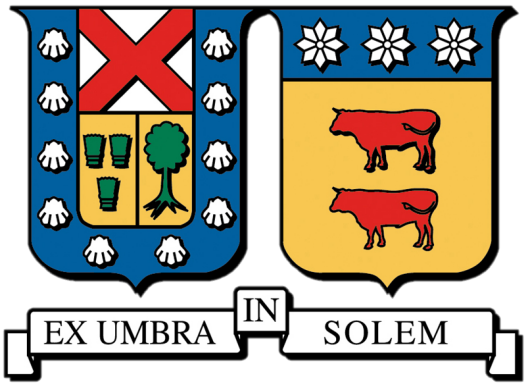
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE FISICA



Characterizing Galaxies Using Non-Parametric Methods Within the IllustrisTNG Cosmological Simulation Framework

DIEGO GERMÁN CORREA HERRERA

MEMORIA PARA OPTAR AL GRADO DE LICENCIADO EN ASTROFISICA

Guía: PhD. Yara Jaffé Ribbi

Co-Guía: PhD. Diego Pallero Astargo

Enero-2025

Resumen

Las galaxias son sistemas fundamentales para el estudio del cosmos, que están compuestas por gas, polvo, estrellas y materia oscura. Las galaxias se presentan en una variedad de morfologías, con diferentes estructuras y propiedades físicas que encapsulan una historia de procesos físicos y evolutivos que han ocurrido desde hace miles de millones de años. Ahora, esto nos lleva que podemos interpretar los distintos tipos de galaxias como parte de una secuencia evolutiva, donde actúan tanto procesos internos como los posibles efectos del medio sobre la galaxia.

En este contexto, la clasificación de galaxias, lejos de ser un mero ejercicio descriptivo, resulta esencial para establecer un orden y estudiar los procesos evolutivos que moldean su estructura y propiedades. Sin embargo, a pesar de la importancia de los sistemas de clasificación, no existe un criterio que capture completamente la riqueza morfológica de las galaxias. Además, existen múltiples formas de caracterizar la morfología galáctica, y no siempre hay consenso entre ellas, lo que deja un margen considerable de incertidumbre.

Es por ello que el objetivo de este estudio es dar un primer paso, y comparar dos métodos no paramétricos de clasificación morfológica; Los índices CAS (Concentración, Asimetría y Suavidad) y Deep Learning (DL), utilizando la simulación cosmológica IllustrisTNG (específicamente TNG100), buscando identificar diferencias o similitudes en la clasificación de galaxias, según el método empleado y proporcionar una perspectiva más profunda sobre la importancia del método de clasificación.

Para realizar el estudio se establecieron dos muestras lo más puras posible para comparar ambos métodos de clasificación; La primera contiene galaxias cuya clasificación morfológica es determinada por el modelo de DL, con 9 veces más probabilidad que otras morfologías. Y la segunda, contiene galaxias que pertenecen a ciertos rangos estrechos de parámetros CAS, que fueron usados para asignarles un tipo morfológico.

Comenzamos explorando cómo las propiedades físicas varían para las clasificaciones morfológicas hechas con CAS y Deep Learning, estableciendo las bases para la comparación posterior. Luego, se realiza una comparación cruzada, analizando cómo se relacionan los criterios de ambos métodos, buscando coherencias o discrepancias. Y finalmente, comparamos directamente las muestras clasificadas por ambos métodos en los planos de propiedades fundamentales de galaxias, como

diagrama color (g-r) vs. masa estelar, sSFR (tasa específica de formación estelar, es decir, la tasa de formación estelar normalizada por la masa estelar) vs. masa estelar y gas fraction vs. masa estelar, revisando así también la coherencia de las morfologías asignadas.

El análisis reveló que ambos métodos son efectivos para separar galaxias de tipo temprano y tardío, con límites claros en color = 0.6 y $sSFR = 10^{-11}[Yr^{-1}]$, lo que concuerda con la literatura. Sin embargo, DL muestra limitaciones con galaxias de baja masa estelar ($M_\star \leq 10^{10.5}[M_\odot]$), cuyo origen, creemos, se debe a la resolución limitada de TNG100. CAS, aunque más detallado estructuralmente, puede excluir galaxias fuera de rangos estrictos, afectando la representatividad. Creemos que ambos métodos son complementarios: CAS supera las limitaciones de DL en galaxias elípticas enanas, cuya probabilidad asignada es baja por el ya mencionado tema de resolución, y DL funciona bien para tipos generales (temprano y tardío), pero le cuesta definir con seguridad tipos morfológicos más específicos como lenticular o espirales tardías. Asimismo, creemos que hay una relación complementaria entre la categoría "Late Spiral" en CAS y "Early Spiral" en DL, que podría aprovecharse para una clasificación de galaxias tardías, aunque se requieren estudios adicionales para obtener un sistema de clasificación más detallado.

Un caso particularmente interesante es la categoría "Early Spiral" en CAS, que agrupa galaxias con propiedades características de morfologías transitorias, como las lenticulares. Esto sugiere que renombrar esta categoría podría ayudar a reflejar con mayor precisión las características de las galaxias incluidas en ella. Ahora, dado que DL enfrenta dificultades para asignar altas probabilidades a morfologías intermedias, consideramos que la categoría "Early Spiral" de CAS podría ser utilizada como un criterio complementario a DL, proporcionando un enfoque más robusto de clasificación.

Como trabajo futuro, se propone incorporar parámetros adicionales, como los índices Gini y M20, el índice de Sérsic, elipticidad, y parámetros dinámicos como momento angular específico, para una clasificación más robusta y precisa. También sería valioso replicar este análisis en simulaciones de mayor resolución, como TNG50, incluyendo una clasificación visual por humanos, y con imágenes sintéticas unificadas para ambos métodos, lo cual aún no es posible, ya que estas últimas no existen como tal, sin embargo, se planea solventar el problema con colaboraciones en un futuro cercano. Finalmente, al efectivamente encontrar diferencias y similitudes, este trabajo resalta la importancia de combinar múltiples técnicas y parámetros para avanzar en la comprensión de la diversidad morfológica de las galaxias.

Abstract

Galaxies are fundamental systems for the study of the cosmos, composed of gas, dust, stars and dark matter. Galaxies come in a variety of morphologies, with different structures and physical properties that encapsulate a history of physical and evolutionary processes that have occurred over billions of years. Now, this leads us to interpret the different types of galaxies as part of an evolutionary sequence, where both internal processes and the possible effects of the environment on the galaxy are at work.

In this context, the classification of galaxies, far from being a mere descriptive exercise, is essential for establishing order and studying the evolutionary processes that shape their structure and properties. However, despite the importance of classification systems, there is no single criterion that fully captures the morphological richness of galaxies. Moreover, there are multiple ways to characterize galactic morphology, and there is not always a consensus among them, leaving considerable room for uncertainty.

This is why the aim of this study is to take a first step, and compare two non-parametric methods of morphological classification; the CAS (Concentration, Asymmetry and Smoothness) and Deep Learning (DL) indices, using the IllustrisTNG cosmological simulation (specifically TNG100), seeking to identify differences or similarities in the classification of galaxies, depending on the method used and to provide a deeper perspective on the importance of the classification method. The first contains galaxies whose morphological classification is determined by the DL model, with 9 times more probability than other morphologies. And the second contains galaxies that belong to certain narrow ranges of CAS parameters, which were used to assign them a morphological type.

We begin by exploring how the physical properties vary for the morphological classifications made with CAS and Deep Learning, laying the groundwork for further comparison. Then, we perform a cross-comparison, analyzing how the criteria of both methods relate to each other, looking for consistencies or discrepancies. And finally, we directly compare the samples classified by both methods in the fundamental galaxy property planes, such as color diagram (g-r) vs. stellar mass, sSFR (specific star formation rate, i.e. the star formation rate normalized by stellar mass) vs. stellar mass and gas fraction vs. stellar mass, thus also checking the consistency of the assigned morphologies.

The analysis revealed that both methods are effective in separating early- and late-type galaxies, with clear limits at color = 0.6 and $sSFR = 10^{-11}[Yr^{-1}]$, which is in agreement with the literature. However, DL shows limitations with low stellar mass galaxies ($M_\star \leq 10^{10.5}[M_\odot]$), whose origin, we believe, is due to the limited resolution of TNG100. CAS, although more structurally detailed, may exclude galaxies outside strict ranges, affecting the representativeness. We believe that the two methods are complementary: CAS overcomes the limitations of DL for dwarf elliptical galaxies, whose assigned probability is low due to the aforementioned resolution issue, and DL works well for general types (early and late), but find it difficult to define with certainty more specific morphological types such as lenticular or late spirals. We also believe that there is a complementary relationship between the 'Late Spiral' category in CAS and 'Early Spiral' in DL, which could be exploited for a classification of late galaxies, although further studies are needed to obtain a more detailed classification system.

A particularly interesting case is the 'Early Spiral' category in CAS, which groups galaxies with characteristic properties of transient morphologies, such as lenticular ones. This suggests that renaming this category could help to more accurately reflect the characteristics of the galaxies included in it. Now, given that DL faces difficulties in assigning high probabilities to intermediate morphologies, we consider that the CAS category 'Early Spiral' could be used as a complementary criterion to DL, providing a more robust classification approach.

As future work, we propose to incorporate additional parameters, such as the Gini and M20 indices, the Sérsic index, ellipticity, and dynamic parameters such as specific angular momentum, for a more robust and accurate classification. It would also be valuable to replicate this analysis in higher resolution simulations, such as TNG50, including a visual classification by humans, and with unified synthetic images for both methods, which is not yet possible, as the latter do not exist as such, however, it is planned to solve the problem with collaborations in the near future. Finally, by effectively finding differences and similarities, this work highlights the importance of combining multiple techniques and parameters to advance the understanding of the morphological diversity of galaxies.

Agradecimientos

A mis compañeros de generación, ahora preciados amigos, cuya compañía, apoyo e inspiración fue vital en el transcurso de la carrera. A mis profesores, quienes también han sido una fuente de inspiración y sabiduría. Y, por último, pero no menos importante, a mi familia y amigos externos a la universidad, quienes, a pesar de no comprender del todo a qué me dedico, siempre me han brindado su confianza y apoyo incondicional.

Esta investigación fue financiada por Proyecto FONDECYT REGULAR #1241426 – etapa 2024, cuya asistencia financiera es profundamente agradecida.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Galaxies are the fundamental building blocks of the universe beyond our galaxy. They are varied, and represent complex astrophysical systems where large-scale phenomena are intertwined, such as the dynamics of the universe itself and that of its constituent elements. Their study offers us a good window to explore the cosmos. Each variety of galaxy, whether distinguished by its morphology (e.g., spiral, elliptical, irregular) or specific physical properties like size, mass, and star formation rate, encapsulates a history of the physical and dynamical processes that have occurred over billions of years, such that when added to other elements, they can even allow us to reconstruct cosmic evolution and revealing interactions between baryonic matter and dark matter. In this context, the morphological classification of galaxies, far from being merely a descriptive exercise, serves as an essential tool to investigate evolutionary processes and the underlying conditions that give rise to different galaxy morphologies.

1.1 Overview of Galaxy Formation and Evolution

According to our cosmological model ΛCDM (universe dominated by cosmological constant and cold dark matter), at the beginning of time there was nothing more than a singularity of infinite density that then expanded suddenly (Big Bang) and accelerated until today. But how did galaxies form? Where did they come from? To answer these questions we can start from almost the beginning of time, and one of the most critical observational discoveries of all time: the Cosmic Microwave Background (CMB), the first light of the universe, and where we can find direct evidence of the first density fluctuations in a baby homogeneous and isotropic universe that evolves in time. These density fluctuations served as the seeds for the formation of all cosmic structures.

Before the emission of the CMB, dark matter, being collisionless and interacting only gravitationally, began collapsing into denser regions of these primordial fluctuations. This process, known as "violent relaxation," brought the perturbations to a quasi-equilibrium state, forming what are now referred

to as "dark matter halos". When the universe cooled enough for baryonic matter to decouple from radiation (the era of the CMB emission), baryons fell into the gravitational potentials created by the dark matter halos. This interplay between dark matter and baryonic matter marked the beginning of galaxy formation. This process followed a hierarchical structure: smaller, low-mass halos formed first, later merging to create more massive structures. This "bottom-up" approach is a hallmark of structure formation in the ΛCDM paradigm, where galaxies and larger cosmic entities developed progressively over cosmic time (White & Rees, 1978; Mo et al., 2010). In other words, following this paradigm, the mass of dark matter halos increases over time through the accretion of surrounding material and mergers with other halos. The baryonic matter (gas) falls into the gravitational potentials created by the dark matter halos in a process which is guided by the densities established by the dark matter distribution and that occurs only after the gas has cooled sufficiently to collapse into these structures.

The gas in the halo collapses toward the center while retaining its angular momentum. This angular momentum prevents the gas from becoming too tightly packed. As the gas cools, frictional forces adjust it into nearly circular orbits, resulting in the formation of a flat disk perpendicular to the halo's angular momentum. In this disk, the gas becomes denser than it would in a spherical distribution. This higher density facilitates star formation.

The process of star formation is later regulated by feedback mechanisms, such as the energy and material released during the death of the first stars. In this way, we can qualitatively explain the formation of disk galaxies (Peebles, 1982; Davis et al., 1985; White et al., 1987; Schneider, 2015). Over time, these disk galaxies undergo a series of changes in their internal properties, giving rise to what are commonly known as spiral galaxies. The morphological variations in spiral galaxies—such as the appearance of pseudo-bulges (boxy substructures), bars, and rings—are a result of internal evolution processes. This gradual transformation of the galaxy's structure is referred to as secular evolution (Conselice, 2020).

Observations of the Universe have revealed that not all galaxies are disk/spiral, but there is a great variety, each one a unique and complex mixture of shapes and properties. Despite their diversity, all galaxies share a common composition: they are primarily made up of gas, dust, stars, and dark matter. The variability is due to the fact that, as perhaps the reader has already noticed, **galaxies (and the cosmos in general) are not static systems** (Aragón-Salamanca et al., 1993), their properties mutate and evolve with time due to different factors that are not only limited to secular (or internal) processes, but also extend to the effects exerted by the environment where the galaxies live or other galaxies (external).

External mechanisms can significantly influence galaxy evolution, often driving dramatic structural changes. Gravitational interactions, including galaxy mergers and harassment, play a major role in transforming disk galaxies into more spheroidal systems, commonly referred to as

"boxy-type" galaxies. Fully boxy systems without significant flattening are generally classified as spheroidal galaxies.

Additionally, hydrodynamical interactions, such as ram-pressure stripping by the intracluster medium, are key processes that can remove gas from galaxies and alter their star formation activity. Gravitational mechanisms like galaxy harassment or tidal interactions can further perturb orbital dynamics and contribute to the morphological evolution of galaxies (Boselli & Gavazzi, 2006; Boselli, A. et al., 2019; Larson et al., 1980; Pasquali et al., 2010; Jaffé et al., 2015)..

In most cases, these processes tend to increase the stellar mass, decrease the gas mass, reduce the rate of star formation, modify the color, and alter the dynamics and shape of galaxies, among other physical properties, thereby transforming them significantly. This suggests that galaxy morphology can act as a fossil record, containing valuable information about their evolutionary history. By studying a galaxy's morphology, we can infer key aspects of the physical processes that shaped it over time.

1.2 Morphological Classifications

During the 1910s, the number of galaxies recorded photographically rose to around 1.000, thanks to contributions such as those by Curtis (1918). As more galaxies were observed, it became evident that there were distinct differences between them, prompting scientists to establish classification systems. The first major and widely adopted system was proposed by Hubble (1926, 1936), which remains a cornerstone of galaxy classification to this day.

The Hubble sequence or "Hubble tuning fork" is originally based on visual impressions and is composed mainly of two types of galaxies, **Spirals (S)** and **Ellipticals (E)**, with a clear division between normal Spirals and **Barred Spirals (SB)**. Also presents a type of galaxy called **Lenticular (S0)** (Fig. 1.1), reflecting an original prejudice on the evolution from one type to another. Furthermore, Ellipticals and S0 galaxies together are often called **early-type galaxies**, whereas spirals are termed **late-type galaxies**, but these names do not imply any evolutionary trend and exist only for historical reasons. Hubble's classification scheme, as well as later contributions that further refine morphological types that do not fully conform to Hubble's original idea, as irregular galaxies (de Vaucouleurs, 1963), have become essential in the extragalactic area, since, as mentioned in previous sections, there is a systematic correlation between morphology and a number of general properties of galaxies.

The classification of galaxies according to their morphology has long been considered a fundamental tool for understanding their formation and evolution. However, quantifying morphology remains a challenge, particularly for visual classifications, due to their inherently subjective nature. Variability between galaxies, the presence of features such as bars, rings, and perturbations

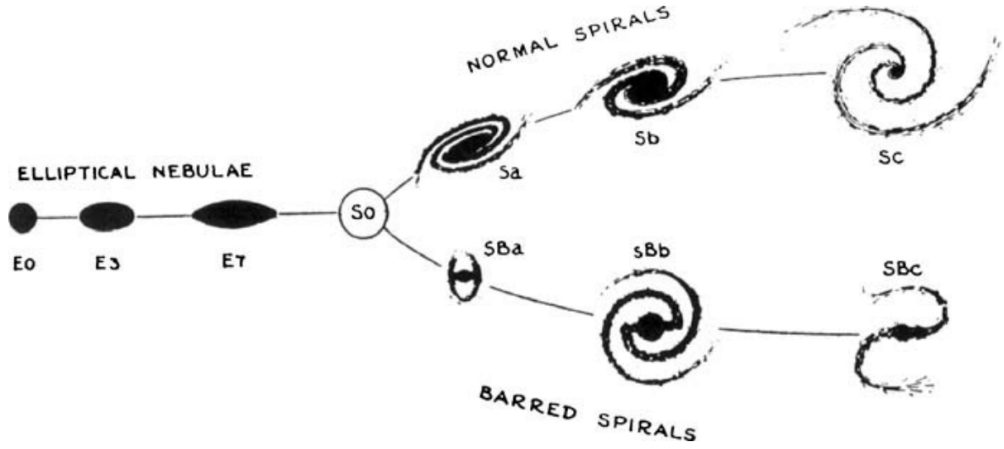


Figure 1.1: Original Hubble Tuning form / Hubble sequence presented in Hubble (1936), where the basic morphologies Spiral (S), lenticular (S0) and elliptical (E) are presented, as well as intermediate phases for the first and last morphology mentioned.

complicate classification labors. Early efforts, such as the Hubble morphological sequence, were intended to describe the diversity of galaxy shapes, but we can not expect galaxies to be characterized by a single linear sequence of properties, and it is unrealistic to expect that the rich phenomenology of galaxies can be summarized so simply. On the other hand, morphological features—such as concentration, asymmetry, and smoothness (commonly referred to as CAS; Conselice (2003))—offer valuable information about processes like star formation and merger history. However, their interpretation is sensitive to observational limitations, including resolution and signal-to-noise ratio. Parametric methods, such as bulge-to-disk (B/D) decompositions, allow structural components to be quantified using models such as Sersic profiles, but these methods have difficulties with irregularities introduced by features such as compact cores and spiral arms. Non-parametric techniques, such as artificial neural networks, provide greater flexibility by avoiding assumptions about analytical light distributions. However, these methods require extensive prior information, including large pre-labeled data sets for training, which can introduce biases in classification. Mitigating these biases requires access to robust computational resources and diverse data to develop efficient and reliable models.

Numerous studies have explored the correlation between galaxy morphology and physical properties, such as stellar populations, gas content, and star formation rates. For example, early-type (boxy) galaxies tend to be massive, red and quiescent, while late-type (disky) galaxies are bluer and actively forming stars. Trends first established in studies such as Holmberg (1958); Morgan & Osterbrock (1969); Bergh (1998); Conselice (2006). But, despite these advances, no single morphological system captures the full diversity of galactic structures over cosmic time. Modern efforts increasingly integrate spectral and photometric data to refine classifications, emphasizing the

need for frameworks that connect morphology to the underlying physics of galaxy formation and evolution. On the other hand, there are also scientists working to build cosmological simulations that allow us to explore galaxy evolution in excellent detail, tracing their evolution and interactions over cosmic time. These simulations, such as IllustrisTNG (Pillepich et al., 2017a,b; Springel et al., 2017; Marinacci et al., 2018; Nelson et al., 2021), EAGLE (Schaye et al., 2015; Crain et al., 2015), and Horizon-AGN (Dubois et al., 2014), incorporate a variety of physical processes, such as gas cooling, star formation, feedback from supernovae, active galactic nuclei and the effects of large-scale structure, recreating realistic galaxy populations. Cosmological simulations allow us to test hypotheses about the various internal parameters of galaxies and their connection to morphology.

In this context, establishing parameter ranges that concentrate galactic morphology becomes a fundamental pillar for unraveling the processes of galaxy formation and evolution. Therefore, in this thesis work, we will use the cosmological simulation IllustrisTNG to compare two non-parametric classification methods: concentration, asymmetry and smoothness (CAS, Conselice 2003), and Deep Learning (described in Ch. 2). This comparison aims to identify differences in the galaxy morphologies depending on the method used, providing a deeper perspective on how these techniques contribute to understanding the structural diversity of galaxies. This thesis work is organized as follows. In chapter 2, we will describe in detail the data catalogs used. In chapter 3, we will describe and explain how we worked the catalogs, including thresholds and considerations for the morphology classification. In chapter 4, we will describe our results. In chapter 5, we will present a discussion to explain the behavior of the results. Finally, in chapter 6 we will conclude the research work by presenting a summary of the findings and a perspective towards the future.

Chapter 2

Data

2.1 IllustrisTNG

We will work mainly with the IllustrisTNG[1] gravo-magneto-hydrodynamic simulation (hereafter TNG) (Pillepich et al., 2017a,b; Springel et al., 2017; Marinacci et al., 2018; Nelson et al., 2021), a project that concentrates the state of the art on cosmology and galaxy formation, and presents a "fictitious universe", performed with the moving-mesh code AREPO, which solve magneto-hydrodynamic continuum equations, coupled with self-gravity (Springel, 2010; Pakmor & Springel, 2013) to calculate the evolution of dark matter, cosmic gas, luminous stars, and supermassive black holes from redshift $z = 127$ to $z = 0$ (organized on Snapshots), all of this taking account the standard cosmological model ΛCDM ($\Omega_{\Lambda,0} = 0.6911$, $\Omega_{m,0} = 0.3089$, $\Omega_{b,0} = 0.0486$, $\sigma_8 = 0.8159$, $n_s = 0.9667$ and $h = 0.6774$). TNG is an improved version of its predecessor, Illustris[2] (described in Weinberger et al., 2016), which did not include magneto-hydrodynamic effects such as those described in Pakmor et al. (2011), Shock Finder (Schaal & Springel, 2014) or AGN (Active Galaxy Nucleus) feedback, which are incorporated in TNG, offering a more comprehensive modeling of astrophysical processes (more details in Nelson et al., 2021). TNG is designed to reproduce key observables: (i) the galaxy stellar mass function, (ii) the stellar-to-halo mass relation, (iii) the total gas mass content within the virial radius (r_{500}) of massive groups, (iv) the stellar mass-size relation and (v) the BH-galaxy mass relation, all at redshift $z = 0$, also (vi) the shape of the general star formation rate density at $z \lesssim 10$. It consists of three main simulations, each employing a different periodic volume box: TNG300 (302.6[Mpc]), TNG100 (110.7[Mpc]), and TNG50 (51.7[Mpc]). All three complement each other, allowing analysis of galaxy formation and evolution on different scales. TNG300 has the largest sample of galaxies, allowing the study of groups and clusters of galaxies, while TNG50 has a smaller, but high-resolution sample that allows detailed study of the structural

[1] https://www.tng-project.org/
[2] https://www.illustris-project.org/

properties of galaxies or the effects on the gas surrounding them. In the middle is TNG100, which provides a balance between number of galaxies and resolution. Each simulation have a catalog that is composed of 100 snapshots, containing information of the simulation at different redshift values. These snapshots contain raw particle data for each type of particle (gas, dark matter, tracers, stars, wind particles, and black holes) as produced by the simulation at specific points in time. This raw data is later organized into catalogs using algorithms. The FoF (Friends-of-Friends) algorithm identifies groups, which correspond to dark matter halos and their associated properties. Within these groups, the Subfind algorithm identifies subgroups or subhaloes, which represent individual overdensities inside halos. These subhaloes, in the appropriate mass ranges, are what we interpret as galaxies. The subhalo catalogs provide processed information for each subhalo, including properties such as "SubhaloStellarPhotometrics," "SubhaloMass," and "SubhaloSFR," among others. Detailed descriptions of these columns can be found online. Additionally, there are extended catalogs, such as the DL (Deep Learning) catalogs and the CAS (Concentration, Asymmetry, Smoothness) catalogs, which offer further insights into subhalo properties.

Most of the data mentioned above is accessible through an online platform based on JupyterLab, which provides access to the primary simulation outputs: snapshots, FoF group catalogs, and Subfind subhalo catalogs. However, post-processed supplementary data, such as the extended catalogs (e.g., DL and CAS), must be downloaded separately. This platform makes accessing and managing the results of these simulations particularly easy and intuitive for researchers.

2.1.1 CAS: Non-parametric optical morphology characterization

For the purposes of this thesis work, one of the supplementary TNG catalogs to be used will be "SKIRT Synthetic Images and Optical Morphologies" (Rodriguez-Gomez et al., 2019), which is composed of morphological measurements, including CAS, MID (Freeman et al., 2013), Ginni-M20 (Lotz et al., 2004), among other statistics, obtained from synthetic images of subhalos in TNG, whose stellar mass exceeds $10^{9.5}$ solar masses. The synthetic images are observed from an angle perpendicular to the x-y plane of the simulation, which results in random orientations for each galaxy (e.g. Fig. 2.1). Then, in addition to predicting the distribution of light produced by stars, the SKIRT radiative transfer code (Baes et al., 2011; Camps & Baes, 2015) or the GALAXEV stellar population synthesis code (Bruzual & Charlot, 2003) is used to model dust attenuation and scattering effects. The latter is used when the fraction of star-forming gas is less than 1/100th of the total baryonic matter. This does not greatly affect the final output, as the differences between SKIRT and GALAXEV are practically indistinguishable below the threshold, but GALAXEV is much cheaper computationally. The result of the above process consists of a 3D data cube for each galaxy, containing a rest-frame SED for each pixel, which is multiplied by the individual SDSS filters (g and i), also, the number of pixels is tuned to match the SDSS pixel scale (0.396

arcsec). This provides idealized synthetic images ($S/N \sim \infty$), which are then post-processed, where atmospheric noise and possible telescope optical effects are modeled and added by convolution with an azimuthally symmetric Gaussian PSF (Point Spread Function), sky background noise is also added as a random Gaussian variable added to each pixel flux value, taking into account a constant standard deviation of $\sigma_{sky} = 1/15$. Furthermore, a weight map is created, which model the uncertainty of each pixel value, including random sky noise and Poisson statistics on electron counting measured by the device (CCD). Also, a segmentation image was created for each synthetic image, using the photometry package "Photoutils[3]".

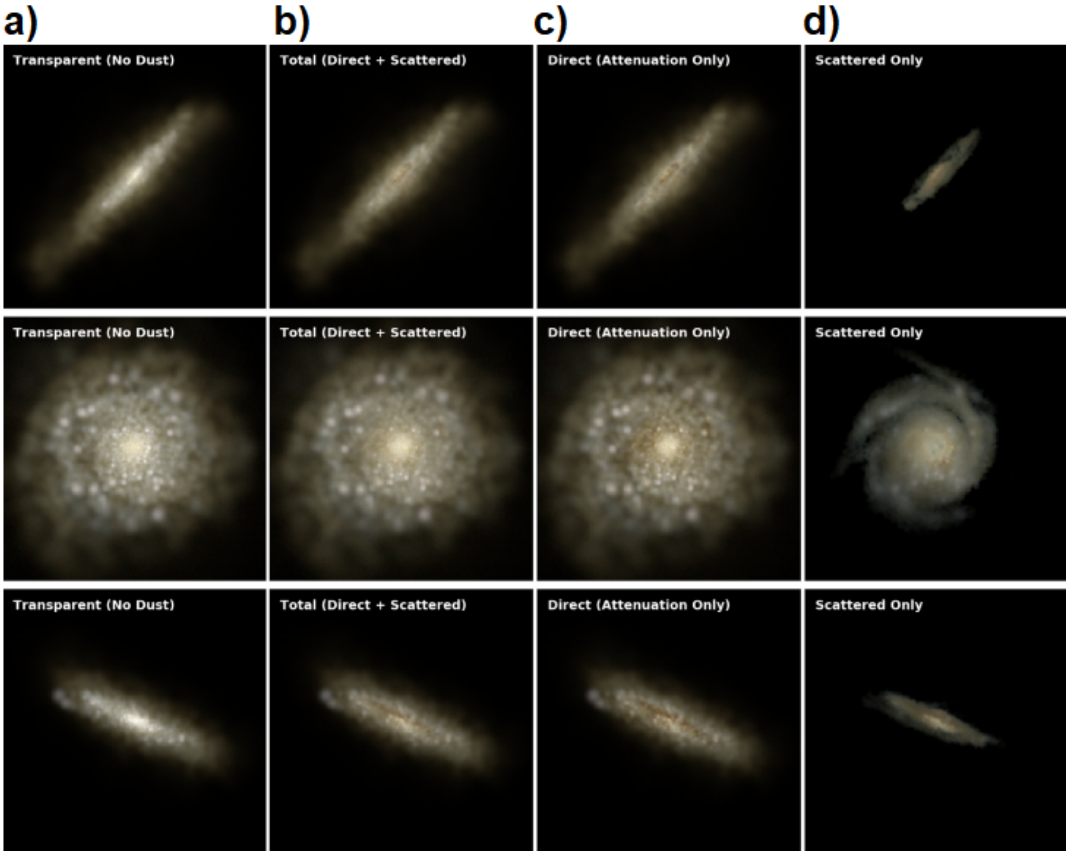


Figure 2.1: Idealized synthetic images (using the PS1 g, r, i filters) of randomly selected galaxies ($M_\star \sim 10^{10.5}$ $[M_\odot]$) from the IllustrisTNG simulation, showing the different light components. From left to right, the different columns show (a) the light distribution that would be observed without any dust effects, (b) the light distribution obtained with the full dust modelling, (c) the contribution of rays that have not been scattered, and (d) the contribution from rays that have been scattered at least once. The field of view in each panel corresponds to 10 stellar half-mass radii. Note that the inclusion of dust tends to reduce the light concentration near the galactic center. Extracted from Rodriguez-Gomez et al. (2019).

[3]https://photutils.readthedocs.io/en/stable/

Now, to analyze the synthetic images, the python package "Statmorph[4]", developed by the same authors of the catalog and also described in Rodriguez-Gomez et al. (2019), was used. This package outputs many structural indices that allow us to characterize and quantify the structure of a galaxy, without making prior assumptions of its shape (unlike parametric methods), through which we can categorize the different galaxies (Fig. 2.2). The results include a flag parameter, where $flag == 0$ indicates a reliable measurement, with $S/N > 2.5$. The rest of the data may be unreliable for a variety of reasons, such as an extra object in the foreground or not being masked correctly. Now, Among the structural indices are those previously mentioned, which form the acronym CAS, defined as follows:

- Concentration:

$$C = 5 \log_{10} \left(\frac{r_{80}}{r_{20}} \right) \tag{2.1}$$

  Where r_{20} and r_{80} are the radio that contains the 20% and the 80% of the galaxy light. The total flux is calculated as that within $2r_{petro}$. Higher concentration indicates that the galaxy's light is more concentrated towards the nucleus, which is typical of elliptical galaxies and early-type galaxies, due to their smooth light distribution, which decreases exponentially with distance from the nucleus, and lower or modest concentration indicates that the light is more distributed towards the outer parts, typical of spiral and late-type galaxies, where there is a more extended light distribution due to the disk and spiral arms with high presence of star formation.

- Asymmetry:

$$A = \frac{\sum_{i,j} |I_{ij} - I_{ij}^{180}|}{\sum_{i,j} |I_{ij}|} - A_{avg} \tag{2.2}$$

  where I_{ij} corresponds to the flux values of the original image, I_{ij}^{180} corresponds to the flux values of the image rotated 180 degrees and A_{avg} is the average asymmetry of the background, which has been obtained from the analysis of a 32x32 pixel region without emission sources.

  The sum is calculated over all pixels within a radius of $2r_{petro}$ from the galaxy's center, which the image is rotated is defined as the one that minimizes the asymmetry. Low asymmetry is related to well-ordered and undisturbed galaxies, such as early-type spiral galaxies and elliptical galaxies, while high asymmetry is indicative of interacting galaxies, intense star formation or irregular morphologies, such as late-type galaxies or those in mergers.

[4]https://statmorph.readthedocs.io/

- Smoothness:

$$S = \frac{\sum_{i,j} I_{ij} - I^{\sigma}_{ij}}{\sum_{i,j} I_{ij}} - S_{avg} \tag{2.3}$$

where I_{ij} corresponds to the flux values of the original image, I^{σ}_{ij} corresponds to the flux of the image smoothed by boxcar filter of width $\sigma = 0.25 r_{petro}$ and S_{avg} is the average smoothness of the background, which, has been obtained from the analysis of a 32x32 pixel region without emission sources.

The sum spans the pixels between the sigma distance and $2r_{petro}$, from the point that minimizes A. Small values of S indicates a galaxy with a uniform, clump-free light distribution, which is typical of evolved galaxies with small star formation, such as ellipticals or early-type galaxies, in contrast, a high S value indicates the presence of substructures, such as star-forming clusters, prominent spiral arms or interaction regions. It is common in spiral, late-type or disturbed galaxies.

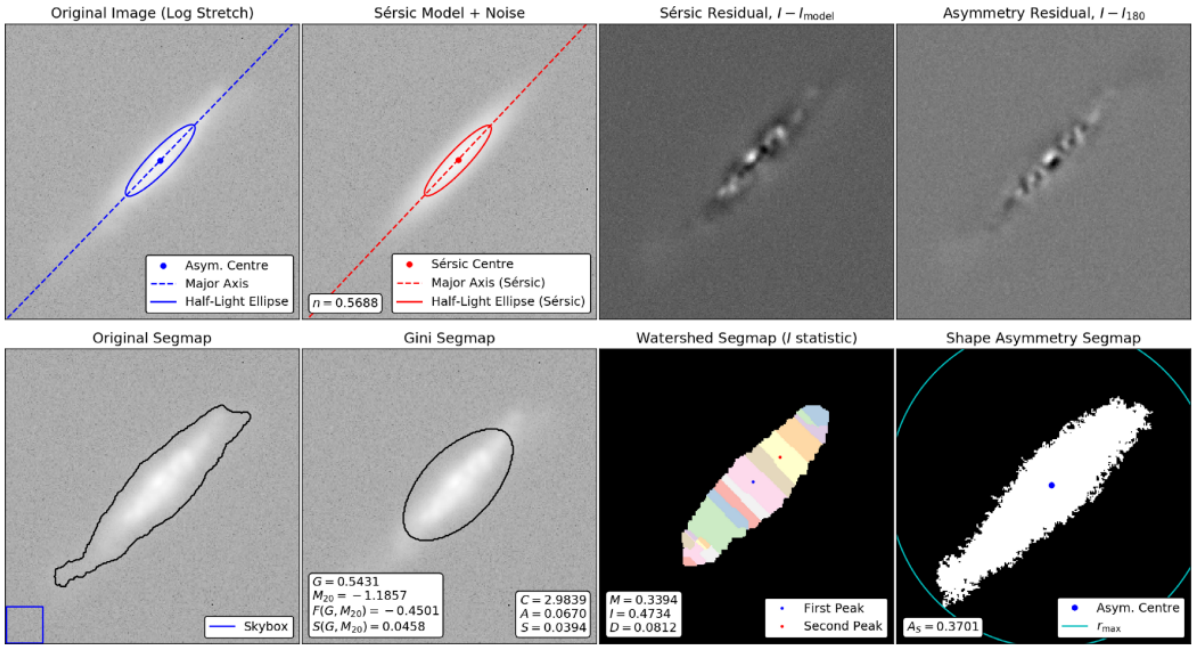


Figure 2.2: Some of the morphological measurements performed by statmorph, shown for the galaxy in the top row of Fig. 2.1 (using the full dust model). Extracted from Rodriguez-Gomez et al. (2019).

2.1.2 The Deep-Learning galaxy morphologies classification

Another useful complementary catalog for the purpose of this work will be the Deep Learning morphologies (DL) from Huertas-Company et al. (2019); Varma et al. (2021). The study/catalog provides the probabilities that subhalos with masses greater than $10^{9.5}$ visually belong to specific morphological groups. These probabilities were obtained by classifying SDSS-like synthetic images (similar to Fig. 2.1) of TNG objects using a Convolutional Neural Network (CNN). The synthetic images were created following the methodology outlined by Rodriguez-Gomez et al. (2019), with the main difference being that only the SDSS r-band filter was utilized in this work. The CNN was trained on the catalog from Nair & Abraham (2010), which contains visual classifications of approximately 14,000 galaxies. Specifically, three binary classification models were developed: the first estimates the probability of a galaxy being late-type, the second determines the probability of being lenticular for galaxies classified as early-type, and the third evaluates the probability of a late-type galaxy being a spiral galaxy.

Chapter 3

Methodology

TNG100 is an ideal choice for this study because it strikes a balance between simulation volume and resolution, offering a sufficiently large sample of galaxies while providing the detail necessary for robust morphological analyses. To ensure reliable results, we restrict the subhalos to those with stellar masses greater than $10^{9.5} M_{\odot}$, consistent with the supplementary catalogs. This selection ensures that the galaxies included in the analysis are resolved with a sufficient number of stellar particles (~ 2500) for meaningful interpretations.

3.1 Development of a morphology criteria based on deep learning classifications

Following the lineaments established by Huertas-Company et al. (2019), we classify galaxies into four morphological categories: Elliptical (E), Lenticular (S0), Early-Spiral (Sab), and Late-Spiral (Scd). The Bayesian probabilities for each category are defined as described in Equation 3.1. These describe how likely it is that the object in question belongs to one of the categories, based on visual resemblance to real galaxies.

$$\begin{aligned} P(E) &= P(Early) \times P(E/Early) \\ P(S0) &= P(Early) \times P(S0/Early) \\ P(Sab) &= P(Late) \times P(Sab/Late) \\ P(Scd) &= P(Late) \times P(Scd/Late). \end{aligned} \tag{3.1}$$

Now, the classification criteria are outlined in Equation 3.2. These criteria incorporate a parameter α, which controls the probability threshold required for a galaxy to be assigned to a specific morphology Specifically, if $\alpha = \frac{1}{x}$, the classification requires at least "x" times more probability that an object belongs to a given morphological class than to others.

$$
\begin{aligned}
\mathrm{E} &= (\alpha \cdot P(\mathrm{E}) > P(\mathrm{S0})) \wedge (\alpha \cdot P(\mathrm{E}) > P(\mathrm{Sab})) \wedge (\alpha \cdot P(\mathrm{E}) > P(\mathrm{Scd})) \\
\mathrm{S0} &= (\alpha \cdot P(\mathrm{S0}) > P(\mathrm{E})) \wedge (\alpha \cdot P(\mathrm{S0}) > P(\mathrm{Sab})) \wedge (\alpha \cdot P(\mathrm{S0}) > P(\mathrm{Scd})) \\
\mathrm{Sab} &= (\alpha \cdot P(\mathrm{Sab}) > P(\mathrm{S0})) \wedge (\alpha \cdot P(\mathrm{Sab}) > P(\mathrm{E})) \wedge (\alpha \cdot P(\mathrm{Sab}) > P(\mathrm{Scd})) \\
\mathrm{Scd} &= (\alpha \cdot P(\mathrm{Scd}) > P(\mathrm{S0})) \wedge (\alpha \cdot P(\mathrm{Scd}) > P(\mathrm{Sab})) \wedge (\alpha \cdot P(\mathrm{Scd}) > P(\mathrm{E}))\,.
\end{aligned}
\tag{3.2}
$$

To assess the reliability of the classifications and to define an appropriate alpha value, we examine how the probability of belonging to each category is distributed in the galaxies of our sample ($M\star > 10^{9.5}\ M_{\odot}$), using a histogram (Figure 3.1), which shows that only the elliptical and early spiral morphologies reach a significant number of objects with really high probability levels. Based on this, in general, we focus our analysis on these two DL morphologies. First, a preliminary test was performed using a lower probability threshold of 3 times the probability ($\alpha = 1/3$), to include more objects, but this only led to an increase of the scatter in the data, reducing its interpretive power. In addition, the number of lenticular galaxies and late spirals remained almost as low as using stricter thresholds. Finally, a 9 times higher probability threshold was considered optimal in order to maintain the quality of the classifications (i.e., $\alpha = 1/9$), whose resulting sample will be referred to as 'high probability'. Although we only retain a small part of the total sample of galaxies present in TNG100, this ensures a purer sample, leading to high interpretability when studying the distributions within the selection.

From the classification carried out with deep learning, we obtain 314 Elliptical galaxies and 641 Early Spiral galaxies (with $\alpha = 1/9$).

3.2 Development of a morphology criteria based on CAS parameters

From the catalogs from Rodriguez-Gomez et al. (2019), we will focus exclusively on cases where flag = 0. Our analysis will primarily involve the concentration, asymmetry, and smoothness parameters. For comparative purposes, we redefine the smoothness as $10 \times S$ following the definition established in Conselice (2003).

As classification criteria, we adopt the ranks provided in Table 6 of Conselice (2003), present in the table 3.1, which contains only those morphologies that we consider relevant for further comparison with DL. We use these ranges as thresholds for concentration (C), asymmetry (A) and smoothness (S) to assign morphological labels to the galaxies in our dataset. This classification is simple, but allows us to label galaxies systematically, using only the CAS structural indices, in a way that is consistent with established empirical criteria. It is worth mentioning that the code used for the classification is designed to ensure that galaxies are uniquely classified into a single morphological

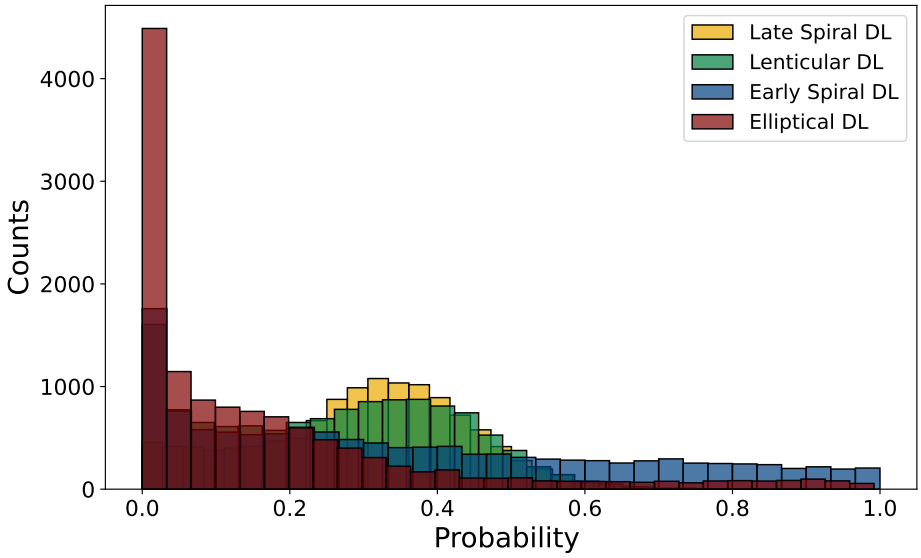


Figure 3.1: Histogram of the probability distributions for the classification of galaxies ($M_\star > 10^{9.5}\ M_\odot$) using Deep Learning models, providing a comparison among the probability distributions for elliptical (red), lenticular (green), early spiral (blue), and late spiral (yellow) galaxies. Here each galaxy have assigned a value of probability for each morphology, here is plotted the four values. In other words, each color and label represent how the probability of being a particular morphological type is distributed across the entire sample. This means that the big amount of low probabilities for Elliptical, correspond to the same galaxies group of galaxies which have high values of early type morphologies.

category, even when the CAS classification ranges of galaxies partially overlap. For this, logical conditions are used to exclude galaxies that could have been classified in multiple categories. It is worth mentioning that this catalog does not include an analysis of synthetic images convolved with the "r" filter for TNG100, so we will use those processed with the "i" filter, the reddest available.

From CAS indices classification we obtain 163 ellipticals, 401 dwarf ellipticals, 236 Early Spirals and 633 Late spirals. Both datasets will be studied in the following sections.

Table 3.1: Averages and 1σ Variations of Structural Parameters for Galaxy Types. Reproduced and adapted from Conselice (2003), table 6. In this work, we use it as classification criteria.

Type	$C(R)$	$A(R)$	$S(R)$
Ellipticals	4.4 ± 0.3	0.02 ± 0.02	0.00 ± 0.04
Early-Type Spirals (Sa - Sb)	3.9 ± 0.5	0.07 ± 0.04	0.08 ± 0.08
Late-Type Spirals (Sc - Sd)	3.1 ± 0.4	0.15 ± 0.06	0.29 ± 0.13
Dwarf Ellipticals	2.5 ± 0.3	0.02 ± 0.03	0.00 ± 0.06

Chapter 4

Results

With the morphological classification with both methods ready, we will analyze and compare the physical properties of the classified galaxies using TNG data. First, we will examine the physical property distributions separately for the deep-learning and CAS-classified galaxies to identify general trends. Also, a cross-histogram comparisons will also be performed to assess the consistency between the two methods. Finally, a direct comparison between the galaxies classified by both methods will be performed on key diagrams, such as stellar mass-color, sSFR-stellar mass, and gas-stellar mass fraction. This approach will allow us to explore differences or similarities in physical parameter spaces that may be related to galaxy morphology.

4.1 Physical property distributions by classification System

First, we analyze how some physical properties are distributed among galaxies and general tendencies, depending on the method by which they have been classified.

A pronounced bimodality is evident in all the plots of Figure 4.1, including mass, star formation rate (SFR), color, and gas content. Early Spiral DL galaxies tend to concentrate lower stellar mass ranges, exhibit high specific star formation rates, show a color close to the blue, and have higher gas fractions. In contrast, Elliptical DL galaxies are predominantly found in higher stellar mass ranges, exhibit lower specific star formation rates, have redder colors, and generally exhibit lower gas fractions. A notable observation is the sharp cut-off in the distribution of elliptical galaxies around $M_\star \sim 10^{10.5}\,[M_\odot]$, unlike early spiral galaxies, whose distribution decreases more gradually toward higher masses. For sSFR, there is a clear separation around $sSFR \sim 10^{-11}\,[Yr^{-1}]$, consistent with the bimodality mentioned earlier. Elliptical galaxies tend to exhibit low star formation rates, aligning with their classification as quiescent in the literature. Regarding color, the division appears marked around a value of 0.6. Interestingly, a significant fraction of elliptical

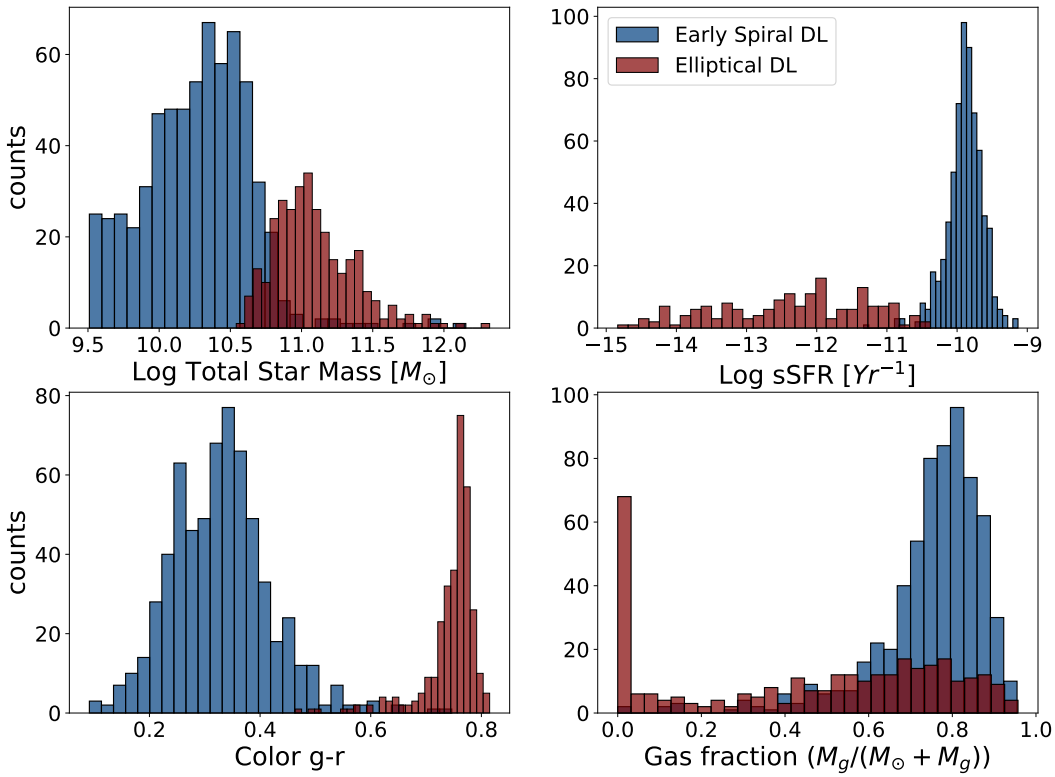


Figure 4.1: Distributions of physical properties in the classifications made with DL, where those classified as Early Spiral are represented in blue, and those that are classified as Elliptical are represented in red. *Top left panel*: Log Stellar mass distribution. *Top right panel*: Log Specific star formation rate (sSFR) distribution. *Bottom left panel*: Color $g - r$ distribution. *Bottom right panel*: Gas fraction distribution. In general, there is a strong bimodality in these distributions, except in the gas fraction, which may be due to how TNG works the gas loss.

galaxies show unexpectedly high gas fractions, deviating from the typical expectations for this morphological type.

Regarding the classifications made using the CAS parameters, Figure 4.2 also shows a bimodality present in all the histograms, although with differences with respect to the classifications made with DL. In terms of stellar mass, the clearest separation is observed between Late Spiral and Elliptical, as well as between the latter and Dwarf Elliptical. In the sSFR plot, the behavior is similar to the stellar mass, with the bimodality being distributed similarly between the morphologies, and a marked separation around $sSFR \sim 10^{-11}[Yr^{-1}]$). In the color histogram, the Dwarf Elliptical no longer clusters with Early Spirals, and clusters mainly with the Ellipticals in the red region, while the Late Spiral is relegated almost alone to the blue region. However, a visible amount of Dwarf Elliptical is still observed in the blue part. This behavior is similar to that observed in the Gas Fraction histogram, with the particularity that, as in the DL classifications, we find a significant amount of Elliptical and Dwarf Elliptical (Early-type galaxies) with unexpectedly high

gas fractions for this type of morphology. The galaxies labelled as Early Spiral CAS appear to be similarly distributed in all the plots, occupying intermediate ranges in mass, sSFR, color and Gas Fraction. Particularly noteworthy is the presence of a peak around $M_\star \sim 10^{10.8}[M_\odot]$, an almost centered distribution at $sSFR \sim 10^{-11}[Yr^{-1}]$ and $color \sim 0.6$, just the points where the separation of the modality is marked.

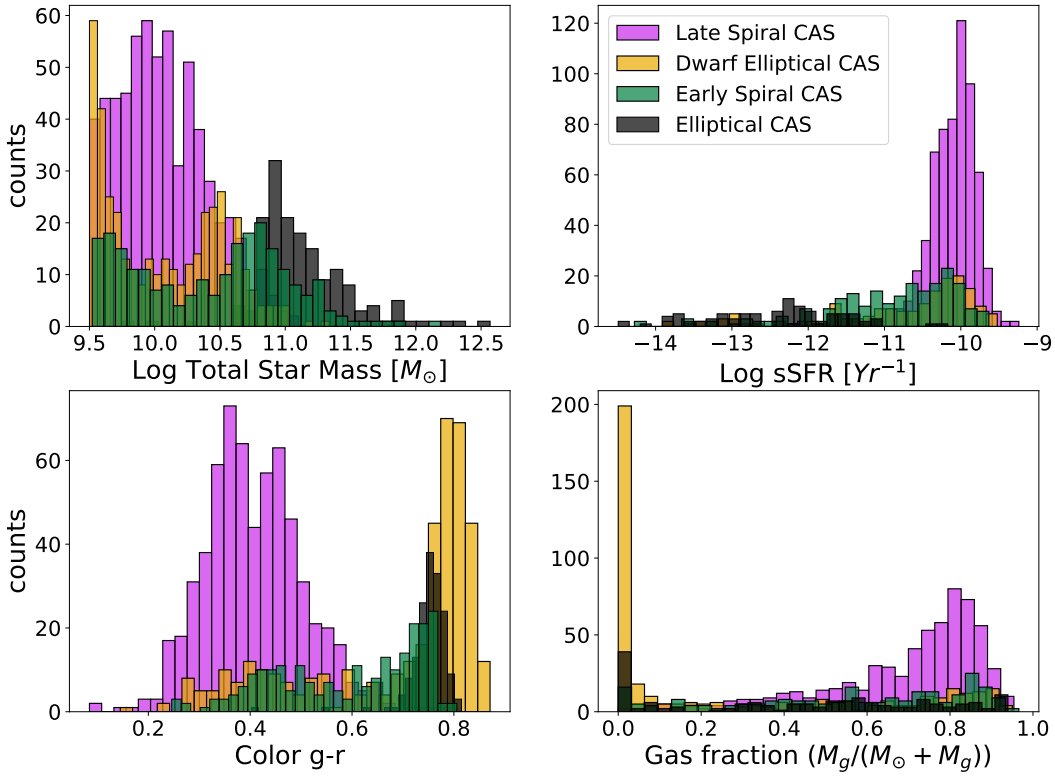


Figure 4.2: Distributions of the physical properties in the CAS classifications, where those classified as Late Spiral are represented in magenta, those classified as Elliptical are represented in black, those classified as Dwarf elliptical are represented in orange and those classified as Early Spiral are represented in green. *Top left panel:* Stellar mass distribution. *Top right panel:* Specific star formation rate distribution. *Bottom left panel:* Color g-r distribution. *Bottom right panel:* Gas fraction distribution.

Analyzing these distributions already gives us indications of similarities and differences between the two classification methods. Both approaches exhibit bimodality in most physical properties, with some differences. To better understand the relationship between the two methods, we can perform a cross-comparison, exploring how the classifications interact, where they align or diverge.

4.2 Cross comparison between CAS and Deep Learning

We explore how the DL model probabilities are distributed in the CAS classifications, as well as how the CAS parameters are distributed in the DL-classified galaxies. This will allow us to analyze the behavior of the parameters of each classification system in both data sets.

On one hand, Figure 4.3 shows a clear bimodality between the two DL classifications in all histograms of the CAS parameters (3 lower panels). Galaxies labelled as Elliptical DL tend to be very smooth ($S \sim 0.041 \pm 0.156$), symmetric ($A \sim 0.022 \pm 0.036$), and with high concentration ($C \sim 4.254 \pm 0.554$), while Early Spiral DL galaxies show more moderate values of smoothness ($S \sim 0.535 \pm 0.275$), asymmetry ($A \sim 0.285 \pm 0.109$), and concentration ($C \sim 2.969 \pm 0.462$), with a practically Gaussian distribution. The galaxies classified as Elliptical DL completely cover the range of parameters used to define the Elliptical CAS label. A similar behavior is observed with the Dwarf Elliptical CAS label, although with one exception in the concentration range; these galaxies seem to lie in a region where the probability, for some reason, was not sufficient to be classified as Elliptical DL, due the number of the latter gradually decreases in the direction of the concentration range for Dwarf Elliptical CAS. On the other hand, galaxies labelled as Early Spiral DL seem to represent better the range of CAS parameters corresponding to the Late Spiral CAS classification than to the Early Spiral CAS classification. Moreover, the latter seems to be always in the area where the division between the Early Spiral DL and Elliptical DL distribution is marked (i.e. in the middle of both). In general, we see that spiral galaxies classified with DL possess greater asymmetry than spirals classified with CAS.

Now, in Figure 4.4, we can see the DL probability distribution for galaxies labeled using CAS criteria and DL criteria. The plot reveals notable trends: The Elliptical CAS classification predominantly contains galaxies with a high probability of being classified as Elliptical DL and low probabilities of being Early Spiral DL, Late Spiral DL or Lenticular DL, which is in good agreement with the distribution of probabilities seen for Elliptical DL (upper right panel). This suggests a possible overlap between galaxies classified as Elliptical DL and those classified as Elliptical CAS. The Early Spiral CAS classification, in general, shows galaxies with low probabilities of belonging to any of the DL classifications, except for a small subset with high probabilities of being classified as Early Spiral DL or Elliptical DL, which suggest that the Early Spiral CAS classification captures a population of galaxies that the DL model struggles to classify confidently into any specific category. The small subset with high probabilities of being Early Spiral DL or Elliptical DL may indicate some overlap with these categories. Interestingly, this distribution seems like a random selection from the overall galaxy sample, as its probability distribution remember that seen in Figure 3.1. This last two pints could imply that the Early Spiral CAS includes, in the majority, a mix of galaxies

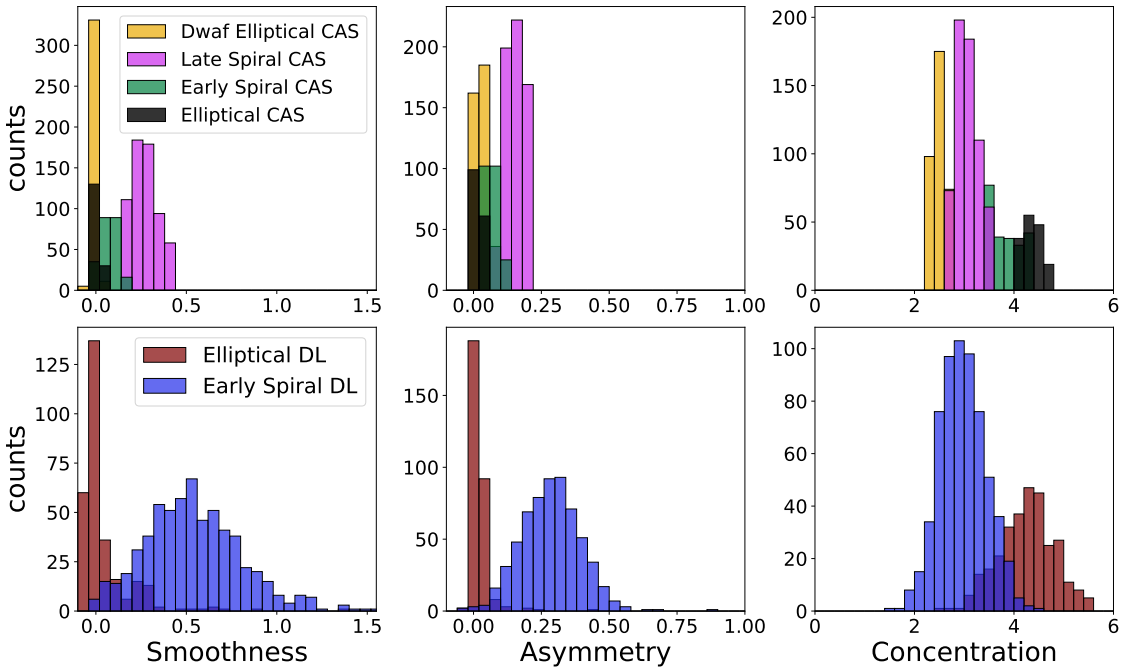


Figure 4.3: CAS parameter distributions for CAS and Deep Learning classifications. The first row (top 3 panels) shows the distribution of CAS parameters for galaxies classified according to the CAS criteria. The second row (bot 3 panels) Shows how CAS parameters (concentration, asymmetry and smoothness) are distributed in Deep Learning classified sample mentioned in section 3.1 as "high probability".

with intermediate or ambiguous properties for DL. The Dwarf Elliptical CAS classification similarly contains galaxies with low probabilities of belonging to any DL classification, with probability values below 80% in most cases, and many galaxies showing a notably low probability of being Early Spiral DL. Unlike Early Spiral CAS, the probability distribution for Dwarf Elliptical CAS does not resemble that of Figure 3.1. This indicates that Dwarf Elliptical CAS is capturing a set of galaxies that is hard to classify for DL, but that remains a restricted region in parameter space, being sufficiently particular to be distinct from the full sample of galaxies, unlike Early Spiral CAS. Finally, the Late Spiral CAS classification includes a significant number of galaxies with a high probability of being Early Spiral DL, while the remaining DL classifications present low probabilities/probability level, which is in good agreement with the distribution of probabilities seen for Early Spiral DL (lower right panel). Similar to the case of the Elliptical morphologies, this suggests a possible overlap between Early Spiral DL and Late Spiral CAS.

Overall, the cross-comparison between the CAS and DL classifications shows that there are notable overlaps and differences in the way morphological categories are assigned to galaxies, such that Elliptical CAS galaxies appear to be closely aligned with their DL counterpart, Late Spiral

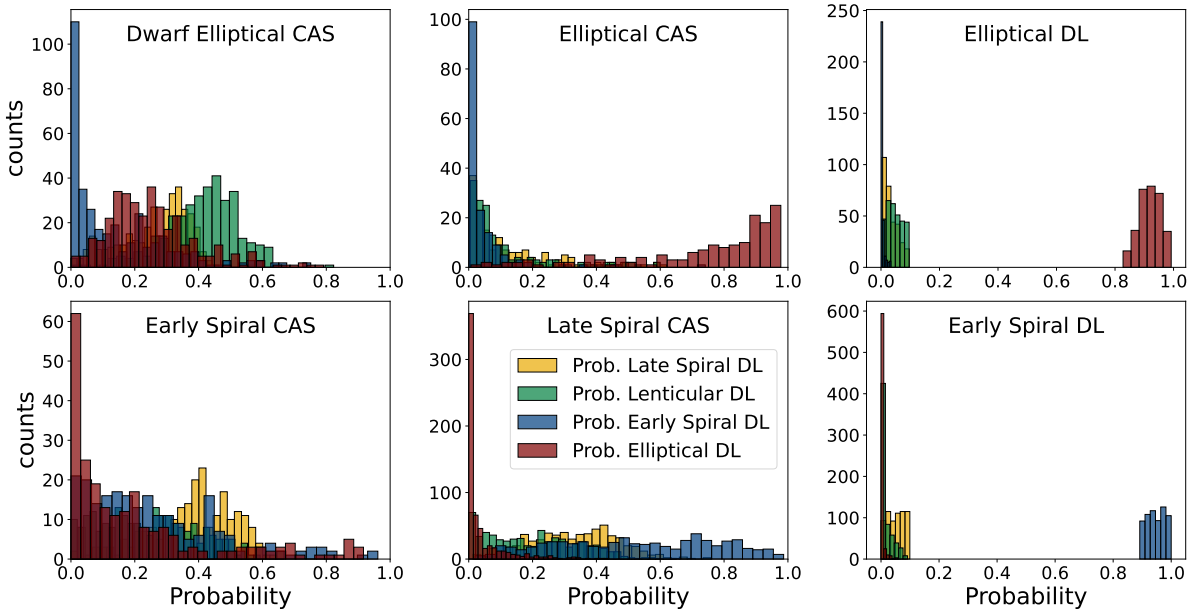


Figure 4.4: Deep Learning probability distributions for the CAS and Deep learning classifications. Left and middle panels shows how the probabilities assigned by the Deep Learning model for the CAS-classified galaxies are distributed. The right panels shows how it is distributed the probability in the DL classifications.

CAS appear to be closer to Early Spiral DL than to their counterpart, and Early Spiral CAS appear to be representing intermediate features with respect to Elliptical DL and Early Spiral DL. In addition, CAS dwarf elliptical galaxies occupy a particular range of concentration that is being neglected by the DL model when assigning an early type category. The next section will focus on a direct comparison in parameter space to unveil these insights.

4.3 Direct comparison in parameter space

The distribution of physical properties and the cross-comparison between the CAS and DL classifications have revealed general trends and possible overlaps in the morphological assignments of galaxies. We will now analyze how galaxies are positioned in specific parameter spaces, which will allow a direct assessment of the differences and similarities between the classifications made by the two methods, as well as the coherence of each one.

The analysis of the color vs. stellar mass plot (Figure 4.5) provides a direct synthesis of

the behaviors and suggestions discussed in the previous sections, offering a unified view of the various properties associated with each morphological classification. One notable observation is the significant overlap between Elliptical CAS and Elliptical DL classifications, occurring specifically in the red sequence region of the diagram, where elliptical galaxies are known to reside. This overlap is so substantial that Elliptical DL galaxies practically encompass the entire Elliptical CAS population. We can also observe that both Elliptical CAS and Elliptical DL classifications exhibit a sharp cut-off around $M_\star = 10^{10.5}[M_\odot]$, indicating similar behavior in this mass range. This suggests that the observed limit is not only due to the classification method, such as the confidence threshold applied to elliptical DL galaxies, but may also be due to TNG100's own characteristics, such as resolution.

We can see how the Dwarf Elliptical CAS complete the red zone, of masses lower than $M_\star = 10^{10.5}[M_\odot]$, which Elliptical CAS and Elliptical DL are not being able to contain. There is also a small density of CAS dwarf elliptical galaxies in the blue part of the graph that could be considered an outlier of the general trend. However, it should be noted that there are Dwarf ellipticals that are star-forming, with all that that entails.

As showed in previous sections, the Early Spiral CAS galaxies are distributed in an area where we expect to find galaxies with transient morphologies, with high concentrations at the lower edge of the red sequence and at the upper edge of what we can identify as the blue cloud, living in the middle of the predominant morphologies, which could be considered as the green valley. Also, there is clearly a closer proximity between Late Spiral CAS and Early Spiral DL than between the latter and its direct CAS counterpart. Although it makes sense that Late Spiral CAS is concentrated in areas of lower mass than Early Spiral DL, it is curious that Late Spiral CAS is in a redder area than Early Spiral CAS.

Reviewing the plot of sSFR versus stellar mass (Figure 4.6), we observe similar trends to those of the color vs. stellar mass plot, as both displays simultaneously the trends presented in the previous sections. The overlap between elliptical CAS and elliptical DL is also particularly evident, which reinforces the convergence of these classifications. In addition, we observe that there are a Dwarf Elliptical CAS galaxies density which have a particularly high sSFR, similar to that of Late Spiral CAS and Early Spiral DL galaxies in some cases. This behavior could be a consequence of the same reason that is generating the bluest isolated group, present in the color vs. stellar mass plot. Now, an interesting finding is that Late Spiral CAS have lower sSFR than Early Spiral DL, which is striking and suggests that the differences observed in the color vs. mass plot are not exclusively due to the difficulties with respect to color recreation in the simulations, but could be a consequence of other characteristic of TNG100 or the classification criteria used. Even though, both classifications together seem to reconstruct the main sequence of star formation, which make sense. Finally, we see that Early Spiral CAS still concentrates galaxies with intermediate properties,

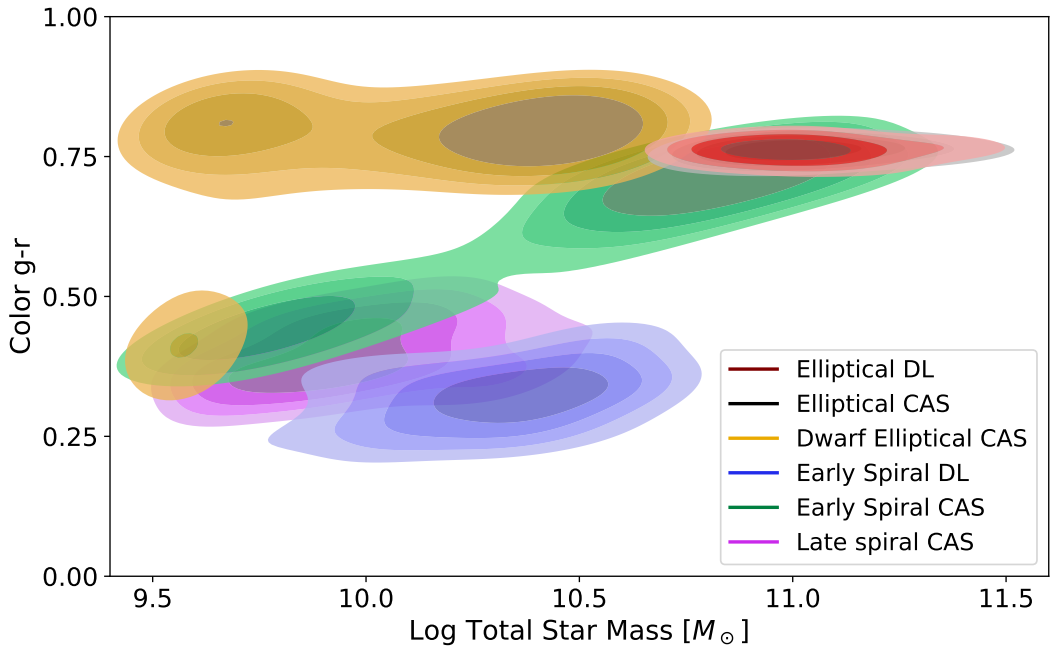


Figure 4.5: Color(g-r) vs. Star mass diagram, including both CAS and Deep Learning classifications. Contours represent two-dimensional density and are generated using Kernel Density Estimation (KDE). The contour levels represent regions of equal density, where the innermost areas correspond to higher concentrations of objects.

which reinforces the idea that this category includes objects whose classification is more ambiguous for DL.

Last but not least, we analyze the gas fraction vs. stellar mass plot (Figure 4.7). In this parameter space, we expect elliptical galaxies to have low gas fractions, consistent with their quiescent nature, while spiral galaxies should exhibit higher gas fractions due to their ongoing star formation. The overlap between the Elliptical CAS and Elliptical DL classifications is once again evident. As noted in previous sections, both classifications display a bimodality in the gas fraction, suggesting that this behavior arises from the treatment of gas in the simulations rather than from the classification criteria themselves. Future work will investigate whether this bimodality is related to the density and temperature of the gas, or possibly to gas associated with the circumgalactic medium (CGM).

No major differences are observed between the gas fraction ranges of the Early Spiral DL and Late Spiral CAS classifications. However, an isolated group of Dwarf Elliptical CAS galaxies with unusually high gas fractions stands out, likely corresponding to the bluest galaxies identified in Figure 4.5. In contrast, most Dwarf Elliptical CAS galaxies have gas fractions near zero, aligning with expectations for this morphological type. As in other parameter spaces, Early Spiral CAS galaxies display a dispersed distribution, spanning a wide range of gas fraction values.

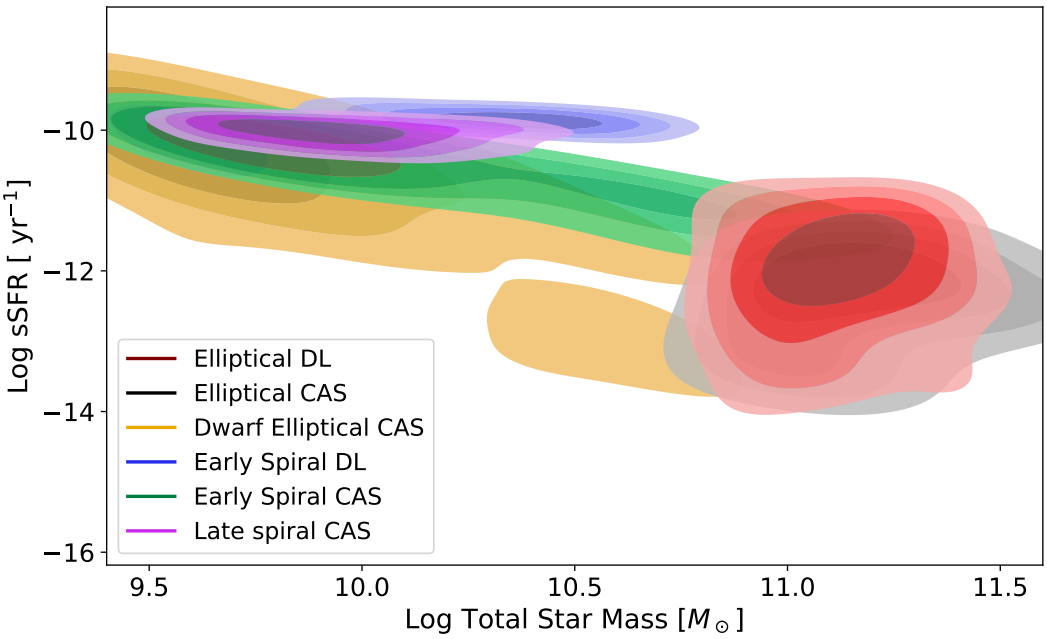


Figure 4.6: sSFR vs. Star mass diagram, including both CAS and Deep Learning classifications. Contours represent two-dimensional density and are generated using Kernel Density Estimation (KDE). The contour levels represent regions of equal density, where the innermost areas correspond to higher concentrations of objects.

Now, referring to numbers; for elliptical galaxies, DL identifies 314 objects and CAS 160, with an overlap of 69 objects, which represents a 43.1% of Elliptical CAS sample. In the case of early spirals, DL classifies 641 galaxies and CAS 229, but only 3 objects coincide in both classifications, which implies an overlap of 1.3%. Finally, for late spirals classified by CAS and early spirals according to DL, the sets include 626 and 641 galaxies, respectively, with 31 common objects, which is equivalent to an overlap of 4.95% of the CAS sample. It is worth noting that the number of exact matches increases when the stringency of the DL classification is reduced. For instance, with $\alpha = 1/3$, DL identifies 722 galaxies as Ellipticals, while CAS classifies 160. Of these, 113 galaxies are common to both classifications, representing 70% of the Elliptical CAS sample. In the case of early spirals, DL categorizes 1881 galaxies and CAS identifies 229, with an overlap of only 12 objects, corresponding to 5.2%. For late spirals classified by CAS and early spirals identified by DL, the respective samples include 626 and 1881 galaxies, with 165 galaxies in common, representing 26.3% of the CAS sample.

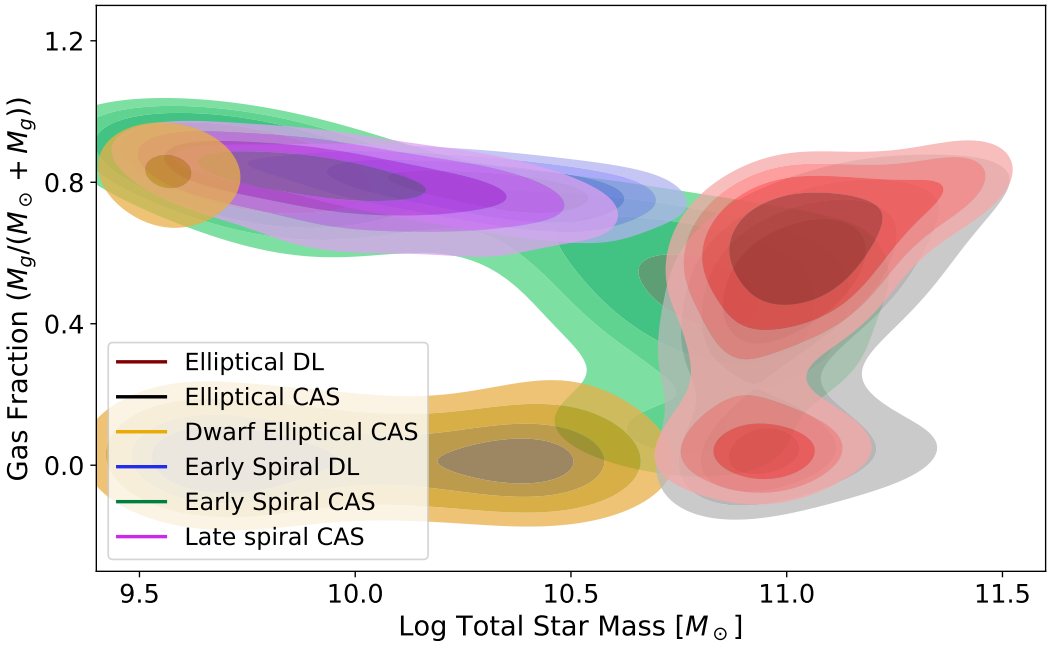


Figure 4.7: Gas Fraction vs. Star mass diagram, including both CAS and Deep Learning classifications. Contours represent two-dimensional density and are generated using Kernel Density Estimation (KDE). The contour levels represent regions of equal density, where the innermost areas correspond to higher concentrations of objects.

Chapter 5

Discussion

Both classification methods, Deep Learning (DL) and the CAS morphological parameters (Concentration, Asymmetry and Smoothness), are able to separate early-type galaxies from late-type galaxies. Furthermore, the sharp separation of morphologies in $sSFR \sim 10^{-11}[Yr^{-1}]$ in both CAS and DL classifications agrees with the literature (Muzzin et al., 2012; Wetzel et al., 2012, 2013), as does the sharp separation at color = 0.6. In fact, this matches the classification criterion selected in Nelson et al. (2017), which gives weight to both classification methods for generally separating between early and late type galaxies. However, when analyzing the performance of both methods in detail, certain inherent limitations appears, which merit further discussion.

In general terms, for the case of the Deep Learning (DL) method, the separation between early- and late-type galaxies is good at relatively high stellar masses, where particle number and resolution are sufficient to allow accurate classification. However, the DL model shows difficulties in discerning fine details, especially at low stellar masses, where particle number and resolution decrease, with the consequence that the model assigns low probability values to the low similarity of the simulated galaxies to real galaxies visually, added to this is the possibility of a bias of human origin, since the model is trained with manually labeled galaxies (Medina-Rosales et al., 2024). This is in agreement with the observations made in Huertas-Company et al. (2019), where the limitations of the DL model to discern fine details are also emphasized.

In addition to what was mentioned in the previous paragraph, let us remember that there is a difference in the filters in the synthetic images used to obtain the CAS parameters (filter i) and to train and use the DL model (filter r), which may be introducing limitations through a chromatic bias in the results, since there are two different regions of the electromagnetic spectrum. However, we rule out that this filter difference significantly biases our results, as the CAS measurements are performed exclusively on i-filter images, while the DL classification relies solely on r-filter images. Since there is no direct crossover of data between the two methods, we largely dismiss the possibility of serious bias effects related to the choice of filters. However, this does not detract from

the fact that there may be limitations in both classifications due to the aforementioned difference in capturing different aspects of galaxies.

Regarding the fact that the DL classifications cover larger ranges of asymmetry and smoothness (Fig. 4.3), we can relate this to a direct consequence of the high probability threshold we are defining. It is logical that DL is assigning higher probabilities to those galaxies with more pronounced features, as well as to galaxies that have a larger number of particles (higher resolution). This means that if we lower the threshold of probability in the DL classification, they should cover CAS values closer to those presented by the galaxies being classified with CAS, as well as a more extended distribution towards lower masses. Now, in respect to CAS classifications, rather narrow ranges of values are being used to ensure the purity of the sample when classifying galaxies. However, as a consequence, many galaxies that do not fall within the CAS classification ranges are excluded from the sample, i.e. the completeness is reduced.

Now, let focus only on the behavior of early-type galaxies. It was already mentioned how the DL classification is limited to assign a high probability level to galaxies with more marked features and with better resolution (higher mass or amount of particles), and how the CAS classification uses a very limited range of values to ensure the purity of the sample. The DL constraint clearly explains the sharp cutoff at $M_\star = 10^{10.5}[M_\odot]$ for the DL elliptical galaxy distribution, present in all parameter space plots (Fig. 4.5, 4.6, 4.7), as well as in the preliminary results (histograms on how physical properties are distributed in the DL classifications, 4.1). Interestingly, this cut-off is also repeated in the CAS elliptical galaxy classification, in addition to a very clear overlap between both classifications as far as parameter space is concerned. Now, if we calculate the exact overlap—where the same objects are classified in exactly the same way—the overlap may not appear particularly large, with 43.1% of the Elliptical CAS sample being the best case. However, when viewed in parameter space, the overlap is more notable. For example, both methods consistently place Elliptical galaxies in the red sequence of the color-stellar mass diagram and the low-sSFR region of the sSFR-stellar mass diagram.

This behavior means, by one side, that CAS elliptical galaxies are covering objects that do not have a sufficient level of probability in the DL method, and on the other hand, that DL is including objects that fall outside the CAS classification ranges, i.e., that do not present values of CAS parameters so distinctive as to be included in a pure sample. This means that, while we can rely on both classifications, it is best to use them to complement each other. By lowering the probability threshold in the DL classification, we observe an increase in the number of direct matches between Elliptical CAS and Elliptical DL galaxies. This result further reinforces the idea discussed above.

Now, with respect to the Dwarf Elliptical CAS, we can clearly see in Figures 4.5, 4.6 and 4.7, what was being suggested from the cross-comparison of classifications (Fig. 4.4); that is, that

Dwarf Elliptical CAS is being able to effectively classify galaxies that have too low probability to be classified according to the DL model, due to the low amount of particles at low masses, which translates into low resolution. This idea is supported by the fact that as the DL probability requirement is reduced, direct overlaps between Elliptical DL and Dwarf Elliptical CAS begin to appear. Additionally, some outliers can be observed within the Dwarf Elliptical CAS population, characterized by noticeably blue colors and unusually high gas fractions. This brings us to the broader discussion of gas content across different classifications. As highlighted throughout the results (Fig. 4.7), early-type galaxies—such as Elliptical CAS, Elliptical DL, and Dwarf Elliptical CAS—consistently show a non-negligible amount of gas, an observation that warrants further attention.

In general, we think this behavior arises because the TNG simulation contains a larger amount of gas than expected, which can even lead to the overly efficient formation of bulges in low-mass galaxies, as demonstrated in Huertas-Company et al. (2019). However, for Dwarf Elliptical CAS galaxies, this behavior is not particularly anomalous, as such galaxies are known to sometimes retain significant gas fractions and exhibit ongoing star formation. That said, we believe caution is needed when interpreting the reasons behind the observed gas fractions, as this requires a more detailed investigation beyond the scope of this work. In future studies, it would be valuable to measure the fractions of hot, warm, and cold gas for each morphological type to better understand these trends.

In the other side, as we saw in the results (Fig. 4.5, 4.6, 4.7), there is a closer connection between late spiral CAS galaxies and early spiral DL galaxies than between early spiral DL galaxies and their direct CAS counterpart. Although it is expected that late spiral CAS galaxies are concentrated in regions of lower stellar mass compared to early spiral DL galaxies, it is intriguing that late spiral CAS galaxies occupy a redder area in the stellar mass-color plot than early spiral CAS galaxies (fig. 4.5). In addition, late spiral CAS galaxies and early spiral DL galaxies show similar gas fractions (fig. 4.7), which reinforces their potential to be used in a complementary way to classify late-type galaxies.

Now, it is worth to discuss the Early Spiral CAS classification separately, due these galaxies are distributed in an area where transient morphologies are expected to be found (Fig. 4.5). They show high concentrations at the lower edge of what we can identify as the red sequence and at the upper edge of what we could call the blue cloud, thus placing them between the predominant morphologies. Therefore, at best, this category could be renamed as lenticular or intermediate class, and be used as a complement to DL.

On the other hand, when considering the positioning of the Early Spiral CAS and Late Spiral CAS categories in the parameter spaces (Fig. 4.5, 4.6, 4.7), we notice that the CAS classification seems to assign "later" morphological labels than what is apparent in the parameter space plots.

This discrepancy may result from how "late" and "early" are defined within the CAS framework, and could suggest a need to refine these thresholds for more accurate comparisons in the future. Additionally, we believe this behavior may also stem from the low resolution at low masses in the simulations, which introduces clumpiness and biases the CAS parameters toward interpreting the morphologies in this way.

We further investigate the fact that, in Deep Learning (DL), the separation between early- and late-type galaxies is highly accurate, though it struggles to distinguish finer subclassifications. This is evident even when grouping probabilities to define "early" as S0 + Ellipticals and "late" as Sab + Scd. Such grouping significantly increases the number of objects with high classification probabilities (Figure 5.1), although it comes at the cost of losing subclassification detail. Given the DL model's limitations in distinguishing between subclasses, it seems more practical to focus on the broader Early/Late classification, as this approach ensures greater reliability and utility for future analyses.

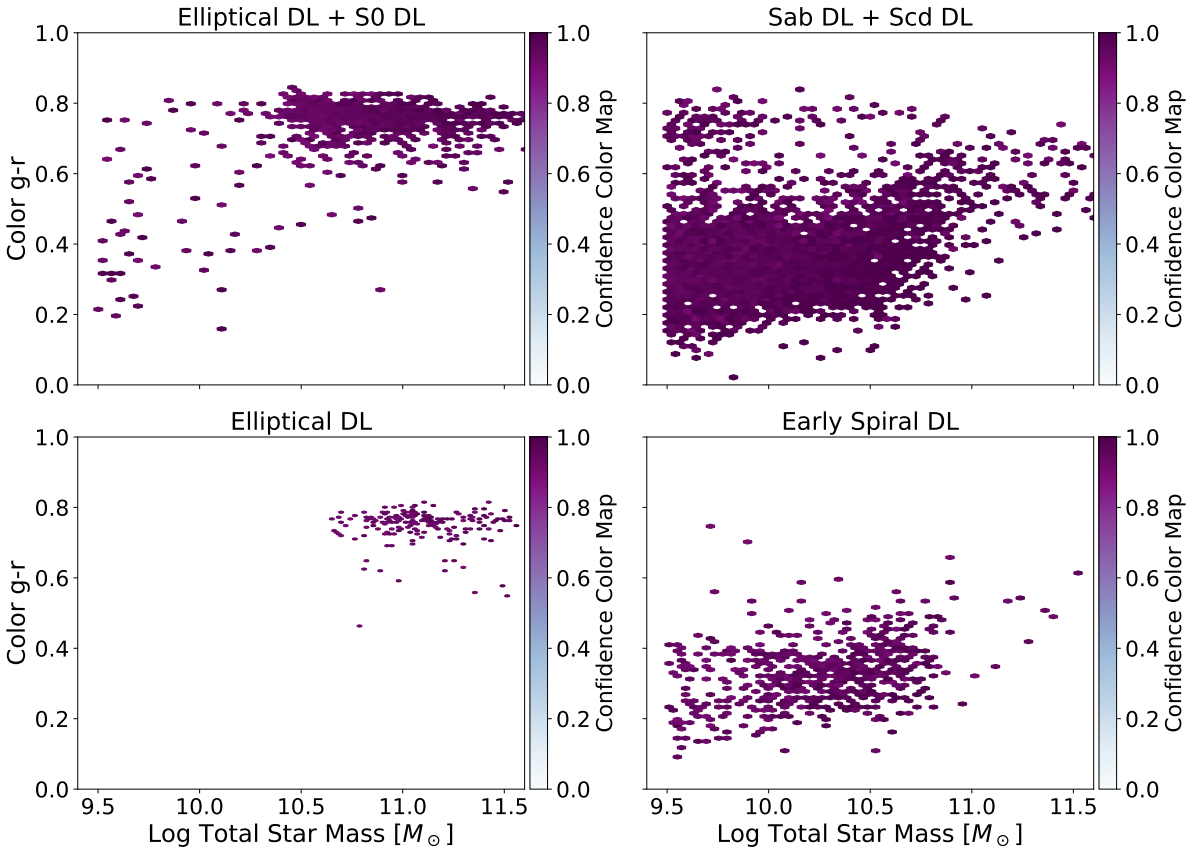


Figure 5.1: Distribution of galaxy populations in the color-mass diagram. Each panel represents a different morphological classification scheme: *Upper left panel:* Ellipticals + S0, *Upper right panel:* Sab + Scd, *Lower left panel:* Ellipticals (early, excluding lenticulars), and *Lower right panel:* Early spirals. The hexagons indicate the object density, and the color intensity reflects the probability in the assigned morphological classification. In all plots, only objects referred previously as "high probability" have been included, improving the visibility of the figures.

Chapter 6

Conclusions and future perspectives

The analysis of the morphological parameters considered in this thesis has shown the strengths and limitations of the classification methods based on Deep Learning (DL) and CAS (Concentration, Asymmetry and Smoothness) indices, showing significant differences in their behavior. Both methods are effective for separating early and late-type galaxies, with clear limits in color = 0.6 and $sSFR = 10^{-11}[Yr^{-1}]$, which is in concordance with the literature (Nelson et al., 2017; Muzzin et al., 2012; Wetzel et al., 2012, 2013); however, DL shows limitations for low stellar mass galaxies due to the limited resolution of the simulations, which reduces the probability in their finer classifications. Nevertheless, when high-probability classifications are obtained, the DL works well for separating general types, although it faces difficulties in distinguishing subclasses such as lenticular and early and late spirals, a fact also remarked in Huertas-Company et al. (2019). In contrast, the CAS method provides a more detailed classification in terms of structural parameters, but its reliance on narrow ranges to ensure purity may exclude galaxies that do not fit perfectly, affecting the completeness of the sample. A key discrepancy between the two methods lies in the classification of low-mass galaxies: while DL tends to classify them with low probability or not at all, CAS may assign later morphological labels, probably due to clumpiness and the impact on asymmetry and smoothness parameters associated with low resolution.

In simple terms, both Deep Learning (DL) and CAS indices can be relied upon to classify elliptical galaxies with $M_\star \geq 10^{10.5}[M_\odot]$, and in fact, it is best to use them in a complementary way. Additionally, the CAS index for classifying lower-mass elliptical galaxies can overcome the limitations of DL, allowing reliable labels to be assigned to low-mass spheroidal galaxies or those with low resolution in the simulation ($M_\star \leq 10^{10.5}[M_\odot]$). The 'Early Spiral' classification criterion in CAS seems to group mainly objects with intermediate properties, which turn out to be too ambiguous for DL; this category could be renamed 'Lenticular' or 'intermediate class' and used as a complement to DL. On the other hand, the classifications of 'Late Spiral' in CAS and 'Early Spiral' in DL can be used in a complementary way to classify late-type galaxies, although more

detailed studies are needed to achieve a more precise classification.

As an improvement of this work, we propose to incorporate additional morpho-kinematic parameters, such as the Gini and M20 indices, the Sersic index, and dynamical parameters such as the specific angular momentum and ellipticity, which would allow a more complete view of the structural and dynamical characteristics of galaxies, and a more accurate classification. By combining these parameters with CAS indices and Deep Learning (DL) techniques, we aim to develop a combined and robust criteria to effectively separate early and late type galaxies, and their respective subclasses. Also, it would be very interesting to perform a similar study, but in a higher resolution simulation, such as TNG50, to refine the proposed classification criteria.

In addition to the above, there is the need to re-test the same classification criteria, this time using synthetic images with the same filter for both CAS and DL. Unfortunately, this data does not exist at present, so this is a future work that remains to be done. The images used to train the DL model, nor the synthetic observations with r filter, are not public either, so it remains pending to review these images for a better understanding of how the model is classifying. A significant improvement would be to ensure that the results of DL models are free from human bias, as suggested by Medina-Rosales et al. (2024). Additionally, incorporating a desbiased-human visual classification as a reliable reference sample could serve as a benchmark to validate automatic classification methods and provide an extra layer of quality control for the results.

Finally, this work also highlights the importance of using multiple complementary methods for the morphological classification of galaxies, combination of different techniques and the inclusion of additional parameters, which is useful for the advance of our understanding of the morphological diversity of galaxies and the underlying physical processes that shape them. All within the framework of the IllustrisTNG cosmological simulation, specifically TNG100.

Bibliography

Aragón-Salamanca A., Ellis R. S., Couch W. J., Carter D., 1993, Monthly Notices of the Royal Astronomical Society, 262, 764

Baes M., Verstappen J., Looze I. D., Fritz J., Saftly W., Pérez E. V., Stalevski M., Valcke S., 2011, The Astrophysical Journal Supplement Series, 196, 22

Bergh S., 1998, Galaxy Morphology and Classification. Cambridge University Press

Boselli A., Gavazzi G., 2006, , 118, 517

Boselli, A. et al., 2019, A&A, 631, A114

Bruzual G., Charlot S., 2003, Monthly Notices of the Royal Astronomical Society, 344, 1000

Camps P., Baes M., 2015, Astronomy and Computing, 9, 20

Conselice C. J., 2003, The Astrophysical Journal Supplement Series, 147, 1

Conselice C. J., 2006, Monthly Notices of the Royal Astronomical Society, 373, 1389

Conselice C. J., 2020, in 2514-3433, The Cosmic Evolution of Galaxy Structure. IOP Publishing, pp 4–1 to 4–23, doi:10.1088/2514-3433/abb602ch4, https://dx.doi.org/10.1088/2514-3433/abb602ch4

Crain R. A., et al., 2015, , 450, 1937

Curtis H. D., 1918, Publications of Lick Observatory, 13, 9

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, The Astrophysical Journal, 292, 371

Dubois Y., et al., 2014, , 444, 1453

Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M., 2013, Monthly Notices of the Royal Astronomical Society, 434, 282

Holmberg E., 1958, Meddelanden fran Lunds Astronomiska Observatorium Serie II, 136, 1

Hubble E. P., 1926, The Astrophysical Journal, 64, 321

Hubble E. P., 1936, Realm of the Nebulae

Huertas-Company M., et al., 2019, Monthly Notices of the Royal Astronomical Society, 489, 1859–1879

Jaffé Y. L., Smith R., Candlish G. N., Poggianti B. M., Sheen Y.-K., Verheijen M. A. W., 2015, Monthly Notices of the Royal Astronomical Society, 448, 1715

Larson R. B., Tinsley B. M., Caldwell C. N., 1980, The Astrophysical Journal, 237, 692

Lotz J. M., Primack J., Madau P., 2004, The Astronomical Journal, 128, 163

Marinacci F., et al., 2018, Monthly Notices of the Royal Astronomical Society, 480, 5113

Medina-Rosales E., Cabrera-Vives G., Miller C. J., 2024, Monthly Notices of the Royal Astronomical Society, 531, 52

Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution

Morgan W. W., Osterbrock D. E., 1969, , 74, 515

Muzzin A., et al., 2012, , 746, 188

Nair P. B., Abraham R. G., 2010, The Astrophysical Journal Supplement Series, 186, 427

Nelson D., et al., 2017, Monthly Notices of the Royal Astronomical Society, 475, 624

Nelson D., et al., 2021, The IllustrisTNG Simulations: Public Data Release (arXiv:1812.05609), https://arxiv.org/abs/1812.05609

Pakmor R., Springel V., 2013, Monthly Notices of the Royal Astronomical Society, 432, 176

Pakmor R., Bauer A., Springel V., 2011, Monthly Notices of the Royal Astronomical Society, 418, 1392

Pasquali A., Gallazzi A., Fontanot F., Van Den Bosch F. C., De Lucia G., Mo H. J., Yang X., 2010, Monthly Notices of the Royal Astronomical Society, 407, 937–954

Peebles P. J. E., 1982, The Astrophysical Journal Letters, 263, L1

Pillepich A., et al., 2017a, Monthly Notices of the Royal Astronomical Society, 473, 4077–4106

Pillepich A., et al., 2017b, Monthly Notices of the Royal Astronomical Society, 475, 648

Rodriguez-Gomez V., et al., 2019, Monthly Notices of the Royal Astronomical Society, 483, 4140

Schaal K., Springel V., 2014, Monthly Notices of the Royal Astronomical Society, 446, 3992

Schaye J., et al., 2015, , 446, 521

Schneider P., 2015, Galaxy evolution. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 521–571, doi:10.1007/978-3-642-54083-7˙10, https://doi.org/10.1007/978-3-642-54083-7_10

Springel V., 2010, Monthly Notices of the Royal Astronomical Society, 401, 791

Springel V., et al., 2017, Monthly Notices of the Royal Astronomical Society, 475, 676

Varma S., et al., 2021, Monthly Notices of the Royal Astronomical Society

Weinberger R., et al., 2016, Monthly Notices of the Royal Astronomical Society, 465, 3291

Wetzel A. R., Tinker J. L., Conroy C., 2012, , 424, 232

Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, , 432, 336

White S. D. M., Rees M. J., 1978, Monthly Notices of the Royal Astronomical Society, 183, 341

White S. D. M., Davis M., Efstathiou G., Frenk C. S., 1987, Nature, 330, 451

de Vaucouleurs G., 1963, The Astrophysical Journals, 8, 31