1024

Lagos, C. d. P., Bravo, M., Tobar, R., et al. 2024, MNRAS, 531, 3551, doi: 10.1093/mnras/stae1024

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

Lee, J., Yi, S. K., Elahi, P. J., et al. 2014, MNRAS, 445, 4197, doi: 10.1093/mnras/stu2039

Lemos, P., Coogan, A., Hezaveh, Y., & Perreault-Levasseur, L. 2023, arXiv e-prints, arXiv:2302.03026, doi: 10.48550/arXiv.2302.03026

Li, C., & White, S. D. M. 2009, MNRAS, 398, 2177, doi: 10.1111/j.1365-2966.2009.15268.x

Lovell, C., & et. al. 2025, In preparation.

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

Moustakas, J., Coil, A. L., Aird, J., et al. 2013, ApJ, 767, 50, doi: 10.1088/0004-637X/767/1/50

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493, doi: 10.1086/304888

Nelson, D., Pillepich, A., Genel, S., et al. 2015, Astronomy and Computing, 13, 12, doi: 10.1016/j.ascom.2015.09.003

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, ApJ, 959, 136, doi: 10.3847/1538-4357/ad022a

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Perez, L. A., Genel, S., Villaescusa-Navarro, F., et al. 2023, ApJ, 954, 11, doi: 10.3847/1538-4357/accd52

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

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

Rodriguez-Gomez, V., Genel, S., Vogelsberger, M., et al. 2015, MNRAS, 449, 49, doi: 10.1093/mnras/stv264

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

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

Schaller, M., Borrow, J., Draper, P. W., et al. 2024, Monthly Notices of the Royal Astronomical Society, 530, 2378, doi: 10.1093/mnras/stae922

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. 2022, arXiv e-prints, arXiv:2209.06843. https://arxiv.org/abs/2209.06843

Shao, H., de Santi, N. S. M., Villaescusa-Navarro, F., et al. 2023, ApJ, 956, 149, doi: 10.3847/1538-4357/acee6f

Sobol', I. 1967, USSR Computational Mathematics and Mathematical Physics, 7, 86, doi: https://doi.org/10.1016/0041-5553(67)90144-9

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51, doi: 10.1146/annurev-astro-082812-140951

Somerville, R. S., Hopkins, P. F., Cox, T. J., Robertson, B. E., & Hernquist, L. 2008, Monthly Notices of the Royal Astronomical Society, 391, 481, doi: 10.1111/j.1365-2966.2008.13805.x

Somerville, R. S., Popping, G., & Trager, S. C. 2015, MNRAS, 453, 4337, doi: 10.1093/mnras/stv1877

Somerville, R. S., & Primack, J. R. 1999, MNRAS, 310, 1087, doi: 10.1046/j.1365-8711.1999.03032.x

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. 2005a, 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., White, S. D. M., Jenkins, A., et al. 2005b, Nature, 435, 629, doi: 10.1038/nature03597

Srisawat, C., Knebe, A., Pearce, F. R., et al. 2013, MNRAS, 436, 150, doi: 10.1093/mnras/stt1545

Steinborn, L. K., Dolag, K., Comerford, J. M., et al. 2016, MNRAS, 458, 1013, doi: 10.1093/mnras/stw316

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

Tornatore, L., Borgani, S., Dolag, K., & Matteucci, F. 2007, MNRAS, 382, 1050, doi: 10.1111/j.1365-2966.2007.12070.x

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

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

Wang, M., & Deng, W. 2018, arXiv e-prints, arXiv:1802.03601, doi: 10.48550/arXiv.1802.03601

Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435, doi: 10.1146/annurev-astro-081817-051756

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

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

Xie, L., De Lucia, G., Hirschmann, M., & Fontanot, F. 2020, MNRAS, 498, 4327, doi: 10.1093/mnras/staa2370

Xie, L., De Lucia, G., Hirschmann, M., Fontanot, F., & Zoldan, A. 2017, MNRAS, 469, 968, doi: 10.1093/mnras/stx889

Xie, L., De Lucia, G., Fontanot, F., et al. 2024, ApJL, 966, L2, doi: 10.3847/2041-8213/ad380a

Yu, J., Zhao, C., Gonzalez-Perez, V., et al. 2024, MNRAS, 527, 6950, doi: 10.1093/mnras/stad3559

Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791, doi: 10.1086/466510

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