Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **522,** 3236–3247 (2023) https://doi.org/10.1093/mnras/stad1186
Advance Access publication 2023 April 24

High-fidelity reproduction of central galaxy joint distributions with neural networks

Natália V. N. Rodrigues [1]★ Natalí S. M. de Santi,[1,2] Antonio D. Montero-Dorta [3] and L. Raul Abramo[1]

[1]*Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, Rua do Matão 1371, CEP 05508-090, São Paulo, Brazil*
[2]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*
[3]*Departamento de Física, Universidad Técnica Federico Santa María, Casilla 110-V, Avenida España 1680, Valparaíso, Chile*

Accepted 2023 April 11. Received 2023 March 20; in original form 2023 January 29

ABSTRACT

The relationship between galaxies and haloes is central to the description of galaxy formation and a fundamental step towards extracting precise cosmological information from galaxy maps. However, this connection involves several complex processes that are interconnected. Machine Learning methods are flexible tools that can learn complex correlations between a large number of features, but are traditionally designed as deterministic estimators. In this work, we use the IllustrisTNG300-1 simulation and apply neural networks in a binning classification scheme to predict probability distributions of central galaxy properties, namely stellar mass, colour, specific star formation rate, and radius, using as input features the halo mass, concentration, spin, age, and the overdensity on a scale of 3 h^{-1} Mpc. The model captures the intrinsic scatter in the relation between halo and galaxy properties, and can thus be used to quantify the uncertainties related to the stochasticity of the galaxy properties with respect to the halo properties. In particular, with our proposed method, one can define and accurately reproduce the properties of the different galaxy populations in great detail. We demonstrate the power of this tool by directly comparing traditional single-point estimators and the predicted joint probability distributions, and also by computing the power spectrum of a large number of tracers defined on the basis of the predicted colour–stellar mass diagram. We show that the neural networks reproduce clustering statistics of the individual galaxy populations with excellent precision and accuracy.

Key words: galaxies: statistics – cosmology: large-scale structure of Universe – methods: data analysis – methods: statistical.

1 INTRODUCTION

Characterizing the connection between the properties of galaxies and those of the underlying population of dark matter (DM) haloes is one of the most crucial aspects to understand the large-scale structure (LSS) of the Universe. This link not only encapsulates fundamental information about the process of galaxy formation, but it is also a crucial step to optimize the extraction of cosmological constraints from galaxy maps.

The halo–galaxy connection is nowadays investigated using a variety of techniques (see e.g. Wechsler & Tinker 2018). On the one hand, empirical methods use DM-only simulations as the basis on top of which different analytical prescriptions are implemented in order to establish that connection. These techniques include subhalo abundance matching (e.g. Conroy, Wechsler & Kravtsov 2006; Behroozi, Conroy & Wechsler 2010; Trujillo-Gomez et al. 2011; Favole et al. 2016, 2022; Guo et al. 2016; Contreras, Angulo & Zennaro 2020a, b; Hadzhiyska et al. 2021), halo occupation distributions (e.g. Berlind & Weinberg 2002; Zehavi et al. 2005, 2018; Artale et al. 2018; Bose et al. 2019; Hadzhiyska et al. 2020a; Xu, Zehavi & Contreras 2021), and empirical forward modelling (e.g. Becker 2015; Moster, Naab & White 2018; Behroozi et al. 2019). On the other hand, it is possible to model, with varying degrees of detail, the physical mechanisms that shape the process of galaxy formation. In this context, hydrodynamical simulations (e.g. Somerville & Davé 2015; Naab & Ostriker 2017; Pillepich et al. 2018a, b; Springel et al. 2018; Villaescusa-Navarro et al. 2021, 2022) are perhaps the most ambitious efforts. These models employ known physics to simulate, at a subgrid level, a variety of processes that are related to galaxy formation such as star formation, radiative metal cooling, and supernova, stellar, and black hole feedback (for reviews on this, see Somerville & Davé 2015; Naab & Ostriker 2017). This modelling can also be approached from a semi-analytical, less computationally demanding, perspective. These semi-analytical models (e.g. White & Frenk 1991; Guo et al. 2013) employ physically motivated recipes to mimic the galaxy formation processes.

In this paper, we investigate the halo–galaxy connection from a machine learning (ML) perspective. The issue of the halo–galaxy connection has been addressed using ML by many works (e.g. Kamdar, Turk & Brunner 2016; Agarwal, Davé & Bassett 2018; Calderon & Berlind 2019; Jo & Kim 2019; Man et al. 2019; Yip et al. 2019; Zhang et al. 2019; Kasmanoff et al. 2020; Delgado et al. 2021; McGibbon & Khochfar 2021; Shao et al. 2021; de Andres et al. 2022; Jespersen et al. 2022; Lovell et al. 2022; Stiskalek et al. 2022; Chittenden & Tojeiro 2023). In de Santi et al. (2022), we provide an ML suite combining some of the most powerful, well-known models in the literature to predict central galaxy properties using host halo properties. All the applied methods, however, are designed to return a single value for each galaxy property, independently of the

★ E-mail: natalia.villa.rodrigues@usp.br

remaining properties. However, there are many complex interrelated processes involved in the formation and evolution of galaxies, and their properties cannot be precisely determined by halo properties alone. Therefore, a model that proposes to map the relation between galaxies and host haloes should encode not only the correlations between galaxy properties, but also the uncertainties due to the stochastic aspects of galaxy formation. In other words, any given halo could host a central galaxy with a variety of properties and, hence, a model should return joint probability distributions for the possible values of those galaxy properties, instead of a single one.

The ML suite from our previous work (Santi et al. 2022) provided encouraging results in terms of single-point estimation metrics, such as the Pearson correlation coefficient between true and predicted values, especially for stellar mass, which is highly correlated with halo mass. However, deterministic models that try to predict individual galaxy properties can be biased towards the most frequent values, and thus fail to recover the overall distributions of the galaxy properties. In that paper, this issue is treated as an imbalanced data problem; i.e. despite of the fact that different output values could be associated with some fixed set of halo properties, the machine tends to assign the most frequent values. To address this problem, we made use of a data augmentation technique to increase the weight of the less represented instances, which allowed us to better recover the underrepresented populations, but still in a way that each halo is assigned a single, individual value for each central galaxy property (Santi et al. 2022).

In this work, we proceed by predicting probability distributions with neural networks (NNs) with a binning classification scheme, which we refer to as $\mathrm{NN_{class}}$, for the same central galaxy properties as Santi et al. (2022), namely stellar mass, $g-i$ colour, specific star formation rate (sSFR), and galaxy radius. This not only enables us to recover the overall distributions of the galaxy properties from the IllustrisTNG300-1 (hereafter, TNG300) sample, but also to capture the intrinsic scatter in the halo–galaxy mapping by providing, for each halo, the probability distributions associated with its central galaxy properties. We also train $\mathrm{NN_{class}}$ to predict the galaxy properties jointly, finding that the joint distributions recover correlations that are lost when predicting univariate distributions independently. ML probability-based descriptions have been used in related contexts, in particular with NNs, such as photometric redshift estimation (e.g. Lima et al. 2022), dynamical mass of galaxy cluster estimation (e.g. Ramanah et al. 2020; Ho et al. 2021), and recently in the halo–galaxy connection (e.g. Stiskalek et al. 2022).

In order to study how $\mathrm{NN_{class}}$ captures the intrinsic stochasticity in the halo–galaxy connection, we analyse the shape of the distributions of individual galaxies, which gives some insights into the contribution of secondary halo properties. Moreover, we analyse how this uncertainty affects clustering statistics, namely the power spectrum. Our technique enables us to define as many galaxy populations as wished, and to analyse to what extent those populations occupy the same types of haloes. We explore this flexibility by computing the power spectrum of a large number of galaxy populations (tracers), selected on the basis of the colour–stellar mass diagram.

The paper is organized as follows: The IllustrisTNG data and the chosen set of halo and galaxy properties are described in Section 2. In Section 3, we explain how we applied NNs to predict joint probability distributions. Section 4 analyses the quality of the results obtained with the NNs by comparing the predictions with the IllustrisTNG catalogue. In Section 5, we present our results in terms of the power spectra of several galaxy populations. Finally, we outline our main conclusions in Section 6, and discuss our plans for future improvements and applications.

2 DATA

Our analysis is based on data from the IllustrisTNG magneto-hydrodynamical cosmological simulation (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018, 2019; Pillepich et al. 2018a, b; Springel et al. 2018). This simulation suite, which was generated using the AREPO moving-mesh code (Springel 2010), is an improved version of the previous Illustris simulation (Genel et al. 2014; Vogelsberger et al. 2014a, b). IllustrisTNG features a variety of updated subgrid models accounting for star formation, radiative metal cooling, chemical enrichment from SNII, SNIa, and AGB stars, and feedback mechanisms (including stellar and supermassive black hole feedback). These models were calibrated to reproduce an array of observational constraints, such as the $z=0$ galaxy stellar mass function and the cosmic star formation rate (SFR) density, to name a few (see the aforementioned references for more information). The IllustrisTNG simulation adopts the standard Lambda cold dark matter cosmology (Planck Collaboration XIII 2016), with parameters $\Omega_\mathrm{m}=0.3089$, $\Omega_\mathrm{b}=0.0486$, $\Omega_\Lambda=0.6911$, $H_0=100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ with $h=0.6774$, $\sigma_8=0.8159$, and $n_\mathrm{s}=0.9667$.

The ML methodology that we developed in this work to reproduce the halo–galaxy connection is applied to galaxy clustering in terms of the power spectrum. For this reason, in order to minimize cosmic variance (CV), we chose to analyse the largest box available in the data base, TNG300, spanning a side length of $205\,h^{-1}$Mpc with periodic boundary conditions. TNG300 contains 2500^3 DM particles of mass $4.0\times10^7\ h^{-1}\,\mathrm{M_\odot}$ and 2500^3 gas cells of mass $7.6\times10^6\ h^{-1}\,\mathrm{M_\odot}$. The adequacy of TNG300 in the context of clustering science has been extensively proven in a variety of analyses (see e.g. Contreras et al. 2020a; Gu et al. 2020; Hadzhiyska et al. 2020b, 2021; Montero-Dorta et al. 2020b, 2021a, b; Shi et al. 2020; Favole et al. 2022; Santi et al. 2022).

In this work, we employ both galaxy and DM halo information from TNG300. DM haloes in the entire IllustrisTNG suite are identified using a friends-of-friends algorithm based on a linking length of 0.2 times the mean of the inter-particle separation (Davis et al. 1985). As in Santi et al. (2022), the following halo properties are used as input features to train the NNs:

(i) *Virial mass* ($M_\mathrm{vir}[h^{-1}\,\mathrm{M_\odot}]$), which is computed by adding up the mass of all gas cells and particles contained within the virial radius R_vir (based on a collapse density threshold of $\Lambda_\mathrm{c}=200$). In order to ensure that haloes are well resolved, we impose a mass cut $\log_{10}(M_\mathrm{vir}[h^{-1}\,\mathrm{M_\odot}])\geq10.5$, corresponding to at least 500 DM particles.

(ii) *Virial concentration* (c_vir), defined in the standard way as the ratio between the virial radius and the scale radius, i.e. $c_\mathrm{vir}=R_\mathrm{vir}/R_\mathrm{s}$. R_s is obtained by fitting the DM density profiles of individual haloes with an NFW profile (Navarro, Frenk & White 1997).

(iii) *Halo spin* (λ_halo), for which we follow the Bullock et al. (2001) definition: $\lambda_\mathrm{halo}=|J|/\sqrt{2}M_\mathrm{vir}V_\mathrm{vir}R_\mathrm{vir}$. Here, J and V_vir are the angular momentum of the halo and its circular velocity at R_vir, respectively.

(iv) *Halo age*, parametrized as the half-mass formation redshift $z_{1/2}$. This parameter corresponds to the redshift at which half of the present-day halo mass has been accreted into a single subhalo for the first time. The formation redshift is measured following the progenitors of the main branch of the subhalo merger tree computed with SUBLINK, which is initialized at $z=6$.

(v) The *overdensity* around haloes on a scale of $3\ h^{-1}$ Mpc (δ_3), defined as the number density of subhaloes within a sphere of radius $R=3\ h^{-1}$ Mpc, normalized by the total number density of subhaloes in the TNG300 box (e.g. Artale et al. 2018; Bose et al. 2019).

On the other hand, subhaloes (i.e. gravitationally bound substructures) are identified in IllustrisTNG using the SUBFIND algorithm (Springel et al. 2001; Dolag et al. 2009). Subhaloes containing a non-zero stellar mass component are labelled as galaxies. Again, following Santi et al. (2022) for consistency, TNG300 galaxies are characterized in this work using the following basic properties:

(i) The *stellar mass* (M_* $[h^{-1}\,\mathrm{M}_\odot]$), which includes all stellar particles within the subhalo. In order to ensure that galaxies are well resolved, we impose a mass cut $\log_{10}(M_*[h^{-1}\,\mathrm{M}_\odot]) \geq 8.75$, corresponding to at least 50 gas cells.

(ii) The *colour* $g - i$, computed from the rest-frame magnitudes, which are obtained in IllustrisTNG by adding up the luminosities of all stellar particles in the subhalo (Buser 1978). Note that the specific choice of colour is rather arbitrary. We have checked that using other combinations (i.e. $g - r$) provides similar results.

(iii) The *specific star formation rate* (sSFR $[\mathrm{yr}^{-1}h]$), which is the SFR normalized by stellar mass. The SFR is computed by adding up the SFRs of all gas cells in the subhalo. Note that around 14 per cent of the galaxies at redshift $z = 0$ in TNG300 have SFR $= 0$. In order to avoid numerical issues, we have adopted the same approach as in Santi et al. (2022), assigning to these objects artificial values of SFR, such that they end up distributed around $\log_{10}(\mathrm{sSFR}[\mathrm{yr}^{-1}\,h]) = -13.5$.

(iv) The *galaxy size*, parametrized as the stellar (3D) half-mass radius ($R^{(*)}_{1/2}[h^{-1}\,\mathrm{kpc}]$) – i.e. the comoving radius containing half of the stellar mass in the subhalo.

3 METHODOLOGY

NNs are designed to learn how to map an instance, which is characterized by some set of input features X, to a set of output features Y, by weighting and combining the input features. These weights are fitted by minimizing a loss function with some optimizer.

In this work, the input features are the halo properties and the outputs are the galaxy properties introduced in Section 2. Starting with a sample where the target value Y is known for all instances (the TNG300 catalogue), we split it into training, validation, and test sets. The training set is used to fit the model parameters (weights). The validation set is used to monitor overfitting, i.e. to ensure that the model is properly generalizing to data outside of the training set, and to fit the model's hyperparameters.[1] The test set remains completely blind to the training and validating procedures, and can thus be used to infer the performance of the model when applied to entirely new instances. The training, validation, and test sets contain, respectively, 48, 12, and 40 per cent of the initial sample of 174 527 objects from the TNG300 catalogue.

Our goal is to predict central galaxy properties from a set of halo properties. In the context of ML, this would in principle fall in the category of a supervised regression problem. However, traditional regression models are designed to output single values, while any given halo could host many different central galaxies (since the set of halo properties that we use as inputs does not determine exactly the outcome of the galaxy formation process in terms of the precise values of the galaxy properties). This is reflected, as an example, in the well-known scatter in the stellar-to-halo mass relation (Wechsler & Tinker 2018; Stiskalek et al. 2022). Therefore, in order to incorporate this uncertainty, we need a model that returns not only a single best-estimate value for each galaxy property, but some proxy for the probability distribution for those properties.

In this paper, we have addressed this issue by converting the regression problem into a classification. The idea is to define K classes by splitting each galaxy property into K intervals, or bins. Just like in the usual classification tasks, the model will return a score associated with each class (bin). These scores add up to one, giving a probabilistic interpretation of the output. This approach has been widely used, as an example, in the context of photometric redshift estimation (Sadeh, Abdalla & Lahav 2016; Pasquet et al. 2019; Lima et al. 2022). We refer to our method, which is based on training NNs classifiers, as $\mathrm{NN_{class}}$.

As a starting point, we train four models to predict each galaxy property individually as univariate distributions; i.e. we have separate models to predict $P(M_*)$, $P(g - i)$, $P(\mathrm{sSFR})$, and $P(R^{(*)}_{1/2})$. As we discuss in Section 4, this approach is sufficient to recover the overall distribution $P(Y)$ for a given sample. However, this does not guarantee, a priori, that the joint distributions are well reproduced. Therefore, we proceed to predict jointly pairs of properties, namely $P(M_*, g - i)$, $P(M_*, \mathrm{sSFR})$, $P(g - i, \mathrm{sSFR})$, and $P(R^{(*)}_{1/2}, M_*)$. Our strategy is similar to the univariate $P(Y)$ case: we make a grid in the $\{Y_1, Y_2\}$ subspace in such a way that the output corresponds to pixels in this grid. Although in this paper we restrict ourselves to only two galaxy properties when predicting joint distributions, a similar approach could be used, in principle, to characterize galaxies and define populations using an arbitrary number of properties. This generalization will be implemented in an upcoming paper.

Unless otherwise stated, for all the results shown here we set $K = 50$ classes for each one of the central galaxy properties, in equally spaced bins. For stellar mass, for example, this corresponds to bins of 0.085 dex. We must draw attention to the fact that this choice of binning is arbitrary. We have tried different numbers of bins, finding similar results in terms of the recovery of the distributions. Note that more refined versions of NNs that output distributions without binning the properties, and thus keeping it as a regression problem, already exist in the literature. In the context of photo-z estimation, Lima et al. (2022), for example, compares different types of NNs that return distributions, such as Mixture Density Networks (Bishop 1994), Bayesian NNs, and also NNs following a similar strategy as in this work, with a binning classification scheme. Ho et al. (2021) estimate the probability distribution of the dynamical mass of galaxy clusters and also compare several types of NNs, including a classifier that is similar to our $\mathrm{NN_{class}}$. In the context of the halo–galaxy connection, Stiskalek et al. (2022) model the stellar-to-halo mass relation scatter with a Gaussian distribution and train an ensemble of NNs that predicts the mean and standard deviation. We found the binned classification to be a simpler approach that works as a proof of concept. A more careful exploration of alternative methods is left as future refinements.

Throughout the analysis, we compare our $\mathrm{NN_{class}}$ method with the deterministic models developed by Santi et al. (2022), which we use as our baseline. In that work, several ML models are combined to return a final, consensus output for the same galaxy properties described in Section 2. The two consensus estimators are built from either the 'Raw' models, which were trained with the original TNG300 sample, or the 'SMOGN' models, which were trained using a data-augmented version of that data set. The SMOGN models were developed because of the difficulty for Raw models to recover the least frequent values of galaxy properties – i.e. to reproduce the tails of the distributions. The SMOGN data augmentation technique is a strategy to handle imbalanced data sets, whereby additional objects are artificially introduced in the training sample in order to force the

[1] In an NN, the model's parameters are the weights to be learned automatically, while the hyperparameters are the number of layers, neurons, number of epochs, etc., which are often chosen manually.

machine to give more importance to less represented objects (Kunz 2019).

The specifications of NN_{class} are described as follows. We use the categorical cross-entropy loss function and the ADAM optimizer (Kingma & Ba 2014) to train the networks. The architecture may change depending on the galaxy properties to be predicted. In general, our developed networks have a single intermediate layer, with a number of neurons that typically depends on whether the output is a univariate or a joint distribution. We use the L2 regularization, which applies a penalty proportional to the square of the model's weights. The number of epochs (iterations) is constrained with an early-stopping criterion based on the validation set loss. In the intermediate layers, we used the rectified linear unit as activation function, while in the output layer we use the Softmax function, which is similar to the Sigmoid function, but it normalizes the output in such a way that the scores of the K classes add up to one. In this way, the NN_{class} output works as a proxy for a probability in bins of galaxy properties.

4 RESULTS

Fig. 1 shows the distributions of the galaxies in the test set. The first column is the truth table, the TNG300 catalogue. The second column is the NN_{class} prediction of univariate distributions, i.e. galaxy properties predicted independently. With the univariate distributions, we can compute the joint distributions as $P(Y_1)\otimes P(Y_2)$, which are shown in the heatmap diagrams. The third column is the NN_{class} prediction for the joint distributions $P(Y_1, Y_2)$, which can be integrated to recover the univariate distributions $P(Y)$ shown in the marginal plots from the third column, i.e.

$$P(Y_i) = \int P(Y_i, Y_j)\mathrm{d}Y_j. \quad (1)$$

The univariate distributions predicted by NN_{class}, shown in black solid lines in the second-column plots of Fig. 1, are in excellent agreement with the true distributions from TNG300, shown in grey shaded regions. They also reproduce fairly well the joint distributions $P(Y_1)\otimes P(Y_2)$ for most cases. The $P(g-i)\otimes P(\mathrm{sSFR})$ joint distribution, however, fails to reproduce the shape of the distribution for redder colours and lower sSFRs. According to this prediction, red galaxies could have virtually any value of sSFR, while what we actually observe in TNG300 is that as galaxies move from the blue to the red peak, their sSFRs decrease. This important feature is recovered when NN_{class} is trained to predict $P(g-i, \mathrm{sSFR})$ jointly (third column in Fig. 1).

The above result indicates that our input halo properties alone are unable to predict accurately the correlations between colour and sSFR. The model would need additional features in order to capture this relation. It is interesting, however, that we can overcome this limitation by predicting the joint distribution directly using only the presented halo properties. This exercise indicates that, in order to robustly assign galaxies to haloes, with all the properties consistently correlated, the properties should be predicted together. Note that, in principle, one could define galaxy populations based on as many parameters as wished. Therefore, in the most general case, we would have an N-dimensional distribution associated with each host halo.

As a complementary analysis, Fig. 2 shows two additional well-known relations in the context of the halo–galaxy connection: the stellar-to-halo mass relation, and the galaxy size–halo mass relation obtained with TNG300 and with $P(M_*)$ and $P(R^{(*)}_{1/2})$ predicted by NN_{class}.

Figs 1 and 2 allow for a visual inspection of the results. In order to quantify the similarity between the distributions, we have performed the Kolmogorov–Smirnov (KS) test (for more details, see Ivezić et al. 2014):

$$\text{KS test values: } \Delta = \max(|F_1 - F_2|), \quad (2)$$

where F_1 and F_2 are cumulative distributions. The results are shown in Table 1. For comparison, we also show the values obtained with our baseline models, Raw and SMOGN, from Santi et al. (2022). Once again, we see that for most cases the independent prediction of univariate distributions reproduces fairly well the joint distributions, except for colour and sSFR. In all cases, NN_{class} provides significantly lower values as compared to Raw and SMOGN.

So far, we have focused on the combined distributions for the entire test sample. We now turn our attention to individual objects and the probability distributions that our ML machinery predicts for them. In particular, Fig. 3 displays, in a similar format to that of Fig. 1, some examples of the joint probability distribution $P(M_*, g-i)$ for three illustrative cases: a red object, a blue object, and an object lying at the so-called green valley region (from left to right). In each panel, the host halo mass is specified on the top, whereas the true TNG300 values of stellar mass and colour are shown as the dashed lines. As a reference, we also include in the marginal plots the distributions of the objects in the test set within a bin of ± 0.1 in halo mass around the values indicated on the top of the plots.

The first thing to notice from Fig. 3 is that the distributions are significantly narrower along the x-axis, as compared to the y-axis. This is of course expected, since stellar mass is the galaxy property that displays a tighter relation with the halo properties (particularly with halo mass), and therefore is the easiest to predict. It is also noteworthy that not all distributions can be well approximated by a Gaussian distribution. Some distributions are significantly skewed or, depending on halo mass, even bimodal, reflecting the well-known colour/sSFR bimodality of the galaxy population (e.g. Baldry et al. 2004).

The red galaxy on the left-hand panel shows very little scatter in colour. This is typically the case for red galaxies hosted by haloes with $\log_{10}(M_{\rm vir}[h^{-1}\,\mathrm{M}_\odot]) \gtrsim 12.5$. By visually inspecting Figs 1 and 2, we can get a sense as to why this happens: massive haloes are typically populated by massive galaxies, since the scatter in the stellar-to-halo mass relation is small. Massive galaxies are almost exclusively very red, which explains why the machine predicts a very narrow distribution of colours from the set of halo properties employed. The situation is very different for the blue galaxy featured in the middle panel. In this case, the predicted colour distribution is much broader than that for the red galaxy. Here, the host halo mass is much smaller, which implies a larger scatter in the stellar-to-halo mass relation. On top of that, blue galaxies intrinsically display a wide range of colours. All this uncertainty is captured by the machine in terms of a wider colour distribution.

Finally, the green-valley galaxy on the right-hand panel of Fig. 3 represents the most extreme case of the three, where the colour degeneracy produces a bimodal distribution. These objects are caught between two intrinsically different populations, i.e. the blue cloud and the red sequence. The analysis of individual distributions reveals that these objects are the ones that display a weaker relation with the properties of their host haloes (at least the ones analysed in this work). As discussed in Santi et al. (2022), these objects exemplify the most clear case where halo properties alone seem insufficient to predict the colour/sSFR, thus emphasizing the advantages of our probability-based methodology.

This probability distribution description on an individual object basis allows us to explore the dependence of galaxy properties on secondary halo properties at fixed halo mass (a dependence that

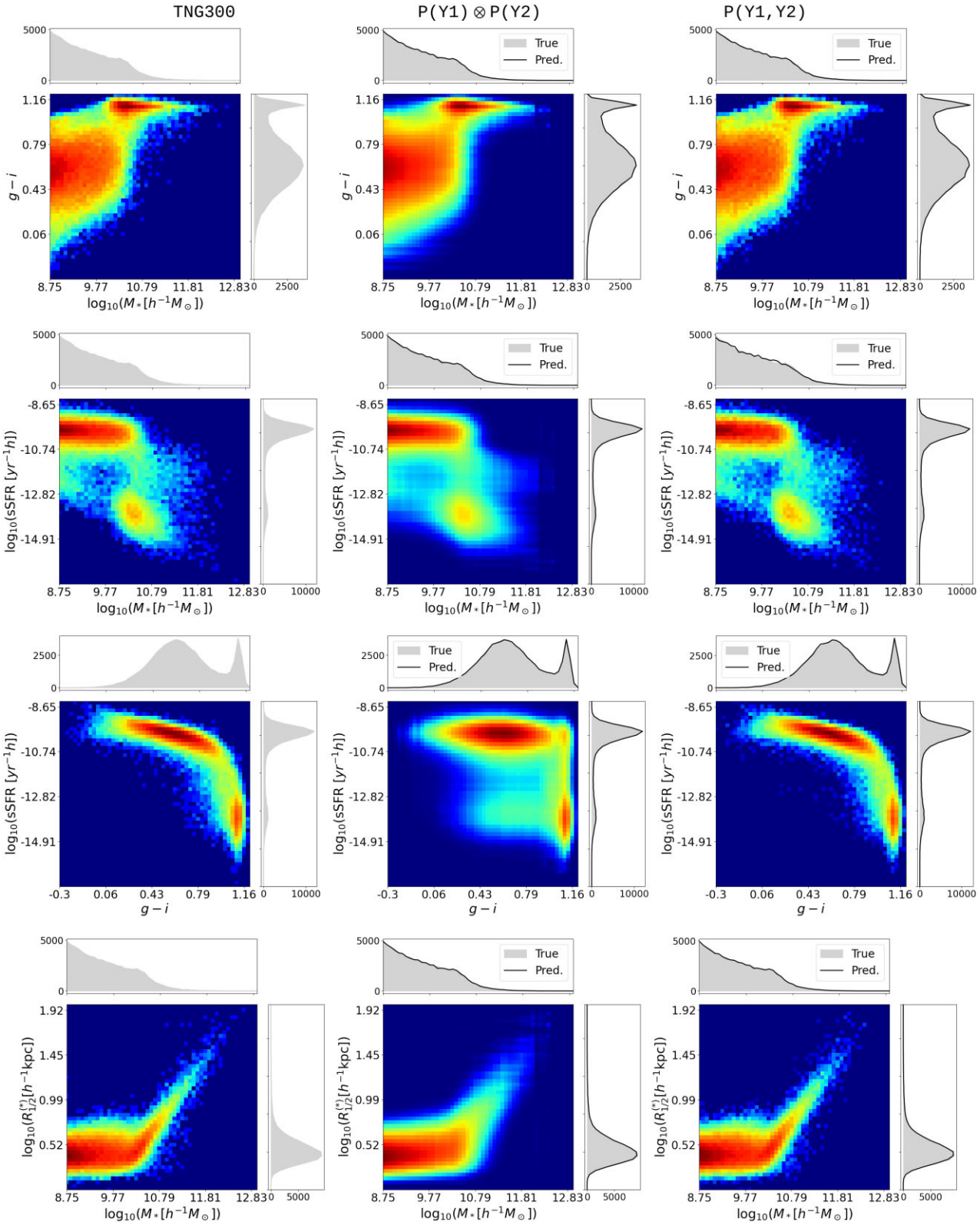

Figure 1. Distributions of galaxy properties. From top to bottom: colour versus stellar mass, sSFR versus stellar mass, sSFR versus colour, and radius versus stellar mass. The first column shows the true distributions from TNG300. The second column shows the distributions computed from the univariate distributions as predicted by NN_{class} – i.e. predicted independently from each other. The third column shows the joint distributions as predicted by NN_{class}. The grey shaded regions in the marginal plots correspond to the TNG300 distributions, while the black solid lines correspond to the NN_{class} predictions. The univariate distributions shown in the third column plots were computed by marginalizing the joint distributions.

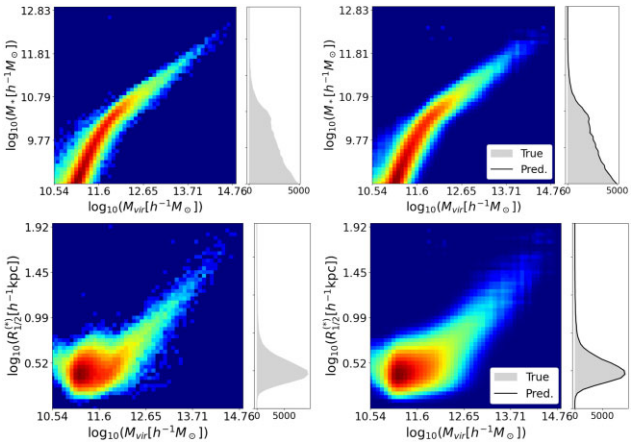

Figure 2. Stellar-to-halo mass relation (top) and galaxy size–halo mass relation (bottom) from the TNG300 catalogue (left) and from $\mathrm{NN_{class}}$ predictions (right).

is closely related to the so-called galaxy assembly bias effect; see e.g. Wechsler & Tinker 2018; Sato-Polito et al. 2019; Montero-Dorta et al. 2021b). In particular, we have analysed the dependence of $P(M_*,\, g-i)$ on halo age at fixed halo mass for green-valley objects. To this end, we selected objects in the test sample with predicted colour within the range $0.80 < g-i \leq 1.05$ and halo masses of $11.8 < \log_{10}(M_{\rm vir}[h^{-1}\,{\rm M}_\odot]) < 12.2$ (we have checked that choosing a narrower halo mass range would not alter our results significantly). This subset was subsequently split by halo age (taking the 15 and 85 per cent quantiles). For younger haloes, a stack of all distributions still reveals some bimodality in colour, albeit with a stronger preference for the blue peak. The predicted probability distribution for green-valley galaxies in older haloes is, conversely, much more skewed towards redder colours. The tail of the distribution for these objects still covers the green valley, which means that in some realizations these host haloes will be populated by a green-valley central galaxy (although the probability for this to happen is low). These results are reassuring in terms of the robustness of our methodology, demonstrating that our probability description is capable of capturing secondary halo dependences.

5 POWER SPECTRUM

With the help of the method presented in this work, we have greater flexibility to define different tracers based on galaxy properties. In this section, we explore the performance of $\mathrm{NN_{class}}$ in terms of the accuracy with which we can reproduce the power spectra of those tracers. We compute spectra for tracers in the test set, using the PYTHON package NBODYKIT (Hand et al. 2018). For the truth TNG300 catalogue, we use the positions of the central galaxies, but for the predictions we use the positions of the host haloes. Once again, we compare $\mathrm{NN_{class}}$ with the baseline models from Santi et al. (2022). As a complementary analysis, in Appendix B we compare the power spectra of tracers defined according to the same criteria of that previous work, which are based on individual galaxy properties.

Since TNG300 is a single box, the uncertainties of the spectrum on each bandpower k_i, for each tracer α, are computed according to the theoretical (Gaussian) covariance (Feldman, Kaiser & Peacock 1994), i.e.

$$\frac{\sigma^2_{\alpha,i}}{P^2_{\alpha,i}} = \frac{2}{V\tilde{V}}\left(\frac{1+\bar{n}_\alpha P_{\alpha,i}}{\bar{n}_\alpha P_{\alpha,i}}\right)^2, \tag{3}$$

with $\tilde{V} = 4\pi k_i^2 \Delta k/(2\pi)^3$, and the residuals are defined as

$$\frac{\left(P^{\rm pred}_{\alpha,i} - P^{\rm TNG300}_{\alpha,i}\right)^2}{\sigma^2_{\alpha,i}}. \tag{4}$$

Our choice of tracers is driven by the fact that the target selection in galaxy surveys often relies on the analysis of colour–magnitude diagrams (see e.g. Eisenstein et al. 2001, 2011; Zhou et al. 2020). One of the most common ways to define galaxy populations is in terms of the red sequence and the blue cloud, which can also be clearly distinguished in the colour–stellar mass diagram, as shown in Fig. 1. They are two distinct populations with different biases, hence their interest for studies of large-scale structure.

In a similar fashion, we defined seven tracers ($\alpha = 1, \ldots, 7$) based on the colour–stellar mass diagram, $P(M_*,\, g-i)$. We split red galaxies ($g-i > 1.05$) into lower ($\alpha = 1$) and higher ($\alpha = 2$) stellar masses. Conversely, 'green-valley' galaxies (defined as $0.80 < g-i \leq 1.05$) are split into three mass bins, leading to populations $\alpha = 3$, 4, 5. Finally, blue galaxies ($g-i \leq 0.8$) are separated into lower ($\alpha = 6$) and higher ($\alpha = 7$) stellar mass bins. This selection is outlined in Table 2, and it is represented in the lower right corner of Fig. 4.

An interesting feature of the probabilistic approach is that each galaxy is generated through a realization of a probability distribution spreading over many bins. As a consequence, we can build many catalogues of central galaxy properties by drawing values y_1, y_2 from $P(Y_1, Y_2)$. We have performed $r = 42$ realizations of $P(M_*,\, g-i)$, leading to as many values of M_* and $g-i$ for each halo. We then compute the spectrum of each of these samples, and from that the mean and variance of the spectra. For the mean spectrum $\bar{P}_{\alpha,i}$, we compute the uncertainties according to equation (3).

Fig. 4 shows the power spectra and residuals of the seven tracers defined in terms of $P(M_*,\, g-i)$ (see Table 2). Tracers $\alpha = 3$, 4 are relatively rare; hence, their corresponding regions in colour–stellar mass space are poorly populated by single-point estimators. Therefore, a model that predicts galaxies in these regimes improves the quality of the fit considerably – i.e. it reduces χ^2. We had already seen an improvement with the SMOGN models, which better recover this region as compared to the Raw models, but with $\mathrm{NN_{class}}$ this improvement is even more pronounced. There are only a few $\alpha = 5$ galaxies in TNG300, which makes this population very sparse. In

Table 1. KS test values for univariate (1D) and joint (2D) distributions computed with the NNs and the baseline models.

1D KS	$P(Y)$	Raw	SMOGN	2D KS	$P(Y_1)\otimes P(Y_2)$	$P(Y_1, Y_2)$	Raw	SMOGN
$P(M_*)$	0.002	0.064	0.064	$P(M_*,\, g-i)$	0.010	0.005	0.183	0.163
$P(g-i)$	0.004	0.181	0.116	$P(M_*,\, \mathrm{sSFR})$	0.012	0.009	0.253	0.209
$P(\mathrm{sSFR})$	0.004	0.213	0.168	$P(g-i,\, \mathrm{sSFR})$	0.110	0.009	0.266	0.176
$P(R^{(*)}_{1/2})$	0.009	0.217	0.110	$P(M_*,\, R^{(*)}_{1/2})$	0.015	0.007	0.217	0.150
	–	–	–	$P(M_{\rm vir},\, M_*)$	0.008	–	0.064	0.064
	–	–	–	$P(M_{\rm vir},\, R^{(*)}_{1/2})$	0.012	–	0.217	0.110

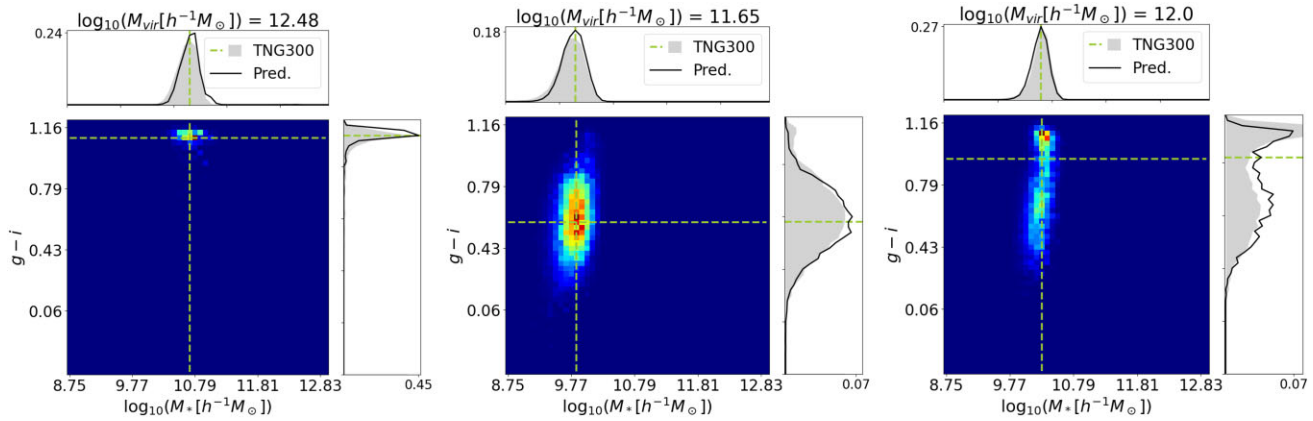

Figure 3. $P(M_*, g-i)$ for individual objects predicted by $\mathrm{NN_{class}}$. The dashed green lines show the true values for stellar mass and colour from TNG300. The shaded regions in the marginal plots are the distributions of objects with similar halo mass as indicated on the top of the corresponding panel.

Table 2. Criteria for splitting central galaxies by stellar mass and colour, in order to define the tracers used in the power spectrum analysis.

Tracer	$\log(M_*[h^{-1}\,\mathrm{M}_\odot])$	$g-i$	# Objects
$\alpha=1$	(9.5, 10.5]	>1.05	4073
$\alpha=2$	>10.5	>1.05	5207
$\alpha=3$	≤ 9.5	(0.80, 1.05]	4786
$\alpha=4$	(9.5, 10.5]	(0.80, 1.05]	5950
$\alpha=5$	>10.5	(0.80, 1.05]	1267
$\alpha=6$	≤ 9.5	≤ 0.80	29 695
$\alpha=7$	(9.5, 10.5]	≤ 0.80	18 432

particular, it has the largest variance over realizations. Conversely, all models are equally good at reproducing the power spectra of tracer populations closer to the peaks of the probability distributions: for $\alpha = 1$, 2, 6, and 7, χ^2 is comparable between all models.

As discussed earlier, we are able to draw multiple samples from the probabilities predicted by $\mathrm{NN_{class}}$. Each realization leads to slightly different power spectra, as can be seen in Fig. 4. By computing the variance of the multiple $P(k)$, we can assess the uncertainties due to the intrinsic stochasticity in the halo–galaxy connection. Fig. 5 compares the relative errors $\sigma^2/P^2_{\mathrm{TNG300}}(k)$ computed using σ^2_{CV}, from equation (3) (which encodes the uncertainty due to CV), with $\sigma^2_{\mathrm{NN_{class}}}$, which encodes the statistical uncertainties in the halo–galaxy connection estimated with $\mathrm{NN_{class}}$. As we already saw in Fig. 4, the CV error bars are typically larger than the scatter in the power spectra due to the multiple realizations of the $\mathrm{NN_{class}}$ probabilities. The contribution of $\sigma^2_{\mathrm{NN_{class}}}$ seems more relevant for the tracer population 5, which is very sparse. However, for all tracers σ^2_{CV} decreases for smaller scales (due to the Fourier bin volume), while $\sigma_{\mathrm{NN_{class}}}$ remains approximately constant. Therefore, the relative contribution of $\sigma_{\mathrm{NN_{class}}}$ for the total error budget of the power spectra appears to become more important at smaller scales.

Even though we see no evidence of a bias associated with this additional source of statistical uncertainties, the stochastic nature of the relationship between galaxies and their haloes may present further challenges for multitracer analyses of LSS (McDonald & Seljak 2009; Seljak 2009). The advantages of the multitracer technique are reliant upon the partial cancellation of CV that results from clustering measurements from different galaxy types that are assumed to reflect the same underlying DM density field (in that respect, see also Abramo & Leonard 2013; Abramo, Secco & Loureiro 2016). The 'stochastic bias' associated with the nature of the galaxy–halo connection can dilute some of the expected CV cancellation. However, that stochastic component seems to affect mostly the power spectra on small scales, where non-linear effects already limit our ability to employ the multitracer technique effectively (see e.g. Montero-Dorta et al. 2020a).

6 DISCUSSION AND CONCLUSIONS

Although there is an obvious relation between the baryonic and DM components of haloes, there is also mounting evidence that the properties of haloes alone are insufficient to reproduce the properties of galaxies, since the latter are shaped by a variety of galaxy formation processes. On the other hand, ML regression models are traditionally designed to reproduce single-value statistics, and thus are ill-equipped to encode the intrinsic scatter in the halo–galaxy connection. Building on the recent work of Santi et al. (2022), here we use the TNG300 hydrodynamical simulation in combination with NNs to map the connection between the properties of central galaxies and the properties of their hosting haloes. As in the aforementioned work, NNs are trained to reproduce the stellar mass, $g-i$ colour, sSFR, and radius of TNG300 galaxies based on a set of halo/environmental properties that include virial mass, concentration, formation redshift, spin, and overdensity (computed over scales of $3\,h^{-1}$ Mpc). In order to alleviate the deficiencies of ML deterministic regression models, we have tested a different approach for the first time in the context of the halo–galaxy connection. The NNs are now trained to predict probability distributions instead of single-value statistics by means of a binning classification scheme. In essence, the distributions of galaxy properties are split into K narrow bins so that the NNs can associate a score to each of the K classes. This is performed in such a way that the output can be used as a proxy for the probability distributions of the central galaxy properties.

We have shown that this approach is in fact capable of producing bivariate distributions of galaxy properties, i.e. $P(Y_1, Y_2)$, in outstanding agreement with those from TNG300 (here, $\{Y_1, Y_2\}$ is any pair of galaxy properties). These joint distributions can be compared with the product of the two 1D (disjoint) distributions, $P(Y_1)$ and $P(Y_2)$. For the joint distributions, we employ 2D $K \times K$ grids, representing the binned galaxy properties, where each pixel on the grid corresponds to a class. In either case, predicting the probability distributions yields significantly better results compared with the deterministic approach (Santi et al. 2022), as both a visual inspection and the 2D KS test reveal. As a reference, our 2D KS

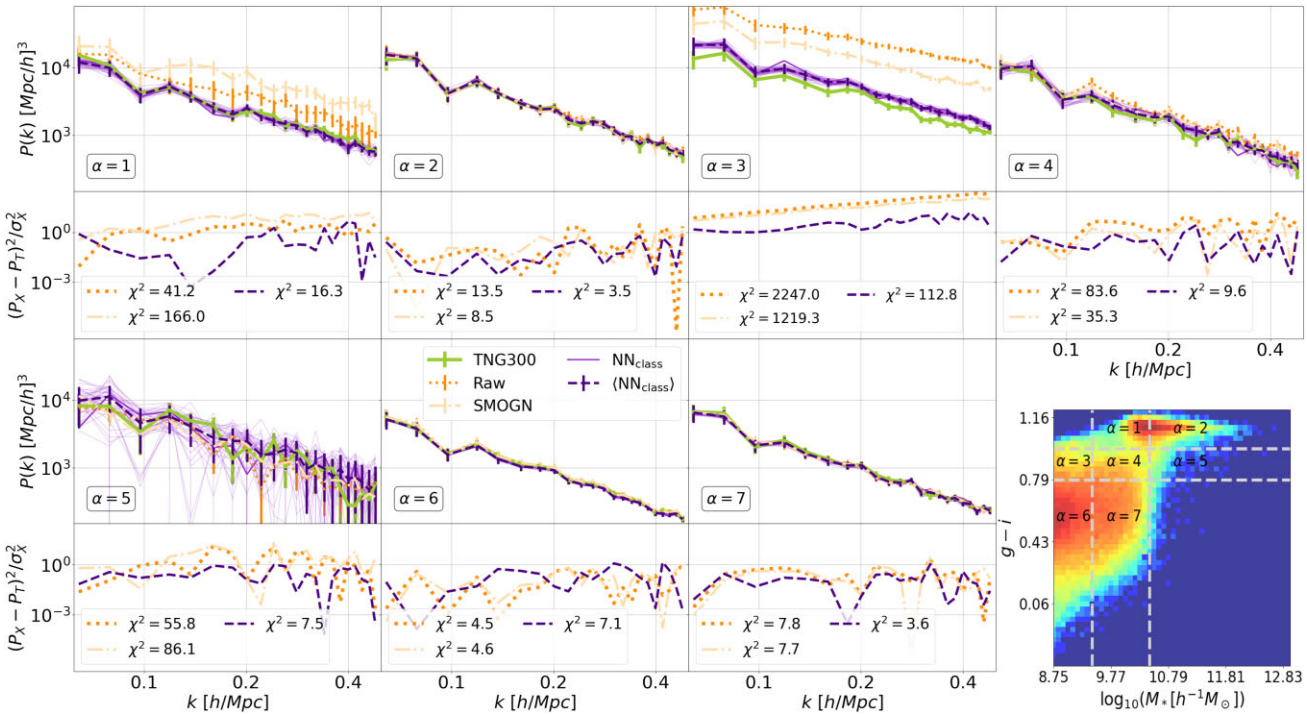

Figure 4. Power spectra and residuals for seven tracers selected on the basis of the colour–stellar mass diagram (bottom right panel). The green solid lines correspond to TNG300, while the light purple solid lines correspond to spectra from $r = 42$ samples drawn from the probabilities predicted by $\mathrm{NN_{class}}$. The dark purple, thick dashed lines correspond to the mean of those realizations. The baseline models are shown in orange: darker dotted lines correspond to the Raw model and lighter dotted–dashed lines correspond to the SMOGN model.

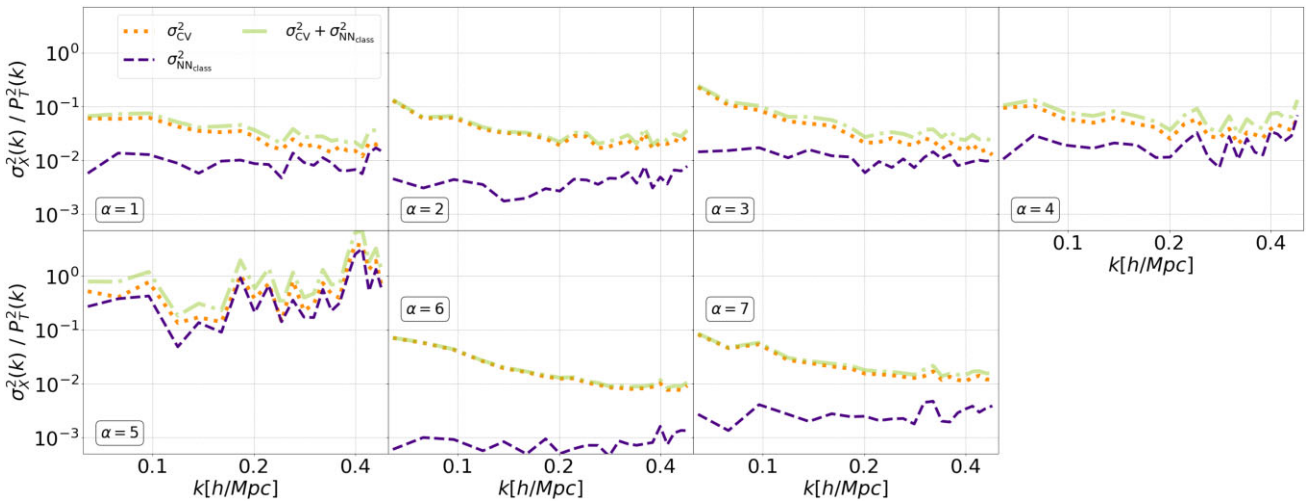

Figure 5. Relative error for seven tracers selected based on the colour–stellar mass diagram. The variances are normalized by the TNG300 spectrum $P_T(k)$ of each tracer α. Orange dotted lines correspond to the relative error computed with equation (3), purple dashed lines correspond to the relative error computed with $\mathrm{NN_{class}}$, and green solid lines correspond to the total relative error.

test for the joint distributions $P(Y_1, Y_2)$ yields performance results that are better by factors of 10–30 as compared to those reported in Santi et al. (2022). We have also checked that predicting galaxy pairs directly is particularly advantageous for the colour–sSFR joint distribution, where the stellar mass, the main anchor of the halo–galaxy connection, is not included.

An important subproduct of our analysis is the joint distributions for individual galaxies, which can be understood as the probability distributions that an object occupies a given location on the 2D diagrams for the galaxy properties. As an illustration, we have analysed the individual joint distributions of stellar mass and colour, and verified that the distributions for red galaxies, particularly for those that live in massive haloes, are significantly more concentrated than those for blue and green-valley objects. For the latter, the individual distributions can even become bimodal in certain halo mass ranges. This is a robustness test for our methodology, showing that these individual distributions are good estimators of the uncertainty that results from attempting to predict galaxy properties from incomplete

(halo) information. The main advantages of our method are that it provides a more complete description of the interconnected relations between galaxy and halo properties, as compared to single-value ML approaches, and that it can be easily implemented in cosmological and galaxy formation models.

As an application of our methodology, we have shown that our predictions are capable of reproducing with unprecedented precision the power spectra of any given number of tracers defined based on the colour–stellar mass diagram (we showed results for seven tracers, but the analysis can be extended to more galaxy populations). We have also checked that the statistical uncertainty in our models (which can be obtained by sampling the distributions several times, creating multiple catalogues) is often small compared with the uncertainty that emanates from CV (particularly on large scales). In this sense, our method is clearly advantageous for cosmological studies employing a high number of tracers and/or underrepresented populations, as compared with the more traditional single-value approaches (see Santi et al. 2022, for comparison). These advantages can be exploited in the context of multitracer cosmological analyses, where clustering information from multiple galaxy population and redshift ranges is combined in order to reduce the uncertainties in the estimation of the power spectrum and thus the bias and cosmological parameters (e.g. Abramo & Leonard 2013; Abramo et al. 2016; Montero-Dorta et al. 2020a; Abramo, Ferri & Tashiro 2022).

One interesting application of our method is to paint galaxies on to haloes in DM-only simulations. As we have discussed in this work, when central galaxy properties are predicted jointly, their correlations are in agreement with those from hydrodynamical simulations. However, in order to extend our analysis to a higher number of dimensions, i.e. to predict joint distributions of three or more properties, or to extend the approach to satellite galaxies, it is necessary to optimize the discretization of the galaxy distributions. Presently, our method can become computationally inefficient for this purpose, as so far we are considering bins of equal size across the galaxy property diagrams. Moreover, by categorizing the galaxy properties and treating the bins as individual classes, one may lose the information that nearby bins are more similar than distant ones. Therefore, using additional metrics and trying alternative loss functions to quantify the difference between the distributions may be helpful, especially when handling higher dimensions (see e.g. Stiskalek et al. 2022). Follow-up work will be devoted to improving this methodology in order to generalize the analysis.

Finally, the flexibility of our method in terms of reproducing both the clustering and internal properties of virtually any galaxy population with precision may have applications in the context of galaxy assembly bias, i.e. the secondary dependences of galaxy clustering at fixed halo mass (see e.g. Lin et al. 2016; Montero-Dorta et al. 2017; Zu et al. 2017; Niemiec et al. 2018; Zentner et al. 2019; Obuljen, Percival & Dalal 2020; Salcedo et al. 2022; Wang et al. 2022). In particular, recent attempts to probe the effect with observations (Salcedo et al. 2022; Wang et al. 2022) have employed forward-modelling techniques using specifically generated galaxy mocks. Our methodology and statistical descriptions seem ideal to be incorporated into these models.

ACKNOWLEDGEMENTS

NVNR, NSMS, and LRA acknowledge Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support. NSMS acknowledges financial support from FAPESP, grants 2019/13108-0 and 2022/03589-4. ADMD thanks FONDECYT for financial support through the FONDECYT Regular 2021 grant 1210612.

DATA AVAILABILITY

The material presented in this paper is available in the repository: https://github.com/nvillanova/central-galaxies-joint-distributions.

REFERENCES

Abramo L. R., Leonard K. E., 2013, MNRAS, 432, 318
Abramo L. R., Secco L. F., Loureiro A., 2016, MNRAS, 455, 3871
Abramo L. R., Ferri J. V. D., Tashiro I. L., 2022, J. Cosmol. Astropart. Phys., 2022, 013
Agarwal S., Davé R., Bassett B. A., 2018, MNRAS, 478, 3410
Artale M. C., Zehavi I., Contreras S., Norberg P., 2018, MNRAS, 480, 3978
Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Becker M. R., 2015, preprint (arXiv:1507.03605)
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bishop C., 1994, Mixture Density Networks. Aston Univ., Birmingham
Bose S., Eisenstein D. J., Hernquist L., Pillepich A., Nelson D., Marinacci F., Springel V., Vogelsberger M., 2019, MNRAS, 490, 2192
Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001, ApJ, 555, 240
Buser R., 1978, A&A, 62, 411
Calderon V. F., Berlind A. A., 2019, MNRAS, 490, 2367
Chittenden H. G., Tojeiro R., 2023, MNRAS, 518, 5670
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Contreras S., Angulo R., Zennaro M., 2020a, MNRAS, 504, 5205
Contreras S., Angulo R., Zennaro M., 2020b, MNRAS, 508, 175
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
de Andres D., Yepes G., Sembolini F., Martí nez-Muñoz G., Cui W., Robledo F., Chuang C.-H., Rasia E., 2022, MNRAS, 518, 111
de Santi N. S. M., Rodrigues N. V. N., Montero-Dorta A. D., Abramo L. R., Tucci B., Artale M. C., 2022, MNRAS, 514, 2463
Delgado A. M., Wadekar D., Hadzhiyska B., Bose S., Hernquist L., Ho S., 2022, MNRAS, 515, 2733
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, 497
Eisenstein D. J. et al., 2001, AJ, 122, 2267
Eisenstein D. J. et al., 2011, AJ, 142, 72
Favole G. et al., 2016, MNRAS, 461, 3421
Favole G., Montero-Dorta A. D., Artale M. C., Contreras S., Zehavi I., Xu X., 2022, MNRAS, 509, 1614
Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23
Genel S. et al., 2014, MNRAS, 445, 175
Gu M. et al., 2020, preprint (arXiv:2010.04166)
Guo Q., White S., Angulo R., Henriques B., Lemson G., Boylan-Kolchin M., Thomas P., Short C., 2013, MNRAS, 428, 1351
Guo H. et al., 2016, MNRAS, 459, 3040
Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., 2021, MNRAS, 501, 1603
Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., Spergel D. N., 2020b, MNRAS, 493, 5506
Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., 2021, MNRAS, 501, 1603
Hand N., Feng Y., Beutler F., Li Y., Modi C., Seljak U., Slepian Z., 2018, AJ, 156, 160
Ho M., Farahi A., Rau M. M., Trac H., 2021, ApJ, 908, 204
Ivezić Ž., Connolly A., VanderPlas J., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton Series in Modern Observational

Astronomy. Princeton Univ. Press, Princeton, NJ, available at: https://books.google.com.br/books?id=h2eYDwAAQBAJ
Jespersen C. K., Cranmer M., Melchior P., Ho S., Somerville R. S., Gabrielpillai A., 2022, ApJ, 941, 7
Jo Y., Kim J.-h., 2019, MNRAS, 489, 3565
Kamdar H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 457, 1162
Kasmanoff N., Villaescusa-Navarro F., Tinker J., Ho S., 2020, preprint (arXiv:2012.00186)
Kingma D. P., Ba J., 2014, Adam: A Method for Stochastic Optimization, preprint (arXiv:1412.6980)
Kunz N., 2019, SMOGN, available at: https://github.com/nickkunz/smogn
Lima E. et al., 2022, Astron. Comput., 38, 100510
Lin Y.-T., Mandelbaum R., Huang Y.-H., Huang H.-J., Dalal N., Diemer B., Jian H.-Y., Kravtsov A., 2016, ApJ, 819, 119
Lovell C. C., Wilkins S. M., Thomas P. A., Schaller M., Baugh C. M., Fabbian G., Bahé Y., 2022, MNRAS, 509, 5046
McDonald P., Seljak U., 2009, JCAP, 2009, 007
McGibbon R., Khochfar S., 2022, MNRAS, 513, 5423
Man Z.-Y., Peng Y.-J., Shi J.-J., Kon X., Zhang C.-P., Dou J., Guo K.-X., 2019, ApJ, 881, 74
Marinacci F. et al., 2018, MNRAS, 480, 5113
Montero-Dorta A. D. et al., 2017, ApJ, 848, L2
Montero-Dorta A. D., Abramo L. R., Granett B. R., de la Torre S., Guzzo L., 2020a, MNRAS, 493, 5257
Montero-Dorta A. D. et al., 2020b, MNRAS, 496, 1182
Montero-Dorta A. D., Artale M. C., Abramo L. R., Tucci B., 2021a, MNRAS, 504, 4568
Montero-Dorta A. D., Chaves-Montero J., Artale M. C., Favole G., 2021b, MNRAS, 508, 940
Moster B. P., Naab T., White S. D. M., 2018, MNRAS, 477, 1822
Naab T., Ostriker J. P., 2017, ARA&A, 55, 59
Naiman J. P. et al., 2018, MNRAS, 477, 1206
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Nelson D. et al., 2018, MNRAS, 475, 624
Nelson D. et al., 2019, Comput. Astrophy. Cosmol., 6
Niemiec A. et al., 2018, MNRAS, 477, L1
Obuljen A., Percival W. J., Dalal N., 2020, J. Cosmol. Astropart. Phys., 2020, 058
Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26
Pillepich A. et al., 2018a, MNRAS, 473, 4077
Pillepich A. et al., 2018b, MNRAS, 475, 648
Planck Collaboration XIII, 2016, A&A, 594, A13
Ramanah D. K., Wojtak R., Ansari Z., Gall C., Hjorth J., 2020, MNRAS, 499, 1985
Sadeh I., Abdalla F. B., Lahav O., 2016, Publ. Astron. Soc. Pac., 128, 104502
Salcedo A. N. et al., 2022, Sci. China Phys. Mech. Astron., 65, 109811
Sato-Polito G., Montero-Dorta A. D., Abramo L. R., Prada F., Klypin A., 2019, MNRAS, 487, 1570
Seljak U., 2009, Phys. Rev. Lett., 102, 021302
Shao H. et al., 2021, ApJ, 927, 85
Shi J. et al., 2020, ApJ, 893, 139
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Springel V., 2010, MNRAS, 401, 791
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Springel V. et al., 2018, MNRAS, 475, 676
Stiskalek R., Bartlett D. J., Desmond H., Anbajagane D., 2022, MNRAS, 514, 4026
Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, ApJ, 742, 16
Villaescusa-Navarro F. et al., 2021, ApJ, 915, 71
Villaescusa-Navarro F. et al., 2022, ApJ, 265, 54
Vogelsberger M. et al., 2014a, MNRAS, 444, 1518
Vogelsberger M. et al., 2014b, Nature, 509, 177
Wang K., Mao Y.-Y., Zentner A. R., Guo H., Lange J. U., van den Bosch F. C., Mezini L., 2022, MNRAS, 516, 4003
Wechsler R. H., Tinker J. L., 2018, Annu. Rev. Astron. Astrophys., 56, 435
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
Xu X., Zehavi I., Contreras S., 2021, MNRAS, 502, 3242
Yip J. H. T. et al., 2019, preprint (arXiv:1910.07813)
Zehavi I. et al., 2005, ApJ, 621, 22
Zehavi I., Contreras S., Padilla N., Smith N. J., Baugh C. M., Norberg P., 2018, ApJ, 853, 84
Zentner A. R., Hearin A., van den Bosch F. C., Lange J. U., Villarreal A., 2019, MNRAS, 485, 1196
Zhang X., Wang Y., Zhang W., Sun Y., He S., Contardo G., Villaescusa-Navarro F., Ho S., 2019, preprint (arXiv:1902.05965)
Zhou R. et al., 2020, Res. Notes AAS, 4, 181
Zu Y., Mandelbaum R., Simet M., Rozo E., Rykoff E. S., 2017, MNRAS, 470, 551

APPENDIX A: SINGLE VALUE ESTIMATION

In this appendix, we discuss the results of the $\mathrm{NN_{class}}$ in terms of single-point estimation scores. Throughout the paper, our analysis focuses on the performance in terms of how well we can recover the distributions. Since we do not have a single value associated with each data set instance, but a distribution, one can sample several times from this distribution in order to estimate the most probable value, and compute single-point estimation metrics with it. Once again, we take the average of $r = 42$ realizations of each predicted galaxy property and calculate the Pearson correlation coefficient (PCC) between the true and estimated values as

$$\mathrm{PCC} = \frac{\mathrm{cov}(\mathrm{y^{pred}}, \mathrm{y^{true}})}{\sigma_{\mathrm{y^{pred}}}\sigma_{\mathrm{y^{true}}}}. \tag{A1}$$

Fig. A1 shows the PCC score as a function of the number of realizations and also the values of the baseline models for the four galaxy properties. In this exercise, we sample from univariate distributions $P(Y)$ instead of joint distributions. $\mathrm{NN_{class}}$ provides results comparable to the single-point estimators Raw and SMOGN as the number of realizations increases, which indicates that $\mathrm{NN_{class}}$ are also good maximum likelihood estimators.

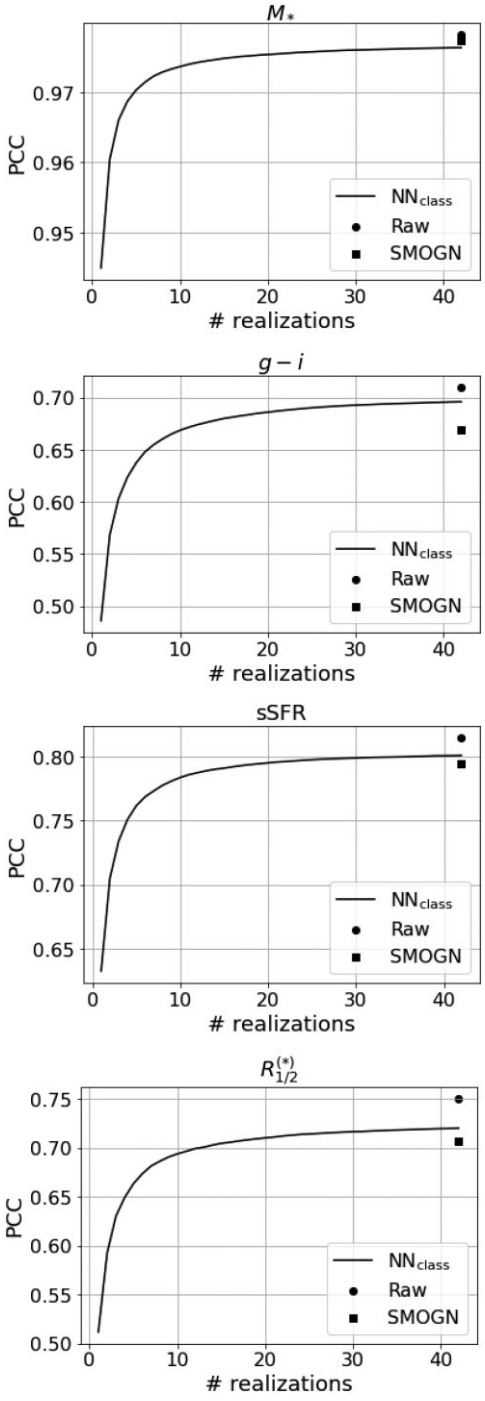

Figure A1. PCC of $\mathrm{NN_{class}}$ (solid lines) as a function of the number of realizations of $P(Y)$, $Y = M_*$, $g-i$, sSFR, and $R^{(*)}_{1/2}$. The PCC values of the baseline models Raw and SMOGN are shown as dotted and squared markers, respectively.

APPENDIX B: POWER SPECTRUM: ADDITIONAL RESULTS

In this appendix, we show the power spectrum of the tracers defined in Santi et al. (2022) (see Fig. B1). The galaxies are divided into two populations based on each of the properties. The univariate distributions can be obtained from different joint distributions, by marginalizing them (see equation 1). Stellar mass can be obtained from $P(M_*, g-i)$, $P(M_*, \mathrm{sSFR})$, and $P(M_*, R^{(*)}_{1/2})$, colour can be obtained from $P(M_*,\ g-i)$ and $P(g-i, \mathrm{sSFR})$, sSFR can be obtained from $P(g-i, \mathrm{sSFR})$ and $P(M_*, \mathrm{sSFR})$, and the radius can be obtained from $P(M_*, R^{(*)}_{1/2})$. Once again, for $\mathrm{NN_{class}}$ we show $r = 42$ realizations as well as the mean of the spectra of all r samples. We see that for these tracers there is no clear advantage of the $\mathrm{NN_{class}}$ over the Raw model: in most cases, $\mathrm{NN_{class}}$ performs similar to the Raw models, although for sSFR the results for $\mathrm{NN_{class}}$ are slightly worse (which is not entirely unexpected, since sSFR is a particularly difficult property to predict based only on the halo properties that we take into account). Note that here we are computing the average of the spectra of many realizations of the predicted distributions, as in Fig. 4. In this way, we can explore the advantage of having a tool that recovers the complete range of possible values. In order to have a more straightforward comparison with the single-point estimators, one can compute the spectrum of the tracers defined based on the maximum likelihood values of galaxy properties, as in Appendix A.

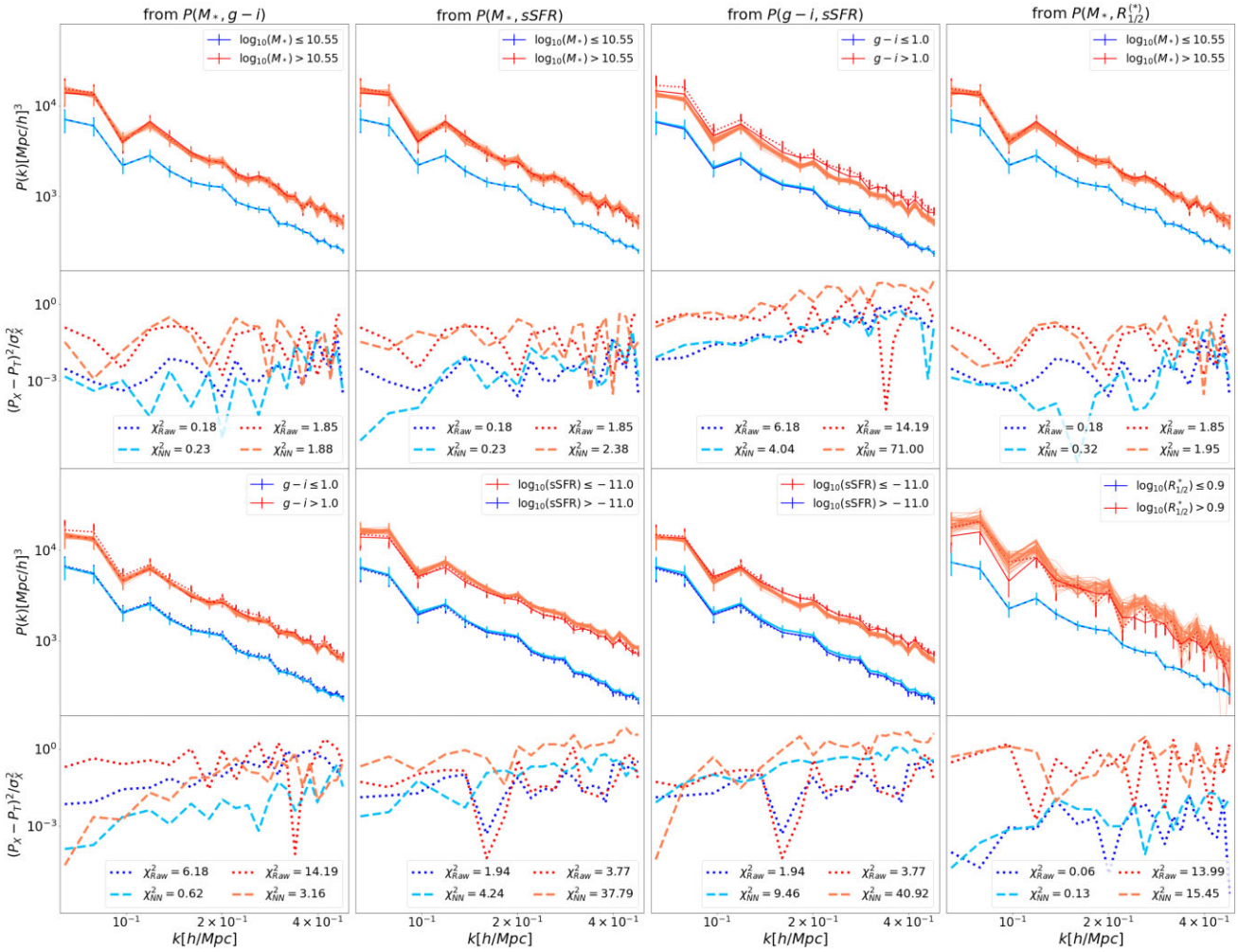

Figure B1. Power spectrum and residuals of two tracers defined by splitting each galaxy property. The higher bias tracers are shown in red, and the lower bias tracers are shown in blue. The properties are obtained by marginalizing the joint distributions and can thus be obtained with more than one distribution. The first column shows the results for stellar mass and colour obtained with $P(M_*, g-i)$. The second column shows the results for stellar mass and sSFR obtained with $P(M_*, \mathrm{sSFR})$. The third column shows the results for colour and sSFR obtained with $P(g-i, \mathrm{sSFR})$. The fourth column shows the results for stellar mass and radius obtained with $P(M_*, R^{(*)}_{1/2})$. The power spectrum of each $\mathrm{NN_{class}}$ realization is shown as solid lines. The mean $\mathrm{NN_{class}}$ spectra are shown as dashed lines and the Raw model spectra are shown as dotted lines.

This paper has been typeset from a TEX/LATEX file prepared by the author.