The Astrophysical Journal, 954:125 (14pp), 2023 September 10 https://doi.org/10.3847/1538-4357/ace96e

Cosmology with One Galaxy? The ASTRID Model and Robustness

Nicolas Echeverri-Rojas[1], Francisco Villaescusa-Navarro[2,3], Chaitanya Chawak[4], Yueying Ni[5,6], ChangHoon Hahn[3], Elena Hernández-Martínez[7], Romain Teyssier[8], Daniel Anglés-Alcázar[2,9], Klaus Dolag[7,10], and Tiago Castro[11,12,13]
[1] Instituto de Fisica, Universidad de Antioquia, A.A.1226, Medellin, Colombia; nicolas.echeverrir@udea.edu.co
[2] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[3] Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA
[4] Indian Institute of Science Education and Research (IISER) Tirupati, Tirupati-517507, India
[5] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[6] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[7] Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstr. 1, D-81679 München, Germany
[8] Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA
[9] Department of Physics, University of Connecticut, 196 Auditorium Road, U-3046, Storrs, CT 06269, USA
[10] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, D-85741 Garching, Germany
[11] INAF-Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, I-34143 Trieste, Italy
[12] INFN, Sezione di Trieste, Via Valerio 2, I-34127 Trieste TS, Italy
[13] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, I-34151 Trieste, Italy
Received 2023 April 17; revised 2023 July 6; accepted 2023 July 19; published 2023 August 29

Abstract

Recent work has pointed out the potential existence of a tight relation between the cosmological parameter Ω_m, at fixed Ω_b, and the properties of individual galaxies in state-of-the-art cosmological hydrodynamic simulations. In this paper, we investigate whether such a relation also holds for galaxies from simulations run with a different code that makes use of a distinct subgrid physics: Astrid. We also find that in this case, neural networks are able to infer the value of Ω_m with a $\sim$10% precision from the properties of individual galaxies, while accounting for astrophysics uncertainties, as modeled in Cosmology and Astrophysics with MachinE Learning (CAMELS). This tight relationship is present at all considered redshifts, $z \leqslant 3$, and the stellar mass, the stellar metallicity, and the maximum circular velocity are among the most important galaxy properties behind the relation. In order to use this method with real galaxies, one needs to quantify its robustness: the accuracy of the model when tested on galaxies generated by codes different from the one used for training. We quantify the robustness of the models by testing them on galaxies from four different codes: IllustrisTNG, SIMBA, Astrid, and Magneticum. We show that the models perform well on a large fraction of the galaxies, but fail dramatically on a small fraction of them. Removing these outliers significantly improves the accuracy of the models across simulation codes.

Unified Astronomy Thesaurus concepts: Cosmology (343)

1. Introduction

Inferring the values of cosmological parameters is one of the most essential tasks in cosmology. Galaxy clustering is commonly used to carry out this task, although other methods (e.g., the cosmic distance ladder) can also be used to estimate the values of some parameters.

Recently, Villaescusa-Navarro et al. (2022a) claimed that the properties of individual galaxies could be used to infer the value of Ω_m. The authors showed that by training neural networks on galaxy properties from individual galaxies to perform likelihood-free inference on the values of cosmological parameters, they were able to constrain Ω_m, at fixed Ω_b, with $\sim$10% precision. The authors presented a potential explanation, stating that galaxy properties exist in a low-dimensional manifold that is affected differently by Ω_m than by astrophysical processes such as supernova and active galactic nucleus (AGN) feedback. In that work, the authors used thousands of state-of-the-art hydrodynamic simulations from the Cosmology and Astrophysics with MachinE Learning (CAMELS) project (Villaescusa-Navarro et al. 2021, 2023). These included simulations performed with the AREPO (Springel 2010) and GIZMO (Hopkins 2015) hydrodynamic codes, implementing the subgrid galaxy formation models of IllustrisTNG (Nelson et al. 2015) and SIMBA (Davé et al. 2019).

In this work, we have made use of a new suite of simulations, CAMELS-Astrid (Ni et al. 2023), run with the MP-Gadget code using the Astrid model (Bird et al. 2022; Ni et al. 2022), which solve the hydrodynamic equations and implement feedback with a yet different method than the CAMELS-IllustrisTNG and CAMELS-SIMBA simulations discussed above. As with CAMELS-IllustrisTNG and CAMELS-SIMBA, the CAMELS-Astrid simulations have different values of cosmological and astrophysical parameters.

We show that neural networks can also infer the value of Ω_m, at fixed Ω_b, from properties of individual galaxies of the CAMELS-Astrid simulations with a $\sim$10% precision. We also investigate what are the most important galaxy properties used by the model to make the inference and show that our results hold at different redshifts. We then focus our attention on the robustness of the different models (i.e., models trained on galaxies from different simulations). We show that the models perform well on most galaxies and that removing outliers helps the model to make unbiased predictions.

This paper is organized as follows. In Section 2, we describe the data and machine-learning methods we use. Next, we present the results of our analysis, in terms of the precision and

accuracy of the models, in Section 3. We then conclude in Section 4.

2. Methods

In this section, we describe the data we use and the machine-learning methods we utilize. We also outline the metrics we consider to quantify the accuracy and precision of the models.

2.1. Data

We train our model using galaxy properties from individual galaxies of the CAMELS hydrodynamic simulations (Villaescusa-Navarro et al. 2021, 2023). These simulations can be classified into four different suites:

1. *IllustrisTNG*. Simulations run with the AREPO code (Springel 2010) using the IllustrisTNG subgrid physics (Weinberger et al. 2017; Pillepich et al. 2018).
2. *SIMBA*. Simulations run with the GIZMO code (Hopkins 2015) using the SIMBA subgrid physics (Davé et al. 2019).
3. *Astrid*. Simulations run with the MP-Gadget code (Feng et al. 2018) using the Astrid subgrid physics (Bird et al. 2022; Ni et al. 2022).
4. *Magneticum*. Simulations run with the Open-Gadget code using a similar but improved subgrid physics model following Fabjan et al. (2011), Hirschmann et al. (2014), Teklu et al. (2015), and Steinborn et al. (2016).

Every suite contains 1000 simulations (except Magneticum, which contains 50 simulations), each of them with different values of $\Omega_{\rm m}$, σ_8, $A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, and $A_{\rm AGN2}$, which are varied in a Latin hypercube with the boundaries:[14]

$$0.1 \leqslant \Omega_{\rm m} \leqslant 0.5, \quad (1)$$

$$0.6 \leqslant \sigma_8 \leqslant 1.0, \quad (2)$$

$$0.25 \leqslant A_{\rm SN1}, A_{\rm AGN1} \leqslant 4.0, \quad (3)$$

$$0.5 \leqslant A_{\rm SN2}, A_{\rm AGN2} \leqslant 2.0\,. \quad (4)$$

The $A_{\rm SN}$ and $A_{\rm AGN}$ parameters control the efficiency of the supernova and AGN feedback, and their specific definition depends on the considered subgrid model. We refer the reader to Villaescusa-Navarro et al. (2021) and Ni et al. (2023) for further details on this. All simulations follow the evolution of 256^3 dark matter particles plus 256^3 initial fluid elements in a periodic box of $(25\ h^{-1}\ {\rm Mpc})^3$ from $z = 127$ down to $z = 0$. We note that in all these simulations, the value of $\Omega_{\rm b}$ is fixed at 0.049.

Halos and subhalos are identified using SUBFIND (Springel et al. 2001; Dolag et al. 2009) from all simulation snapshots. In this work, we consider galaxies[15] with stellar masses $M_* \geqslant 5 \times 10^8\, h^{-1} M_\odot$. Villaescusa-Navarro et al. (2022a) considered galaxies with smaller stellar masses (e.g., galaxies with stellar masses $M_* \geqslant 2 \times 10^8\, h^{-1} M_\odot$). We note that the mass of a baryonic fluid element in the initial conditions is $\sim 1.3 \times 10^7\, h^{-1} M_\odot$, so in codes where the mass of a star is similar to its progenitor fluid element, the stellar mass threshold would correspond to roughly 40 star particles. However, in the case of Astrid, the masses of the star particles can be significantly smaller, given the star formation and feedback model used. We have checked that our conclusions do not change if we consider galaxies with smaller stellar masses. For each galaxy, we consider 14 properties, computed by SUBFIND:

1. $M_{\rm g}$: the gas mass of the subhalo hosting the galaxy, including the contribution from the circumgalactic medium.
2. $M_{\rm BH}$: the total mass of the black holes in the galaxy.
3. M_*: the stellar mass of the galaxy.
4. M_t: the total mass of the subhalo hosting the galaxy.
5. $V_{\rm max}$: the maximum circular velocity of the subhalo hosting the galaxy: $V_{\rm max} = \max(\sqrt{GM(<R)/R})$.
6. σ_v: the mass-weighted velocity dispersion of all the particles contained in the galaxy's subhalo.
7. Z_g: the mass-weighted gas metallicity of the galaxy.
8. Z_*: the mass-weighted stellar metallicity of the galaxy.
9. *SFR*: the galaxy star formation rate.
10. J: the modulus of the galaxy's subhalo spin vector.
11. V: the modulus of the galaxy's subhalo peculiar velocity.
12. R_*: the radius containing half of the galaxy's stellar mass.
13. R_t: the radius containing half of the total mass of the galaxy's subhalo.
14. $R_{\rm max}$: the radius at which $\sqrt{GM(<R)/R} = V_{\rm max}$.

We train independently three different models: (1) using IllustrisTNG galaxies; (2) using SIMBA galaxies; and (3) using Astrid galaxies. We then test the models on galaxies from all four suites. We note that we do not train a model on Magneticum galaxies, as this suite only contains 50 simulations; not enough to train the models.

We follow Villaescusa-Navarro et al. (2022a) and first split the simulations into training (900 simulations), validation (50 simulations), and testing (50 simulations) sets. We then take individual galaxies from the training and validation sets and pass them to the neural networks during training. We do this because we want to avoid a situation where galaxies from the same simulation are used during training, validation, and testing. The reason behind this is that those galaxies may be correlated, since they belong to the same simulation, and there may be some information leakage underneath. Using our procedure, we ensure that the galaxies in the test set come from simulations whose galaxies have never been seen during training.

Unless stated explicitly, we only train our models on galaxies from a single simulation suite. However, we also investigate the behavior of our models when trained on galaxies from two different simulation suites (e.g., IllustrisTNG and Astrid).

2.2. Neural Networks

We train neural networks to infer the values of the cosmological and astrophysical parameters from the above 14 properties of individual galaxies. Our models consist of a series of fully connected layers with dropout and LeakyReLU activation functions. The number of layers, the number of neurons per layer, the value of the dropout, the weight decay, and the learning rate are hyperparameters that we optimize using Optuna (Akiba et al. 2019).

Our models take as input a vector of 14 dimensions (representing the individual galaxy properties) and return $2N$ numbers, where N is the number of parameters considered; $N = 6$ when inferring all parameters and $N = 1$ when only inferring $\Omega_{\rm m}$. For each parameter, the models predict the mean (μ_i) and standard deviation (σ_i) of the marginal posterior for each parameter. To achieve this, we minimize the loss function of Jeffrey & Wandelt (2020) with the modifications described in Villaescusa-Navarro

[14] In the case of the Astrid simulations, the parameter $A_{\rm AGN2}$ varies from 0.25 to 4.

[15] We define a galaxy as a subhalo with at least one star particle.

et al. (2022b). We emphasize that by construction, the output of our models represents the posterior mean and standard deviation, without making any assumption about the shape of the posterior.

We perform more than 100 Optuna trials[16] minimizing the value of the validation loss.

2.3. Accuracy and Precision Metrics

From the properties of a single galaxy, our models predict two numbers for the considered parameter i: the marginal posterior mean (μ_i) and the standard deviation (σ_i). We denote by θ_i the true value of the parameter i. In order to quantify the accuracy and precision of a given model, we make use of four different statistics:

1. The rms error (RMSE), defined as

$$\mathrm{RMSE}_i = \sqrt{\langle (\theta_i - \mu_i)^2 \rangle} \ . \tag{5}$$

   The smaller the RMSE value, the more accurate the model is.
2. The mean relative error (ϵ), defined as

$$\epsilon_i = \left\langle \frac{\sigma_i}{\mu_i} \right\rangle . \tag{6}$$

   The smaller the mean relative error, the more precise the model is.
3. The coefficient of determination (R^2), defined as

$$R_i^2 = 1 - \frac{\sum_i (\theta_i - \mu_i)^2}{\sum_i (\theta_i - \bar{\theta}_i)^2}. \tag{7}$$

   The closer the value to 1, the more accurate the model is.
4. The reduced chi-squared (χ^2), defined as

$$\chi_i^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\theta_i - \mu_i}{\sigma_i} \right)^2 . \tag{8}$$

   The value of the reduced χ^2 is used to quantify the reliability of the errors (posterior standard deviation for us). Values close to 1 indicate that the errors are properly quantified, while values larger/smaller than 1 show that the errors are underestimated/overestimated.

3. Results

In this section, we first present the results obtained by training the models on Astrid galaxies. We then study the robustness of this and the models trained on IllustrisTNG and SIMBA galaxies.

3.1. Astrid Galaxies

Villaescusa-Navarro et al. (2022a) showed that neural networks were able to infer the value of Ω_m from individual galaxies from either IllustrisTNG or SIMBA simulations. Here we investigate whether this claim holds for individual galaxies generated by a different code (MP-Gadget) that uses an independent and different subgrid model (Astrid). We emphasize that while Arepo, Gizmo, MP-Gadget, and Open-Gadget may share the same (or a similar) gravity solver, the methods used to solve the hydrodynamic simulations and to implement the subgrid physics model can be substantially different.

We train our model on properties from individual galaxies of the Astrid simulations at $z = 0$ to infer the values of all parameters (Ω_m, σ_8, A_SN1, A_SN2, A_AGN1, and A_AGN2). We then test the model on individual galaxies from Astrid simulations.[17] We show the results of this test in Figure 1. Each point in the figure represents a single random galaxy. Our model predicts both the posterior mean and the standard deviation, which are shown as points and error bars, respectively.

As can be seen, the model is able to infer the value of Ω_m from the properties of individual galaxies with high accuracy and precision. Similar to Villaescusa-Navarro et al. (2022a), we find that the model is unable to infer the values of σ_8 and A_AGN1. On the other hand, our model seems to be able to infer the values of A_SN1, A_SN2, and A_AGN2, although with large error bars. The accuracy and precision metrics for each parameter are reported in the bottom right of each panel.

In the case of Ω_m, the mean relative error is $\sim$11% and $R^2 = 0.842$, indicating good precision and accuracy. The value of χ^2 is close to 1, showing that the error bars are accurately estimated. We note that some parameters have a very large value of χ^2 (e.g., A_SN1). This is due to a few outliers that contribute largely. Overall, the accuracy and precision metrics indicate that the model is well trained.

From now on, we will focus our attention on inferring the value of Ω_m and leave all other parameters aside. In order to improve the precision and accuracy of the model, we retrain it to predict only the posterior mean and standard deviation of Ω_m. We do this because this is usually an easier task than inferring several parameters at the same time—a more complicated task, prone to degeneracies and local minima.

It is interesting to visualize the average results for all galaxies in a given simulation, rather than individual galaxies. In this way, we are less sensitive to outliers and we can detect biases more easily. To carry out this task, we compute

$$\bar{\mu}_i = \frac{1}{N_s} \sum_{j \in s} \mu_{i,j} \qquad \bar{\sigma}_i = \frac{1}{N_s} \sum_{j \in s} \sigma_{i,j} \ , \tag{9}$$

where i denotes the considered parameter (e.g., Ω_m) and j runs over all N_s galaxies of a given simulation s. We show the average results for all 50 simulations in the test set in the left panel of Figure 2.

From the metrics, we can see that this model is indeed slightly more accurate and precise than the one used to infer all six properties. Overall, we see that the model is able to infer the value of Ω_m with a small bias in most of the cases. From these results, we can already conclude that the Astrid galaxies also exhibit a tight relationship between Ω_m and their individual properties. We emphasize that this relation, as determined by the networks, already accounts for changes in supernova and AGN parameters as modeled in CAMELS.

3.1.1. Redshift Dependence

We now study whether the tight relation between Ω_m and the galaxy properties holds at redshifts other than $z = 0$. For this, we train our models on Astrid galaxies at redshifts 1, 2, and 3 and compute the mean values of all the galaxies in a given simulation according to Equation (9). We then show the results in Figure 2.

As can be seen, our models perform well at all considered redshifts. Our results indicate a tighter relation between Ω_m and the galaxy properties at higher redshifts. This could be due to astrophysics effects being less severe on the galaxy properties.

[16] A trial represents a particular combination of the hyperparameter values.

[17] We emphasize that the galaxies we use to test the model come from simulations whose galaxies were not used to train the model.

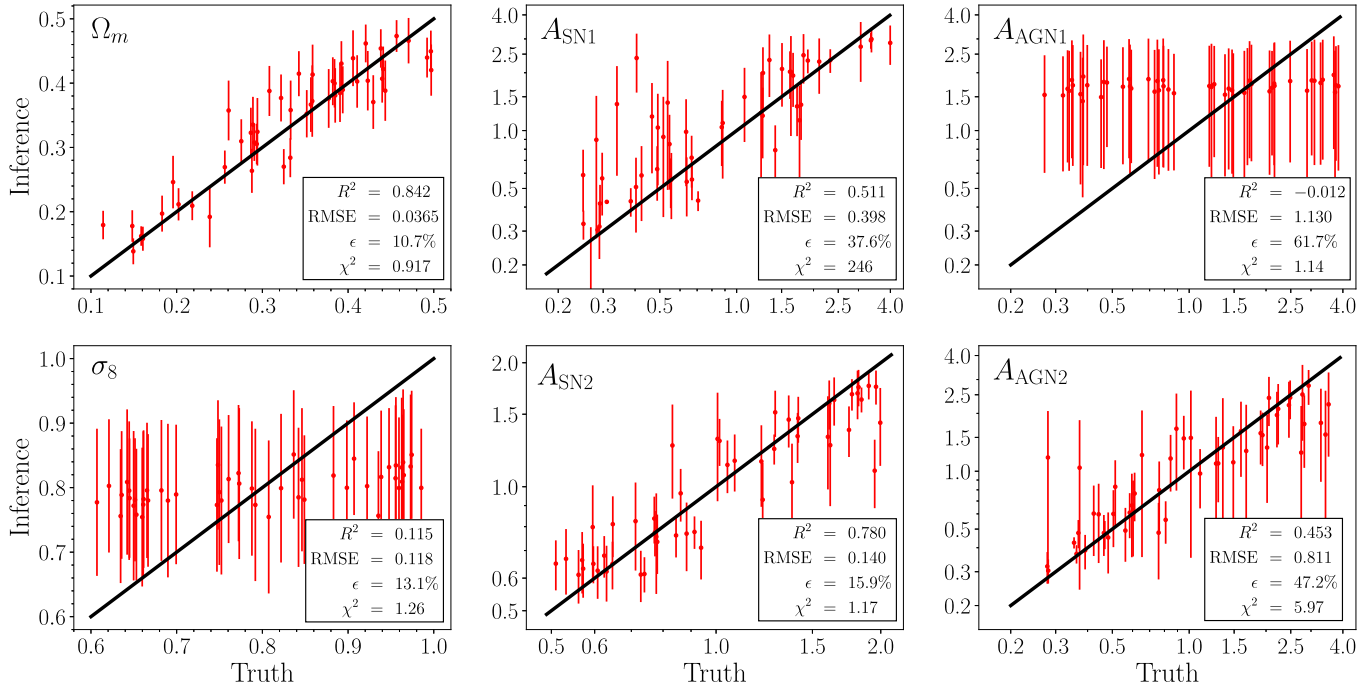

Figure 1. We have trained a neural network to perform likelihood-free inference on the values of the cosmological (Ω_m and σ_8) and astrophysical (A_{SN1}, A_{SN2}, A_{AGN1}, and A_{AGN2}) parameters using as input 14 properties of individual galaxies from the Astrid simulations at $z = 0$. Once the network is trained, we test it using individual galaxies from the test set. The different panels show the posterior means (points) and standard deviations (error bars) predicted by the network versus the true values. Every point with its error bar represents a single galaxy chosen randomly from each simulation of the test set. We find that our model is able to infer the value of Ω_m from the properties of individual galaxies with a ∼10% precision.

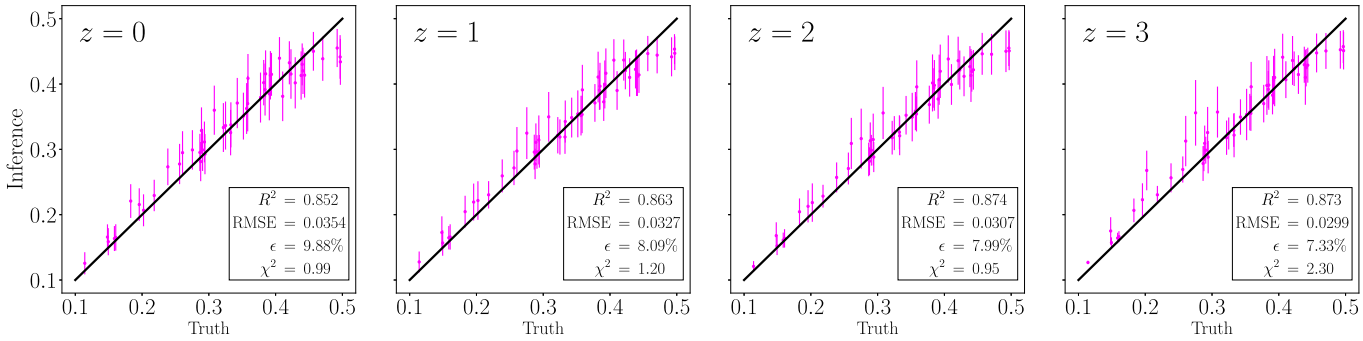

Figure 2. Redshift dependence: we have trained neural networks to infer the value of Ω_m using the properties of individual galaxies at different redshifts, for galaxies of the Astrid simulations. For each galaxy of each simulation of the test set, we compute the posterior mean and standard deviation for Ω_m. Next, we compute the means of those two numbers (Equation (9)) and plot them in the figure for the 50 different simulations in the test set. We show the results at redshifts 1, 2, and 3. As can be seen, our networks can infer the value of Ω_m from individual galaxies at redshifts higher than $z = 0$ with an accuracy similar to the one achieved by the models at $z = 0$.

We note that these results are in agreement with those of Villaescusa-Navarro et al. (2022a), who performed a similar analysis for IllustrisTNG and SIMBA galaxies. We note that a model trained on galaxies at a given redshift will not work if it is tested on galaxies at a different redshift, also in agreement with the results of Villaescusa-Navarro et al. (2022a).

3.1.2. Relevant Features

We now investigate what are the most relevant galaxy properties used by the models to infer the value of Ω_m. For this, we follow the procedure utilized in Villaescusa-Navarro et al. (2022a), which we briefly describe here. First, a gradient-boosted tree regressor[18] is used to predict the value of Ω_m from the 14 properties of the individual galaxies. Next, one of the galaxy properties is removed from the input, the regressor is retrained, and its accuracy is saved. This procedure is repeated for all 14 properties. The set with 13 properties that achieves the highest accuracy is kept for the next phase, and the property outside that set is discarded. The above procedure is then repeated by removing one property at a time, until the set only contains one property.

The above procedure[19] allows us to identify sets of variables that carry different fractions of the information. In Figure 3, we show the loss in the accuracy of the model as we discard the galaxy properties. For instance, keeping all galaxy properties but peculiar velocities and R_t has a negligible effect on the

[18] We use this method instead of neural networks, as this task would be too computationally expensive to carry out with neural networks.

[19] We note that this procedure is not meant to be optimal, and other sets of variables may yield similar results.

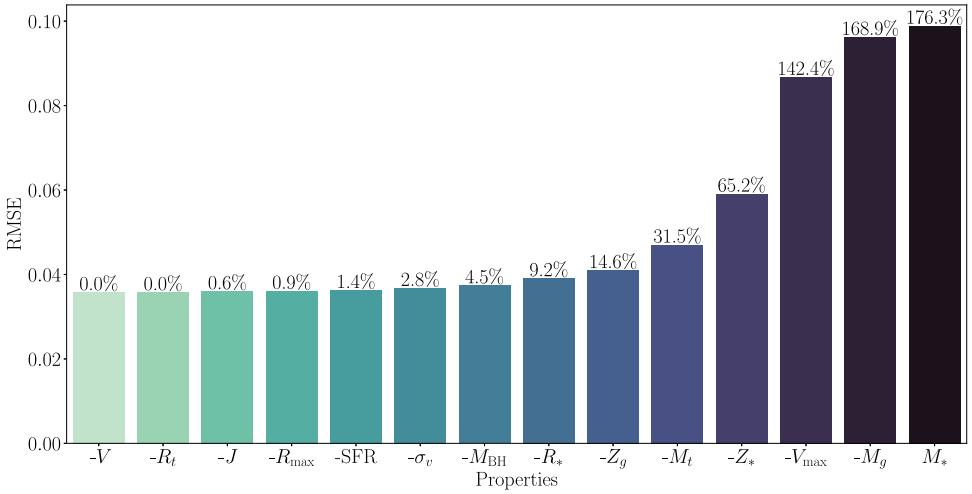

Figure 3. We rank order the galaxy properties for Astrid, such that the variables contributing the most to the model accuracy are on the right, while the features contributing the least are on the left (see the text for details of the procedure used). The vertical bars indicate the accuracy (in terms of RMSE) achieved by the considered variables, cumulatively from left to right, and the black numbers on top of them show the loss in accuracy with respect to a model trained using all variables. For instance, a model that only uses M_* achieves an RMSE of ∼0.1 and performs 176.3% worse than the model trained on all 14 properties (with an RMSE of ∼0.04). We emphasize that this ordering was derived when training gradient boosting tree models to perform regression to the value of Ω_m.

accuracy of the model. On the other hand, the set $\{M_*, M_g, V_{max}, Z_*, M_t\}$ achieves an RMSE value that is ∼30% lower than the one of the model trained on all the galaxy properties. Adding one more variable, $\{M_*, M_g, V_{max}, Z_*, M_t, Z_g\}$, further improves the accuracy: only ∼15% worse than using all the galaxy properties.

It is very interesting to see that three of the most relevant galaxy properties are: (1) the stellar mass (M_*); (2) the maximum circular velocity (V_{max}); and (3) the stellar metallicity (Z_*). Those variables are also among the most relevant for galaxies in the IllustrisTNG and SIMBA models (Villaescusa-Navarro et al. 2022a).

3.2. Robustness

It is important to investigate how well the different models generalize; i.e., whether the networks are able to infer the value of Ω_m from galaxies of simulations run with different codes to the ones used for training. Villaescusa-Navarro et al. (2022a) showed that the models trained on IllustrisTNG galaxies were not able to infer the correct value of Ω_m when tested on SIMBA galaxies (and the other way around).

We have trained three different models using $z = 0$ galaxies from (1) IllustrisTNG, (2) SIMBA, and (3) Astrid simulations. We then test the models on galaxies from all four codes: IllustrisTNG, SIMBA, Astrid, and Magneticum. To simplify the analysis, we compute the mean results from all galaxies using Equation (9). We show the results in Figure 4. We emphasize that the performance metrics shown in the different panels represent the results of taking the average over all the galaxies in the test set; for instance, $\chi^2 = (\sum_i^N \chi_i^2)/N$.

First, we are able to reproduce the results of Villaescusa-Navarro et al. (2022a), as the models trained on IllustrisTNG/SIMBA galaxies do not perform well when tested on SIMBA/IllustrisTNG galaxies.[20] On top of this, we find that those models do not perform well when tested on galaxies from the Astrid and Magneticum simulations. Similarly, we find that the model trained on Astrid galaxies does not perform well when tested on the IllustrisTNG, SIMBA, and Magneticum galaxies. It is interesting to see that the models trained on IllustrisTNG (Astrid) galaxies do not perform that badly when tested on the Astrid (IllustrisTNG) galaxies, perhaps signaling similarities between these two simulations.

We note that, from Figure 4, we can only reach conclusions about the mean behavior of the models. Thus, there are different possibilities that can explain our results. First, it could be that the models fail because the galaxies from different codes are very different; in this case, we would expect a generic failure of the model. In other words, the networks should infer wrong values of Ω_m for all (or the majority of the) galaxies. Second, it could be that the mean of the models is off due to the presence of some outliers where the models fail catastrophically. In order to shed light on this, we have computed, for each individual galaxy i in the test set, the value of its reduced chi-squared:

$$\chi_i^2 = \frac{(\theta_i - \mu_i)^2}{\sigma_i^2}, \tag{10}$$

where θ_i is the value of Ω_m of the galaxy, while μ_i and σ_i are the posterior mean and standard deviation predicted by the network.

In Figure 5, we show the distribution of the χ^2 values for the individual galaxies of the test sets of the different simulations. We find that most galaxies have low χ^2 values in all cases. For instance, 83%, 67%, 98%, and 62% of the IllustrisTNG, SIMBA, Astrid, and Magneticum galaxies, respectively, have χ^2 values below 5 when tested on the model trained on Astrid galaxies. However, the χ^2 distribution for galaxies tested on models different from the ones used for training display long tails with large χ^2 values; 27%, 5%, and 38% of the SIMBA, Astrid, and Magneticum galaxies, respectively, have χ^2 values larger than 10 when tested on the model trained on IllustrisTNG galaxies.

This indicates that the failure of the models is due to the presence of outliers. To verify this, we have removed all galaxies with $\chi_i^2 > 7$ from the test sets. We emphasize that our models are trained using all galaxies in the considered suite (e.g., all galaxies in the IllustrisTNG training set). We then

[20] Note that in this case we are using a slightly different cut in stellar mass when selecting the galaxies, so the results are similar but not identical.

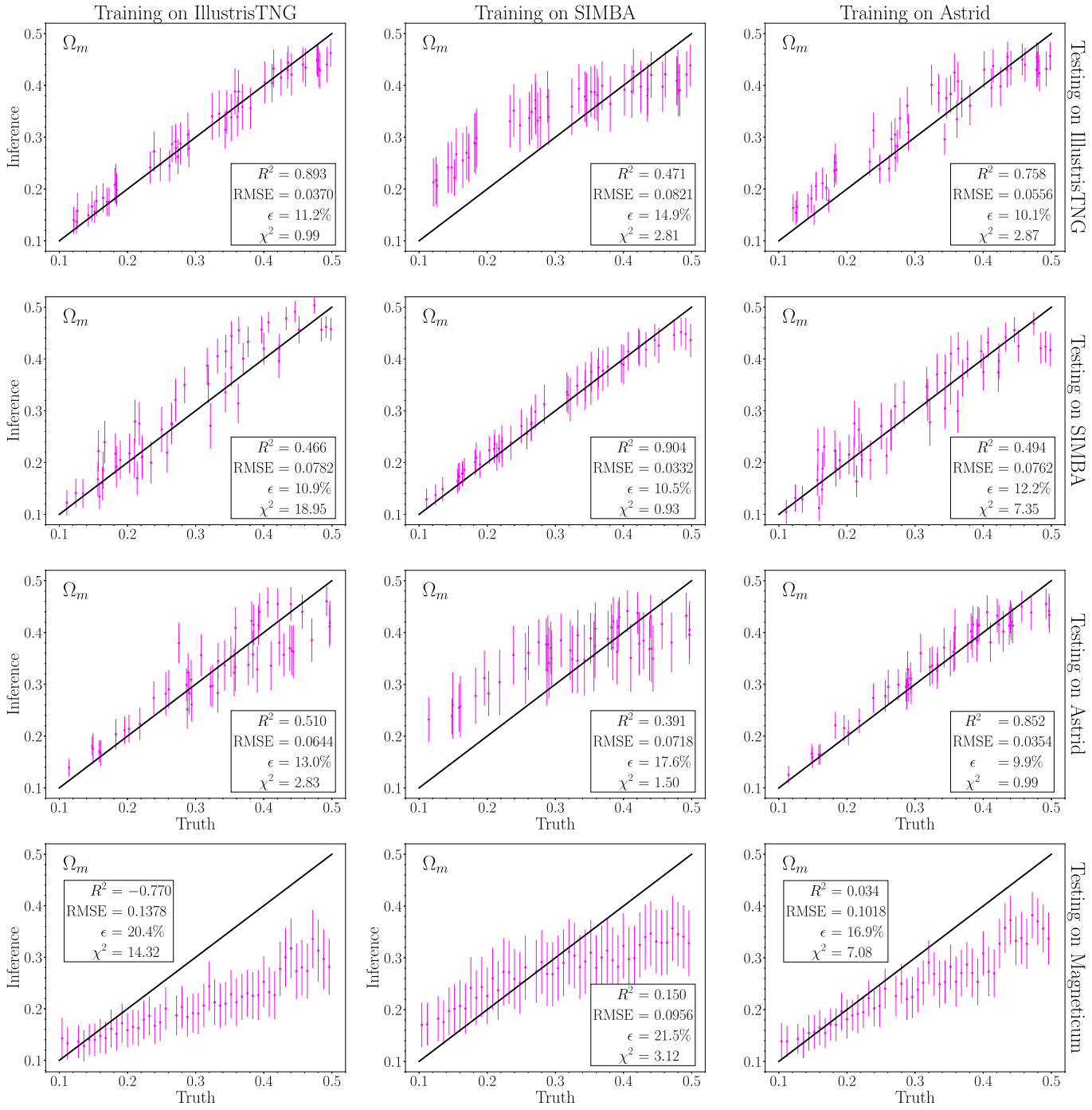

Figure 4. Robustness test. We test models trained on individual galaxies from IllustrisTNG (left column), SIMBA (middle column), and Astrid (right column) on galaxies from the IllustrisTNG (first row), SIMBA (second row), Astrid (third row), and Magneticum (fourth row) suites. Each point represents the average result of all the galaxies in that simulation (see Equation (9)). We find that none of the models are robust. On the other hand, the model trained on the Astrid galaxies performs relatively well when tested on the IllustrisTNG and SIMBA galaxies.

compute the mean values of all the remaining galaxies using Equation (9) and show the results in Figure 6.

As can be seen, the performance metrics of all the models significantly improve. In particular, the models trained on IllustrisTNG or Astrid galaxies perform well on galaxies from all simulations, with the exception of the Magneticum galaxies with large values of $\Omega_{\rm m}$. It is also interesting to note that although the model trained on SIMBA galaxies exhibits low χ^2 values, it is not robust: for low values of $\Omega_{\rm m}$, it systematically overpredicts the true value. On the other hand, this model performs better on Magneticum galaxies than the other two models.

We emphasize that removing the outliers will naturally lead to better predictions overall, so it is not surprising that the models become more robust when using this method. On the other hand, it is important to note that only a relatively small fraction of the galaxies behave as outliers. For instance, for the model trained on Astrid galaxies, we find that only 11%, 26%, <1%, and 30% of the IllustrisTNG, SIMBA, Astrid, and Magneticum galaxies, respectively, have $\chi^2_i \geqslant 7$.

Overall, the model trained on Astrid galaxies seems to be the one with the best generalization properties once the outliers have been removed. The fact that it fails for Magneticum

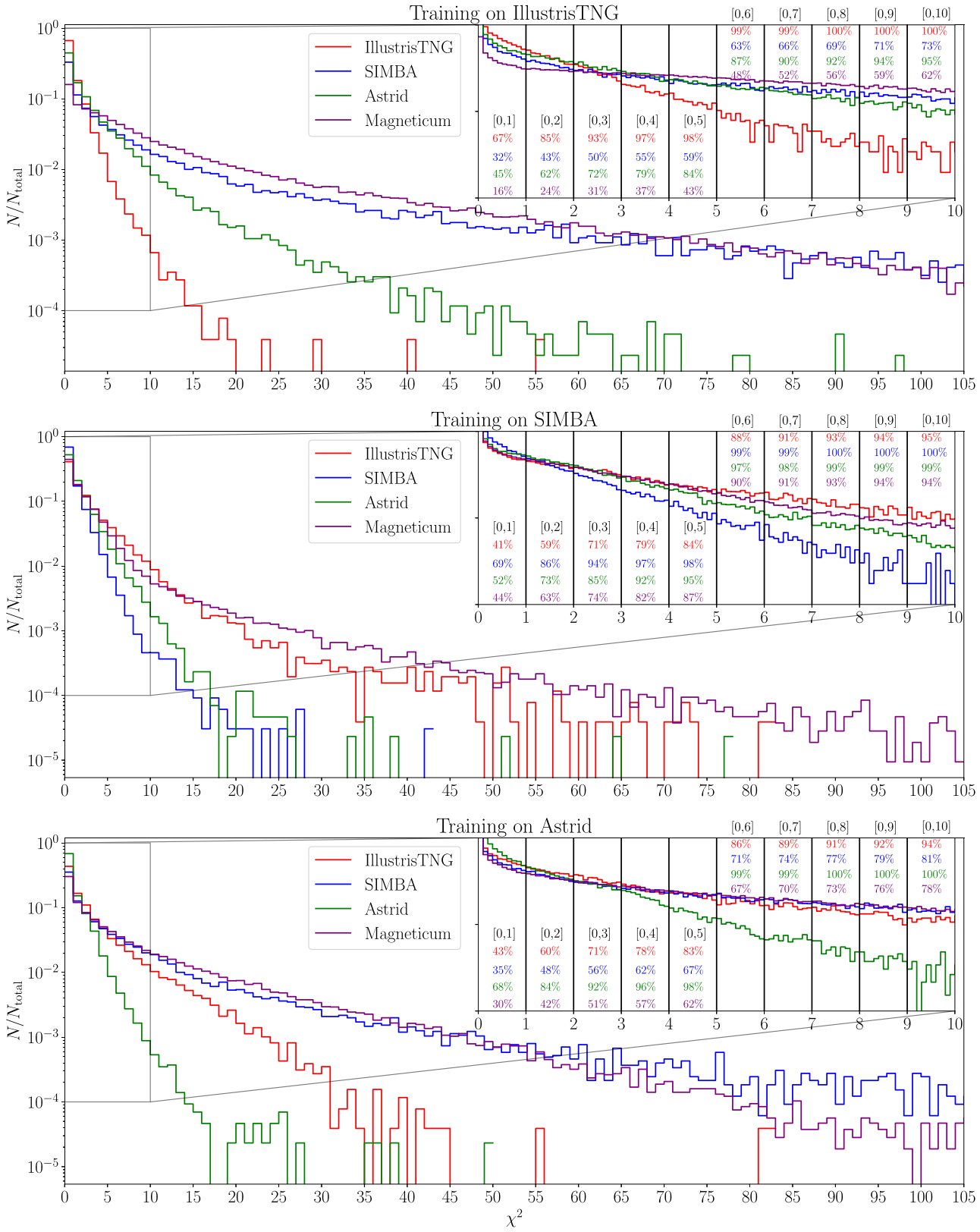

Figure 5. We have trained three different models using IllustrisTNG (top), SIMBA (middle), and Astrid (bottom) galaxies at $z = 0$. We test each model on galaxies from the IllustrisTNG, SIMBA, Astrid, and Magneticum simulations. For each galaxy i, we compute $\chi^2_i = (\theta_i - \mu_i)^2/\sigma^2_i$. The different lines show the distribution of χ^2 in the different setups. As can be seen, the χ^2 distribution changes if the simulation is tested on galaxies from a different code than the one used for training. On the other hand, most of the galaxies have low χ^2 values. The numbers in the subpanels indicate the fractions of galaxies with values of χ^2 smaller than the indicated threshold.

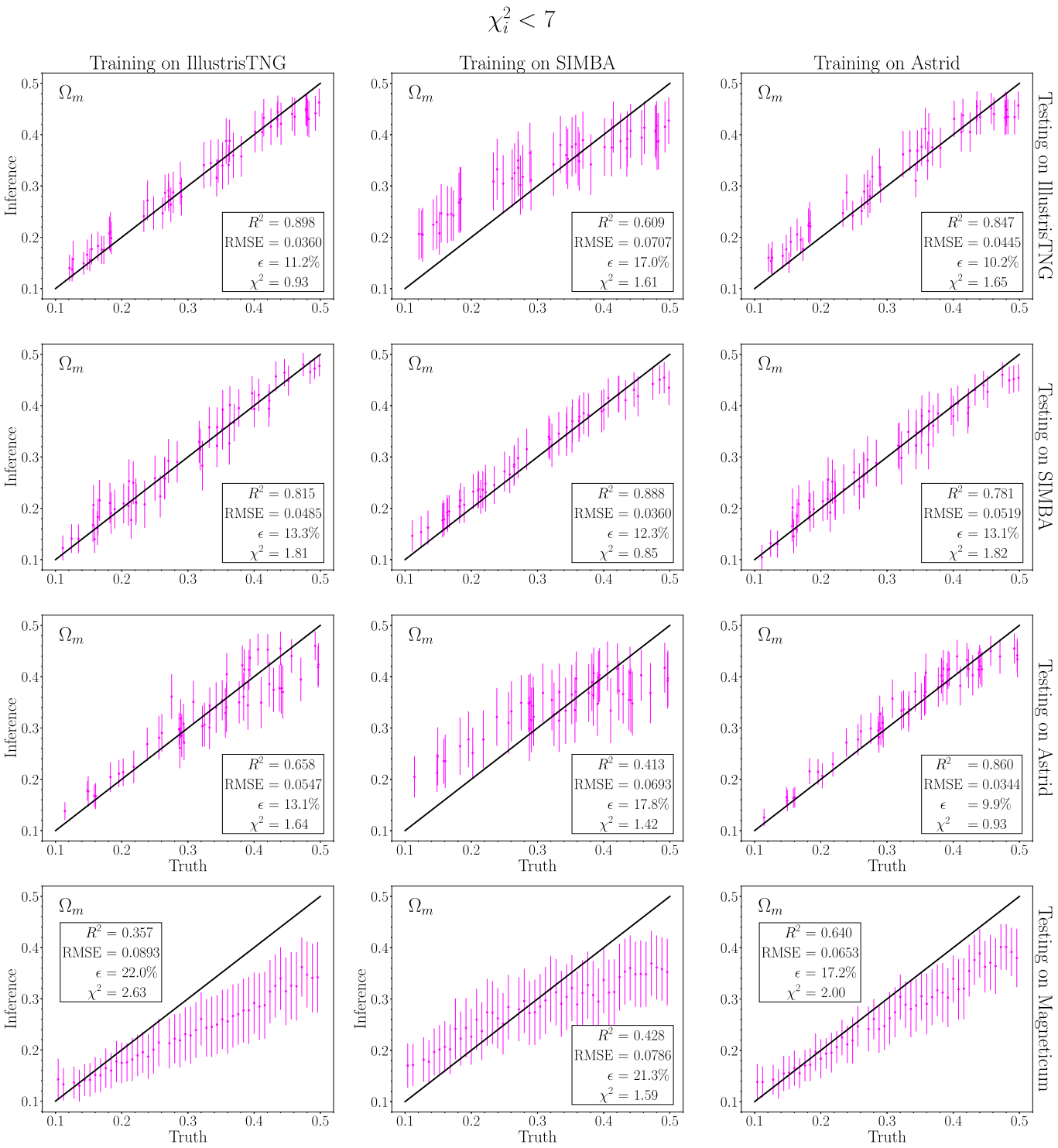

Figure 6. The same as Figure 4, but removing all galaxies whose individual χ^2 values are higher than 7. As can be seen, the models trained on the Astrid and IllustrisTNG galaxies are very robust (with the exception of the Magneticum galaxies with $\Omega_m \gtrsim 0.3$), while the model trained on the SIMBA galaxies is not.

galaxies with large values of Ω_m may be due to intrinsic differences between the Magneticum simulations and the other models (see, e.g., N. S. M. de Santi et al. 2023, in preparation). We have checked that more aggressive cuts in the χ_i^2 lead to less unbiased predictions, as expected (see Appendix A). Similar conclusions can be reached for SIMBA galaxies: by applying more aggressive cuts in the χ^2, the models become more and more robust across simulations. On the other hand, we also know that Astrid galaxies exhibit the most diverse set of properties (Ni et al. 2023). Thus, it is perhaps expected that they should perform best when tested on galaxies from other simulations.

These results indicate that we could develop robust models simply by knowing their range of validity; in other words, by not using them in cases where they will not perform well. A natural question arises: how do we identify model outliers

a priori? In other words, if we are not able to compute the χ^2 of a galaxy (e.g., with a real galaxy, we do not know the true value of $\Omega_{\rm m}$), how can we flag it as an outlier? While we do not provide a rigorous answer to this question in this paper, we have investigated the distribution of the galaxy properties for these outliers. We find evidence suggesting that those outliers correspond to galaxies whose properties are far away from the distribution of the galaxies used for training. We provide details of this test in Appendix B. These results indicate that the outliers of the model may simply correspond to outliers in the distribution.

We conclude this section by noting that training models on galaxies from two different codes does not seem to help in improving the robustness of the model. This was one of the hypotheses suggested by Villaescusa-Navarro et al. (2022a) for improving the generalization capabilities of the models. We provide further details of this in Appendix C.

4. Conclusions

We now summarize the main takeaways from this work:

1. Villaescusa-Navarro et al. (2022a) claimed that it is possible to infer the value of $\Omega_{\rm m}$, at fixed $\Omega_{\rm b}$, from the properties of individual simulated galaxies. They showed that their models were able to perform that task using galaxies generated by IllustrisTNG and SIMBA simulations. In this work, we have shown that it is also possible to infer $\Omega_{\rm m}$, at fixed $\Omega_{\rm b}$, with a $\sim$10% precision from the properties of individual galaxies generated by Astrid simulations, which employ a different method to solve the hydrodynamic equations and utilize a completely different subgrid model than the IllustrisTNG and SIMBA simulations.
2. The properties of the Astrid galaxies seem to be more sensitive to the values of the astrophysical parameters than the IllustrisTNG and SIMBA galaxies. Because of this, our models are able to infer the values of $A_{\rm SN2}$ and $A_{\rm AGN2}$ (although with large error bars); this was not possible with the IllustrisTNG and SIMBA galaxies.
3. The tight relation between $\Omega_{\rm m}$ and the properties of individual Astrid galaxies is present at all redshifts considered in this work: $z = 0$, 1, 2, and 3. Models trained at higher redshifts are able to infer $\Omega_{\rm m}$ slightly more accurately.
4. The five most important properties used by the model to infer $\Omega_{\rm m}$ from Astrid galaxies are $\{M_*, M_g, V_{\rm max}, Z_*, M_t\}$. By using only these properties, our models are able to infer $\Omega_{\rm m}$ with an accuracy that is only 30% worst than when using all 14 galaxy properties. Interestingly, the stellar mass, the maximum circular velocity, and the stellar metallicity are the top properties for the models trained on IllustrisTNG, SIMBA, or Astrid simulations.
5. The model trained on Astrid galaxies is not robust, and it fails when tested on IllustrisTNG, SIMBA, and Magneticum galaxies. The models trained on IllustrisTNG and SIMBA galaxies also perform badly when tested on the Astrid and Magneticum galaxies.
6. An important factor behind the lack of robustness of our models is the presence of outliers. As expected, removing those outliers significantly improves the robustness of the models. We note that the fraction of outliers is relatively small. For instance, for the model trained on Astrid galaxies, only 11%, 26%, <1%, and 30% of the IllustrisTNG, SIMBA, Astrid, and Magneticum galaxies, respectively, have $\chi^2 \geqslant 7$.
7. We note that all models exhibit a bias when tested on Magneticum galaxies with $\Omega_{\rm m} \gtrsim 0.3$; even after applying the $\chi^2_i > 7$ cut. This bias can be due to the fact that those simulations exhibit systematic differences with respect to other models (N. S. M. de Santi et al. 2023, in preparation) and therefore may require a more aggressive cut. We have checked that more aggressive cuts improve the performance of the model.
8. Our results indicate that model outliers (defined as galaxies with $\chi^2_i \geqslant 7$) tend to correspond to galaxies with properties either outside or in the tails of the galaxy distribution (see Appendix B).
9. Training on galaxies from two different simulations (e.g., IllustrisTNG and SIMBA) does not make the model robust and it still fails when it is tested on a third simulation.

It is important to emphasize that we identify outliers as galaxies having large χ^2 values ($\chi^2_i \geqslant 7$). Removing these outliers will naturally decrease the mean χ^2 value of the whole population, so it is not surprising that our models become more robust after performing this task. The important thing to note is that those outliers only represent a relatively small fraction of the galaxies.

Identifying outliers as galaxies with large χ^2_i values can only be done if the true value of $\Omega_{\rm m}$ is known. Thus, this method cannot be used with real galaxies. On the other hand, we have some hints that outliers tend to correspond to galaxies whose properties reside on the outskirts of the distribution used for training (see Appendix B). Therefore, it may be possible to identify outliers by finding galaxies whose properties are far away from the manifold that contains galaxy properties. Being able to identify and discard outliers will improve the robustness of the model, as we have shown in this paper. This opens the door to being able to apply our method to real data if the galaxies we consider are not outliers with respect to those we train the models on. We will investigate this avenue in detail in future work.

Finally, we emphasize that a model that is able to infer the value of $\Omega_{\rm m}$ from galaxies of simulations run with three different codes is not guaranteed to perform well on simulations from a new simulation. This is clearly illustrated with the Magneticum galaxies; even after removing the outliers, the model trained on Astrid galaxies does not perform well on the Magneticum galaxies. In this case, we may need to be even more aggressive in the way we define outliers to improve the robustness of the models. It is thus important to test the models on galaxies from as many diverse simulations as possible.

Overall, in this work, we have shown that galaxies from three different types of simulations (run with different codes and employing different subgrid physics models) exhibit a tight relationship between $\Omega_{\rm m}$ and their individual internal properties. This relation is not affected by astrophysics (at least not in the way we model them in CAMELS), since our simulations vary the astrophysical parameters controlling the efficiency of supernova and AGN feedback. While our models are not robust yet, in this work we have shown that identifying and removing outliers seems a promising way of addressing this issue. This method may also make the models robust to effects from

supersample covariance and changes in astrophysics and cosmological parameters not covered in CAMELS.

Acknowledgments

We have made use of the XGB,[21] PyTorch, and Optuna packages. We thank Marc Huertas-Companys, Arnab Lahiry, Natali de Santi, and Helen Shao for useful conversations. We thank the RECA[22] (Red Estudiantes Colombianos en Astronomía) Summer Internship Program for providing an avenue to initiate the discussions that led to this work. N.E. thanks the Simons Foundation for support while carrying out this work. The work of F.V.N. is supported by the Simons Foundation. The CAMELS project is supported by the Simons Foundation and the NSF grant AST 2108078. D.A.A. acknowledges support from NSF grants AST-2009687 and AST-2108944, CXO grant TM2-23006X, Simons Foundation Award CCA-1018464, and Cottrell Scholar Award CS-CSA-2023-028 by the Research Corporation for Science Advancement. K.D. acknowledges support from the COMPLEX project from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement ERC-2019-AdG 882679, as well as support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2094-390783311. Details on the CAMELS simulations can be found at https://www.camel-simulations.org.

Appendix A
χ^2 Cut

In Figure 6, we saw that all models failed when tested on Magneticum galaxies with $\Omega_m \gtrsim 0.3$, even after removing the galaxies with $\chi^2_i \geqslant 7$. We suggested that more aggressive cuts could improve the robustness of the model. In order to verify that, we have tested the model trained on Astrid galaxies on Magneticum galaxies, after removing galaxies with χ^2_i greater than 3, 5, and 7. We show the results in Figure 7. As expected, the model becomes more robust, the more aggressive our cuts are. We reach similar conclusions in other scenarios, e.g., training on SIMBA galaxies and testing on IllustrisTNG and Astrid galaxies.

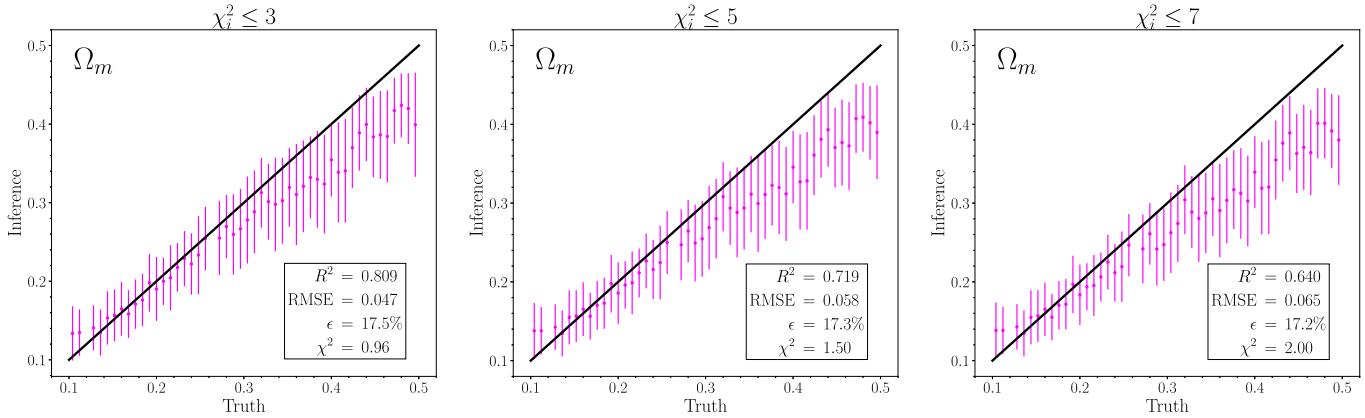

Figure 7. We have tested the model trained on Astrid galaxies on Magneticum galaxies after removing galaxies with $\chi^2_i \geqslant 3$ (left), $\chi^2_i \geqslant 5$ (middle), and $\chi^2_i \geqslant 7$ (right). As expected, the more aggressive we are in removing outliers, the better the model works.

[21] https://xgboost.readthedocs.io
[22] https://www.astroreca.org/en

Appendix B
Outliers

Here we study whether the model outliers, defined as galaxies with $\chi_i^2 \geqslant 7$, correspond to outliers in the galaxy properties. In other words, whether the outliers represent galaxies whose properties are on the tails (or outside) of the distribution. For this, we first consider all the galaxies in the Astrid training set. Next, for each galaxy, we take its stellar mass, gas mass, maximum circular velocity, and stellar metallicity (the four most important properties, according to Figure 3). We then project these properties into 2D maps and show their distribution in Figures 8–11 with hexagons. The color of a hexagon indicates how many galaxies are in that region, as indicated by the color bar.

We then test the model, trained on Astrid galaxies, on galaxies from the IllustrisTNG, SIMBA, Astrid, and Magneticum test sets. For each galaxy, we compute the values of the χ_i^2. We then select the ten galaxies with the highest χ^2 value for each suite. Finally, we project the properties of these galaxies into the 2D plots. We show the results in Figures 8–11 with colored points. The sizes of the points indicate their χ^2 values. In the case of IllustrisTNG, SIMBA, and Magneticum galaxies,

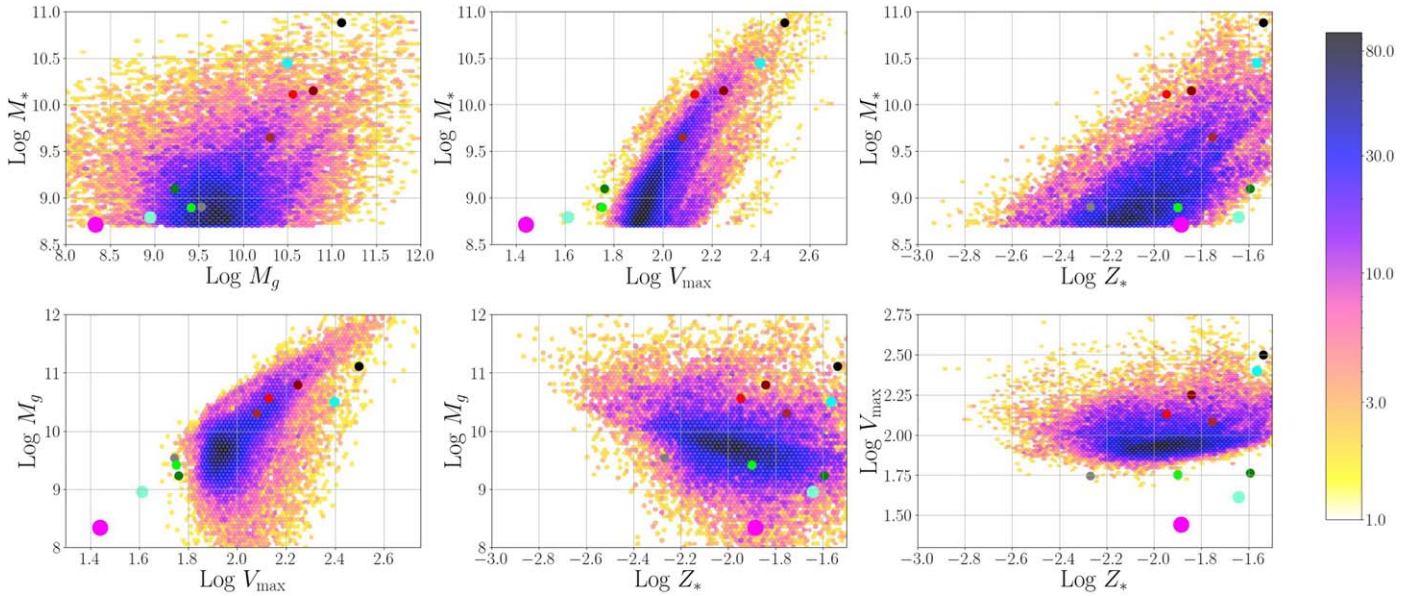

Figure 8. For each galaxy in the training set of the Astrid suite, we have considered four properties: (1) stellar mass, M_*; (2) gas mass; M_g, (3) maximum circular velocity, V_{max}; and (4) stellar metallicity, M_*. The different panels show the 2D distributions of these properties with hexagons. The color of a hexagon indicates the number of galaxies in that region of parameter space (see the color bar). We then test our model, trained on Astrid galaxies, on IllustrisTNG galaxies. For each of those galaxies, we compute their χ_i^2 values and select the 10 galaxies with the highest values. The χ_i^2 of these galaxies ranges from 40 to 144. We then project these galaxies into the different 2D properties. The colors of the points are used to match the galaxies across panels, and their sizes indicate the χ^2 values: larger points represent higher values. As can be seen, these galaxies tend to reside in regions in parameter space with very low density in one or several dimensions. This indicates that these outliers tend to correspond to galaxies whose properties are on the tails (or outside) of the Astrid distribution.

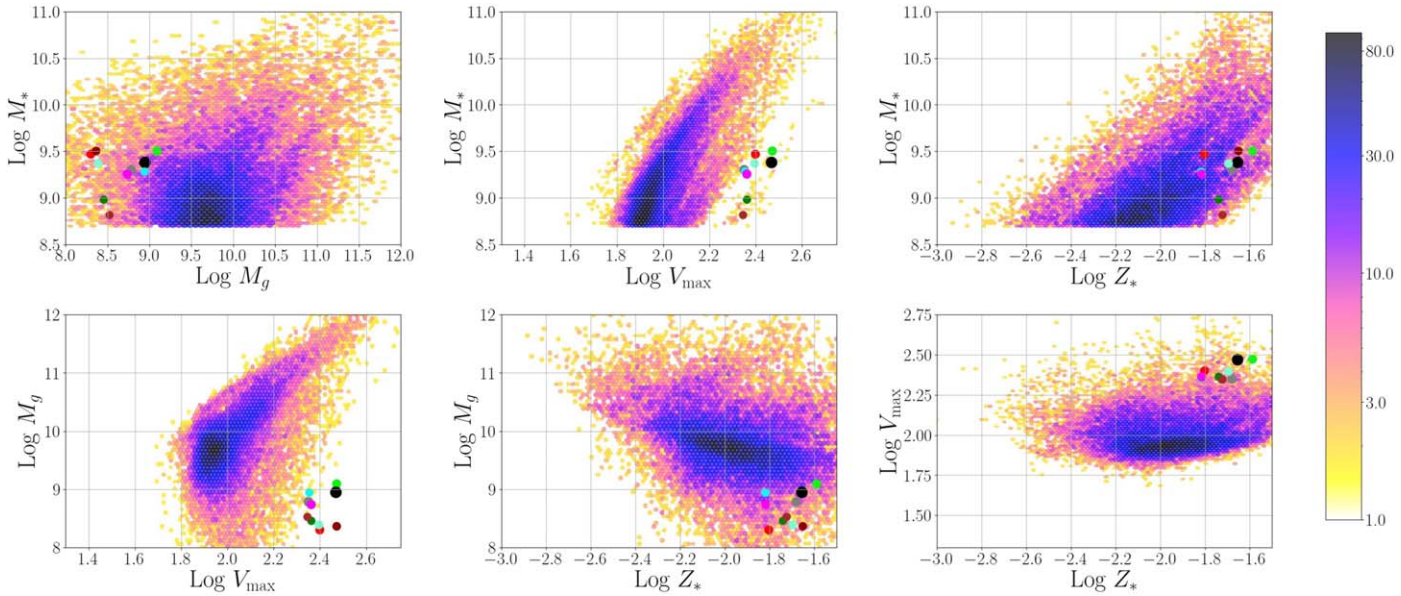

Figure 9. The same as Figure 8, but testing the model on SIMBA galaxies. The χ_i^2 of these galaxies range from 252 to 531.

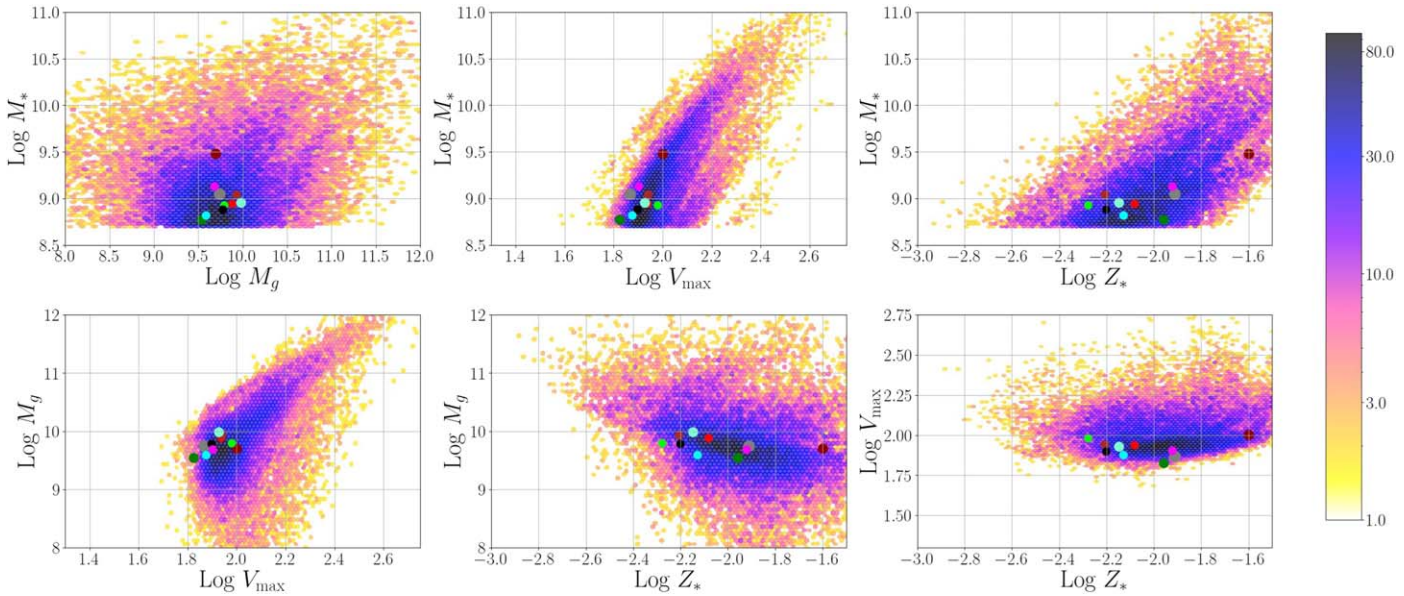

Figure 10. The same as Figure 8, but testing the model on Astrid galaxies. The χ_i^2 of these galaxies ranges from 24 to 49.

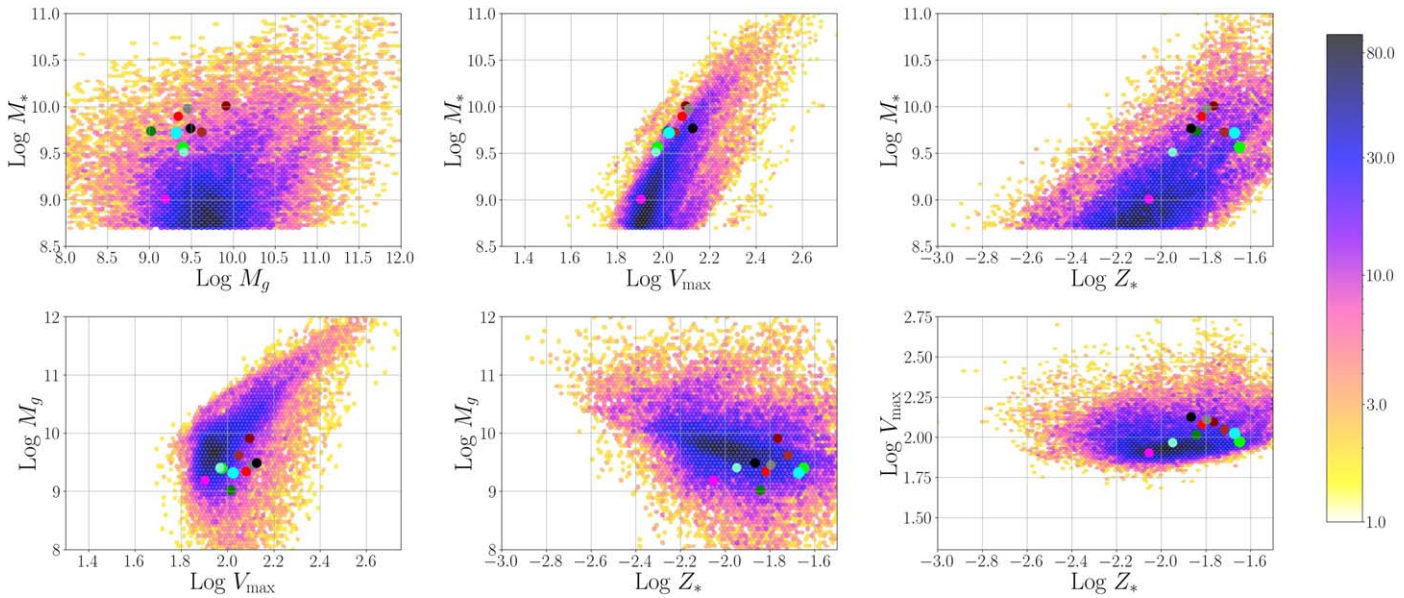

Figure 11. The same as Figure 8, but testing the model on Magneticum galaxies. The χ_i^2 of these galaxies ranges from 111 to 180.

it is clear that those galaxies are far away from the main distribution of Astrid galaxies (sometimes mostly along one or two directions). In the case of Astrid, the galaxies are instead located within the distribution, with some of them touching the tails. This is reflected in their smaller χ^2 values (see the captions to Figures 8–11).

To investigate whether the distribution of galaxies with large χ_i^2 values is different from the ones with smaller values, we have repeated the above exercise with galaxies randomly taken from the different test sets. We find that a larger fraction of galaxies occupy regions in parameter space more densely covered by the Astrid galaxies. However, a random sampling of the IllustrisTNG, SIMBA, and Magneticum galaxies also selects galaxies that are located on the tails of the Astrid distribution. This is expected, as the fractions of galaxies with $\chi_i^2 \geqslant 4$ can be 22%, 38%, and 43% in the case of IllustrisTNG, SIMBA, and Magneticum, respectively.

While this is not a rigorous analysis, our results indicate that the model outliers represent galaxies whose properties are located in the tails of the distribution. This fact can be exploited to increase the robustness of the models. In future work, we plan to make use of more sophisticated machine-learning tools, like normalizing flows, to address this question in a more rigorous manner.

Appendix C
Robustness with Two Suites

One of the reasons behind the lack of robustness of our models may be that galaxies from the different suites are just too different and their properties exist in different regions. Training models on galaxies from two (or more) suites may improve the robustness of the model, by forcing it to learn common features across models.

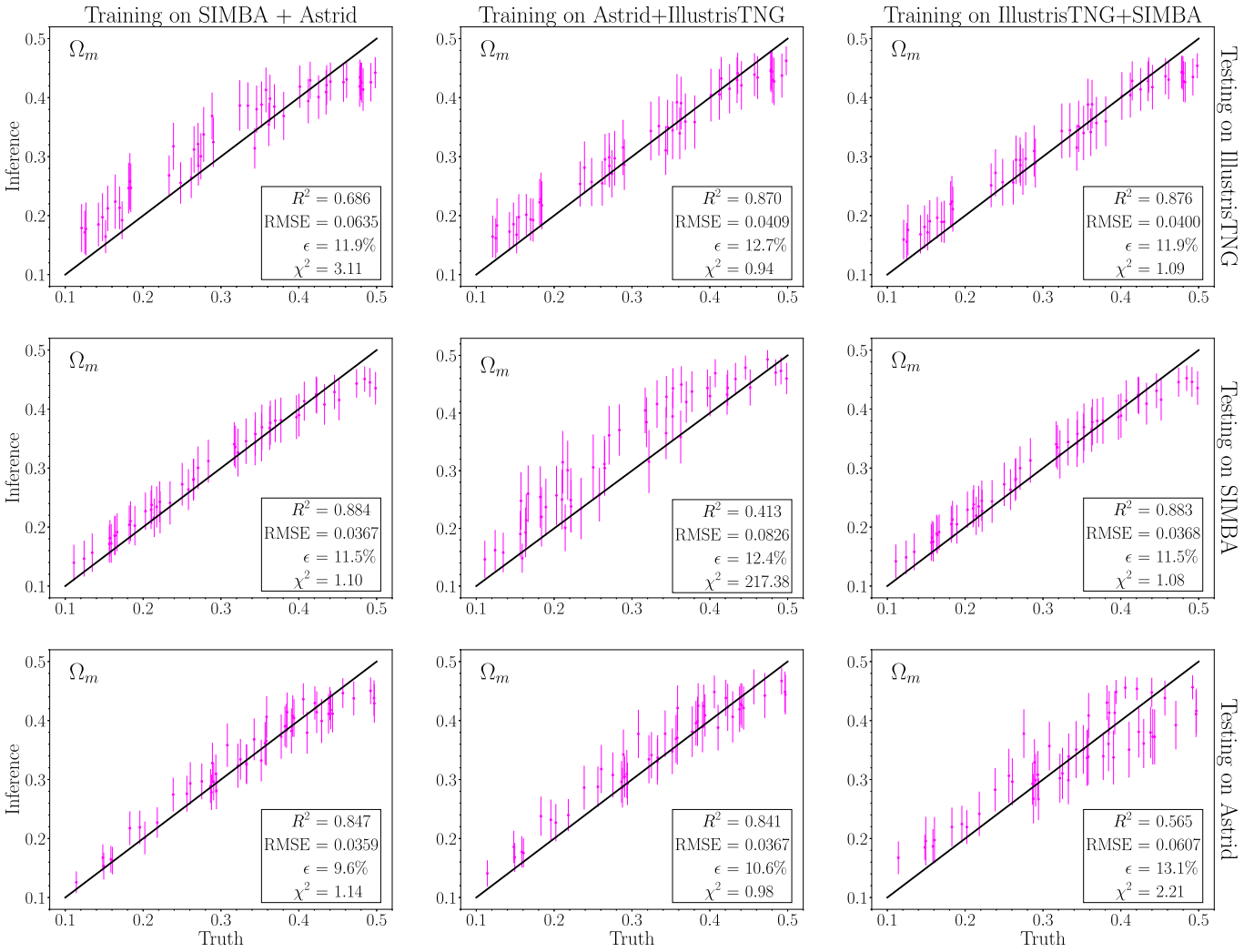

Figure 12. The same as Figure 4, but training on SIMBA and Astrid galaxies (left column), Astrid and IllustrisTNG galaxies (middle column), and IllustrisTNG and SIMBA galaxies (right column). As can be seen, training on galaxies from two suites does not make the model robust.

On the other hand, training on such a configuration may make the model first classify the galaxy (e.g., recognizing it is a SIMBA galaxy) before performing the usual task. If so, the model will not generalize well.

In order to test this, we have trained models using galaxies from two suites, e.g., Illustris and SIMBA galaxies. We then test the model on galaxies from all the different suites. We show the results in Figure 12. We find that models trained on galaxies from two suites work well when tested on galaxies from those suites, but fail when tested on a third suite. This indicates that we cannot build robust models by training networks on galaxies from different simulations. Given the fact that the number of different subgrid physics models is really small, we believe this statement may hold in general.

ORCID iDs

Francisco Villaescusa-Navarro https://orcid.org/0000-0002-4816-0455
Yueying Ni https://orcid.org/0000-0001-7899-7195
ChangHoon Hahn https://orcid.org/0000-0003-1197-0902
Elena Hernández-Martínez https://orcid.org/0000-0002-1329-9246
Romain Teyssier https://orcid.org/0000-0001-7689-0933
Daniel Anglés-Alcázar https://orcid.org/0000-0001-5769-4945
Tiago Castro https://orcid.org/0000-0002-6292-3228

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in KDD '19: Proc. 25th ACM SIGKDD Int.Conf. on Knowledge Discovery & Data Mining (New York: ACM), 2623
Bird, S., Ni, Y., Di Matteo, T., et al. 2022, MNRAS, 512, 3703
Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827
Dolag, K., Borgani, S., Murante, G., & Springel, V. 2009, MNRAS, 399, 497
Fabjan, D., Borgani, S., Rasia, E., et al. 2011, MNRAS, 416, 801
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
Hirschmann, M., Dolag, K., Saro, A., et al. 2014, MNRAS, 442, 2304
Hopkins, P. F. 2015, MNRAS, 450, 53
Jeffrey, N., & Wandelt, B. D. 2020, arXiv:2011.05991
Nelson, D., Pillepich, A., Genel, S., et al. 2015, A&C, 13, 12
Ni, Y., Di Matteo, T., Bird, S., et al. 2022, MNRAS, 513, 670
Ni, Y., Genel, S., Anglés-Alcázar, D., et al. 2023, arXiv:2304.02096
Pillepich, A., Springel, V., Nelson, D., et al. 2018, MNRAS, 473, 4077
Springel, V. 2010, MNRAS, 401, 791
Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726

Steinborn, L. K., Dolag, K., Comerford, J. M., et al. 2016, MNRAS, 458, 1013
Teklu, A. F., Remus, R.-S., Dolag, K., et al. 2015, ApJ, 812, 29
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021, ApJ, 915, 71
Villaescusa-Navarro, F., Ding, J., Genel, S., et al. 2022a, ApJ, 929, 132
Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. 2022b, ApJS, 259, 61
Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., et al. 2023, ApJS, 265, 54
Weinberger, R., Springel, V., Hernquist, L., et al. 2017, MNRAS, 465, 3291