VICERRECTORÍA DE INVESTIGACIÓN Y POSTGRADO
DIRECCIÓN DE POSTGRADOS Y POSTÍTULOS

FACULTAD DE CIENCIAS
DEPARTAMENTO DE FÍSICA Y ASTRONOMÍA

A STATISTICAL STUDY OF LOPSIDED GALAXIES IN THE ILLUSTRIS TNG SIMULATION USING MACHINE LEARNING ALGORITHMS

Tesis presentada para optar al Grado Académico de Magíster en Astronomía.

AUTOR: VALENTINA FONTIRROIG ROJAS

LA SERENA, CHILE, MARZO 2025

CONSTANCIA

Don Facundo A. Gómez & Don Marcelo Jaque Arancibia.

HACEN CONSTAR:

Que el trabajo correspondiente a la presente Tesis de Magíster, titulada "A Statistical Study of Lopsided Galaxies in the IllistrisTNG Simulations Using Machine Learning Algorithms", ha sido realizada por Doña Valentina Fontirroig Rojas.

Para que conste y en cumplimiento de las normativas vigentes de la Universidad de la Serena, Chile, firmo el presente documento en La Serena, Chile, Marzo de 2025.

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN ASTRONOMÍA

TÍTULO : A STATISTICAL STUDY OF LOPSIDED GALAXIES IN THE ILLUSTRISTNG SIMULATION USING MACHINE LEARNING ALGORITHMS.
PRESENTADA POR : VALENTINA FONTIRROIG ROJAS
DIRECTOR DE TESIS : FACUNDO A. GÓMEZ & MARCELO JAQUE ARANCIBIA

TRIBUNAL CALIFICADOR

El tribunal de tesis, conformado por:

PRESIDENTE :
MIEMBROS DEL TRIBUNAL: *
*
*

ACUERDAN OTORGARLE LA CALIFICACIÓN DE:

La Serena, Chile, **Marzo de 2025**

AGRADECIMIENTOS

Esta tesis fue llevada a cabo no solo a base de mi esfuerzo, pero también gracias al apoyo y confianza de mis seres queridos.

En primera instancia, me gustaría agradecer a mis supervisores, Facundo y Marcelo. Les agradezco el apoyo, sabiduría y confianza que me han brindado. Sin ustedes no podría haber seguido este camino que tanto me gusta.

También me gustaría agradecer a mi familia, a Sandra y a José. Sin su apoyo incondicional no podría haber logrado esto. También a mi pareja Aarón, sin tu apoyo, cariño y (mayoritariamente) paciencia este proceso se hubiera hecho aún más difícil. Me gustaría en esta parte también incluir a mis gatitas, Moca y Michi. Sin su cariño y amor no hubiera soportado estos años tan dificiles.

Por último, agradecer a mis amigos y compañeros por todo el apoyo y cercanía que me han dado. En especial a Javier y a Nicolás, los cuales me han otorgado su amistad y comprensión desde el primer año de licenciatura hasta ahora. Espero podamos continuar en este camino juntos.

Resumen

Las Galaxias Lopsided son galaxias de tipo tardío, las cuales tienen un disco galáctico asimétrico. Este fenómeno es causado por una distribución irregular de su masa estelar o en su luz. A pesar de ser una perturbación relativamente común, aun hay varias preguntas sin responder, en especial con respecto a su origen y la información que se puede extraer con respecto a la historia de formación de galaxias de tipo tardío.

Con la llegada de varios surveys fotométricos de multi-banda, se podrá estudiar estadisticamente esta perturbación, con información que no estaba disponible previamente. Considerando la fuerte correlación entre *lopsidedness* y las propiedades estructurales de galaxias de tipo tardío, esta tesis busca en desarrollar un método de clasificación automática entre galaxias que muestran esta perturbación y galaxias con un disco mas simétrico. Ademas, buscamos explorar si esta clasificación se puede obtener considerando solo propiedades internas, sin información con respecto al ambiente en donde se ubican estas galaxias.

Para esto, seleccionamos una muestra de aproximadamente 8,000 galaxias de tipo tardío de la simulación de IllustrisTNG, TNG50. Realizamos una decomposicion de Fourier a la densidad superficial de la masa estelar para previamente catalogar nuestra muestra entre *lopsided* o simétrica. Con esto, entrenamos y testeamos un clasificador obtenido de el algoritmo de aprendizaje automati-

zado, Random Forest, en donde solo utilizamos información con respecto a las propiedades internas de las galaxias, sin información sobre el ambiente. Exploramos distintos algoritmos para poder lidiar con el desbalanceo de nuestra muestra (65% son galaxias lopsided), seleccionando el mejor basándonos en métricas seleccionadas.

Mostramos que el algoritmo seleccionado provee una clasificacion de galaxias lopsided bastante precisa y rápida. Estos excelentes resultados obtenidos utilizando solo las propiedades internas de las galaxias, están de acuerdo con la hipótesis de que esta perturbación es un tracer de la estructura interna de la galaxia. Ademas, mostramos que resultados similares pueden ser obtenidos considerando como input características observables de estas galaxias, obtenidas mediante *surveys* fotométricos de multi-banda.

Nuestros resultados muestran que estos algoritmos permiten una clasificación rápida y certera de galaxias lopsided, incluso con información rápidamente obtenida de surveys fotométricos, permitiéndonos explorar si esta perturbación del disco pueda estar conectada con las historias de evolución especificas de estas galaxias.

Summary

Lopsided galaxies are late-type galaxies that feature a non-axisymmetric disc caused by an uneven distribution of their stellar mass, or light. Despite being a relatively common perturbation, several questions regarding its origin, and the information that can be extracted from them about the evolutionary history of late-type galaxies.

The advent of several large multi-band photometric surveys will allow us to statistically analyze this perturbation, with information that was not previously available. Given the strong correlation between lopsidedness and the structural properties of the galaxies, this thesis aims to develop a method to automatically classify late-type galaxies between lopsided and symmetric. We seek to explore whether an accurate classification can be obtain by only considering their internal properties, without additional information regarding the environment inhabited by the galaxies.

We select a sample of $\approx 8,000$ late type galaxies from the Illustris TNG50 simulation. A Fourier decomposition of their stellar mass surface density is used to label galaxies as lopsided and symmetric. We trained a Random Forest classifier to rapidly and automatically identify this type of perturbations, exclusively using galaxies internal properties. We explore different algorithm to deal with the imbalance nature of our data, and select the most suitable approach based on the

considered metrics.

We show that our trained algorithm can provide a very accurate and rapid classification of lopsided galaxies. The excellent results obtained by our classifier, trained with features that do not account for the galaxies environment, strongly supports the hypothesis that lopsidedness is mainly a tracer of galaxies internal structures. We also show that similar results can be obtained when considering as input features observable quantities that are readily obtainable from multi-bad photometric surveys.

Our results show that algorithms such as those considered allow a rapid and accurate classification of lopsided galaxies from large multi-band photometric surveys, allowing us to explore whether lopsidedness in present-day disc galaxies is connected to galaxies specific evolutionary histories.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Understanding the formation and evolution of galaxies are key aspects to constrain the current standard cosmological model, **Lambda Cold Dark Matter (ΛCDM). This model is a theoretical framework that describes the formation and evolution of the Universe (e.g. see Peebles, 1998). In this, Λ represents the cosmological constant, correlated with dark energy and the expansion of the Universe. CDM denotes cold dark matter, which is constituted by weakly interacting particles with low (non-relativistic) velocity (Belén Barreiro, 2000).**

The ΛCDM model describes that the formation of galaxies occurs by a hierarchical growth (e.g. see White and Rees, 1978; Frenk et al., 1996; Ratra and Vogeley, 2008), where small overdensities collapse, forming haloes of dark matter (DM haloes). The baryons are then "trapped" into their potential well, heated by shock and causing the gas to reach virial temperatures of the halo. At the same time this is happening, the haloes are merging with more clumps, growing in size and mass. As the gas cools down by photon emission and losing angular momentum, the gas condenses at the center of the halo, forming what we know today as galaxies.

Studying the morphology of galaxies is an important topic that gives insight into the formation and evolution of galaxies. Some examples consist of understanding the effects of the environment, origins of bars, driving mechanisms for spiral structure, among others (e.g. see Buta, 2011). Among these morphological

features, lopsidedness in late-type galaxies present a unique paradigm in studying the evolution and formation of spiral galaxies. In comparison with more symmetrical late-type galaxies, lopsided galaxies exhibit different internal properties and evolutionary paths, and can affect the dynamics of the hosting disk differently. However, the origin of lopsidedness is not as well understood as other common perturbations, such as bars and spiral arms. Thus, further understanding lopsidedness in late-type galaxies provides a powerful tool used to constrain the current cosmological model and to better understand how galaxies form and evolve over time.

In this thesis, we focus on studying lopsided galaxies in a large sample from the IllustrisTNG simulation by comparing their internal properties, excluding any information about the environment whatsoever with those from late-type galaxies that feature a more symmetrical disk. We have divided the introduction section into two parts. In the first half, we define lopsided galaxies and highlight their key differences from more symmetrical late-type galaxies. Additionally, we will trace the evolution of interest in lopsided galaxies, from the earliest studies on the topic to the most recent, illustrating why they have become such interesting objects to study the formation and evolution of spiral galaxies. In the second part, we discuss the application of machine learning (ML) algorithms in astronomy, supported by relevant examples, and we dive deeper into how Random Forest algorithms work, as it is the main algorithm we use in this thesis.

1.1 Lopsided Galaxies

Lopsided galaxies are late-type galaxies that feature a non-axisymmetric disk caused by an uneven distribution of stellar mass or light. A clear example of this asymmetry in galaxies is shown in Fig. 1.1, where it shows the M101 galaxy, or pinwheel galaxy, image obtained from the Hubble Space Telescope. M101 is a face-on spiral galaxy that shows a clear non-axisymmetric distribution on its galactic disk, where the lower right side of the galaxy is more extended than the upper left side, and the galactic center seems to be off-centered from the galaxy's center of mass.

Figure 1.1: M101 galaxy. Example of a lopsided galactic disk. Source: Hubble Space Telescope/NASA.

Disk asymmetry is a phenomenon that has been addressed in quite early studies, where it has been known that not all spiral galaxies have a "perfect" or completely circular disk. For instance, Sandage (1961) presented an atlas for the images of 176 galaxies obtained from blue-sensitive plates, aiming to explain their morphologies and differences based on Hubble's galaxy classification (commonly known as the tuning fork diagram). This atlas showed that not all spiral galaxies exhibit a symmetric disk structure. Considering that the structure of galaxies can be also linked with the distribution of neutral hydrogen (HI), different early studies regarding HI in spiral galaxies have also noted an asymmetrical distribution in their galactic disk. An example of this is by Beale and Davies (1969), where they studied the HI distribution of M101 obtained from the 21-cm line from the Mark I radio telescope. In this study, they found that there is almost two times more HI on the north side than on the south of the nucleus, which clearly represents an uneven distribution in the galactic disk. **An example of this can be seen in Fig. 1.2, which shows the optical and HI feature of the M101 galaxy.** This asymmetry can also be seen in the integrated spectrum over the whole galaxy, or also called global HI profile, where the flux density is

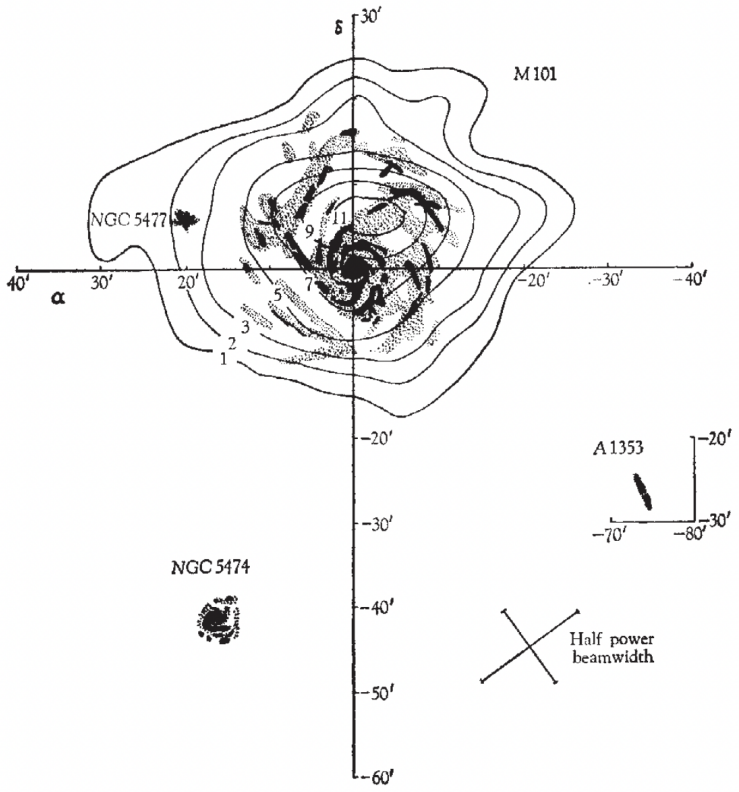

Figure 1.2: Optical and neutral hydrogen (HI) features of M101. Three nearby companions are also shown. Image obtained from Beale and Davies (1969).

much higher in the northeast part of the galaxy in comparison with the southwest side, leading to a higher difference in both halves of the velocity profile. They also found that optical features of the galaxy follow a similar distribution of the asymmetry in HI. This was also suggested by Rogstad (1971). Thus, it can be said that lopsidedness is a global phenomenon affecting the overall galactic disk, where the gas and stars are being affected by the same lopsidedness potential (Jog, 1997).

Although up until this point the asymmetry was known, not many studies were made on the topic due to the complexity in comparison with studying more symmetrical galaxies. In particular, it was not until Baldwin et al. (1980) work that lopsidedness was given a "formal" definition. In this study, they examined the HI distribution of 20 galaxies, where they defined lopsidedness as the ratio of the HI gas density from the two sides of the galaxy to be 2:1 and that the asymmetry extends over a large radial interval. Considering this definition, 6 galaxies

had high evidence for large asymmetries, especially at their outskirts, 4 galaxies did present a certain degree of asymmetry, but less than the previous 6, and the rest of the galaxies had a more disturbed HI distribution due to possible tidal interactions or no large-scale asymmetries, which leads to difficulties regarding the determination of lopsidedness.

After this, lopsided galaxies seemed to be forgotten for a few years, aside from the typical HI distribution studies for some selected galaxies. In 1994, Richter and Sacisi rekindled this topic by studying lopsidedness in the largest observational sample of late-type galaxies, up to that day. Their sample consisted of 1,738 isolated disk-like galaxies, obtained from 6 different single-dish HI surveys. With this data, they aimed to study the frequency of the asymmetries in galaxies, and to estimate the number of galaxies that exhibit it. Their main objective was to analyze the origin and evolution of lopsidedness. To define lopsidedness, they considered a similar definition as Baldwin et al. (1980), where the two halves of the global HI profile were compared with the following criteria: (a) significant peak flux differences ($\gtrsim 8\sigma$ or $\gtrsim 20\%$) between the two horns, (b) total flux differences ($\gtrsim$ 55:45%) between the low- and high-velocity halves, or (c) width differences ($\gtrsim$ 4 velocity channels or $\gtrsim$ 50[km/s]) between the two horns. This led to classifying the asymmetry in the global HI profile of galaxies as Strong, Weak, or No (insignificant) asymmetries. This classification resulted in 113 Strong galaxies, 206 Weak galaxies, and 281 galaxies with No asymmetries. A few considerations had to be made in regard to the quality of the data obtained from the surveys, e.g. excluding galaxies with distorted, peculiar looking profiles, and profiles that are plotted before baseline removal, among others. This led to a decrease in classified galaxies, where only 600 out of 1,738 galaxies were able to be classified. Nonetheless, the authors concluded that asymmetries are a common phenomenon in spiral galaxies, considering that it has an incidence of 50%. This was then confirmed by Haynes et al. (1998), where they studied a sample of 104 isolated spiral galaxies with high spectral resolution. By employing two different methods, one of which is similar to Richter and Sacisi (1994), they concluded that their significant asymmetry is also present in about 50% of the global HI profiles they analyzed.

Regarding the global HI profiles, it is important to note that the comparison between both halves of the HI profile is made by visual inspection, which could

add uncertainty in determining asymmetries. It also depends on the galaxy's projection or the "direction" of the asymmetry, reducing the overall projected perturbation of the galaxy. Thus, it is believed that this method could have underestimated the asymmetry degree in these galaxies (Bournaud et al., 2005; Jog and Combes, 2009).

Following works have also studied the number of lopsided galaxies in their samples with other types of methods, such as a Fourier decomposition of the stellar mass or light distribution, 180° radial rotation of the galaxy image, and a sector-based asymmetry analysis. These three methods are explained by the following papers:

- Zaritsky and Rix (1997) studied the asymmetries in images of 60 field spiral galaxies in the I and K' band, both obtained from the 1-m Swope telescope in Las Campanas Observatory. To measure lopsidedness, they performed a Fourier Decomposition on the disk's light distribution. The Fourier decomposition yields the A_1 parameter, referred as $< A_1 >$ and defined as the average of the ratio of the m=1 and m=0 Fourier amplitudes between a certain radial interval, which in this work is 1.5 and 2.5 scale lengths. **These Fourier modes are defined as the amplitude of the first and zeroth term of the summation, respectively.** We further explain how this decomposition analysis works in Chapter 3.1. In this case, galaxies with $< A_1 >$ values greater than 0.2 were considered to be significantly lopsided. The Fourier decomposition is highly sensitive to inclination, as it depends on the light or stellar mass distribution of the galaxy. Thus, the galaxies were considered to have kinematic inclinations of less than 32° (or $\cos i = 0.85$, with i the inclination degree of the galaxy). In total, 16 out of 60 galaxies presented significant lopsidedness, translating to a approximately 30% of the total sample.

- Conselice et al. (2000) studied the rotational asymmetry in a sample of 113 late-type galaxies images from Frei et al. (1996), to develop an unambiguous method to measure lopsidedness. These were nearby, high brightness surface galaxies, considering early elliptical and S0s to late-type spiral galaxies, irregulars and galaxies with peculiar features. The images of 31 spiral galaxies were obtained from the Palomar Observatory in the Thuan-Gunn g,r, and i

photometric bands. For the remaining 82 spiral and elliptical galaxies, the images were obtained from the Lowell Observatory in the B_j and R bands. To measure lopsidedness, they tested two similar methods which considered the normalized subtraction of the light distribution between the original image and its rotated version by the angle $\phi = 180°$. The first method, which they named "rms" asymmetry method, was defined following Conselice (1997):

$$A^2_{rms} = \frac{\sum (I_o - I_\phi)^2}{2 \sum I_o^2}$$

The second method, named as "abs" asymmetry method, was defined following Schade et al. (1995) and Abraham et al. (1996):

$$A^2_{abs} = \frac{\sum |I_o - I_\phi|}{2 \sum |I_o|}$$

In both cases, I_o is the intensity distribution of the original image for each pixel and I_ϕ the intensity distribution of the image rotated by ϕ for each pixel. The goal by using these methods is to obtain, for each galaxy, a normalized residual value between 0 and 1, where values closer to 1 represent completely asymmetrical galaxies, whereas values closer to 0 represent completely symmetrical galaxies. However, the abs-asymmetries have a better correlation with color (which gives important information about morphologies, star formation, and interactions), thus it is considered to be the main method for their work.

- Kornreich et al. (1998) studied the asymmetry in the R-band images of a sample of 32 face-on spiral galaxies, obtained from the Kitt Peak National Observatory and the Cerro Tololo Inter-American Observatory. To measure global lopsidedness they considered a geometrically-based method, in which the image of the galaxy is divided into a certain number of trapezoids, or "wedges". To do so, they subdivided each galaxy by a certain number n of equal area triangles, where their apex, defined as the vertex located between two equal sized sides and opposite to the unequal side, is located at the center of light of the galaxy. This is often considered as the pixel with the highest brightness. Each wedge is then truncated at a predetermined radius to avoid

the galactic bulge and bars, as they can introduce local asymmetries that do not count for the large-scale non-axisymmetry. Then, the magnitude for each sector was obtained. They were used to get a quantitative measure of lopsided, obtained from the maximum difference of magnitude for all wedges, ΔM_n^{max}. Considering their respective magnitude error σ and that ΔM_n^{max} probably overestimated the asymmetry, the authors reported that approximately 30% of the sample, or 10 out of 32 galaxies, were optically lopsided.

In particular, Kornreich et al. (1998) reported that the "wedge" method has a few advantages over the others. This advantages consisted of not depending on dominant even spiral modes and the inclination of the galaxy. Instead, it would be reflected as a decrease in magnitude, which is almost negligible for most galaxies. It could also measure other types of asymmetries, aside from lopsidedness, such as "boxy" or "triangular" shapes. Moreover, the authors claimed that this method is applicable in galaxies when the Fourier Decomposition method fails, mostly due to its definition, where symmetry is only dependent on the disk's radius. Such cases can be one-armed spiral galaxies and m=1 spiral galaxies that are qualitatively symmetric but are classified as lopsided, probably due to its A_1 being adjacent to the A_1 threshold that separates both types of galaxies. However, this particular issue was addressed by other studies, such as Jog (1997), Angiras et al. (2006), and van Eymeren et al. (2011), where the Fourier decomposition is only considered to be a representative measure of the global lopsidedness if the phase ϕ_1 of the m=1 mode remains constant over large radii, even if the magnitude of A_1 has a noisy behavior.

As different methods of measuring lopsidedness have been proposed, it is important to ensure a similar method to thoroughly compare the results of different studies. In regard to this, Jog and Combes (2009) proposed the use of the Fourier decomposition analysis to do so, as it gives a quantitative measurement, it is defined within the galactic radius, it avoids further assumptions on the light or mass distribution of the galaxy, and it is less computationally expensive on larger samples.

Considering the studies mentioned up until now, it is clear that lopsidedness is a common phenomenon in the nearby Universe, where a high percentage of galaxies in different samples present different degrees of this non-axisymmetry. More recent studies have also continued measuring lopsidedness in newer data samples obtained from more recent surveys and telescopes. For instance, Bournaud et al. (2005) measured lopsidedness in the NIR images of 149 spiral galaxies, selected from the Ohio State University Bright Spiral Galaxy Survey (OSUBSGS; Eskridge et al. 2002), in which they performed a Fourier decomposition analysis on the light distribution within 1.5 to 2.5 disk scale lengths. Considering that the mean of A_1 was 0.11, 34% of the sample presented values higher than this, i.e. 34% of the galaxies were considered to be lopsided. Zaritsky et al. (2013) also measured lopsidedness with a Fourier decomposition of the light distribution of 167 nearby galaxies from the Spitzer Survey of Stellar Structure in Galaxies (S^4G; Sheth et al. 2008), between 1.5 to 2.5 scale lengths (inner radius) and 2.5 to 3.5 scale lengths (outer radius). The latter was considered to check the behavior of lopsidedness in even larger radii. Considering the average strength of m=1 within the inner radius, $< A_1 >_i$, the authors claimed that there are many tens of percent of lopsided galaxies depending on the selected threshold to differentiate between both types of galaxies. **On that same survey, Laine et al. (2014) created a catalog of galaxies with different visual features, one of them being lopsidedness. To classify this asymmetry, they visually inspected the near-infrared images of the complete sample of S^4G, consisting of 2,352 galaxies. They considered a galaxy "asymmetric" if the outermost isophote were not elliptical. In total, 506 galaxies were considered asymmetric, or 22± 1% of the total sample. Kruk et al. (2017) studied the offsets between the stellar bar and their disks of a galaxy sample selected from the Sloan Digital Sky Survey (SDSS; York et al., 2000) data release 7 (Strauss et al., 2002; Abazajian et al., 2009). To consider a galaxy having an offset between the bar and disks, the measured offset between the photometric centers of bar and disk components has to be larger than the galaxy's full width at half-maximum of the point spread function (which describes the intensity distribution of the point source). Considering this, the resulting "offset" sample consisted of 271 galaxies. As previous works suggested (e.g. Pardy et al., 2016), bars can be correlated to lopsidedness. To check this, they measure**

this asymmetry using a Fourier decomposition. This resulted in 90% of the offset sample showing A_1 values larger than 0.05 (thus considered lopsided) and 63% of those were considered strong lopsided with A_1 values larger than 0.1. These thresholds were considered following Bournaud et al. (2005).

Simulations are an important tool to study the different processes happening in the Universe, including lopsidedness. As in this thesis we make use of the IllustrisTNG simulation to select a large sample of late-type galaxies, we select a few works that describe lopsidedness also using IllustrisTNG. An in-depth description of IllustrisTNG and their simulations is in Chapter 2.1. In particular, Watts et al. (2020) studied the HI distribution of galaxies in TNG100, as it has been previously shown that the asymmetries in the HI distributions are a common occurrence. Their final sample consisted of 10,699 galaxies, which are able to replicate the HI gas fraction scaling relations from GALEX Arecibo SDSS Survey (Catinella et al., 2018). To measure asymmetries, they used a method that quantifies the differences between the two horns in the integrated HI profile. This method is called the areal flux ratio parameter $A_{\rm fr}$ (Haynes et al., 1998), which is defined by the integrated flux in each half of the spectrum, bounded by the limits $V_{\rm max}$ and $V_{\rm min}$ and divided by the middle velocity $V_{\rm M} = 0.5(V_{\rm min} + V_{\rm max})$. The resulting $A_{\rm fr}$ yields a value from 1 onward. Values closer to 1 represent symmetrical galaxies and the larger the value, higher is the asymmetry. By considering the threshold 1.1, 1.2, and 1.3 to separate between symmetric and lopsided galaxies, 62%, 39%, and 25% of the galaxies in the sample are considered asymmetric, respectively. Łokas (2022) also make use of TNG100 to study the asymmetry of the disk's stellar component on a sample of 1,912 disk-like galaxies. However, instead of using $A_{\rm fr}$, Łokas used the Fourier decomposition on the surface brightness distribution of the stellar particles within $(1-2)R_{50}$, which is an equivalent to the radial interval used in observational studies, particularly in Rix and Zaritsky (1995) and Bournaud et al. (2005). This resulted in 161 lopsided galaxies with $A_1 > 0.1$, or 8% of the total sample.

This last work raises an intriguing question: are simulations and observational data comparable when studying lopsidedness? By selecting disk-like galaxies in TNG100, Łokas found that only 8% of their sample have a certain degree of lopsidedness. If they were using a radial interval similar to observations, how come they obtained such a low quantity of lopsided galaxies? Łokas suggested this could be caused by two possible reasons. First of all, although their sample follow observational trends (e.g. Reichard et al., 2009) where lopsided galaxies tend to be more star forming than symmetric galaxies, the overquenching effect (i.e more rapid quenching in green valley galaxies than in observations; e.g. Angthopo et al., 2021) known to happen in IllustrisTNG could be affecting the low number of lopsided galaxies. In other words, if too many galaxies stop star formation, they would be less likely to generate a lopsided perturbation on their stellar disk. Secondly, the limited resolution of TNG100 could also play a role, as it has been shown that the simulated galaxy's disk is thicker than in observations and thus, their dynamics are affected (Haslbauer et al., 2022). This causes that subtle effects on the galactic disk might not be considered, affecting the reproducibility of the lopsided disks in IllustrisTNG.

Several works have also found that lopsided galaxies show differences in their structural properties with respect to more symmetrical late-type galaxies. In particular, Reichard et al. (2008) studied a sample of 25,155 low redshift ($z < 0.06$) galaxies (including early-type galaxies) in SDSS (York et al., 2000; Stoughton et al., 2002), and showed that lopsided galaxies tend to have lower concentration and stellar mass density within their half light radius than symmetrical galaxies. This suggests that there is a correlation between lopsidedness and the structural properties of the galaxies. **More recent studies have make use of the Illustris TNG50 to study lopsidedness in the nearby Universe. TNG50 is the simulation with the smallest volume, thus it has the highest resolution out of the three simulations. This can help with the issues previously mentioned. In particular, Varela-Lavin et al. (2023) studied lopsidedness in a sample a of 240 late-type galaxies at $z = 0$. They measured lopsidedness by applying a Fourier decomposition to the stellar mass of the particles within the radial interval $(0.5 - 1.1)R_{\rm opt}$ (i.e. the radius**

where the superficial brightness profile in the V-band falls to a magnitude of 26.5 mag arcsec^{-2}). Varela-Lavin et al. (2023) found a similar strong correlation between lopsidedness and the internal properties of galaxies. Specifically, they found an anti-correlation between lopsidedness and the tidal force exerted by the inner regions on the outskirts of their galactic disk. This result indicates that less gravitationally cohesive disk galaxies are more susceptible to developing this asymmetry when exposed to external perturbations. Dolfi et al. (2023) extended this study by considering a larger sample of $z = 0$ TNG50 disk-like galaxies, located in different environments. They showed that, independently of the environment, while symmetric galaxies are typically assembled at early times ($\sim$ 8 to 6 Gyr ago), with a relatively short and intense burst of central star formation, lopsided galaxies assembled over a longer time period, with less prominent initial bursts and a subsequent milder and constant star formation rate up to $z = 0$. This results suggest that even if there are differences between simulations and observational data (e.g. different sample characteristics and radial interval when measuring A_1), their structural properties (e.g. concentration and stellar mass density, among others) follow a similar trend and, thus, they can be comparable.

Interestingly, it has also been shown that lopsidedness in the galactic disks can have a significant impact in the dynamics and evolution of the host galaxy. This effect can cause enhanced star forming regions, fueling the central active galactic nucleus, redistributing matter, among others (e.g. Jog and Combes, 2009).

Despite lopsided galaxies being an ubiquitous object in the nearby universe and showing significant structural differences in comparison with more symmetrical galaxies, this asymmetry has received less attention than other commonly studied perturbations (e.g. Sellwood, 2013; Conselice, 2014; Erwin, 2019). Moreover, the origin of this asymmetry is not quite well understood, as both galaxies in the field and in denser environments present lopsidedness. Different mechanisms have been proposed as the main driver of this asymmetry, raising the question of whether lopsidedness is a consequence of, for example, internal processes in the

galactic disk, or if it is caused by interactions between galaxies in denser environments. The following mechanisms have been proposed as some of the possible main drivers of this asymmetry:

Lopsided elliptical orbits. To explain the asymmetry in the HI distribution of their sample, Baldwin et al. (1980) proposed that lopsidedness is associated with a lopsided pattern of elliptical orbits. This hypothesis emerged because alternative mechanisms (e.g. tidal interactions or gas accretion) failed to account for the longevity of asymmetries in isolated galaxies. Considering previous studies of spiral arms formation, Baldwin et al. proposed a different mechanism as the differential rotation of the galactic disk would wind up the asymmetries in one or two rotation periods, which do not account for the observations of lopsided galaxies. To explain it, they considered elliptical orbits, as these orbits create density variations in gas and stars due to them moving closer and farther from the galactic center. If they are aligned in a specific pattern, a stable lopsided distribution emerges in their apogalactica (farthest point in the orbit from the galactic center). As the precession rate in elliptical orbits $\Omega - \kappa$ is negative (where Ω is the angular velocity of material orbiting the center of the galaxy and κ is the epicyclic frequency of said material, **defined as the oscillation frequency of a perturbed material or, in other words, how fast they oscillate in and out from the galactic center)**, the lobes of the elliptic orbits rotate backwards in comparison with the rotation of the galaxy, which resulted in a much more slower precession than the overall galaxy's differential rotation. Considering this, the resulting wind-up time between two points (near the apocentre and pericentre) is given by $2\pi/\Delta(\kappa - \Omega)$, where $\Delta(\kappa - \Omega)$ is the difference in precession rates across radii. In other words, this is the time scale where the differential rotation disrupts the asymmetric structures. For a flat rotation curve (constant velocity), the differential shearing of the lopsided pattern have an epicyclic frequency of $\kappa = 1.414\Omega$, resulting in a winding-up time of $\approx 5(2\pi/\Delta\Omega)$. This means that the time it takes for the asymmetric pattern to wind-up is 5 times slower than the material arms. For the outer parts, this would take 5.410^9yr, which is way more than what lopsidedness lasts caused by tidal interactions. However, this resulting time was not sufficient to account for the origin of some observed asymmetries, as it is still less than the life time of the galaxy. Furthermore, the origin of these orbits is not clear. Even so, it gives an insight on the longevity of lopsidedness.

Tidal encounters. Galaxies in denser environments, like groups or clusters, are more likely to interact with each other. These interactions could be in the form of distant interactions, such as flybys, or closer interactions, such as minor and major mergers. However, all types interactions can influence the dynamics and evolution of the galactic disk. In the first paper addressing lopsidedness, Beale and Davies (1969) proposed that the non-axisymmetry observed in M101 was caused by nearby companions, a pattern also seen in other similar systems, such as the Milky Way and M31. However, Zaritsky and Rix (1997) argued that lopsidedness is not necessarily caused by nearby companions, but could instead result from mergers or an interaction that resulted in the companion receding far from the host galaxy, enough to not be detected in the observed images. An intriguing finding from mergers, proposed by Zaritsky and Rix, is that there is a correlation between lopsidedness and enhanced star forming regions triggered by mergers. To investigate this, they checked correlation between the A_1 parameter and the mass-normalized color ΔB **, which is a proxy of star formation activity.** ΔB **is defined as the difference between the observed and predicted blue magnitude, both calculated by the Tully-Fisher relation (Tully and Fisher, 1977) which uses the HI line width. For the predicted blue magnitude, however, it is calculated by considering pure Hubble flow (motion of galaxies due to the expansion of the Universe) with** $H_o = 75\mathrm{km\,s^{-1}Mpc^{-1}}$**.** A_1 **and** ΔB were found to be correlated by the Spearman rank correlation test with a confidence of 96%. This can be explained by the mechanism responsible for lopsidedness also affecting the stellar populations. Walker et al. (1996) also proposed a similar result, suggesting that asymmetries could be induced by satellite accretion in a minor merger process. These interactions can also affect the morphology of a galaxy, leading to the creation of spiral arms and asymmetries in the galactic disk. Taking this into account, Rudnick et al. (2000) considered minor mergers to be the possible main mechanism for lopsidedness in their sample. Furthermore, following a similar path to Zaritsky and Rix (1997), they also proposed an overall correlation between lopsidedness (and thus interaction) and recent (≤ 0.5Gyr) star formation histories, as well as current ($\leq 10^7$yr) star formation rates.

Tidal encounters were thought to be the most accepted explanation for the origin of lopsidedness. Yet in some cases, galaxies do not seem to have significant companions capable of triggering such asymmetries. Bournaud et al. (2005) reached this conclusion by studying the effect of tidal forces from companions on each galaxy of their sample. As mentioned beforehand, this sample consisted on 149 spiral galaxies from OSUBSGS. Galaxies were considered to be companions if they had a radial velocity within 500km s^{-1} and they were within 2.5 degrees on the sky from the main galaxy. This information was obtained from the NED database. By comparing the companions' tidal effects exerted to the main galaxy and the main galaxy's A_1 parameter, it was clear that there was no correlation. **This was calculated using a tidal parameter[1], which quantifies their effects of tidal forces by considering the mass of the host galaxy (M_o), its scale length (R_o), the sum of the companions' mass (M_i), and their respective projected distances (D_i).** As mentioned by Bournaud et al., this conclusion is also in agreement with Wilcots and Prescott (2004), where they study the HI distribution of 14 Magellanic spiral galaxies. From observations, only 4 of the 14 galaxies showed companions. However, only 2 of them have companions that are interacting and causing lopsidedness. They concluded that lopsidedness is rather long-lived and not related to the environment these galaxies reside in. It is important to note, however, that even if not a high percentage of lopsidedness in galaxies is caused by any form of tidal interactions, it can not be ruled out completely. There are still some cases that can explain it, as we have mentioned in the previous works.

Disk response to the distorted dark matter halo. A consequence of a distant tidal interaction between galaxies is the response of the galactic disk to their interacting DM haloes. As the haloes of two distant galaxies interact, the distorted halo applies a lopsided potential to the disk, causing it to warp or triggering non-axisymmetrical structures. This phenomenon can explain the asymmetric velocity profile in the HI gas distribution. As the gas becomes unstable in an overdense

[1]A word of caution: although named the same, the tidal parameter used in Bournaud et al. (2005) is not the same as the one we use in this thesis. In Bournaud et al.'s case, T_P is defined as $\log(\sum_i \frac{M_i}{M_o}(\frac{R_o}{D_i})^3)$, which quantifies the tidal forces exerted from the satellites to the host galaxy. In our case, the parameter we use to train and test the classifiers is the tidal parameter T_P, which is a proxy of the force exerted from the inner regions of the galaxy to its outskirts, indicating how gravitationally cohesive the galaxy is. The definition is given in Chapter 2.2.

region of a lopsided potential due to the increase in surface density, it enhances the star formation rate and creates the disparity between velocities in both halves (Jog, 1997). This phenomenon can also be attributed to a relatively weak interaction between a host galaxy and a satellite (Weinberg, 1998). As an example, Gómez et al. (2016) studied the vertical structure of a Milky Way-like simulated galaxy, finding that it develops a strong vertical pattern capable of forming a Monoceros ring-like structure. Gómez et al. proposed that this pattern is driven by an offset, or displacement, of the halo's center of mass and its density cusp (which resulted from the torque of an overdensity wake caused by the response of the DM halo to the passage of the satellite), rather than direct tidal forces from the satellite. In this case, the satellite undergoes a slow flyby, with a relatively small mass ($\sim 4\times10^{10}\mathrm{M}_{\odot}$) and low pericentric velocity ($\sim 215\mathrm{km\,s}^{-1}$ at $\sim 80\mathrm{kpc}$), making it insufficient to directly cause such distortions. However, due to its slow motion, the DM halo resonantly interacts with the satellite, inducing the formation of a density wake that amplifies the satellite perturbation (see also Laporte et al., 2018). This perturbation is then transmitted to the inner regions of the halo, affecting the embedded stellar disk and resulting in a vertical oscillation. In a similar note, Varela-Lavin et al. (2023) studied if there is a correlation between the offset of haloes and the lopsided perturbations in the disk in a much larger sample. In this, they found that a correlation between the occurrence of lopsided patterns and perturbations in the DM density field. However, there is also a large number of lopsided galaxies that have smaller DM density perturbations, similar to those of symmetric galaxies. Thus, it is important to note that the response of the galactic disk to a perturbed DM halo is not necessarily the main driver of the asymmetry in some cases.

Asymmetric gas accretion. To further investigate the origins of unusual spiral morphology, Phookun et al. (1993) studied the possible reasons for the $m = 1$ spiral structure of NGC 4254, a spiral galaxy with one arm more prominent than the others, also defined as a one-armed spiral galaxy. This particular galaxy is shown to not be interacting with any of its companions. However, it has prominent $m = 1$, $m = 3$, and $m = 5$ modes in comparison with $m = 2$ (Iye et al., 1982), and it is a photometrically normal Sc spiral galaxy with an unusual strong one-armed spiral structure in the stellar component (Schweizer, 1976) and a flat HI rotation curve (Guhathakurta et al., 1988). Aside from interactions, gas infall

in the galactic disk was the strongest candidate to produce a prominent $m = 1$ mode, most likely from a tidally disrupted gas cloud or dwarf galaxy. The possible scenario depicted by the authors is as follows: a cloud of gas orbiting NGC 4254 falls within the galaxy's tidal radius, which disrupts it and causes some of the gas to spread over the galactic disk. This causes a perturbation on the disk, resulting in a more prominent spiral arm and, thus, a prominent $m = 1$ mode. Nonetheless, for this scenario to work another amplification mechanism is needed, as asymmetric gas accretion alone is a rather weak interaction to cause such asymmetry. The authors suggested to be swing amplification, a mechanism where self-gravity and differential rotation amplify spiral density waves. The conclusions from this work show an interesting result: external subtle influences can shape the morphology of some galaxies. Bournaud et al. (2005) also reached a similar conclusion regarding asymmetric gas accretion, proposing that lopsidedness in their sample is probably caused by it. This conclusion was reached by studying a sample of simulated galaxies with a N-body simulation, which consists of galaxies described by particles of stars, gas, and DM. In this, an asymmetric gas accretion was proposed to be the driving mechanism that results in a lopsided disk, as it accounts for the previous observational properties and creates strong $m = 1$ asymmetries, fueling the gaseous disk and, thus, enhancing star formation. To explain their reasoning, Bournaud et al. considered ideal scenarios where gas accretion is fueled by one, two, and three cosmological filaments, although in reality galaxies are fueled by many more. These resulted in a long-lived strong $m = 1$ mode. In another study, Łokas (2022) also reached to the conclusion that the asymmetry in their sample was caused by an asymmetric star formation, most probably thanks to an asymmetric gas accretion. This conclusion was supported by the fact that the galaxies that present such asymmetry in their sample have, on average, more gas, higher star formation rate, lower metallicity, and bluer colors.

In summary, lopsided galaxies are a key aspect to understand the different dynamical processes that are affecting late-type galaxies in the nearby Universe. **They are an ubiquitous object in the Universe (i.e. $\sim$ 30% - 50% of late-type galaxies in observational samples have shown some degrees of this asymmetry) and,** in comparison with more symmetrical galaxies, lopsided galaxies have different internal properties, evolutionary paths, and affect the dynamics of the hosting disk. Furthermore, studying the asymmetry in a large

sample of galaxies can give important insight into the physical origin of lopsidedness in disk-like galaxies, and even further understanding the formation and evolution of spiral galaxies.

1.2 Automation in today's context

Machine Learning (hereafter ML) algorithms have revolutionized data analysis by enabling automated pattern recognition and prediction from large datasets. These algorithms, ranging from supervised learning methods (models that use labeled data to compare with the predictions) to unsupervised learning techniques (models that identify patterns in unlabeled data), provide powerful tools for uncovering internal relations from complex systems. Thanks to their versatility, they have been used in a wide range of fields, such as medicine, biology, technology, among others.

These algorithms can be subdivided between classification and regression tasks. Classifiers are algorithms that group, or categorize, data with certain characteristics into a set of classes or categories. Some examples consist of epileptic activity classification in electroencephalogram signals (Rajendra Acharya et al., 2012), classification of cancerous and non-cancerous skin lesions for early detection (Masood et al., 2014), and object detection for self-driving cars (Gupta et al., 2021).

On the other hand, regression tasks capture the underlying relation between variables, predicting a numerical value based on the inputs. A few examples of this algorithms are forecasting daily stock market return (Zhong and Enke, 2017), grouping customers based on their annual income and spending score (Nandapala and Jayasena, 2020), and anomaly detection in streamed data (Degirmenci and Karal, 2022).

In the field of astronomy, the use of ML algorithms has grown significantly due to the rapid increase in available data. However, as the volume of data grows, so does the complexity of studying diverse astronomical sources. In the following subsections, we explore into the use of ML algorithms in astronomy, describing a few popular algorithms and examples. We then describe and summarize the key

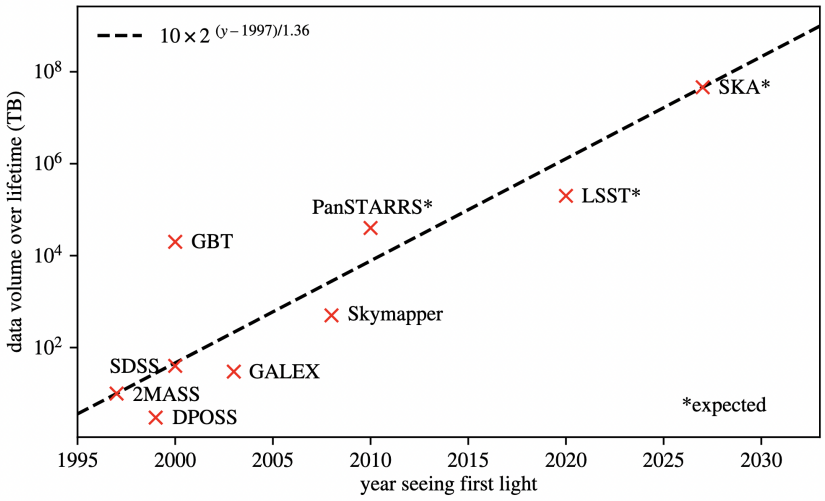

Figure 1.3: Increment of the volume of data obtained from current and future telescopes and surveys with respect their launched date. Fig. obtained from Smith and Geach (2023).

characteristics of Random Forests, algorithm central to this thesis for analyzing our sample.

1.2.1 Automation in the Astronomy Field

In the recent decades, the quantity of data provided by some of the current surveys and telescopes, such as SDSS (York et al., 2000), GAIA (Gaia Collaboration et al., 2016), Zwicky Transient Facility (ZTF; Bellm et al., 2019), James Webb Space Telescope (JWST; Gardner et al., 2006), and next-generation surveys, such as Large Synoptic Space Telescope (LSST; Ivezić et al., 2019), Square Kilometer Array (Dewdney et al., 2009), among others, has increased significantly. A clear example of the increment of data expected in the following years is shown in Fig. 1.3, obtained from Smith and Geach (2023).

As the volume of data increases, using traditional approaches to study different sources (e.g visual inspection) can be a daunting task which could result in missing important information or discoveries. To avoid this, ML algorithms have been gaining more popularity over the years in the astronomy field, which has resulted in almost a necessity to study the different processes occurring in our Universe.

The use of ML algorithms in the field of astronomy dates from the '90s, approximately. Boroson and Green (1992) studied the correlation of selected properties (e.g absolute magnitude, peaks, full width half maximum, equivalent widths of certain lines, among others) from the spectra of 89 quasi-stellar objects in the Bright Quasar Survey catalog (Schmidt and Green, 1983). This was achieved by applying a Principal Component Analysis (PCA; Hotelling, 1936; Shlens, 2014) to the properties, where they found that there is indeed a strong (anti-)correlation between the FeII and [OIII] measurements and an inverse correlation between the strength of HeII λ4686 and the optical luminosity. Odewahn et al. (1992) developed an automated classification of stars and galaxies for the Palomar Sky Survey's automated plate scanner. In this, each plate could result in $\approx$ 250,000 images, thus it was necessary to develop an automated method to classify the sources in those images. To do so, they developed and tested two different Artificial Neural Networks (ANN), a perceptron[2] and Backpropagation Neural Network (Rumelhart et al., 1986). The resulting success rate (defined as correct classification / total number of the sample $\times$100) for stars and galaxies is above 90%, which fluctuates depending on the magnitude of the sources. Galaz and de Lapparent (1997) classified the stellar spectra from the ESO-Sculptor Survey (de Lapparent et al., 1993) using PCA on their spectra. This algorithm reconstructs the spectra in a continuous spectral sequence, accounting for $\sim$ 97% of the total flux of each spectrum, where it is highly correlated to the Hubble type morphology and the galaxy's stellar populations.

From 2000 onward, thanks to the incremental in computational power, machine based algorithms started to be more considered to study different processes in the Universe. A few examples in this era consist of galaxy/star classification in wide-field images (Andreon et al., 2000) and predict galaxy morphology by using characterizing and easily ready features from SDSS (Ball et al., 2004) using Neural Networks, automatically classifying periodic variable stars from All-Sky Automated Survey 1 to 2 using an unsupervised Bayesian classifier on their light

[2]The perceptron is the simplest form of a neural network, introduced by Rosenblatt (1958). It consists an input layer directly connected to the output layer, with no in between hidden layers. The information passes through the network in a strictly forward manner. Thus, it is limited to solving linearly separable problems, failing to model complex or non-linear relationships in data.

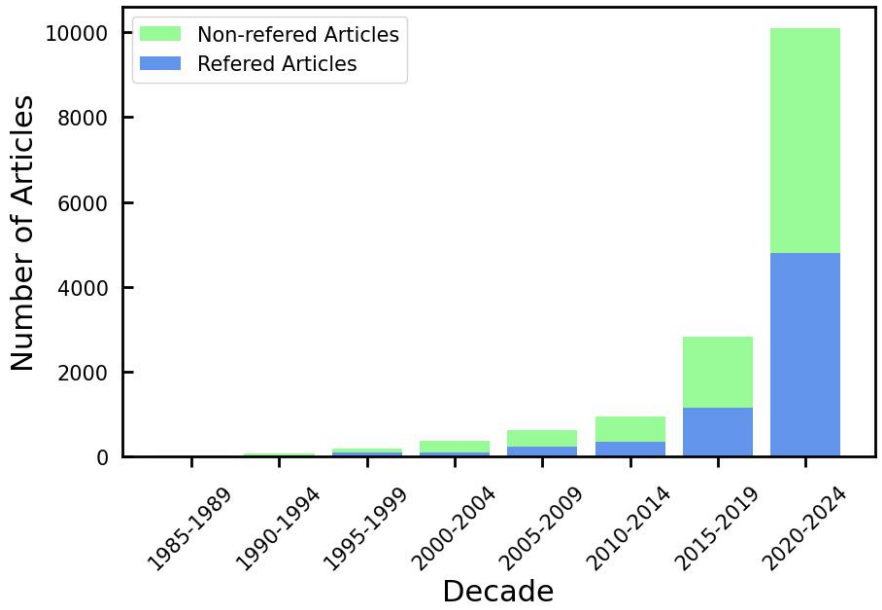

Figure 1.4: Number of referred (blue) and non-referred (green) articles in the astronomy subfield that use machine learning algorithms. Data obtained from NASA Astrophysics Data System.

curve (Eyer and Blake, 2005), and an unsupervised classification of stellar spectra using the clustering algorithm k-means (Sánchez Almeida and Allende Prieto, 2013). However, a significant shift towards the use of automation algorithms began in 2015. An example of the increment on the number of referred and non-referred articles that have used ML algorithms in the astronomy field is shown in Fig. 1.4. Data was obtained from NASA Astrophysics Data System [3]. This increment started thanks to Tensorflow (Abadi et al., 2015) and Pytorch (Paszke et al., 2017) to be more available to general public (Fotopoulou, 2024). These are hubs, or libraries, of codes for developing ML and artificial intelligence algorithms, based on python interfaces to make it more user-friendly while still being able to do computationally intensive tasks using the GPU and a C++ backend for speed. Furthermore, it was the first time an image classification algorithm was able to surpass human classification using a pre-established large dataset of daily images (He et al., 2015).

To further comprehend the use of ML algorithms in astronomy, we present a few papers as examples in the galactic/extragalactic subfield, divided by the

[3] https://ui.adsabs.harvard.edu

scientific objectives they address:

Source Classification: In this case, different sources can be classified by using observed or simulated information, such as the pixels from an image or features that are able to characterize the source. Huertas-Company et al. (2015) classified the morphology of $\sim$ 50,000 galaxies in the H-band, within $1 < z < 3$. The galaxies were selected from the five fields of the Cosmic Assembly Near-IR Deep Extragalactic Legacy Survey (CANDELS; Koekemoer et al., 2011). Using a Convolutional Neural Network (CNN; O'Shea and Nash, 2015), they were able to identify objects from galaxies, and distinguish irregular and spheroidal from disk galaxies with less than 1% in misclassification. Dieleman et al. (2015) also morphologically classified the image of galaxies using the CNN architecture. This was part of the Galaxy Challenge (AstroDave et al., 2013) from Kaggle[4], where the main goal of this challenge was to develop an algorithm capable of classifying the morphology of galaxies for upcoming (at that time) surveys. In this challenge, there were two sets of galaxies: a training set and the evaluation set (in other words, a testing set). Each set contained galaxies with a wide variety of characteristics, such as morphology, color, and size, obtained from the Galaxy Zoo 2 project (GZ2; Willett et al., 2013). By rotating and cropping the galaxies' images, they obtained an accuracy of $\sim$ 99% for the different questions of GZ2 (smoothness, edge-on, bar, spiral, bulge, anything odd, roundness, odd feature, among others).On the other hand, Sánchez-Sáez et al. (2021) developed a fast classifier for the light curves of 15 transient and variable objects from LSST, including stochastic objects such as quasi-stellar objects, blazars, and host-dominated active galactic nuclei. To classify between the different classes, they developed a two-level (algorithm explained in the following subsection) using 152 features, which included information about periodicity, colors, galactic coordinates, morphological classification, among others. The first level consisted of subdividing the data between transient, periodic, and stochastic. The second level consisted of three classifiers further subdividing the previous three main classes into subclasses, such as supernovaes, eclipsing binaries, pulsating stars, among others. The resulting metrics showed that the first level classifier had a good overall performance, with a score of 0.96, 0.99, and 0.97 for macro-averaged precision, recall, and F1-score,

[4] https://www.kaggle.com

respectively. The second level, however, the scores decreased to 0.57, 0.76, and 0.59, respectively. As discussed by the authors, this decrease in performance could be due to the imbalanced nature of their dataset, finding suitable features to separate classes, and, in some cases, the similarity in light curves posed a challenge in subdividing one class (e.g AGNs).

Redshift Estimation: Estimating photometric redshift is an important aspect to delimitate the large-scale structure of the Universe and to constrain the cosmological model (Blake and Bridle, 2005). Photometric redshift can be calculated by comparing the observed Spectral Energy Distributions (energy vs frequency or wavelength; hereafter SED) with SED templates at different redshift. This templates can be empirical, which are obtained by observations, or theoretical, simulated using stellar population synthesis models. The difference is then calculated by a χ^2-fitting, where the best fit is the one that minimizes it. A few examples on this methodology are Lanzetta et al. (1998), Bolzonella et al. (2000), and Salvato et al. (2011). However, considering the increase of multi-wavelength photometric data from surveys in previous years, other methods based on ML algorithms have been developed to estimate redshift. There are many examples of this estimation on the early 2000s. In particular, Collister and Lahav (2004) were one of the first works to estimate photometric redshift using ANNs. To achieve this, they developed a multi-layer perceptron network, trained and tested with photometric data and spectroscopic redshift from SDSS. The resulting photometric redshift predictions ($z_{\rm phot}$) were then compared with the spectroscopic redshift ($z_{\rm spec}$). By using the root mean square deviation (defined as $\sigma_{\rm rms} = \langle (z_{\rm phot} - z_{\rm spec})^2 \rangle^{1/2}$), they obtained a value of 0.0229, which is the lowest deviation in comparison with other previous similar methods. They also tested this network in fainter targets, where the main motivation of obtaining their $z_{\rm phot}$ is to avoid the difficulty (and thus expensiveness) in obtaining $z_{\rm spec}$. The resulting deviation had a slight decrease ($\sigma_{\rm rms} = 0.0327$). Vanzella et al. (2004) also developed a multi-layer perceptron to predict $z_{\rm phot}$. However, they consider a mix between observed data from SDSS (SEDs and $z_{\rm spec}$) and theoretical SEDs, resulting in an improved prediction. More current works have also measure $z_{\rm phot}$ by using more complex regression algorithms (e.g. Zhang et al., 2013; D'Isanto and Polsterer, 2018; Pasquet et al., 2019; Henghes et al., 2022).

Physical Parameters Estimation: Unsupervised algorithms can be employed in estimating physical properties of different sources. Frontera-Pons et al. (2017) applied both denoising autoencoders (DAE) and PCA to galaxies' SED to derive a data-driven diagram and thus, study their formation and evolution. They reported that DAEs were able to recover galaxy bi-modality (clear separation between star-forming and quiescent galaxies), it provides a continuous evolution on the galaxy population with respect redshift, and it shows a clear separation between distributions in regards mass (higher mass galaxies at higher redshift) by plotting the first and second DAE component. A similar behavior is seen by plotting the first and second principal components, specially regarding their mass and specific star formation rate. Mahor et al. (2023) presents an application of a CNN to estimate important parameters of interacting galaxies using images from the GalMer database (Chilingarian et al., 2010). This is a library of merger simulations. The desired parameters are the spin, the relative inclination of the galaxy, viewing angle, and the azimuthal angle, which are fundamental parameters to understand tidal formations and building dynamic models. They obtain an overall good results (R^2-score of 0.9986 and a mean absolute error of 0.4348), which demonstrates the ability of the model to generalize with this simulated data. A similar result was obtained when using real data from SDSS, with a R^2-score of 0.899.

There is no doubt that ML algorithms are an important tool that are positively impacting the astronomy field, and will continue to do as technology evolves. This mainly due to their application to different sources in large volumes of data, obtaining important information about their underlying relations. With this in mind, in this thesis we make use of the Random Forest (hereafter RF; Breiman, 2001) algorithm to achieve our goal. This is a supervised algorithm that poses a great advantage in the automation of different classification and regression tasks. For instance, it can describe different complexity relations between the parameters, or features, of a sample considering their assigned label. It can also works with a wide variety of different datasets and sizes, among other advantages. In the case of astronomy, it is clear that its use have grown as a result of the rapidly increase of data. As application examples, RF poses a great alternative to classify different sources in different wavelengths (Gao et al., 2009), estimation of photometric redshifts (Carliles et al., 2010), perform automatic classification of

light curves of variable stars (Sánchez-Sáez et al., 2021), predict underlying gas conditions of the circumgalactic medium (Appleby et al., 2023), identify galaxy mergers (Guzmán-Ortega et al., 2023) and estimate different galaxies' physical properties (Mucesh et al., 2021), among other applications.

1.2.2 Random Forests

RF is a type of supervised algorithm that poses a great advantage in the automation of different classification and regression tasks. For instance, it can describe different complexity relations between the parameters, or features, of a sample considering their assigned label. It can also works with a wide variety of different datasets and sizes, among other advantages. In the case of astronomy, it is clear that the use of ML algorithms, such as RF, have grown as a result of the significant increase of data with the current and next-generation surveys and telescopes.

RF consists of an ensemble, or collection, of decision trees. A decision tree is a tree-like predictive model composed of nodes, where the sample is recursively divided by conditions in the form of $x_i^{(j)} < \boldsymbol{X}_j$, the latter being the j-th feature and $x_i^{(j)}$ a certain threshold based on the j-th feature. In other words, decision trees divide the input space, which depends on the selected feature/s by the decision tree, to create subspaces that are able to differentiate between the different classes. A visual example of the previous description is shown in Fig. 1.5 and Fig. 1.6, following Breiman et al. (1984). Fig. 1.5 shows a diagram of how a decision tree works. In this case, we consider an initial condition in the form of $x_1 \leq 0.7$ in the first node $n_{1,1}$, where x_1 is a selected feature by the model. If data from this subsample fulfill or not this condition, it is partitioned between "Yes", to the left node $n_{2,1}$, or "No", to the right node $n_{2,2}$. In the case of node $n_{2,1}$, the data is again partitioned with the condition $x_2 \leq 0.5$, where x_2 is another feature aside from x_1. This follows the same partitioning process as before, where if the data fulfills the condition, it is spread to the node $n_{3,1}$ and if not, it is spread to node $n_{3,2}$. The nodes $n_{2,2}$, $n_{3,1}$, and $n_{3,2}$ do not continue partitioning the sample, but instead give a classification considering a probability of the samples being a certain class, 0 or 1 (or prediction in the case of a regression task). This final nodes are called leaves or terminal nodes, represented as rectangles in this case,

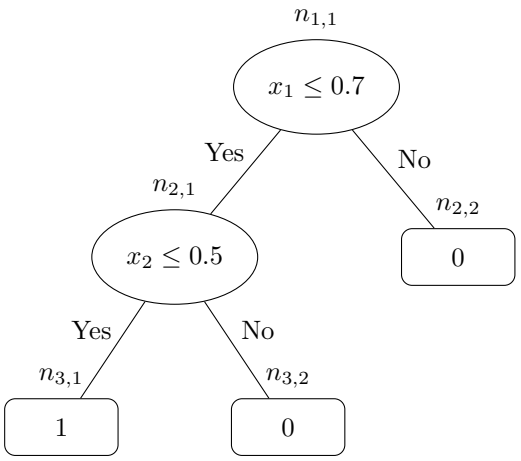

Figure 1.5: Decision Tree structure, following the example of Breiman et al. (1984).

which result by either fully partitioning the sample or until all leaves have less than the minimum quantity to split a node, which is determined by a certain parameter in the tuning process, discussed below. On the other hand, Fig. 1.6 shows another way to understand the partitions made by a decision tree. Considering the same conditions as before, it shows the final divided input space. The quality of the division, or partitioning, is measured by the "purity" of the subspace, where the purer it is, the more datapoints from the same class are assigned. This can be calculated by the Information Gain or the Gini impurity function. Both functions have the same goal, which is to determine the best split in a node, but they are calculated differently. In the case of the Information Gain function, it considers the difference between the entropy of the dataset before and after the split. Considering that the entropy is linked to the "pureness" of a dataset, where values closer to 0 are represented by data points from the same class and values closer to 1 are represented by an even distribution of classes, the information gain function can discriminate how good a split using a particular feature by checking how much the entropy is reduced. On the other hand, the Gini impurity function is calculated by the summation of the difference between each class probability, obtained from the model, where the best split is consider to be the one that minimizes the impurity. In our case, we use the Gini impurity index since it takes less computational time, and is also the default function in the selected classifiers.

Overall, this process is done firstly on a learning set, or training set, where then a new unseen dataset is propagated over the tree to predict the correspond-

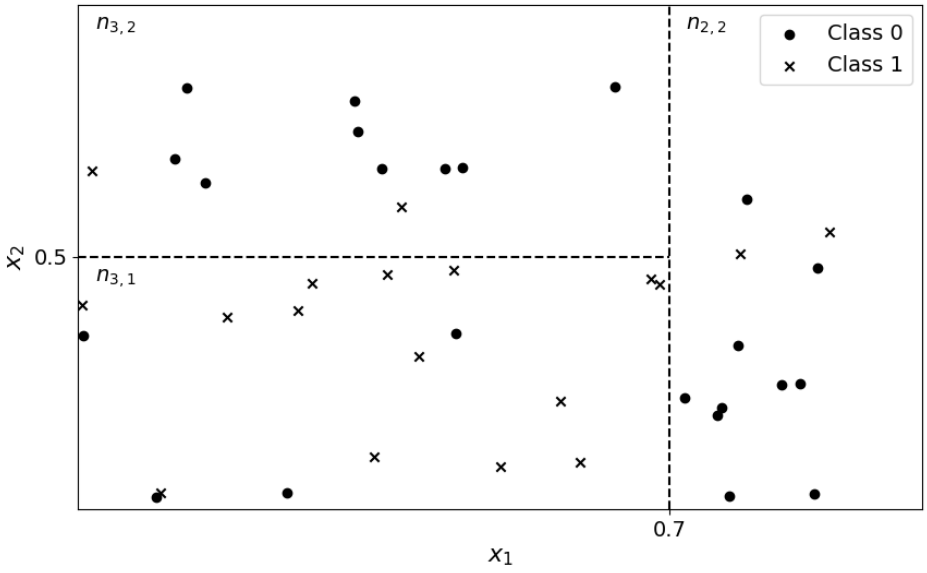

Figure 1.6: Separation of the input space, done by the conditions selected in the Decision Tree. Example obtined from Breiman et al. (1984).

ing class or numeric value.

Although decision trees have numerous advantages due to their intrinsic nature (e.g. they can be used by any kind of sample, they have an easy hyperparameter customization or tuning, and they also estimate the feature importance aside from class predictions), they are easy to overfit. This means that a decision tree may be less accurate when predicting unseen data during testing, as the model tends to overly fit to the training set. RF avoids this issue by training non-correlated decision trees, each on a subsample with replacement of the training set, thus reducing the variance while maintaining high accuracy. For a binary classification task, which is our focus, each decision tree classifies the data as either positive or negative class. Then, the final prediction of the RF is the class predicted by more than half of the trees. This method is called bagging or *bootstrap aggregating* (Breiman, 1996).

1.3 Scope of this Thesis

Current and upcoming large observational surveys, such as the Southern Photometric Local Universe Survey (S-PLUS; Mendes de Oliveira et al., 2019), Javalambre Photometric Local Universe Survey(J-PLUS; Cenarro et al., 2019), Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; Benitez et al., 2014), and LSST, will enable to identify and characterize lopsidedness in a very large number of well-resolved galaxies in the local Universe. This will be crucial to further study the connection between such perturbation and the galaxy internal properties, and to test current model predictions and understand the origin of lopsidedness considering their star formation history in relation with the environment. However, as the volume of data increases, using traditional approaches to study and characterize this non-axisymmetry (e.g., visual inspection, identification of surface brightness residuals with respect to unperturbed distributions, and Fourier decomposition) can become a limiting task. All these techniques require human supervision and intervention and thus, they result in cumbersome and slow approach to study lopsidedness in larger volumes of data, which could also result in missing important information or discoveries.

Given the previously reported strong correlation between lopsidedness and the structural properties of galaxies, this thesis aims to automatically classify galaxies between lopsided and symmetric by only using their internal properties. We also seek to explore whether an accurate classification of this asymmetry can be obtained without including any direct information regarding the environment inhabited by the galaxies. In a first step, we train and test the selected ML algorithms using late-type galaxies obtained from the cosmological simulation IllustrisTNG. In general, cosmological simulations prove to be an excellent tool to use with our classifiers, as they can model the properties and characteristics of galaxies and their environment, avoiding the need to make additional estimations to obtain them as in the case of observations. Furthermore, using simulations ensures a general framework to interpret lopsidedness in the aforementioned observational surveys and telescopes, **especially in LSST, J-PAS, or J-PLUS**, with which we can then directly apply these trained models to observational data. As a second step, we will determine the key parameters that allow the correct classification of lopsided galaxies and thorough study the different classification cases.

Chapter 2

Data

In this section, we present the criteria to select the necessary dataset to train and test our selected classification models, discussed in Sect. 3.2. In particular, we use galaxy models extracted from the fully cosmological simulation, Illustris TNG50 (Nelson et al., 2019a; Pillepich et al., 2019). For each galaxy model, we compute internal parameters that are commonly measured in observational studies to classify galaxies' morphology.

2.1 The IllustrisTNG simulations

IllustrisTNG, successor of the Illustris project (Genel et al., 2014; Vogelsberger et al., 2014; Nelson et al., 2015), is a set of cosmological, gravo-magneto-hydrodynamical simulation, ran with the moving-mesh code AREPO (Springel, 2010). IllustisTNG builds upon its predecessor model (Genel et al., 2014) by incorporating an updated physical model (Pillepich et al., 2018) which accounts for stellar evolution, gas cooling, feedback and growth from supermassive black holes, among others. In particular, the improved model for the feedback of the low accretion mode in super massive black holes resulted in a reduction of the discrepancies with observational constraints identified in the original Illustris simulations, such as the galaxy color bimodality (Nelson et al., 2018). These improvements make IllustrisTNG a powerful tool for comparisons with observational data.

IllustrisTNG consists of three simulations with different volumes: $\sim 50^3$Mpc, $\sim 100^3$Mpc and $\sim 300^3$Mpc, referred as TNG50, TNG100, and TNG300, respec-

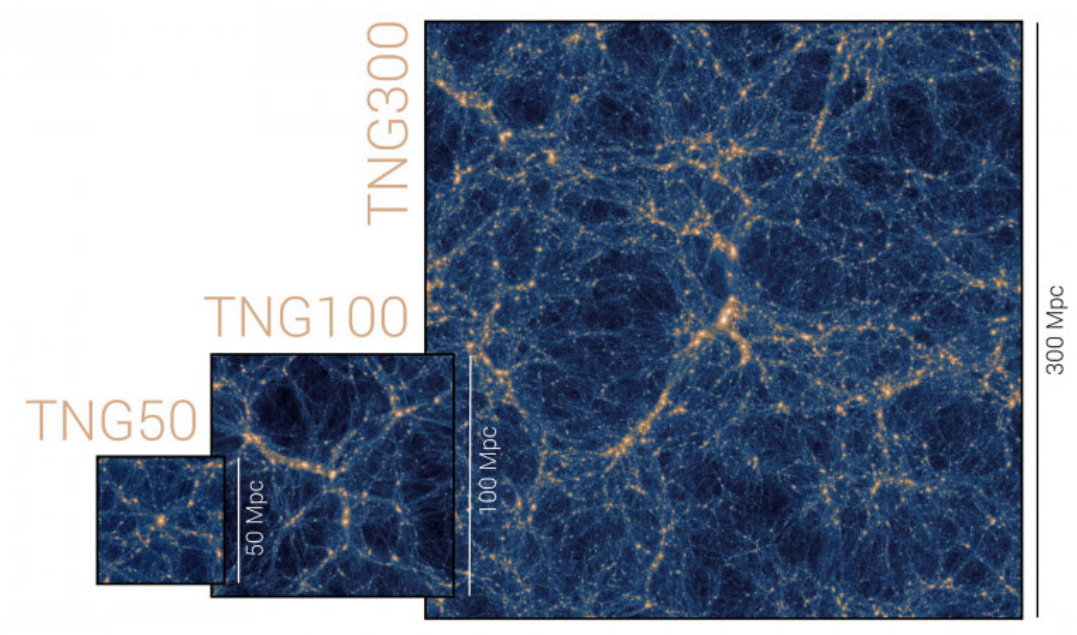

Figure 2.1: Illustration of the volumes of the three simulation of IllustrisTNG. Obtained from IllustrisTNG webpage.

tively. Each simulation was run with different mass and spatial resolution. Fig. 2.1 shows the difference in volume and resolution between the three boxes. As a result in mass and resolution, the three realizations complement each other. For example, the largest simulation box, TNG300, enables the study of galaxy clustering and provides the largest statistical galaxy sample. On the other hand, TNG50 provides the smallest galaxy sample at the high mass end, but it has the highest mass resolution overall. Therefore, it enables a more detailed look at the morphology of galaxies and its structural properties. TNG-100 falls somewhere in between these two other simulations.

The scientific goals of IllustrisTNG are to understand the physical processes that drive the evolution and the structural formation of galaxies, and to make predictions and to compare current and future observational data to further understand the physics around galaxies. A few examples of the use of IllustrisTNG are modeling the formation and evolution of globular clusters to study their kinematics (Chen and Gnedin, 2022), studying the nature of low brightness galaxies (Pérez-Montaño et al., 2022), creating mock galaxy surveys of James Webb Space Telescope and Hubble Space Telescope using the three IllustrisTNG simulations

(Snyder et al., 2022), and training a machine learning algorithm to study the importance of the central massive black hole in quenched galaxies in the early universe (Bluck et al., 2024), among others.

In this thesis, due to its mass and spatial resolution, we make use of the publicly available TNG50-1 model (Pillepich et al. 2019; Nelson et al. 2019a). Having a dark matter, baryonic mass resolution of $4.5 \times 10^5 M_\odot$, and $8.5 \times 10^4 M_\odot$, respectively, TNG50-1 allow us to resolve the structure of $10^9 M_\odot$ stellar disk with at least 10^4 stellar particles, enabling a better characterization of their morphology (Nelson et al., 2019b). Furthermore, as we are employing a machine learning algorithm, we need to make use of a large number of available galaxies to test and train the classifier. TNG50-1 is a great cosmological simulation to achieve that, as it contains a large number of distinct galaxy models.

The cosmological model adopted in IllustrisTNG is a flat ΛCDM universe with the following parameters: Hubble constant $H_0 = 67.8 \mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$, total matter density $\Omega_m = 0.3089$, dark energy density $\Omega_\Lambda = 0.6911$, baryonic matter density $\Omega_b = 0.0486$, rms of mass fluctuations at a scale of 8 h^{-1}Mpc $\sigma_8 = 0.8159$, and a primordial spectral index $n_s = 0.9667$ (Planck Collaboration et al., 2016).

2.2 Galaxy Selection

We will focus our study on central and satellite disk-like galaxies, identified within the redshift range $z = 0$ to $z = 0.5$. The z range considered allow us to obtain a large number of galaxy models to train our classification algorithm. Note that, even though a given galaxy will be present at different snapshots of the simulation, their detailed structure will evolve (see e.g. Varela-Lavin et al., 2023) and, thus, it will serve as input for the training process.

Following Dolfi et al. (2023) based on our selection criteria, we consider galaxies with:

- $N_{\mathrm{tot,stars}} \geq 10^4$, where $N_{\mathrm{tot,stars}}$ represents the number of bound stellar particles. This is used to make sure that galaxies have enough stellar particles to be reasonably well resolved. Considering that the baryonic mass resolution

is $\sim 10^5 M_\odot$, as mentioned before, the minimum stellar mass considered is $\gtrsim 10^9 M_\odot$.

- $f_e > 0.4$, where f_e represents the circularity fraction, defined as the fractional mass of the stellar particles with circularity $\epsilon > 0.7$. The latter has previously shown to reliably select orbits confined to a disk (Aumer et al., 2013). f_e kinematically quantifies the disk's shape, thus ensuring that the galaxies selected are considered "discy" (Joshi et al., 2020).

- $R_{90} \geq 3$kpc. This ensures that the structure of the galactic disk is clearly resolved.

These criteria result in a sample of 7,919 late-type galaxies. The following parameters, measured from each galaxy, are later used to train and test our classification models:

Effective Radius (R_{50}): Also known as half-mass radius. Defined as the radius of the galaxy containing 50% of the stellar mass.

Disk Extension (R_{ext}): Defined as $1.4 \times R_{90}$, where R_{90} is the radius of the galaxy containing 90% of the stellar mass. **Both R_{50} and R_{ext} were calculated considering as the center of the galaxy, the particle with the minimum gravitational potential energy.**

Concentration (C): Ratio between R_{90} and the effective radius R_{50}. Expressed as:

$$C = R_{90}/R_{50}$$

Minor-to-major axis (c/a): Ratio between the minor axis c and major axis a. **Obtained from the eigenvalues of the stellar component's mass tensor within** $2R_{50}$. c/a describes the shape of the inner galactic regions. The values range from 0 to 1, where values closer to 0 indicate flatter inner galactic regions, while values closer to 1 indicate rounder inner galactic regions. In our case, we obtain galaxies between 0.2 to 0.8, as those are the values describing the galactic disk of late-type galaxies.

Disk-to-total mass (D/T): Ratio between the disk's mass and the total mass of the galaxy. **The disk's mass is obtained by selecting particles with** $\epsilon \geq 0.4$**.** D/T is also used to select central and satellite galaxies in the selection criteria.

Star Formation Rate (SFR): Total stellar mass created from gas and dust, per year.

Half-mass (M_{50}): Total (baryonic and dark matter) mass of the galaxy enclosed within R_{50}.

Central Stellar mass Density (μ_*): Density of the stellar mass contained inside R_{50}. Defined as:

$$\mu_* = M^\star_{50}/\pi R^2_{50}$$

Here $M^\star_{50}$ represents the stellar mass of the galaxy enclosed within R_{50}.

Tidal Parameter ($\mathrm{T_P}$): Represents the tidal force applied by the inner galaxy regions ($R < R_{50}$) to the materials located at distances equal to R_{90}. **Defined following Varela-Lavin et al. (2023) as:**

$$T_\mathrm{P} = M_{50}/R^3_{90}$$

Spin Parameter ($\Lambda(R)$): Defined as the galactic disk stellar spin, which is a proxy of the apparent stellar angular momentum. **Calculated following Lagos et al. (2017), which defines the spin parameter as:**

$$\Lambda(R) = \frac{\sum_{i=1}^{N(r)} m_{\star,i} r_i V_{rot}(r_i)}{\sum_{i=1}^{N(r)} m_{\star,i} r_i \sqrt{V^2_{rot}(r_i) + \sigma^2_{\mathrm{1D},\star}(r_i)}}$$

This is calculated in N(r) radial bins. $\sigma^2_{\mathrm{1D},\star}(r_i)$ **represents the 1D velocity dispersion of star perpendicular to the disk's plane,** $V_{rot}(r_i)$ **the rotational velocity of the galaxy, and** $m_{\star,i}$ **the stellar mass enclosed within the** i**-th radial bin** r_i**.**

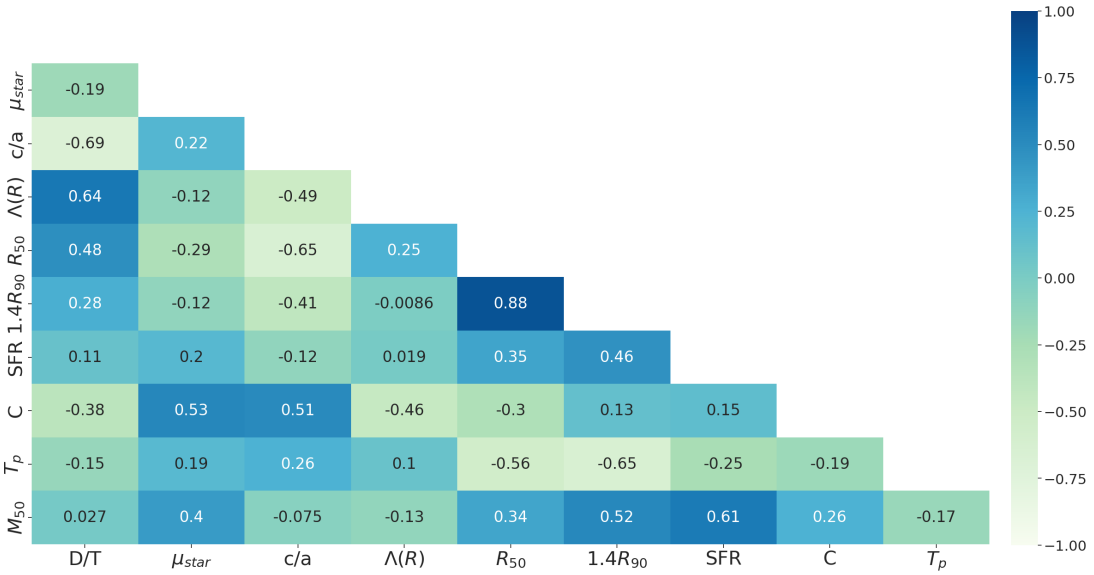

Figure 2.2: Pearson Coefficient Correlation heatmap of the galaxies' features obtained from the IllustrisTNG simulation.

These parameters are computed as described in Dolfi et al. (2023). Note that all selected parameters characterize galaxies internal properties and do not explicitly account for the environment in which the galaxies are located. Moreover, previous works have shown that some of these parameters, such as the disk central stellar density, μ_*, and its extension, $R_{\rm ext}$ are expected to be strongly linked to the occurrence of lopsided perturbations. In Fig. 2.2 we quantify the Pearson Correlation Coefficient between the listed parameters. Checking the parameters' correlation is an important first step to ensure an accurate representation of the classifier's results, as having highly correlated data (Pearson correlation values of 1 and -1) can lead to a misinterpretation of the importance of some parameters. In our case, we note that our parameters do not show a strong correlation, with the exception of R_{50} and $R_{\rm ext}$, which have a score of 0.88. However, this suggests that there is no issue in applying all the selected parameters in our classifier.

Chapter 3

Methodology

In this thesis we make use of RFs and its variations to study our selected dataset. Since we deal with a supervised algorithm, it is necessary to count with a training and testing set where galaxies are already labeled as lopsided or symmetric galaxies. A Fourier Decomposition of the light/mass distribution is often used quantify asymmetries (e.g. Zaritsky and Rix, 1997; Reichard et al., 2008; Varela-Lavin et al., 2023; Dolfi et al., 2023). We will use the radial distribution of the $m = 1$ mode to label our dataset. To prepare our data before applying it to the models, we partition the dataset into a training set and a testing set comprising 70% and 30% of the total sample, respectively. To do so, we employ SCIKIT-LEARN[1]'S STRATIFIEDSHUFFLESPLIT.

In this section we first discuss how lopsidedness is measured in our models, and then describe the two different variations of RFs used. We also discuss our particular application and the metrics used to measure its performance.

3.1 Measuring Lopsidedness

To label the galaxies in our sample between lopsided and symmetric, we apply a Fourier Decomposition. To do so, we measure the amplitude of the first mode $m = 1$ of the stellar disk density distribution, A_1, which quantifies the asymmetry of the stellar mass distribution. Before doing so, we have taken into account a

[1] https://scikit-learn.org

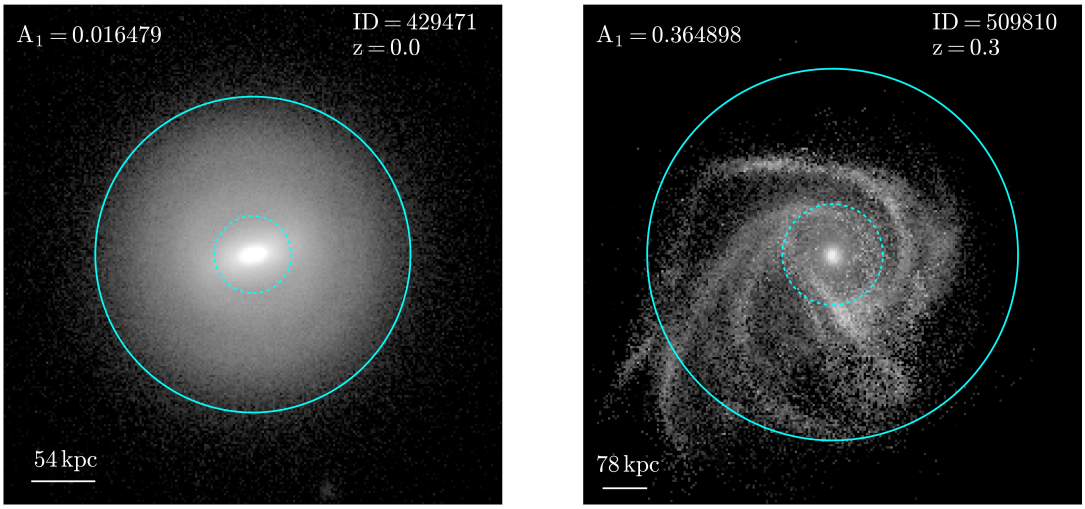

Figure 3.1: V-band face-on projected surface brightness distribution of a symmetric *(left)* and lopsided *(right)* galaxy, considered as examples of the classification made by A_1. Their respective A_1 value, ID (as in TNG50-1), and redshift snapshot are plotted on the upper side. On the lower left, the box size considered for each galaxy is also plotted. For both images, the dashed cyan line represents the radius R_{50} and the solid cyan line represents the radius $1.4R_{90}$, which are the limits of the radial interval used in the Fourier decomposition.

few considerations. First, it is crucial to ensure that each galaxy is projected face-on, as the Fourier Decomposition is highly sensitive to the disk inclination. To do so, we rotate each galaxy such as the z-axis is aligned with the disk angular momentum vector. Secondly, to focus our analysis on stellar discs, we consider only stellar participles located within a cylinder of width equal to $1.4R_{90}$, and a height equal to $2h_{90}$. Here, h_{90} is defined as the vertical distance above and below the disk plane enclosing 90% of the total galaxy stellar mass. The adopted definition for the disk extent allow us to reach their outer regions without introducing contamination from the stellar halo. We have tested several definitions for the disk extent, and found that the, overall, results are not significantly affected by our definition.

The Fourier decomposition for the stellar mass distribution is calculated as follows (Grand et al., 2016) :

$$C_m(R_j, t) = \sum_i M_i e^{(-im\phi_i)}$$

where M_i and ϕ_i are the mass and the azimuthal coordinate of the i-th stellar

particle. The A_1 radial profile is then calculated as follows:

$$A_1(R_j, t) = \frac{B_1(R_j, t)}{B_0(R_j, t)}$$

where $B_1(R_j, t)$ and $B_0(R_j, t)$ are the amplitude or strength of the $m = 1$ and $m = 0$ mode, respectively, within a certain radius R_j and a certain snapshot t. In general, the amplitude of the Fourier decomposition is given by:

$$B_m(R_j, t) = \sqrt{a_m^2(R_j, t) + b_m^2(R_j, t)},$$

where $a_m(R_j, t)$ and $b_m(R_j, t)$ are defined as the real and imaginary values of $C_m(R_j, t)$ for the m-th mode, respectively.

This is firstly done in concentric radial annuli of 0.5 kpc. Then, the averaged value of $A_1(R, t)$ at a given time, t, and over a certain radial interval (hereafter A_1) is used as the global or large-scale lopsidedness indicator. In general, if $A_1 > 0.1$, the galaxy is considered lopsided. For values of $A_1 < 0.1$ galaxies are considered symmetric. This threshold has been widely adopted in the literature, where both large observational and simulated galaxies were considered (e.g. Jog and Combes 2009; Reichard et al. 2008; Varela-Lavin et al. 2023; Dolfi et al. 2023). The radial interval considered to calculate the global A_1 parameter has varied between different works. For instance, Zaritsky and Rix (1997) studied the lopsidedness distribution of a sample of 60 field spiral galaxies, using the radial interval of $(1.5 - 2.5)$ disk scale lengths. On the other hand, Reichard et al. (2008) measured the lopsidedness of a sample obtained from SDSS in the radial interval R_{50}-R_{90}. van Eymeren et al. (2011) reached distances up to 4 to 5 disk scale lengths to study the asymmetries of the discs' outer regions. In our case, we use $\mathrm{R_{50}} - 1.4\mathrm{R_{90}}$, as we find that this radial interval best represent the non-axisymmetry of **our sample. In particular, using R_{50} as the lower limit avoids adding additional information of the galaxies' inner regions (e.g. bars and bulge) to the asymmetry characterization. For the upper limit, we tested different radius, such as R_{opt}, R_{90}, $1.1R_{90}$, and $1.4R_{90}$, for the training and testing of the classifiers. However, we chose $1.4R_{90}$ as it allow us to better approximate the disk extension of the simulated galaxies in our sample.**

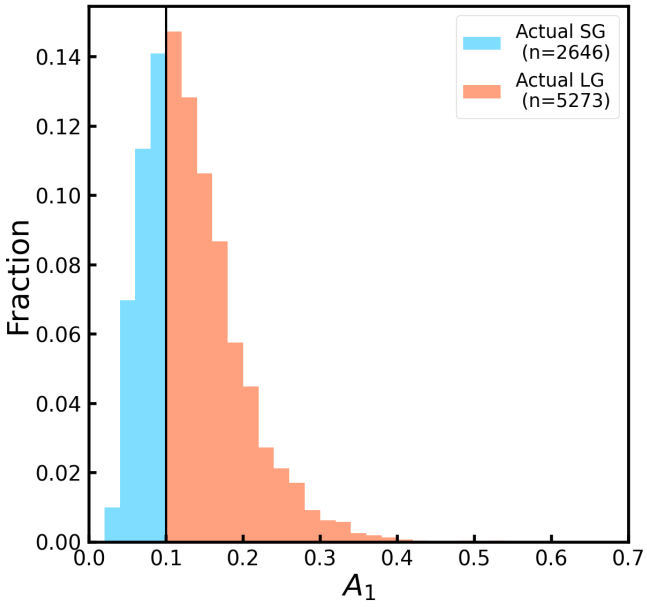

Figure 3.2: A_1 distribution of our total sample obtained by the averaged strength of the m=1 mode of the Fourier Decomposition for each stellar particle within the radial range $R_{50} - 1.4R_{90}$. The black line represents the threshold used to distinguish between lopsided and symmetric galaxies. The orange distribution represents lopsided galaxies (Actual LG) with a total of 5,273 galaxies and the blue distribution represents symmetric galaxies (Actual SG) with a total of 2,646 galaxies.

As an example of the classification made by A_1, Fig. 3.1 shows the face-on projections of the surface brightness distribution in the V-band of two clearly classified cases. Here the dashed and cyan lines indicate the lower and upper radial limits, respectively, considered to compute A_1. Considering their respective A_1 values, the galaxy on the left is classified as a strong symmetric example with $A_1 = 0.02$, while the galaxy on the right is classified as a strong lopsided example with a value of $A_1 = 0.36$.

The resulting A_1 distribution of our sample is shown in Fig. 3.2. The light blue and orange shaded areas indicate the distribution for symmetric and lopsided classified galaxies, based on the selected A_1 threshold (black line). Notably, our sample is imbalanced; i.e. we have a higher quantity of lopsided galaxies with respect to the symmetric cases. Out of the total sample size of 7,919 galaxies, 5,273 (i.e. 65%) are classified as lopsided, while 2,646 (i.e. 35%) as symmetric. We note that we find a larger fraction of lopsided galaxies than observations in

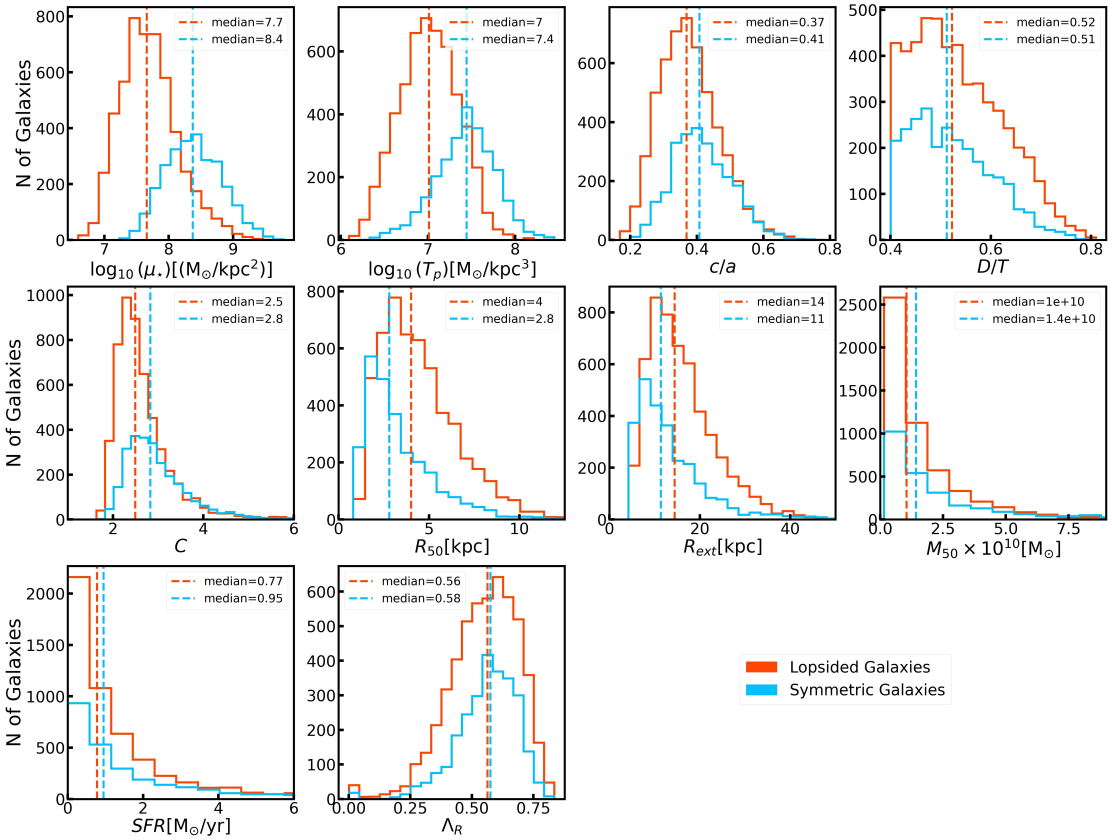

Figure 3.3: Distribution of parameters selected to characterize our galaxy sample. These parameters are used as features by the Random Forest classifier. The orange and blue distributions represent lopsided and symmetric galaxies, respectively. The colored dashed lines represent their respective median.

the local Universe (i.e. 30%; Zaritsky and Rix 1997; Reichard et al. 2008). As previously discussed in Dolfi et al. (2023), this difference can be likely attributed to the different radial interval used to measure the global lopsidedness A_1. For this reason, we are finding a larger fraction of lopsided galaxies than observations, due to the fact that we are reaching out to larger galactocentric radii where the lopsided amplitude is stronger (see also Varela-Lavin et al. 2023). The resulting imbalance imposes a great challenge for the training and testing of our selected machine learning algorithms. In the following section, we dive deeper into this issue and describe the methods we use to address it.

Lastly, Fig. 3.3 shows the distribution of our selected parameters, subdividing both types of galaxies to stress their differences. The dashed lines indicate the median of the corresponding distributions. **In particular,** the first and sec-

ond top panels show the distributions of μ_*, **which is the density of the stellar mass contained inside R_{50}, and T_P, which is the tidal force applied by the inner regions of the galaxy to its outskirts**. It is evident that the two galaxy types show the largest differences in these two parameters. As expected, lopsided galaxies typically show significantly smaller μ_* than their symmetric counterparts. Similarly, lopsided galaxies exhibit smaller values of T_P. This trends are in agreement with previous results (Reichard et al., 2008; Zaritsky et al., 2013; Varela-Lavin et al., 2023) that highlighted that both types of galaxies are indeed characterized by different internal structures.

3.2 Automatic Classification: Random Forests

Due to our dataset being imbalanced, as previously seen in Fig. 3.2, using a RF classifier could lead to an inaccurate classification. The training and testing of the RF are performed considering bootstrapped samples of the corresponding data sets. As each sample follows the same distribution as the original dataset, the majority class would have more predictions in favor, thus having more accurate results than the minority class. To avoid this issue affecting our results, we employ two different algorithms. The first one consists on oversampling the minority class of the training set and then apply it to a RF classifier. To do the oversampling, we use IMBALANCED-LEARN[2]'s SMOTE (Bowyer et al., 2011) method. This creates new "synthetic" data by interpolation between two close datapoints in the multidimensional feature space; in our case a 10 dimensional feature space. The second algorithm consists of using Balanced Random Forests (hereafter BRF; Chen and Breiman 2004)), where we use IMBALANCED-LEARN's BALANCEDRANDOMFORESTCLASSIFIER method. In this case, the bootstrapped sample is only considered for the minority class, whereas the majority class is randomly sampled with replacement, matching the size of the minority class. This avoids manually oversampling the dataset and it is directly performed by each decision tree.

To have an optimal performance of both classifiers using our datasets, we perform an *hyperparameter tuning*, which involves finding the best combination

[2] https://imbalanced-learn.org

of parameters from the models to yield the best results. The parameters involved in the fitting of the RF classifiers are the following:

- n_estimators: Number of decision trees in a RF algorithm. The following number of trees are considered: 2, 4, 8, 16, 32, 64, 128, 256, 500, 1000, or 1500. Although having more than 128 trees is expected to not have higher spike in accuracy, and even 128 trees is expected to be an optimal number of trees (Oshiro et al., 2012), we still consider a higher number due to RFs not consuming as much computational process as other algorithms.

- min_samples_split: Minimum required amount of data points in an internal node to split into further nodes. The minimum amount of data considered is: 5, 10, 15, 20, 30, or 10% of the total data.

- min_sample_leafs: Minimum required amount of data points to be in a terminal node. The minimum number of data points considered is: 1, 2, 3, or 4.

- max_features: Number of features necessary in an internal node to create the best split. The considered methods to calculate the maximum features are:

  - sqrt: Defined as $max\_features = \sqrt{n\_features}$
  - log2: Defined as $max\_features = \log_2(n\_features)$

- max_depth: Maximum depth of a tree, represented as the maximum path from the first node to a terminal node made by each split. The possible depth are: None, 3, 5, 7, 10, 20, 50, 75, 100, 150, or 200. In particular, None causes the tree to keep expanding until all leaves are pure **(i.e terminal nodes)** or by having less data points than min_samples_split.

- sampling_strategy: Sampling strategy to resample the selected class to handle class imbalance. The strategies considered are:

  - majority class: under-samples only the majority class to match the minority class.
  - not majority: under-samples all classes but the majority class.
  - all: under-samples all classes to match the size of the smallest sample.

Table 3.1: Results of the hyperparameter tuning using RANDOMIZEDSEARCHCV for each model.

Hyperparameters	SMOTE+RF	BRF
n_estimators	1500	128
min_samples_split	5	5
min_sample_leafs	2	2
max_features	sqrt	log2
max_depth	75	200
sampling_strategy	-	all

Both classifiers, RANDOMFORESTCLASSIFIER and BALANCEDRANDOMFOREST-CLASSIFIER , use the same hyperparameters, except for sampling_strategy, which is only used by the latter.

To tune both models, we use RANDOMIZEDSEARCHCV with number of iterations $n_{iter} = 10$ and, as cross-validation, REPEATEDSTRATIFIEDKFOLD with number of repeats $n_{repeat} = 10$ and number of splits $n_{splits} = 5$. In both cases, the n_{iter}, n_{repeat}, and n_{splits} are the default values of the parameters. To avoid unnecessary complexity in the calculations, we retain the default values for the current and following analysis. For the tuning process, **we select an arbitrary range of possible values we think each hyperparameter could have** and then apply it to the randomized search. This generates random combinations of hyperparameters and selects the combination that yields the best performance based on a chosen metric, which in our case is balanced accuracy. It is worth highlighting the significant difference in the number of trees between both classifiers,where SMOTE+RF has 1,500 trees in comparison with BRF, which has 128. This discrepancy in the number of trees might be attributed to the added complexity and variability introduced to the minority class by the SMOTE oversampling process. As it creates synthetic data, the complexity and variability of the sample increases, requiring SMOTE+RF to utilize a larger ensemble of trees to effectively generalize the data and achieve robust results.

3.3 Metrics

To measure the performance of both SMOTE+RF and BRF, we use the following metrics considering the use of binary classifiers:

- Precision: Ratio of the number of correctly predicted positive class to the total number of predicted positive class. Expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

where TP represents True Positives (i.e. actual symmetric galaxies classified as symmetric) and FP False Positives (i.e. actual lopsided galaxies classified as symmetric).

- TPR: Ratio of the number of correctly predicted positive class to the number of actual positive class. Expressed as:

$$\text{TPR} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

where FN represents False Negatives (i.e. actual symmetric galaxies classified as lopsided).

- F1-score: Harmonic mean of precision and TPR. Expressed as:

$$\text{F1} - \text{score} = 2 \times \frac{\text{precision} \times \text{TPR}}{\text{precision} + \text{TPR}}$$

- True Negative Rate (TNR) or specificity: Ratio of the correctly predicted negative class to the total number of the actual negative class. Expressed as:

$$\text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

where TN represents True Negatives (i.e. lopsided galaxies classified as lopsided)

- Balanced Accuracy: Average of the recall obtained for each class. Expressed as:

$$\text{balanced accuracy} = \frac{1}{2}\left(\text{TNR} \times \text{TPR}\right)$$

- Geometric Mean (G-mean): Square root of TNR and *TPR*. Expressed as:

$$\text{G} - \text{mean} = (\text{TNR} \times \text{TPR})^{1/2}$$

- ROC-AUC: Calculates the area under the Receiver Operating Characteristic (ROC) curve, by using the trapezoidal rule, which approximates the area under the curve (AUC) as a series of trapezoids. Considering a series of points in the ROC curve, in the form of $(x_1, y_i), (x_2, y_2), ..., (x_N, y_N)$, the area under the curse is expressed as:

$$\text{ROC} - \text{AUC} = \sum_{i=1}^{N-1} \frac{(x_{i+1} - x_i)(y_i + y_{i+1})}{2}$$

The selected metrics are used to evaluate the results of our classifiers. In particular, precision, TPR, and F1-score are important metrics to evaluate the performance of any type of model. However, these metrics are all sensitive to imbalanced dataset. As a result, they could mislead the algorithm during the training and validation process. To avoid this, we focus the analysis of our classifiers to TNR, balanced accuracy, and G-mean. This metrics are selected following Chen and Breiman (2004) work, which ensure a correct analysis due to the imbalanced nature of our dataset. Lastly, we also consider ROC-AUC for the analysis, as it gives us an important insight on how the model is performing without any effect of the imbalance. Still, we present the values for TPR, precision, and F-score, as a reference.

Chapter 4

Results and Analysis

4.1 Classification Results

In this section we introduce and analyze the results of the algorithms for the automatic classification between lopsided and symmetric galaxies. As a brief outline of our classification pipeline, we train the classifiers mentioned in Sect. 3.2 with 5,542 galaxies, constituting 70% of the total sample. This enables the algorithm to obtain important underlying patterns and/or relationships between the galaxies and their features, which are then used for the prediction in the final step. The remaining galaxies are consider for the testing set, which compromises a total of 2,377 galaxies, or 30% of the remaining sample. For each galaxy, these decision trees produce a class prediction—either lopsided or symmetric—and the class that is predicted in more than half of the decision trees is taken as the final prediction for that galaxy. Due to the imbalanced nature of our dataset, we define lopsided as the negative class and symmetric galaxies as the positive class. Usually, the majority class is better represented and naturally favorable by the algorithm over the minority class. To avoid this problem, we designate the minority class as the positive class, which helps with the interpretability of metrics, such as TPR, precision, and ROC-AUC for rare cases. Since we also obtain a proxy of the probability of a galaxy being in the positive/negative class, we test different thresholds, or *cut-off*, to classify the samples and to explore how such threshold can affect our results. As a default, this threshold is set at 0.5, i.e galaxies with probabilities equal or greater than this value are labeled as the positive class or, in our case, symmetric galaxies. Galaxies with probabilities lesser than

Table 4.1: Metric scores of the classifiers, SMOTE+RF and BRF, applied to the testing set. Each score is obtained by averaging the iterations of a cross-validation with $n_{iter} = 5$ and taking into consideration its standard deviation.

	SMOTE+RF	BRF
Metric	Score	Score
Precision	0.702±0.013	0.675±0.007
TPR	0.797±0.019	0.833±0.014
F1-score	0.746±0.012	0.746±0.007
TNR	0.830±0.011	0.799±0.00
G-mean	0.813±0.010	0.816±0.006
Balanced-Accuracy	0.813±0.010	0.816±0.006

this value are labeled as the negative class; i.e. lopsided galaxies. Our analysis showed that differences in the results obtained between the different *cut-offs* is negligible. Therefore the following analysis was performed with the default value, 0.5, for SMOTE+RF and BRF.

The results of each model's performance for the testing set are listed in Table 4.1. Each value of the metrics is obtained by averaging the result scores of each iteration of a cross-validation with number of iterations $n_{iter} = 10$, which is the default value, and taking into consideration its standard deviation. It is clear that both classifiers provide similar results, with comparable values in most metrics. Based on this, we select as our classifier SMOTE+RF since it results in better TNR metric. As previously discussed, we are working with a unbalanced data set, with more than 70% of the data belonging to the negative class (lopsided objects). Thus, a high TNR indicates a better performance for the most populated class of our sample.

Fig. 4.1 shows the confusion matrix (CM) for SMOTE+RF. The x-axis indicates the predicted class or predicted label, obtained from the classifier, and the y-axis show the actual class or actual label, obtained from the A_1 parameter. In general, a CM allows us to visually inspect the fractions of correct and incorrect classification we have obtained. In our testing sample, and based on our A_1 classification criteria, we count with 1,578 true lopsided and a total 799 true symmetric galaxies. Interestingly our classifier is able to correctly classify 81% of the lopsided objects and approximately the same amount for their symmetric

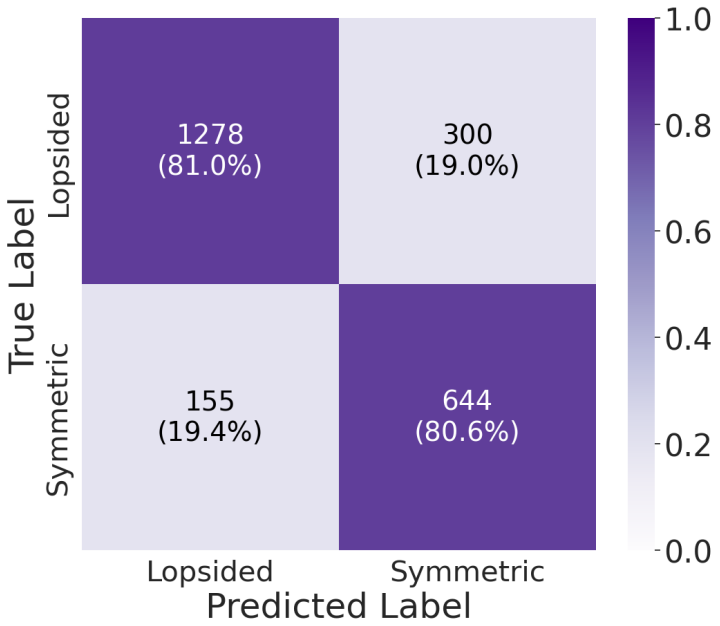

Figure 4.1: Confusion matrix for the testing set of the best model, SMOTE+RF. The x-axis is the predicted class or predicted label, and the y-axis is the actual class or actual label. The percentage with respect each type of galaxy set is on parenthesis.

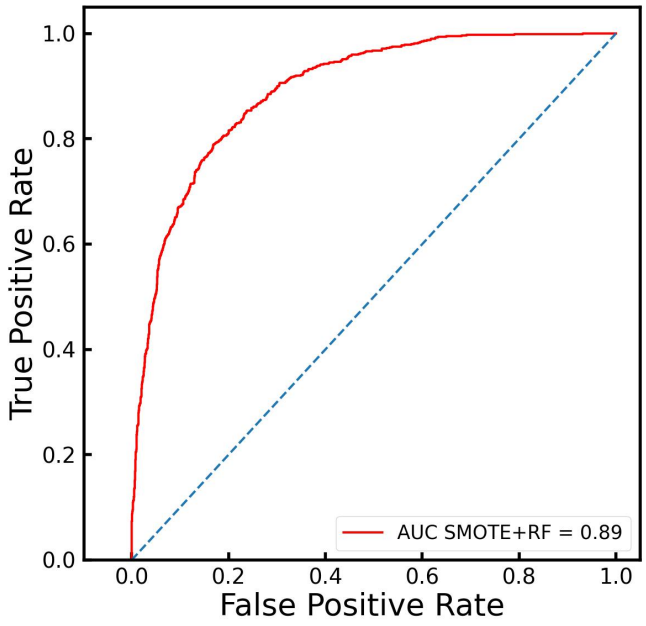

Figure 4.2: Receiver Operating Characteristic (ROC) plot considering all the classification thresholds of the testing set.

counterparts. In absolute number, we obtain a total of 1,922 correctly classified galaxies, against 455 wrongly classified objects. It is worth highlighting the very good performance of the SMOTE+RF classifier, which has been purely obtained based on features that are related to our simulated galaxies internal properties. No information about environments has been introduced during the training process. This is illustrated by obtaining the value of the area under the curve of the ROC plot, as shown in Fig. 4.2. By combining both axis, the resulting area under the curve yields important information about the performance of the classifier. This value is shown in the legend of the plot, where in our case we obtain a value of 0.89. The closer the value is to 1, the better. If the value is closer to 0.5, as shown as blue dashed line, the classifier has a bad performance.

4.1.1 Interpretation of the Random Forest classification

Supervised algorithms, including RFs, suffer from interpretability of the decisions leading to the classification. This is often called the "black box" problem. In RFs, it arises due to the high quantity of decision trees added to the ensemble. In this section, we interpret and analyze the decisions lead by the model to subdivide the galaxies between lopsided and symmetric by ranking the importance of the features used in the classification process.

We use the *permutation_importance_* attribute from RANDOMFORESTCLASSIFIER. There are various methods for ranking feature importance, but given the continuous nature of our dataset—where no categorical features are used for training or testing— we rely solely on *permutation_importance_*. This attribute works by permuting, or shuffling, the values of each feature and calculating the resulting decrease of a specified metric, which by default is accuracy, defined as the fraction or count of the correct predictions. The decrease in the score is then used to rank each feature: the higher the score, the more it affects the model's performance, thus making the feature important for the model to maintain a higher accuracy. However, since our dataset is imbalanced, using accuracy would not return an accurate representation of the importance of our features. To address this issue, we use balanced-accuracy instead. As discussed in Sec. 3.3, this metric represents the averaged fraction of correct classified galaxies for both the negative and positive class. In this, each class contribute equally to the final score, regardless of its

Table 4.2: Feature Importance of each parameter calculated by *permutation_importance* for SMOTE+RF. The score is obtained averaging each iteration of a cross-validation with $n_{iter} = 5$, which is the default value, and taking into consideration its standard deviation.

Rank	Feature	Score
1	μ_*	0.242930±0.010621
2	T_P	0.072285±0.003824
3	SFR	0.047444±0.005302
4	DT	0.009005±0.002047
5	$\Lambda(R)$	0.006586±0.002012
6	M_{50}	0.006234±0.000864
7	c/a	0.004205±0.000934
8	C	0.003042±0.003434
9	R_{50}	0.001760±0.002815
10	R_{ext}	0.000721±0.000990

size. Considering that the accuracy metric disproportionately favors the majority class in imbalanced datasets due to its over-representation, balanced accuracy is an alternative to avoid inaccurate results.

The results of this procedure are shown in Table 4.2, where lists the rank of each feature obtained by *permutation_importance_*. Considering that we want to focus on the performance of SMOTE+RF with unseen data, we only calculate the feature importance for the testing set. We obtain each score by averaging the iterations of a cross-validation with $n_{iter} = 5$ and taking into consideration its standard deviation. This analysis clearly sows that both μ_* and T_P are the highest-ranked parameters, with μ_* ranked first T_P ranked second. As a way to better visualize this, Fig. 4.3 also shows the variation of balanced-accuracy with a box plot. Each box represents the distribution of the score value for each iteration. The dotted line inside each box is the median of the distribution, and each whisker represents the first and last score value. Indeed, we note that μ_* is the top- ranked parameter overall, indicating that it is the most important parameter to consider in the classification process made by SMOTE+RF. As we previously mentioned, and as seen in Fig. 3.3, lopsided and symmetric galaxies are characterized by different μ_* distributions. This is in agreement with previous results (Reichard et al., 2008; Zaritsky et al., 2013; Varela-Lavin et al., 2023; Dolfi et al., 2023), where lopsided galaxies tend to show significantly lower a densities in the

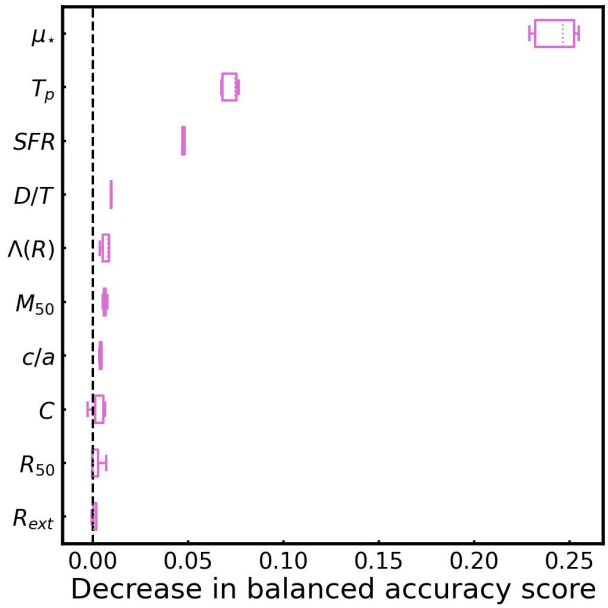

Figure 4.3: Box plot of each feature from the testing set, ranked by their importance as determined by the *feature_permutation_* attribute from SMOTE+RF. Each box represents the range of the different scores obtained from a cross-validation with $n_{iter} = 5$. The inner dashed line represents the median value of each distribution. The whiskers on each box represent the minimum and maximum value of each distribution.

inner regions (as defined by their R_{50}) with respect to the symmetric counterparts.

Although not as important as μ_*, T_P and *SFR* also play an important role in the classification process in comparison with the rest of the features. This is also in agreement with previous results, where an (anti-) correlation between lopsidedness and T_P (e.g. Gómez et al., 2016) and a correlation between lopsidedness and *SFR* (e.g. Conselice et al., 2000) have been reported. In particular, T_P represents a proxy of the tidal force exerted by the inner galactic regions on the outer disk material. In other words, it indicates how gravitationally cohesive a galaxy is. The relevance of this parameter is clearly reflected in the separation between the distribution of both types of galaxies, as previously seen in Fig. 3.3, where lopsided galaxies tend to have lower values of T_P than symmetric galaxies. These findings align with the conclusions of Varela-Lavin et al. (2023) and Dolfi et al. (2023), which propose that lopsided perturbations serve as indicators of intrinsic galaxy properties, rather than being predominantly driven by environmental processes. In other words, galaxies with low central stellar densities are weakly gravitationally cohesive and, thus, are more susceptible to lopsided per-

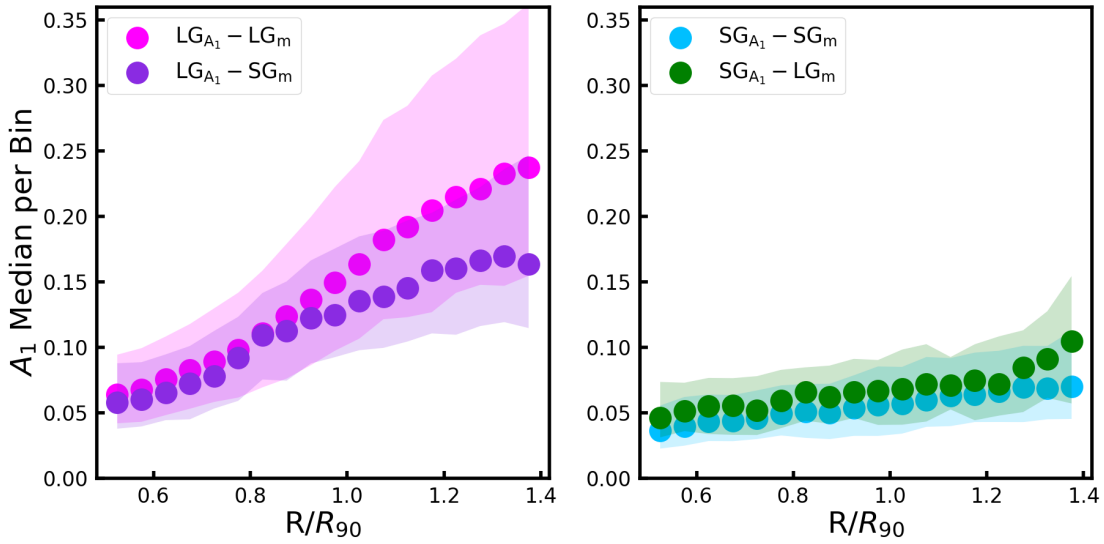

Figure 4.4: Radial profiles of A_1 for our four classification cases, calculated as the median of A_1 for each bin with respect to R_{90}. The fuchsia and blue distributions represent the correctly classified lopsided galaxies ($LG_{A_1} - LG_m$) and symmetric galaxies ($SG_{A_1} - SG_m$), respectively. The green distribution represents symmetric galaxies classified as lopsided ($SG_{A_1} - LG_m$) and the purple distribution represents lopsided galaxies classified as symmetric ($LG_{A_1} - SG_m$). The shaded areas represent the 25th and 75th percentiles of each sample.

turbations, independently of the particular perturbing agent. On the other hand, *SFR* ranking third place is an interesting result, as it is been shown that there is a correlation between A_1 and current *SFR* (Zaritsky and Rix, 1997; Rudnick et al., 2000; Reichard et al., 2009). As discussed by Łokas (2022), some internal properties of lopsided and symmetric galaxies can be linked with their current *SFR*, e.g. lopsided galaxies having bluer colors, larger gas fractions, and lower metallicity than symmetric galaxies. Moreover, Dolfi et al. (2023) showed that lopsided galaxies tend to be, on average, significantly more star forming than symmetric galaxies at later times. Symmetric galaxies, on the contrary, have an earlier assembly with shorter and more intense star forming bursts. As a result, and considering galaxies with similar stellar masses at the present-day, while symmetric galaxies tend to develop a more pronounce central region at earlier times, lopsided galaxies tend to form at larger fraction of their stellar populations later, typically developing a more extended stellar disk and less dense inner regions. Lastly, Fig. 4.3 shows the relative importance of the remaining 7 features. It is clear that they have a minimal impact on the classification procedure.

To analyze the classification made by SMOTE+RF, we plot in Fig. 4.4 the median of A_1 as a function of radius for the four cases defined by the classifier. To generate this figure, the radial extension of each simulated galaxy was normalized by its corresponding R_{90}. We focus on the radial interval $(0.5 - 1.4)\mathrm{R}_{90}$, as it is the considered interval for the Fourier decomposition. The shaded areas represent the 25th and 75th percentiles of the distribution. In the left plot, the fuchsia distribution represents correctly classified lopsided galaxies, defined as ($\mathrm{LG}_{\mathrm{A}_1}$ − LG_{m}), and the purple distribution represent lopsided galaxies classified by our model as symmetric, defined as ($\mathrm{LG}_{\mathrm{A}_1} - \mathrm{SG}_{\mathrm{m}}$). The right plot is the same as the left one but for symmetric galaxies. Here the cyan color represent the distribution of correctly classified symmetric galaxies, defined as ($\mathrm{SG}_{\mathrm{A}_1} - \mathrm{SG}_{\mathrm{m}}$), while in green we show symmetric galaxies classified as lopsided, defined as ($\mathrm{SG}_{\mathrm{A}_1} - \mathrm{LG}_{\mathrm{m}}$). Note that incorrectly classified cases do not follow the same trend as the correctly classified distributions. In the case of ($\mathrm{LG}_{\mathrm{A}_1} - \mathrm{SG}_{\mathrm{m}}$) on the left plot, from $0.5R_{90}$ to $0.9R_{90}$ the magnitude of A_1 starts increasing at the same rate than the correctly classified sample. However, from $0.9R_{90}$ onward, the slope is less steep, meaning that the magnitude of A_1 does not increase as much as in ($\mathrm{LG}_{\mathrm{A}_1} - \mathrm{LG}_{\mathrm{m}}$). In other words, while incorrectly classified galaxies have indeed an outer perturbed region, the strength of the perturbations is typically weaker with respect to correctly classified galaxies. On the right panel we show that the A_1 profile of both correctly ($\mathrm{SG}_{\mathrm{A}_1} - \mathrm{SG}_{\mathrm{m}}$) and incorrectly classified galaxies ($\mathrm{SG}_{\mathrm{A}_1} - \mathrm{LG}_{\mathrm{m}}$) remains below the 0.1 threshold chosen to classify lopsided galaxies based on the A_1 parameter. Nonetheless, wrongly classified symmetric galaxies tend to have a larger A_1 value at all radii and they do cross the threshold at the outermost edge. In the following section we explore in detail the main reasons that drove the SMOTE+RF method to misclassify these galaxies.

4.1.2 Interpretation of the Misclassified Cases

In the previous section, we analyzed the results of applying RFs algorithms to the internal parameters of our selected sample of lopsided and symmetric galaxies. In particular, we find that the μ_* and T_{P} parameters are the primary features used by the classifier to subdivide galaxies as either lopsided or symmetric, consistent with previous observational studies. However, there are 455 galaxies in the testing set that are misclassified. In this section, we focus on the misclassified cases,

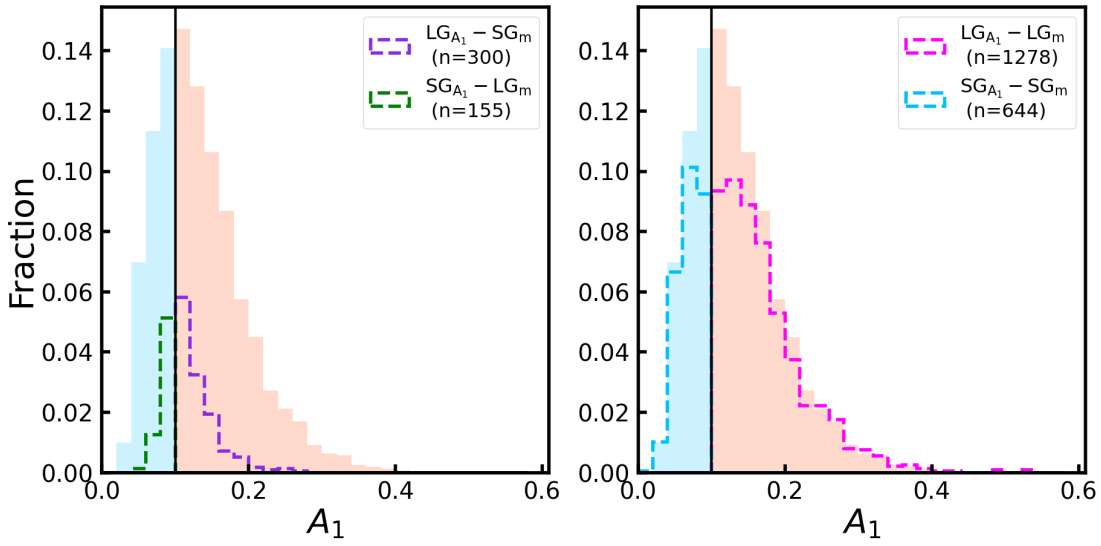

Figure 4.5: A_1 distributions of the four classification cases made by SMOTE+RF applied to the testing set. *(left)* Misclassified cases. The purple dashed distribution represents the actual symmetric galaxies classified by the model as lopsided galaxies ($SG_{A_1} - LG_m$), and the green dashed distribution represents the actual lopsided galaxies classified by the model as symmetric galaxies ($LG_{A_1} - SG_m$). *(right)* Correctly classified cases. The cyan distribution represents symmetric galaxies classified as symmetric ($SG_{A_1} - SG_m$). The magenta distribution represents lopsided galaxies classified as lopsided ($LG_{A_1} - LG$). Each distribution has in parenthesis their respective number.

($LG_{A_1} - SG_m$) and ($SG_{A_1} - LG_m$), to investigate the underlying reasons behind the misclassification.

To further study the incorrectly classified galaxies, in Fig. 4.5 we highlight the A_1 distribution of the four classification cases in comparison with the A_1 distribution of the total sample, as seen in Fig. 3.2. Focusing on the left figure, the purple dashed distribution represents lopsided galaxies classified as symmetric galaxies ($LG_{A_1} - SG_m$) with a median of ~ 0.09, and the green dashed distribution represents symmetric galaxies classified as lopsided galaxies ($SG_{A_1} - LG_m$) with a median of ~ 0.12. The cyan and magenta distributions represent the lopsided galaxies classified as lopsided ($LG_{A_1} - LG_m$) and symmetric galaxies classified as symmetric ($SG_{A_1} - SG_m$), respectively. It is clear that all misclassified galaxies are adjacent to the threshold $A_1 = 0.1$ and, thus, represent challenging cases for our classification models. In Fig. 4.6 we show the distribution of our selected features for all the four different classification cases, following the same color coding as in Fig. 4.5. Each dashed line represents the median of the corresponding distri-

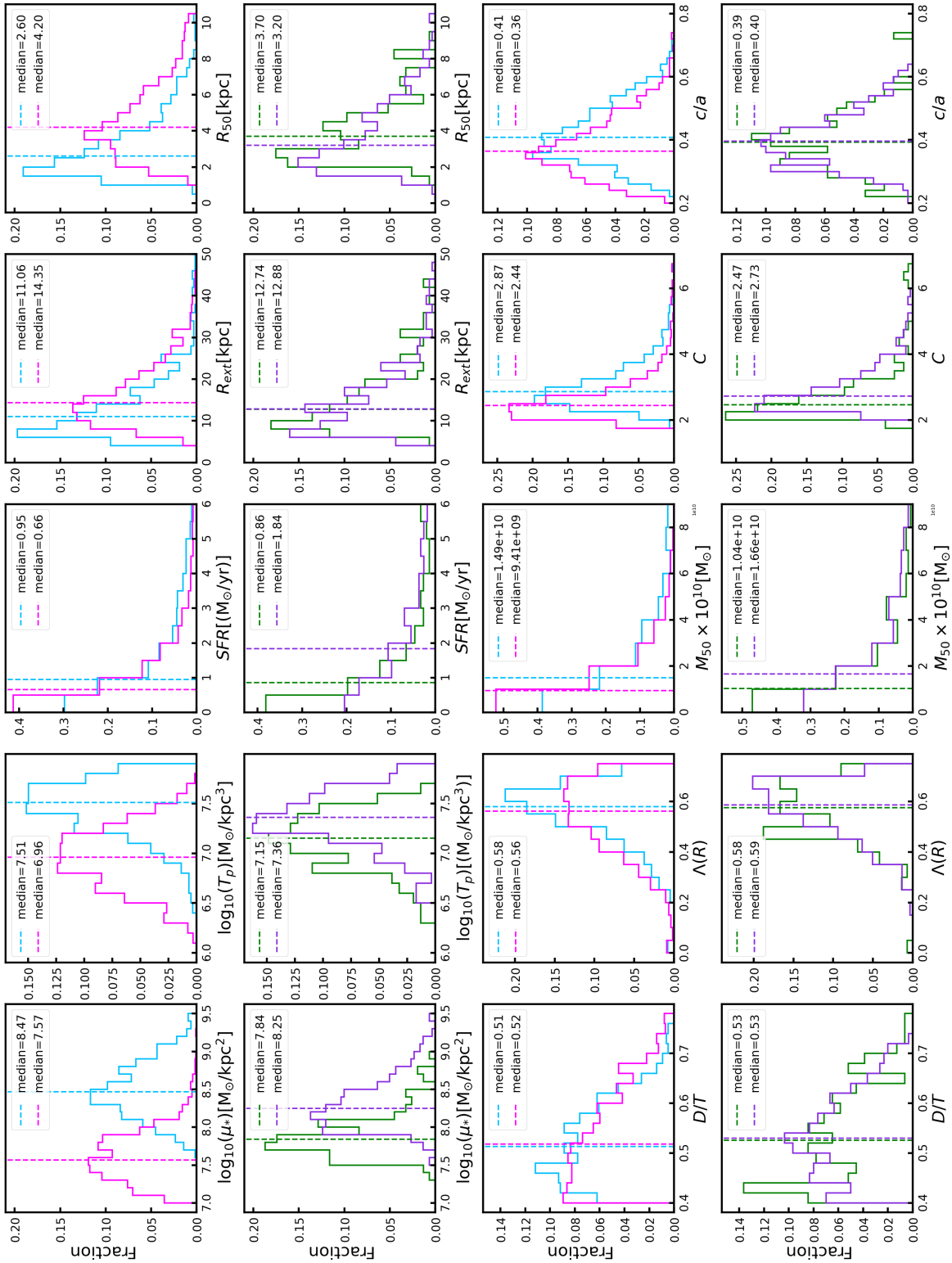

Figure 4.6: Normalized distribution of all the selected features, considering the correct (upper) and incorrect (bottom) classification made by SMOTE+RF. Each distribution has been normalized by their corresponding number of galaxies of each subsample. Their respective number of galaxies is in parenthesis. The fuchsia and blue distributions represent the correctly classified lopsided galaxies ($LG_{A_1} - LG_m$) and symmetric galaxies ($SG_{A_1} - SG_m$), respectively. The green distribution represents symmetric galaxies classified as lopsided ($SG_{A_1} - LG_m$) and the purple distribution represents lopsided galaxies classified as symmetric ($LG_{A_1} - SG_m$). The dashed lines represent the median of their respective distribution.

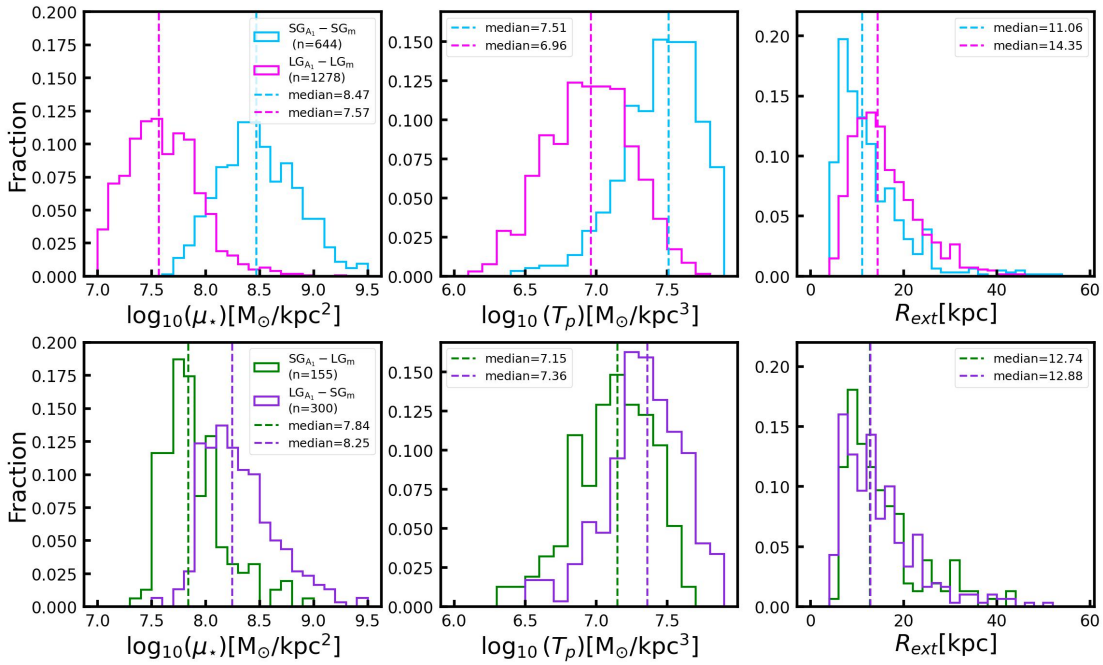

Figure 4.7: Normalized distribution of the central stellar mass density μ_* (*left*), tidal parameter T_P (*middle*), and the disk extension R_{ext} (*right*). Same format and color coding as Fig. 4.6.

bution. The upper panels represent the correctly classified cases and the bottom panels represent the incorrectly classified cases. Note that the distributions differ significantly in all of the three most important parameters by the classifier, μ_*, T_P, and SFR. And as expected, the largest differences are found in μ_* and T_P. However, even the radial distributions, R_{ext} and R_{50}, show differences. For the following analysis, we selected μ_*, T_P, and R_{ext} as they are consider to show the largest difference between the distributions. **For an easier interpretation, we added Fig. 4.7, which only contains the distributions of μ_*, T_P, and R_{ext}. It follows the same format and color coding as Fig. 4.6, where the correct cases are in the upper panels and the incorrect classification cases are in the bottom panels.** Two important things stand out. First, the incorrect distributions of the three inspected parameters show more significant overlap with respect to the correctly classified sample. The medians are, in all cases, closer to the median of the overall sample. This is most clear in the R_{ext} distributions, where both symmetric and lopsided nearly perfectly overlap with each other. Second, and most importantly, we find that galaxies classified as lopsided by our global A_1 parameter, but identified as symmetric by our model ($LG_{A_1} - SG_m$), have values of μ_* and T_P that are consistent with the distribution

of correctly classified symmetric galaxies. In other words, they have relatively large central surface density and T_P values. Upon closer inspection of their images, we observe that such galaxies typically display a symmetric overall disk, but a significant asymmetry in their outermost region. An example of such galaxy is show in the top right panel of Fig. 4.8. These localized asymmetries, captured by the global A_1 parameter, not necessarily reflect the overall structure of the disk and can be caused by recent episodes of gas accretion or very recent strong interactions. On the other hand, galaxies classified as symmetric by the global A_1 parameter but **asymmetric** by our model ($SG_{A_1} - LG_m$) show low μ_* and T_P values. Such galaxies display internal properties of typical lopsided galaxies, but simply the morphological perturbation has not yet been triggered. The top left panel of Fig. 4.8 shows an example of such situation.

To further explore the two examples of misclassified galaxies, in the second and third row of Fig. 4.8 we show their radial A_1 and density profiles, respectively. The cyan regions in the second row highlight the radial interval (0.5 - 1.4)R_{90}, considered to measure A_1. It is worth noting that both galaxies were selected by considering extreme values of μ_* and T_P while having similar stellar mass. For ($SG_{A_1} - LG_m$), the galaxy shows consistently low $A_1(R)$, even up to the disk outermost regions. Interestingly, its inner stellar density is notably lower than expected for a symmetric galaxy. Even its μ_*, highlighted with a red star, falls below the mean of the overall sample (dashed magenta line). On the other hand, for the ($LG_{A_1} - SG_m$), while the $A_1(R)$ shows values consistent with 0 within most of the considered radial range, it shows a very strong rise in the disk outskirts. We note that this galaxy has a denser inner stellar region, highlighted by its large μ_* value which significantly surpass the median of the overall distribution.

Lastly, to understand these unexpected behavior, we explore on the two lower rows the time evolution of the lopsided parameter and the orbital histories. Interestingly, we find that the ($LG_{A_1} - SG_m$) galaxy (right panels) became a satellite of a larger host approximately 1.5 Gyr ago. Previous to the pericentric passage, this galaxy showed A_1 values below the threshold. After the close interaction, the A_1 value rapidly grows as a result of the tidal perturbation of its outer disk. Indeed, we find that this galaxy has internal properties consistent with the symmetric sample, but the strong recent interaction forced an outer tidal disruption,

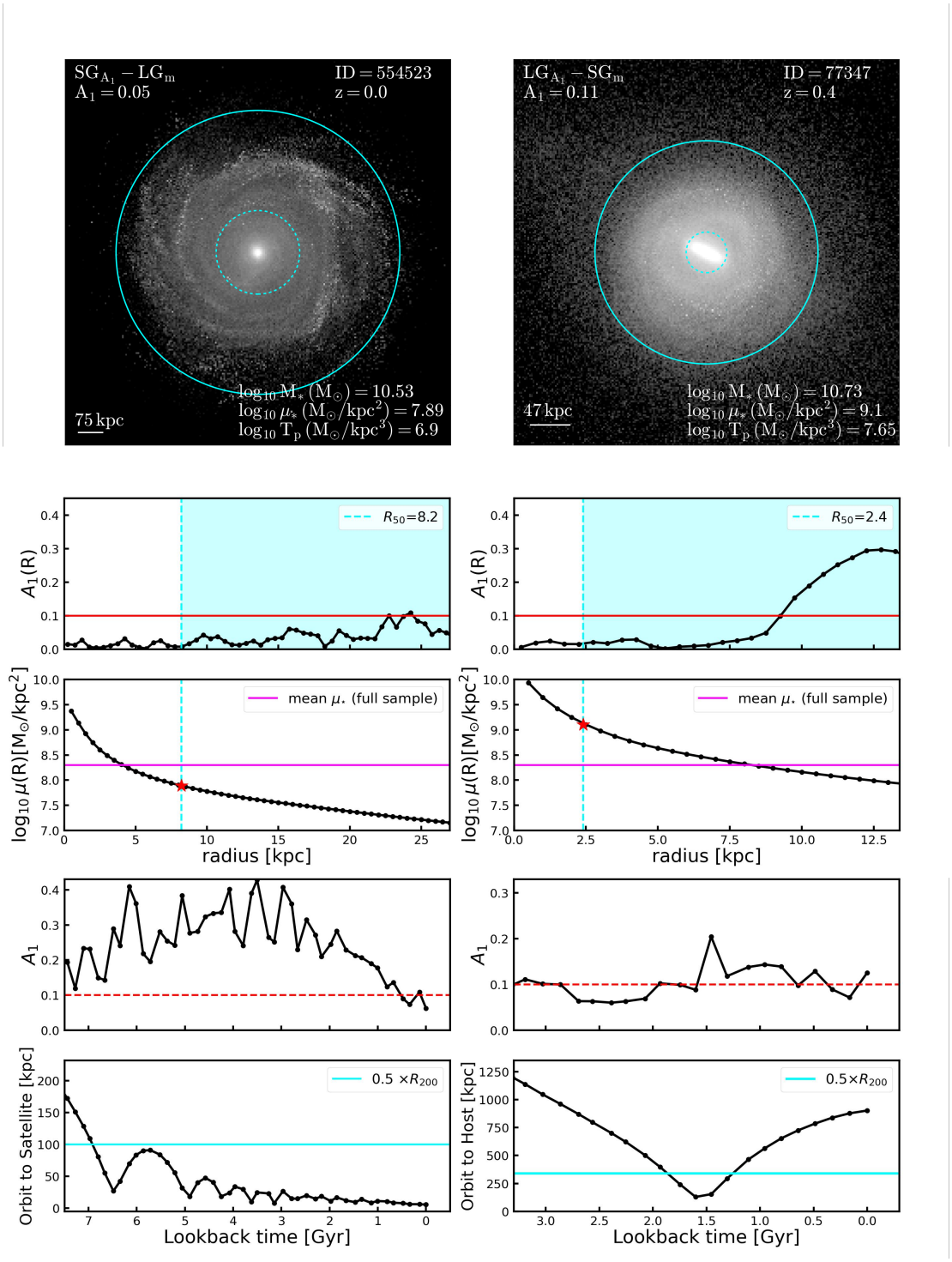

Figure 4.8: *(Top panels)* V-band face-on projected surface brightness distribution of a *(left)* symmetric galaxy classified as lopsided ($SG_{A_1} - LG_m$) and a *(right)* lopsided galaxy classified as symmetric ($LG_{A_1} - SG_m$), considered as examples of the misclassification made by SMOTE+RF. On the upper side, their respective A_1 value and classification case are plotted on the left, and their ID and redshift z on the right. On the bottom right, the values of the stellar mass (M_*), central stellar mass density (μ_*), and tidal parameter (T_P) are plotted. The dashed cyan lines represent the inner radius R_{50} and the solid cyan lines represent the outer radius $1.4R_{90}$, which are the limits of the radial interval used in the Fourier decomposition.

Continuation of Fig. 4.8. *(Middle panels)* Lopsidedness and stellar density profiles with respect to the radius, up to $1.4R_{90}$. In both cases, the cyan lines represent the start of the radial interval, R_{50}. The pink dashed lines represent the average central stellar mass density (μ_*) of the full sample, with a value of 8.3, while the red stars represent the value of the cental stellar mass density of the galaxy, μ_*, within R_{50}. *(Bottom panels)* Lopsidedness and the respective orbit of the most massive satellite with respect to lookback time. The red dashed line represents the A_1 threshold to classify lopsided and symmetric galaxies. The horizontal cyan line represents $0.5 \times R_{200}$, where R_{200} is defined as the virial radius of the central galaxy.

captured by the A_1 parameter. In the case of the ($\mathrm{SG_{A_1}} - \mathrm{LG_m}$) (left panels), the time evolution of A_1 shows that, over most of its evolution, this galaxy was indeed strongly lopsided. The initial perturbations was likely induced by significant interaction with a massive satellite galaxy ($\sim$ 1:10) 6.5 Gyr ago (first pericentric passage). After this point, the galaxy suffered no other interaction with satellite of mass ratios $<$ 1:100. Thus, the lopsided perturbation gradually relaxed, reaching a present-day A_1 value below the considered threshold. Even though its internal structure make this galaxy susceptible to lopsided perturbations, the lack of significant external perturbation during its late evolution resulted on a symmetric configuration at the present-day.

We note that recent interactions cannot explain all the misclassified cases. Indeed, only 76 of the 300 ($\mathrm{LG_{A_1}} - \mathrm{LG_m}$) cases are satellite galaxies of a more massive host. Eight (8) additional galaxies have suffered significant interactions ($>$ 1:20) as centrals during the last 3 Gyr. Thus, important interaction can be attributed to this misclassified class in only 28% of the cases. Nonetheless, as previously discussed, other mechanisms such as gas accretion, instability in a counter-rotating disk and torques from an off-centered dark matter halo could be at play in the remaining cases (Jog and Combes, 2009). Among these mechanisms, asymmetric gas accretion has been proposed as a common driver of lopsidedness. As shown by Bournaud et al. (2005), interactions and mergers can trigger strong lopsidedness in some cases, but they do not account for all the observed statistical properties, such as a correlation between lopsidedness and the Hubble Type, or a correlation between $m = 1$ and $m = 2$ asymmetries, among others. In a follow up study we will focus on the misclassified cases to further the origin of lopsidedness in galaxies with internal properties common to symmetric discs.

Chapter 5

Classification with observational Parameters

Several parameters considered in this work as features require additional modeling to be estimated. Thus, they cannot be directly obtained from observation based on, e.g., photometric data. For example, the calculation of μ_* involves the application of additional stellar population models. Indeed, Reichard et al. (2008) calculated the stellar surface mass density following Kauffmann et al. (2003) definition, which considers the stellar mass and the Petrosian half-light radius in the z-band. Their stellar massed where estimated using a method that combines spectral diagnostics of star formation histories with photometric data. Additionally, the tidal parameter $\mathrm{T_P}$, requires an estimation of the total mass enclosed within R_{50}, which involves dynamical modeling of the galaxy.

Despite their importance of in the classification process of such parameters, in this section we explore wether it is still possible to obtain a reliable classification of lopsided and symmetric galaxies using parameters that are more readily obtainable from photometric data. We follow the same pipeline mentioned earlier, but we train and test the SMOTE+RF classifier with a subset of features that could be estimated from multi-band photometric surveys such as S-Plus (Mendes de Oliveira et al., 2019) and J-PAS (Benitez et al., 2014). In particular, we replace the parameter M_{50} by the galaxies r-band luminosity within R_{50}, L_{50}, thus avoiding the need of stellar population models. In addition to L_{50}, we consider as features R_{50}, R_{ext}, c/a, and SFR. The later can be obtained from narrow band

Table 5.1: Scores of the SMOTE+RF model on the testing set, using only observational parameters. Each score is obtained by averaging the iterations of a cross-validation with $n_{iter} = 5$ and taking into consideration its standard deviation.

Metric	Score
Precision	0.700 ± 0.013
Recall	0.788 ± 0.020
F1-score	0.741 ± 0.011
ROC-AUC	0.809 ± 0.009
TNR	0.830 ± 0.011
G-mean	0.809 ± 0.009
Balanced-Accuracy	0.809 ± 0.009

photometry around the H_α line through the Kennicut relation (Kennicutt, 1998). We keep the same hyperparameters listed in Table 3.1, along with the same training and testing sets.

The results of this test are presented in Table 5.1, where list the metrics obtained from the testing set. Interestingly, we find very good results, with a performance of the SMOTE+RF algorithm that is only very mildly affected by the limited number of features considered. Indeed, most scores are not significantly affected. Compared to our previous results we find a negligible decrease of 0.4% for balanced accuracy and no change for TNR. Additionally, ROC-AUC has a score of 80.9%, which reflects on how well the model is able to differentiate between both classes. As expected, substituting M_{50} by L_{50} did not introduced a significant drop in the performance. To further characterize our classification, the left panel of Fig. 5.1 shows the resulting CM. Note that we obtain a total of 1,927 correctly classified galaxies and only 535 incorrectly classified cases, which represent a 15% increase. Compared to our previous results, this model improves in the identification of actual lopsided galaxies, but performs slightly worse in classifying actual symmetric galaxies as symmetric. The feature importance ranking is shown on the right panel of Fig. 5.1, generated with the *feature_importance* attribute. We find that the most important parameters are now L_{50}, R_{50} and SFR. As before, c/a provides no significant information for the RF classifier.

Our results show that using readily available observational parameters offers a simpler and reliable approach to classify lopsidedness in large observational sam-

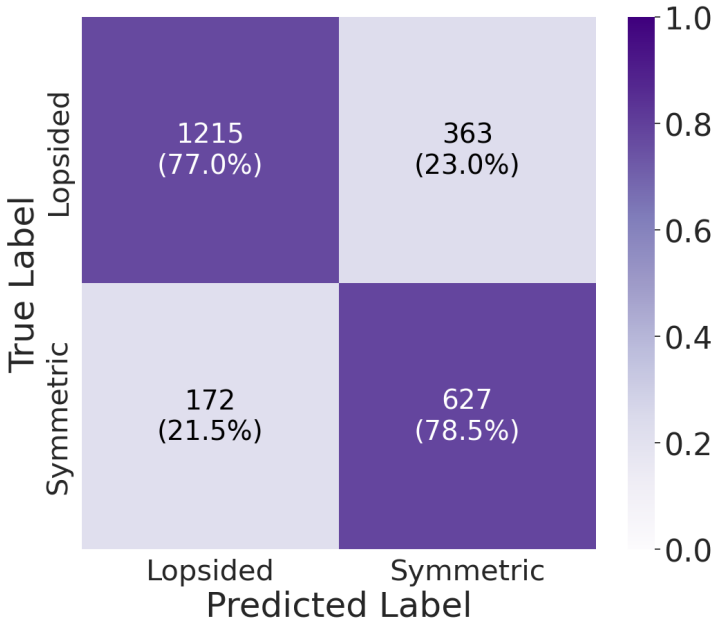

Figure 5.1: Confusion matrix of the testing set using SMOTE+RF with only observational parameters. The x-axis is the predicted class or predicted label, and the y-axis is the actual class or actual label. The percentage with respect each class is on parenthesis.

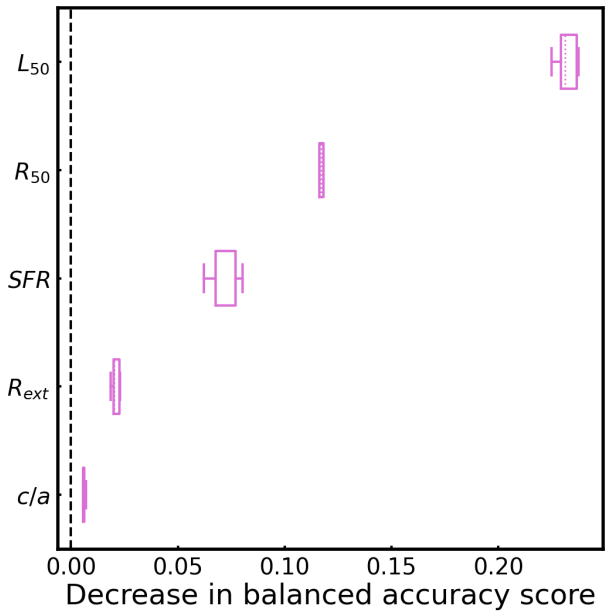

Figure 5.2: Box plot of each observational feature from the testing set, ranked by their importance as determined by the *feature_permutation_* attribute from SMOTE+RF. Each box represents the range of the different scores obtained from a cross-validation with $n_{iter} = 5$. The inner dashed line represents the median value of each distribution. The whiskers on each box represent the minimum and maximum value of each distribution.

ples of galaxies, without the need of parameters that required additional modeling to be estimated, such as μ_* and T_P. This approach could be particularly valuable in large-scale surveys such as those soon will be provided by LSST (Ivezić et al., 2019).

Chapter 6

Discussion and conclusions

In this thesis we selected a large sample of disk-like galaxies from the IllustrisTNG simulation to develop an algorithm capable of automatically classifying galaxies between lopsided and symmetric. Our main goal was to explore whether this classification can be accurately performed using only internal galactic parameters, thus neglecting information about their present-day environment. This notion was concluded thanks to the strong correlation between lopsidedness and the structural properties of galaxies.

To achieve this we employed the Random Forest algorithm, a machine learning approach that involves a supervised training process. To label our data as lopsided and symmetric galaxies we employed a Fourier decomposition of the galaxies' stellar density distribution over the radial interval $R_{50} - 1.4R_{90}$. We computed a radially average power of the $m = 1$ mode, A_1 within this range. Galaxies with $A_1 > 0.1$ were classified as lopsided, and the remaining as symmetric. Our sample resulted in a total 5,273 lopsided and 2,646 symmetric galaxies. **The total sample was then divided into two datasets, a training set and a testing set. The training set consisted of 5,542 galaxies (70% of the total sample) and the testing set 2,377 galaxies (30% of the total sample).** To avoid problems in the classification process due to the imbalanced an nature of the dataset, we employed two variations of the RF algorithm: i) we used SMOTE to oversample symmetric galaxies in the training set, thus evening both classes, and ii) we used the BRF algorithm, which balances both classes on each tree by only bootstrapping the minority class while undersampling the majority.

Based on the considered metrics, we selected SMOTE+RF as the best model. The classification resulted in a total of 1,922 and 455 correctly and incorrectly classified galaxies, respectively. This translates in a balanced accuracy of accurate classification rate of $\approx 80\%$ of both classes. To interpret and understand the different decisions leading the RF to the classification, we used a method to quantify "features importance". In particular we utilized and algorithm that randomly permutes features' values and calculates the decrease in a certain metric; which in our case we choose balanced-accuracy. We found that, to distinguish between both classes, the three most important parameters for the model are μ_*, $T_{\rm P}$, and *SFR*. The excellent results obtained by our classifier, trained with features that do not account for the galaxies environment, strongly supports the hypothesis that lopsidedness is mainly a tracer of galaxies internal structures.

Even though our classifier demonstrated a very good performance, we find that $\approx 20\%$ of the galaxies were misclassified. To study the misclassified cases, we first explore the distribution of the main parameters used by the RF. First, we find that the A_1 value of the misclassified cases lies very close to the threshold used to label galaxies as lopsided or symmetric. As a result, these cases are typically associated to "borderline classifications" by A_1. Interestingly, we find that the distribution of the most important parameters, such as μ_* and $T_{\rm p}$ are in good agreement with class they have been associated to by the RF algorithm. In other words, galaxies classified by A_1 as lopsided, but as symmetric by the RF, have large μ_* and $T_{\rm p}$ values. Conversely, galaxies classified by A_1 as symmetric, but as lopsided by the RF, have low μ_* and $T_{\rm p}$ values.

To further explore why galaxies with large central surface density and strongly cohesive present perturbed outer disk region, we selected a representative case. We find that the selected galaxy became a satellite of a more massive host ≈ 1.7 Gyr ago. Previous to the crossing of host virial radius, the galaxy had a symmetric configuration. However, shortly after its first pericentric passage its outer regions become perturbed due to the strong tidal interaction. Such strong and recent interaction induced a temporary lopsided perturbation on this galaxy. We find that 28% of this misclassified class are either satellites of a more massive host, or have had a very recent strong tidal interactions with a massive companion (>1:20). For the other misclassified cases, other mechanism, such as asymmetric

gas accretion, must be considered to explain the classifications. We will further explore this in a follow up analysis. In the case of galaxies with low μ_* and $T_{\rm P}$ misclassified as symmetric by the RF algorithm, we find that, typically, they have not experienced recent significant interactions with massive companions. Thus, even though the are susceptible to develop a lopsided perturbation, no external interaction have trigger its onset.

Several parameters considered in this work as features require additional modeling to be estimated. Considering the advent of several surveys such as S-PLUS (Mendes de Oliveira et al., 2019), J-PAS (Benitez et al., 2014), and the LSST (Ivezić et al., 2019), we explored whether the performance of our classifier significantly drops when considering features that can be readily obtained from multiband photometric surveys. In particular, we replace stellar mass estimates with their corresponding luminosity in the r-band, and dropped parameters such as $T_{\rm p}$ that involve dynamical modeling to estimate the total galaxy mass within R_{50}. Interestingly, we find the performance of our modeling is very mildly affected, with recovery rates of $\sim 78\%$. These results are very promising, as our algorithm could allow us to rapidly extract samples of lopsided galaxies from large surveys, allowing us to explore whether lopsidedness in present-day disk galaxies is connected to their specific evolutionary histories, which shaped their distinct internal properties (Dolfi et al., 2023).

Chapter 7

Future work

A previous part of this thesis was to study the images of lopsided and symmetric galaxies, also obtained from TNG50, using a CNN (see Sect.1.2 for a brief description). The main goal is this case was to study the morphological differences of their galactic disk with the information obtained from the neural network, with which both types of galaxies could be then characterized and classified. This was done in a CNN used to classify stellar stream in Milky Way-like galaxies from the Auriga simulation, part of the doctorate thesis of Alex Casanova. As both types of galaxies' images were similar, we did not have any problem in using our sample to train and test that classifier. We tested different hyperparameters (such as batch size, regularization, among others) and performed an oversampling and undersampling to the images due the imbalanced nature of our dataset. With this, the resulting accuracy score was of $\sim$80%. However, the misclassified cases had a similar behavior as the ones from SMOTE+RF, where they were also adjacent to the A_1 threshold. Now that we have deeper knowledge in these cases thanks to studying the internal properties of these galaxies, we would like to come back to using CNNs to study lopsidedness as images can also be easily obtained in telescopes and surveys.

Considering the myriad types of machine learning algorithms, we would also like to continue studying and characterizing this asymmetry using new and up-to-date models. For instance, transformers are an interesting algorithm, shown to be one of the most popular and with the highest accuracy among image classi-

fiers (e.g. see Papers with Code[1], a community-based webpage which gathers and ranks open-source deep learning models).

Lastly, and as previously stated, we would like to test our classifiers with observational data from photometric surveys, e.g. from S-PLUS, J-PLUS, S-PAS, and SDSS.

[1] https://paperswithcode.com/about

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., Bailer-Jones, C. A. L., Barentine, J. C., Bassett, B. A., Becker, A. C., Beers, T. C., Bell, E. F., Belokurov, V., Berlind, A. A., Berman, E. F., Bernardi, M., Bickerton, S. J., Bizyaev, D., Blakeslee, J. P., Blanton, M. R., Bochanski, J. J., Boroski, W. N., Brewington, H. J., Brinchmann, J., Brinkmann, J., Brunner, R. J., Budavári, T., Carey, L. N., Carliles, S., Carr, M. A., Castander, F. J., Cinabro, D., Connolly, A. J., Csabai, I., Cunha, C. E., Czarapata, P. C., Davenport, J. R. A., de Haas, E., Dilday, B., Doi, M., Eisenstein, D. J., Evans, M. L., Evans, N. W., Fan, X., Friedman, S. D., Frieman, J. A., Fukugita, M., Gänsicke, B. T., Gates, E., Gillespie, B., Gilmore, G., Gonzalez, B., Gonzalez, C. F., Grebel, E. K., Gunn, J. E., Györy, Z., Hall, P. B., Harding, P., Harris, F. H., Harvanek, M., Hawley, S. L., Hayes, J. J. E., Heckman, T. M., Hendry, J. S., Hennessy, G. S., Hindsley, R. B., Hoblitt, J., Hogan, C. J., Hogg, D. W., Holtzman, J. A., Hyde, J. B., Ichikawa, S.-i., Ichikawa, T., Im, M., Ivezić, Ž., Jester, S., Jiang, L., Johnson, J. A., Jorgensen, A. M., Jurić, M., Kent, S. M., Kessler, R., Kleinman, S. J., Knapp, G. R., Konishi, K., Kron, R. G., Krzesinski, J., Kuropatkin, N., Lampeitl, H., Lebedeva, S., Lee, M. G., Lee,

Y. S., French Leger, R., Lépine, S., Li, N., Lima, M., Lin, H., Long, D. C., Loomis, C. P., Loveday, J., Lupton, R. H., Magnier, E., Malanushenko, O., Malanushenko, V., Mandelbaum, R., Margon, B., Marriner, J. P., Martínez-Delgado, D., Matsubara, T., McGehee, P. M., McKay, T. A., Meiksin, A., Morrison, H. L., Mullally, F., Munn, J. A., Murphy, T., Nash, T., Nebot, A., Neilsen, Jr., E. H., Newberg, H. J., Newman, P. R., Nichol, R. C., Nicinski, T., Nieto-Santisteban, M., Nitta, A., Okamura, S., Oravetz, D. J., Ostriker, J. P., Owen, R., Padmanabhan, N., Pan, K., Park, C., Pauls, G., Peoples, Jr., J., Percival, W. J., Pier, J. R., Pope, A. C., Pourbaix, D., Price, P. A., Purger, N., Quinn, T., Raddick, M. J., Re Fiorentin, P., Richards, G. T., Richmond, M. W., Riess, A. G., Rix, H.-W., Rockosi, C. M., Sako, M., Schlegel, D. J., Schneider, D. P., Scholz, R.-D., Schreiber, M. R., Schwope, A. D., Seljak, U., Sesar, B., Sheldon, E., Shimasaku, K., Sibley, V. C., Simmons, A. E., Sivarani, T., Allyn Smith, J., Smith, M. C., Smolčić, V., Snedden, S. A., Stebbins, A., Steinmetz, M., Stoughton, C., Strauss, M. A., SubbaRao, M., Suto, Y., Szalay, A. S., Szapudi, I., Szkody, P., Tanaka, M., Tegmark, M., Teodoro, L. F. A., Thakar, A. R., Tremonti, C. A., Tucker, D. L., Uomoto, A., Vanden Berk, D. E., Vandenberg, J., Vidrih, S., Vogeley, M. S., Voges, W., Vogt, N. P., Wadadekar, Y., Watters, S., Weinberg, D. H., West, A. A., White, S. D. M., Wilhite, B. C., Wonders, A. C., Yanny, B., and Yocum, D. R. (2009). The Seventh Data Release of the Sloan Digital Sky Survey. *Astrophys. J. Suppl.*, 182(2):543–558.

Abraham, R. G., van den Bergh, S., Glazebrook, K., Ellis, R. S., Santiago, B. X., Surma, P., and Griffiths, R. E. (1996). The Morphologies of Distant Galaxies. II. Classifications from the Hubble Space Telescope Medium Deep Survey. *Astrophys. J. Suppl.*, 107:1.

Andreon, S., Gargiulo, G., Longo, G., Tagliaferri, R., and Capuano, N. (2000). Wide field imaging - I. Applications of neural networks to object detection and star/galaxy classification. *Mon. Not. R. Astron. Soc.*, 319(3):700–716.

Angiras, R. A., Jog, C. J., Omar, A., and Dwarakanath, K. S. (2006). Origin of disc lopsidedness in the Eridanus group of galaxies. *Mon. Not. R. Astron. Soc.*, 369(4):1849–1857.

Angthopo, J., Negri, A., Ferreras, I., de la Rosa, I. G., Dalla Vecchia, C., and Pillepich, A. (2021). Evaluating hydrodynamical simulations with green valley galaxies. *Mon. Not. R. Astron. Soc.*, 502(3):3685–3702.

Appleby, S., Davé, R., Sorini, D., Lovell, C. C., and Lo, K. (2023). Mapping circumgalactic medium observations to theory using machine learning. *Mon. Not. R. Astron. Soc.*, 525(1):1167–1181.

AstroDave, AstroTom, Winton, C. R. ., joycenv, and Willett, K. (2013). Galaxy zoo - the galaxy challenge. https://kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge. Kaggle.

Aumer, M., White, S. D. M., Naab, T., and Scannapieco, C. (2013). Towards a more realistic population of bright spiral galaxies in cosmological simulations. *Mon. Not. R. Astron. Soc.*, 434(4):3142–3164.

Baldwin, J. E., Lynden-Bell, D., and Sancisi, R. (1980). Lopsided galaxies. *Mon. Not. R. Astron. Soc.*, 193:313–319.

Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., and Brunner, R. J. (2004). Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. *Mon. Not. R. Astron. Soc.*, 348(3):1038–1046.

Beale, J. S. and Davies, R. D. (1969). Neutral Hydrogen Asymmetry in the Galaxy M101 as Evidence for Tidal Effects. *Nature*, 221:531–533.

Bellm, E. C., Kulkarni, S. R., Graham, M. J., Dekany, R., Smith, R. M., Riddle, R., Masci, F. J., Helou, G., Prince, T. A., Adams, S. M., Barbarino, C., Barlow, T., Bauer, J., Beck, R., Belicki, J., Biswas, R., Blagorodnova, N., Bodewits, D., Bolin, B., Brinnel, V., Brooke, T., Bue, B., Bulla, M., Burruss, R., Cenko, S. B., Chang, C.-K., Connolly, A., Coughlin, M., Cromer, J., Cunningham, V., De, K., Delacroix, A., Desai, V., Duev, D. A., Eadie, G., Farnham, T. L., Feeney, M., Feindt, U., Flynn, D., Franckowiak, A., Frederick, S., Fremling, C., Gal-Yam, A., Gezari, S., Giomi, M., Goldstein, D. A., Golkhou, V. Z., Goobar, A., Groom, S., Hacopians, E., Hale, D., Henning, J., Ho, A. Y. Q., Hover, D., Howell, J., Hung, T., Huppenkothen, D., Imel, D., Ip, W.-H., Ivezić, Ž., Jackson, E., Jones, L., Juric, M., Kasliwal, M. M., Kaspi, S., Kaye, S., Kelley,

M. S. P., Kowalski, M., Kramer, E., Kupfer, T., Landry, W., Laher, R. R., Lee, C.-D., Lin, H. W., Lin, Z.-Y., Lunnan, R., Giomi, M., Mahabal, A., Mao, P., Miller, A. A., Monkewitz, S., Murphy, P., Ngeow, C.-C., Nordin, J., Nugent, P., Ofek, E., Patterson, M. T., Penprase, B., Porter, M., Rauch, L., Rebbapragada, U., Reiley, D., Rigault, M., Rodriguez, H., van Roestel, J., Rusholme, B., van Santen, J., Schulze, S., Shupe, D. L., Singer, L. P., Soumagnac, M. T., Stein, R., Surace, J., Sollerman, J., Szkody, P., Taddia, F., Terek, S., Van Sistine, A., van Velzen, S., Vestrand, W. T., Walters, R., Ward, C., Ye, Q.-Z., Yu, P.-C., Yan, L., and Zolkower, J. (2019). The Zwicky Transient Facility: System Overview, Performance, and First Results. *Pub. Astron. Soc. Pacific*, 131(995):018002.

Belén Barreiro, R. (2000). The cosmic microwave background. *New Astronomy Reviews*, 44(3):179–204.

Benitez, N., Dupke, R., Moles, M., Sodre, L., Cenarro, J., Marin-Franch, A., Taylor, K., Cristobal, D., Fernandez-Soto, A., de Oliveira, C. M., Cepa-Nogue, J., Abramo, L. R., Alcaniz, J. S., Overzier, R., Hernandez-Monteagudo, C., Alfaro, E. J., Kanaan, A., Carvano, J. M., Reis, R. R. R., Gonzalez, E. M., Ascaso, B., Ballesteros, F., Xavier, H. S., Varela, J., Ederoclite, A., Ramio, H. V., Broadhurst, T., Cypriano, E., Angulo, R., Diego, J. M., Zandivarez, A., Diaz, E., Melchior, P., Umetsu, K., Spinelli, P. F., Zitrin, A., Coe, D., Yepes, G., Vielva, P., Sahni, V., Marcos-Caballero, A., Kitaura, F. S., Maroto, A. L., Masip, M., Tsujikawa, S., Carneiro, S., Nuevo, J. G., Carvalho, G. C., Reboucas, M. J., Carvalho, J. C., Abdalla, E., Bernui, A., Pigozzo, C., Ferreira, E. G. M., Devi, N. C., au2, C. A. P. B. J., Campista, M., Amorim, A., Asari, N. V., Bongiovanni, A., Bonoli, S., Bruzual, G., Cardiel, N., Cava, A., Fernandes, R. C., Coelho, P., Cortesi, A., Delgado, R. G., Garcia, L. D., Espinosa, J. M. R., Galliano, E., Gonzalez-Serrano, J. I., Falcon-Barroso, J., Fritz, J., Fernandes, C., Gorgas, J., Hoyos, C., Jimenez-Teja, Y., Lopez-Aguerri, J. A., Juan, C. L.-S., Mateus, A., Molino, A., Novais, P., OMill, A., Oteo, I., Perez-Gonzalez, P. G., Poggianti, B., Proctor, R., Ricciardelli, E., Sanchez-Blazquez, P., Storchi-Bergmann, T., Telles, E., Schoennell, W., Trujillo, N., Vazdekis, A., Viironen, K., Daflon, S., Aparicio-Villegas, T., Rocha, D., Ribeiro, T., Borges, M., Martins, S. L., Marcolino, W., Martinez-Delgado, D., Perez-Torres, M. A., Siffert, B. B., Calvao, M. O., Sako, M., Kessler, R., Alvarez-Candal, A., Pra, M. D., Roig, F., Lazzaro, D., Gorosabel, J., de Oliveira, R. L., Lima-Neto, G. B., Irwin, J., Liu, J. F.,

Alvarez, E., Balmes, I., Chueca, S., Costa-Duarte, M. V., da Costa, A. A., Dantas, M. L. L., Diaz, A. Y., Fabregat, J., Ferrari, F., Gavela, B., Gracia, S. G., Gruel, N., Gutierrez, J. L. L., Guzman, R., Hernandez-Fernandez, J. D., Herranz, D., Hurtado-Gil, L., Jablonsky, F., Laporte, R., Tiran, L. L. L., Licandro, J., Lima, M., Martin, E., Martinez, V., Montero, J. J. C., Penteado, P., Pereira, C. B., Peris, V., Quilis, V., Sanchez-Portal, M., Soja, A. C., Solano, E., Torra, J., and Valdivielso, L. (2014). J-pas: The javalambre-physics of the accelerated universe astrophysical survey.

Blake, C. and Bridle, S. (2005). Cosmology with photometric redshift surveys. *Mon. Not. R. Astron. Soc.*, 363(4):1329–1348.

Bluck, A. F. L., Conselice, C. J., Ormerod, K., Piotrowska, J. M., Adams, N., Austin, D., Caruana, J., Duncan, K. J., Ferreira, L., Goubert, P., Harvey, T., Trussler, J., and Maiolino, R. (2024). Galaxy Quenching at the High Redshift Frontier: A Fundamental Test of Cosmological Models in the Early Universe with JWST-CEERS. *Astrophys. J.*, 961(2):163.

Bolzonella, M., Miralles, J. M., and Pelló, R. (2000). Photometric redshifts based on standard SED fitting procedures. *Astron. Astrph.*, 363:476–492.

Boroson, T. A. and Green, R. F. (1992). The Emission-Line Properties of Low-Redshift Quasi-stellar Objects. *Astrophys. J. Suppl.*, 80:109.

Bournaud, F., Combes, F., Jog, C. J., and Puerari, I. (2005). Lopsided spiral galaxies: evidence for gas accretion. *Astron. Astrph.*, 438(2):507–520.

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Buta, R. J. (2011). Galaxy morphology.

Carliles, S., Budavári, T., Heinis, S., Priebe, C., and Szalay, A. S. (2010). Random Forests for Photometric Redshifts. *Astrophys. J.*, 712(1):511–515.

Catinella, B., Saintonge, A., Janowiecki, S., Cortese, L., Davé, R., Lemonias, J. J., Cooper, A. P., Schiminovich, D., Hummels, C. B., Fabello, S., Geréb, K., Kilborn, V., and Wang, J. (2018). xGASS: total cold gas scaling relations and molecular-to-atomic gas ratios of galaxies in the local Universe. *Mon. Not. R. Astron. Soc.*, 476(1):875–895.

Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., Marín-Franch, A., Ederoclite, A., Varela, J., López-Sanjuan, C., Hernández-Monteagudo, C., Angulo, R. E., Vázquez Ramió, H., Viironen, K., Bonoli, S., Orsi, A. A., Hurier, G., San Roman, I., Greisel, N., Vilella-Rojo, G., Díaz-García, L. A., Logroño-García, R., Gurung-López, S., Spinoso, D., Izquierdo-Villalba, D., Aguerri, J. A. L., Allende Prieto, C., Bonatto, C., Carvano, J. M., Chies-Santos, A. L., Daflon, S., Dupke, R. A., Falcón-Barroso, J., Gonçalves, D. R., Jiménez-Teja, Y., Molino, A., Placco, V. M., Solano, E., Whitten, D. D., Abril, J., Antón, J. L., Bello, R., Bielsa de Toledo, S., Castillo-Ramírez, J., Chueca, S., Civera, T., Díaz-Martín, M. C., Domínguez-Martínez, M., Garzarán-Calderaro, J., Hernández-Fuertes, J., Iglesias-Marzoa, R., Iñiguez, C., Jiménez Ruiz, J. M., Kruuse, K., Lamadrid, J. L., Lasso-Cabrera, N., López-Alegre, G., López-Sainz, A., Maícas, N., Moreno-Signes, A., Muniesa, D. J., Rodríguez-Llano, S., Rueda-Teruel, F., Rueda-Teruel, S., Soriano-Laguía, I., Tilve, V., Valdivielso, L., Yanes-Díaz, A., Alcaniz, J. S., Mendes de Oliveira, C., Sodré, L., Coelho, P., Lopes de Oliveira, R., Tamm, A., Xavier, H. S., Abramo, L. R., Akras, S., Alfaro, E. J., Alvarez-Candal, A., Ascaso, B., Beasley, M. A., Beers, T. C., Borges Fernandes, M., Bruzual, G. R., Buzzo, M. L., Carrasco, J. M., Cepa, J., Cortesi, A., Costa-Duarte, M. V., De Prá, M., Favole, G., Galarza, A., Galbany, L., Garcia, K., González Delgado, R. M., González-Serrano, J. I., Gutiérrez-Soto, L. A., Hernandez-Jimenez, J. A., Kanaan, A., Kuncarayakti, H., Landim, R. C. G., Laur, J., Licandro, J., Lima Neto, G. B., Lyman, J. D., Maíz Apellániz, J., Miralda-Escudé, J., Morate, D., Nogueira-Cavalcante, J. P., Novais, P. M., Oncins, M., Oteo, I., Overzier, R. A., Pereira, C. B., Rebassa-Mansergas, A., Reis, R. R. R., Roig, F., Sako, M., Salvador-Rusiñol, N., Sampedro, L., Sánchez-Blázquez, P., Santos, W. A., Schmidtobreick, L., Siffert, B. B., Telles, E., and

Vilchez, J. M. (2019). J-PLUS: The Javalambre Photometric Local Universe Survey. *Astron. Astrph.*, 622:A176.

Chen, C. and Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*.

Chen, Y. and Gnedin, O. Y. (2022). Modeling the kinematics of globular cluster systems. *Mon. Not. R. Astron. Soc.*, 514(4):4736–4755.

Chilingarian, I. V., Di Matteo, P., Combes, F., Melchior, A.-L., and Semelin, B. (2010). The galmer database: galaxy mergers in the virtual observatory. *Astronomy and Astrophysics*, 518:A61.

Collister, A. A. and Lahav, O. (2004). ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *Pub. Astron. Soc. Pacific*, 116(818):345–351.

Conselice, C. J. (1997). The symmetry, color, and morphology of galaxies. *Publications of the Astronomical Society of the Pacific*, 109(741):1251.

Conselice, C. J. (2014). The Evolution of Galaxy Structure Over Cosmic Time. *Annual Review of Astronomy & Astrophysics*, 52:291–337.

Conselice, C. J., Bershady, M. A., and Jangren, A. (2000). The Asymmetry of Galaxies: Physical Morphology for Nearby and High-Redshift Galaxies. *Astrophys. J.*, 529(2):886–910.

de Lapparent, V., Bellanger, C., Arnouts, S., Mathez, G., Mellier, Y., and Mazure, A. (1993). Mapping the large-scale structure with the ESO multi-slit spectrographs. *The Messenger*, 72:34–38.

Degirmenci, A. and Karal, O. (2022). Efficient density and cluster based incremental outlier detection in data streams. *Information Sciences*, 607:901–920.

Dewdney, P. E., Hall, P. J., Schilizzi, R. T., and Lazio, T. J. L. W. (2009). The square kilometre array. *Proceedings of the IEEE*, 97(8):1482–1496.

Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Mon. Not. R. Astron. Soc.*, 450(2):1441–1459.

D'Isanto, A. and Polsterer, K. L. (2018). Photometric redshift estimation via deep learning. Generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astron. Astrph.*, 609:A111.

Dolfi, A., Gómez, F. A., Monachesi, A., Varela-Lavin, S., Tissera, P. B., Sifón, C., and Galaz, G. (2023). Lopsidedness as a tracer of early galactic assembly history. *Mon. Not. R. Astron. Soc.*, 526(1):567–584.

Erwin, P. (2019). What determines the sizes of bars in spiral galaxies? *Mon. Not. R. Astron. Soc.*, 489(3):3553–3564.

Eskridge, P. B., Frogel, J. A., Pogge, R. W., Quillen, A. C., Berlind, A. A., Davies, R. L., DePoy, D. L., Gilbert, K. M., Houdashelt, M. L., Kuchinski, L. E., Ramírez, S. V., Sellgren, K., Stutz, A., Terndrup, D. M., and Tiede, G. P. (2002). Near-infrared and optical morphology of spiral galaxies*. *The Astrophysical Journal Supplement Series*, 143(1):73.

Eyer, L. and Blake, C. (2005). Automated classification of variable stars for All-Sky Automated Survey 1-2 data. *Mon. Not. R. Astron. Soc.*, 358(1):30–38.

Fotopoulou, S. (2024). A review of unsupervised learning in astronomy. *Astronomy and Computing*, 48:100851.

Frei, Z., Guhathakurta, P., Gunn, J. E., and Tyson, J. A. (1996). A Catalog of Digital Images of 113 Nearby Galaxies. *Astron. J.*, 111:174.

Frenk, C. S., Baugh, C. M., and Cole, S. (1996). Galaxy formation and evolution: What to expect from hierarchical clustering models. In Bender, R. and Davies, R. L., editors, *New Light on Galaxy Evolution*, pages 247–254, Dordrecht. Springer Netherlands.

Frontera-Pons, J., Sureau, F., Bobin, J., and Le Floc'h, E. (2017). Unsupervised feature-learning for galaxy SEDs with denoising autoencoders. *Astron. Astrph.*, 603:A60.

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., Babusiaux, C., Bailer-Jones, C. A. L., Bastian, U., Biermann, M., Evans, D. W., Eyer, L., Jansen, F., Jordi, C., Klioner, S. A., Lammers, U., Lindegren, L., Luri, X., Mignard, F., Milligan, D. J., Panem, C., Poinsignon, V., Pourbaix,

D., Randich, S., Sarri, G., Sartoretti, P., Siddiqui, H. I., Soubiran, C., Valette, V., van Leeuwen, F., Walton, N. A., Aerts, C., Arenou, F., Cropper, M., Drimmel, R., Høg, E., Katz, D., Lattanzi, M. G., O'Mullane, W., Grebel, E. K., Holland, A. D., Huc, C., Passot, X., Bramante, L., Cacciari, C., Castañeda, J., Chaoul, L., Cheek, N., De Angeli, F., Fabricius, C., Guerra, R., Hernández, J., Jean-Antoine-Piccolo, A., Masana, E., Messineo, R., Mowlavi, N., Nienartowicz, K., Ordóñez-Blanco, D., Panuzzo, P., Portell, J., Richards, P. J., Riello, M., Seabroke, G. M., Tanga, P., Thévenin, F., Torra, J., Els, S. G., Gracia-Abril, G., Comoretto, G., Garcia-Reinaldos, M., Lock, T., Mercier, E., Altmann, M., Andrae, R., Astraatmadja, T. L., Bellas-Velidis, I., Benson, K., Berthier, J., Blomme, R., Busso, G., Carry, B., Cellino, A., Clementini, G., Cowell, S., Creevey, O., Cuypers, J., Davidson, M., De Ridder, J., de Torres, A., Delchambre, L., Dell'Oro, A., Ducourant, C., Frémat, Y., García-Torres, M., Gosset, E., Halbwachs, J. L., Hambly, N. C., Harrison, D. L., Hauser, M., Hestroffer, D., Hodgkin, S. T., Huckle, H. E., Hutton, A., Jasniewicz, G., Jordan, S., Kontizas, M., Korn, A. J., Lanzafame, A. C., Manteiga, M., Moitinho, A., Muinonen, K., Osinde, J., Pancino, E., Pauwels, T., Petit, J. M., Recio-Blanco, A., Robin, A. C., Sarro, L. M., Siopis, C., Smith, M., Smith, K. W., Sozzetti, A., Thuillot, W., van Reeven, W., Viala, Y., Abbas, U., Abreu Aramburu, A., Accart, S., Aguado, J. J., Allan, P. M., Allasia, W., Altavilla, G., Álvarez, M. A., Alves, J., Anderson, R. I., Andrei, A. H., Anglada Varela, E., Antiche, E., Antoja, T., Antón, S., Arcay, B., Atzei, A., Ayache, L., Bach, N., Baker, S. G., Balaguer-Núñez, L., Barache, C., Barata, C., Barbier, A., Barblan, F., Baroni, M., Barrado y Navascués, D., Barros, M., Barstow, M. A., Becciani, U., Bellazzini, M., Bellei, G., Bello García, A., Belokurov, V., Bendjoya, P., Berihuete, A., Bianchi, L., Bienaymé, O., Billebaud, F., Blagorodnova, N., Blanco-Cuaresma, S., Boch, T., Bombrun, A., Borrachero, R., Bouquillon, S., Bourda, G., Bouy, H., Bragaglia, A., Breddels, M. A., Brouillet, N., Brüsemeister, T., Bucciarelli, B., Budnik, F., Burgess, P., Burgon, R., Burlacu, A., Busonero, D., Buzzi, R., Caffau, E., Cambras, J., Campbell, H., Cancelliere, R., Cantat-Gaudin, T., Carlucci, T., Carrasco, J. M., Castellani, M., Charlot, P., Charnas, J., Charvet, P., Chassat, F., Chiavassa, A., Clotet, M., Cocozza, G., Collins, R. S., Collins, P., Costigan, G., Crifo, F., Cross, N. J. G., Crosta, M., Crowley, C., Dafonte, C., Damerdji, Y., Dapergolas, A., David, P., David, M., De Cat, P., de Felice, F., de Laverny, P., De Luise, F., De March, R., de Martino, D., de Souza, R.,

Debosscher, J., del Pozo, E., Delbo, M., Delgado, A., Delgado, H. E., di Marco, F., Di Matteo, P., Diakite, S., Distefano, E., Dolding, C., Dos Anjos, S., Drazinos, P., Durán, J., Dzigan, Y., Ecale, E., Edvardsson, B., Enke, H., Erdmann, M., Escolar, D., Espina, M., Evans, N. W., Eynard Bontemps, G., Fabre, C., Fabrizio, M., Faigler, S., Falcão, A. J., Farràs Casas, M., Faye, F., Federici, L., Fedorets, G., Fernández-Hernández, J., Fernique, P., Fienga, A., Figueras, F., Filippi, F., Findeisen, K., Fonti, A., Fouesneau, M., Fraile, E., Fraser, M., Fuchs, J., Furnell, R., Gai, M., Galleti, S., Galluccio, L., Garabato, D., García-Sedano, F., Garé, P., Garofalo, A., Garralda, N., Gavras, P., Gerssen, J., Geyer, R., Gilmore, G., Girona, S., Giuffrida, G., Gomes, M., González-Marcos, A., González-Núñez, J., González-Vidal, J. J., Granvik, M., Guerrier, A., Guillout, P., Guiraud, J., Gúrpide, A., Gutiérrez-Sánchez, R., Guy, L. P., Haigron, R., Hatzidimitriou, D., Haywood, M., Heiter, U., Helmi, A., Hobbs, D., Hofmann, W., Holl, B., Holland, G., Hunt, J. A. S., Hypki, A., Icardi, V., Irwin, M., Jevardat de Fombelle, G., Jofré, P., Jonker, P. G., Jorissen, A., Julbe, F., Karampelas, A., Kochoska, A., Kohley, R., Kolenberg, K., Kontizas, E., Koposov, S. E., Kordopatis, G., Koubsky, P., Kowalczyk, A., Krone-Martins, A., Kudryashova, M., Kull, I., Bachchan, R. K., Lacoste-Seris, F., Lanza, A. F., Lavigne, J. B., Le Poncin-Lafitte, C., Lebreton, Y., Lebzelter, T., Leccia, S., Leclerc, N., Lecoeur-Taibi, I., Lemaitre, V., Lenhardt, H., Leroux, F., Liao, S., Licata, E., Lindstrøm, H. E. P., Lister, T. A., Livanou, E., Lobel, A., Löffler, W., López, M., Lopez-Lozano, A., Lorenz, D., Loureiro, T., MacDonald, I., Magalhães Fernandes, T., Managau, S., Mann, R. G., Mantelet, G., Marchal, O., Marchant, J. M., Marconi, M., Marie, J., Marinoni, S., Marrese, P. M., Marschalkó, G., Marshall, D. J., Martín-Fleitas, J. M., Martino, M., Mary, N., Matijevič, G., Mazeh, T., McMillan, P. J., Messina, S., Mestre, A., Michalik, D., Millar, N. R., Miranda, B. M. H., Molina, D., Molinaro, R., Molinaro, M., Molnár, L., Moniez, M., Montegriffo, P., Monteiro, D., Mor, R., Mora, A., Morbidelli, R., Morel, T., Morgenthaler, S., Morley, T., Morris, D., Mulone, A. F., Muraveva, T., Musella, I., Narbonne, J., Nelemans, G., Nicastro, L., Noval, L., Ordénovic, C., Ordieres-Meré, J., Osborne, P., Pagani, C., Pagano, I., Pailler, F., Palacin, H., Palaversa, L., Parsons, P., Paulsen, T., Pecoraro, M., Pedrosa, R., Pentikäinen, H., Pereira, J., Pichon, B., Piersimoni, A. M., Pineau, F. X., Plachy, E., Plum, G., Poujoulet, E., Prša, A., Pulone, L., Ragaini, S., Rago, S., Rambaux, N., Ramos-Lerate, M., Ranalli, P., Rauw, G., Read, A., Regibo,

S., Renk, F., Reylé, C., Ribeiro, R. A., Rimoldini, L., Ripepi, V., Riva, A., Rixon, G., Roelens, M., Romero-Gómez, M., Rowell, N., Royer, F., Rudolph, A., Ruiz-Dern, L., Sadowski, G., Sagristà Sellés, T., Sahlmann, J., Salgado, J., Salguero, E., Sarasso, M., Savietto, H., Schnorhk, A., Schultheis, M., Sciacca, E., Segol, M., Segovia, J. C., Segransan, D., Serpell, E., Shih, I. C., Smareglia, R., Smart, R. L., Smith, C., Solano, E., Solitro, F., Sordo, R., Soria Nieto, S., Souchay, J., Spagna, A., Spoto, F., Stampa, U., Steele, I. A., Steidelmüller, H., Stephenson, C. A., Stoev, H., Suess, F. F., Süveges, M., Surdej, J., Szabados, L., Szegedi-Elek, E., Tapiador, D., Taris, F., Tauran, G., Taylor, M. B., Teixeira, R., Terrett, D., Tingley, B., Trager, S. C., Turon, C., Ulla, A., Utrilla, E., Valentini, G., van Elteren, A., Van Hemelryck, E., van Leeuwen, M., Varadi, M., Vecchiato, A., Veljanoski, J., Via, T., Vicente, D., Vogt, S., Voss, H., Votruba, V., Voutsinas, S., Walmsley, G., Weiler, M., Weingrill, K., Werner, D., Wevers, T., Whitehead, G., Wyrzykowski, Ł., Yoldas, A., Žerjal, M., Zucker, S., Zurbach, C., Zwitter, T., Alecu, A., Allen, M., Allende Prieto, C., Amorim, A., Anglada-Escudé, G., Arsenijevic, V., Azaz, S., Balm, P., Beck, M., Bernstein, H. H., Bigot, L., Bijaoui, A., Blasco, C., Bonfigli, M., Bono, G., Boudreault, S., Bressan, A., Brown, S., Brunet, P. M., Bunclark, P., Buonanno, R., Butkevich, A. G., Carret, C., Carrion, C., Chemin, L., Chéreau, F., Corcione, L., Darmigny, E., de Boer, K. S., de Teodoro, P., de Zeeuw, P. T., Delle Luche, C., Domingues, C. D., Dubath, P., Fodor, F., Frézouls, B., Fries, A., Fustes, D., Fyfe, D., Gallardo, E., Gallegos, J., Gardiol, D., Gebran, M., Gomboc, A., Gómez, A., Grux, E., Gueguen, A., Heyrovsky, A., Hoar, J., Iannicola, G., Isasi Parache, Y., Janotto, A. M., Joliet, E., Jonckheere, A., Keil, R., Kim, D. W., Klagyivik, P., Klar, J., Knude, J., Kochukhov, O., Kolka, I., Kos, J., Kutka, A., Lainey, V., LeBouquin, D., Liu, C., Loreggia, D., Makarov, V. V., Marseille, M. G., Martayan, C., Martinez-Rubi, O., Massart, B., Meynadier, F., Mignot, S., Munari, U., Nguyen, A. T., Nordlander, T., Ocvirk, P., O'Flaherty, K. S., Olias Sanz, A., Ortiz, P., Osorio, J., Oszkiewicz, D., Ouzounis, A., Palmer, M., Park, P., Pasquato, E., Peltzer, C., Peralta, J., Péturaud, F., Pieniluoma, T., Pigozzi, E., Poels, J., Prat, G., Prod'homme, T., Raison, F., Rebordao, J. M., Risquez, D., Rocca-Volmerange, B., Rosen, S., Ruiz-Fuertes, M. I., Russo, F., Sembay, S., Serraller Vizcaino, I., Short, A., Siebert, A., Silva, H., Sinachopoulos, D., Slezak, E., Soffel, M., Sosnowska, D., Straižys, V., ter Linden, M., Terrell, D., Theil, S., Tiede, C., Troisi, L., Tsalmantza, P., Tur, D., Vaccari,

M., Vachier, F., Valles, P., Van Hamme, W., Veltz, L., Virtanen, J., Wallut, J. M., Wichmann, R., Wilkinson, M. I., Ziaeepour, H., and Zschocke, S. (2016). The Gaia mission. *Astron. Astrph.*, 595:A1.

Galaz, G. and de Lapparent, V. (1997). The eso-sculptor survey: Spectral classification of galaxies with z ¡ 0.5.

Gao, D., Zhang, Y.-X., and Zhao, Y.-H. (2009). Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, 9(2):220–226.

Gardner, J. P., Mather, J. C., Clampin, M., Doyon, R., Greenhouse, M. A., Hammel, H. B., Hutchings, J. B., Jakobsen, P., Lilly, S. J., Long, K. S., Lunine, J. I., McCaughrean, M. J., Mountain, M., Nella, J., Rieke, G. H., Rieke, M. J., Rix, H.-W., Smith, E. P., Sonneborn, G., Stiavelli, M., Stockman, H. S., Windhorst, R. A., and Wright, G. S. (2006). The James Webb Space Telescope. *Space Science Reviews*, 123(4):485–606.

Genel, S., Vogelsberger, M., Springel, V., Sijacki, D., Nelson, D., Snyder, G., Rodriguez-Gomez, V., Torrey, P., and Hernquist, L. (2014). Introducing the Illustris project: the evolution of galaxy populations across cosmic time. *Mon. Not. R. Astron. Soc.*, 445(1):175–200.

Gómez, F. A., White, S. D. M., Marinacci, F., Slater, C. T., Grand, R. J. J., Springel, V., and Pakmor, R. (2016). A fully cosmological model of a Monoceros-like ring. *Mon. Not. R. Astron. Soc.*, 456(3):2779–2793.

Grand, R. J. J., Springel, V., Gómez, F. A., Marinacci, F., Pakmor, R., Campbell, D. J. R., and Jenkins, A. (2016). Vertical disc heating in Milky Way-sized galaxies in a cosmological context. *Mon. Not. R. Astron. Soc.*, 459(1):199–219.

Guhathakurta, P., van Gorkom, J. H., Kotanyi, C. G., and Balkowski, C. (1988). A VLA H I Survey of the Virgo Cluster Spirals. II. Rotation Curves. *Astron. J.*, 96:851.

Gupta, A., Anpalagan, A., Guan, L., and Khwaja, A. S. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057.

Guzmán-Ortega, A., Rodriguez-Gomez, V., Snyder, G. F., Chamberlain, K., and Hernquist, L. (2023). Morphological signatures of mergers in the TNG50 simulation and the Kilo-Degree Survey: the merger fraction from dwarfs to Milky Way-like galaxies. *Mon. Not. R. Astron. Soc.*, 519(4):4920–4937.

Haslbauer, M., Banik, I., Kroupa, P., Wittenburg, N., and Javanmardi, B. (2022). The high fraction of thin disk galaxies continues to challenge cdm cosmology. *The Astrophysical Journal*, 925(2):183.

Haynes, M. P., Hogg, D. E., Maddalena, R. J., Roberts, M. S., and van Zee, L. (1998). Asymmetry in High-Precision Global H i Profiles of Isolated Spiral Galaxies. *Astron. J.*, 115(1):62–79.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv e-prints*, page arXiv:1502.01852.

Henghes, B., Thiyagalingam, J., Pettitt, C., Hey, T., and Lahav, O. (2022). Deep learning methods for obtaining photometric redshift estimations from images. *Mon. Not. R. Astron. Soc.*, 512(2):1696–1709.

Hotelling, H. (1936). Relations between two sets of variates*. *Biometrika*, 28(3-4):321–377.

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., Pérez-González, P. G., Kartaltepe, J. S., Barro, G., Bernardi, M., Mei, S., Shankar, F., Dimauro, P., Bell, E. F., Kocevski, D., Koo, D. C., Faber, S. M., and Mcintosh, D. H. (2015). A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning. *Astrophys. J. Suppl.*, 221(1):8.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod, T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Bechtol, E., Bechtol, K., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Booth, M. T., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett,

C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Carlin, J. L., Carlson, E. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., de Peyster, R., de Val-Borro, M., Digel, S. W., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eifler, T., Eracleous, M., Emmons, B. L., Fausti Neto, A., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J. B., Gibson, R. R., Gilmore, D. K., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Howard, J. D., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, M. J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W. G., Jones, R. L., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff, K. S., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J. B., Le Guillou, L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz, P. J., Lupton, R. H., Lust, N. B., MacArthur, L. A., Mahabal, A., Mandelbaum, R., Markiewicz, T. W., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., McKercher, R., McQueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Moolekamp, F. E., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Morrison, C. B., Mueller, F., Muller, G. P., Muñoz Arancibia, F., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J. M., Peterson, J. R., Petravick, D. L., Libby Petrick, M. E., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R.,

Plate, S., Plutchak, J. P., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A. P., Regnault, N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J. A., Smith, R. C., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thornton, A. J., Thukral, V., Tice, J., Trilling, D. E., Turri, M., Van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wu, X., Xin, B., Yoachim, P., and Zhan, H. (2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. *Astrophys. J.*, 873(2):111.

Iye, M., Okamura, S., Hamabe, M., and Watanabe, M. (1982). Spectral analysis of the asymmetric spiral pattern of NGC 4254. *Astrophys. J.*, 256:103–111.

Jog, C. J. (1997). Dynamics of Orbits and Local Gas Stability in a Lopsided Galaxy. *Astrophys. J.*, 488(2):642–651.

Jog, C. J. and Combes, F. (2009). Lopsided spiral galaxies. *Phys. Rept.*, 471(2):75–111.

Joshi, G. D., Pillepich, A., Nelson, D., Marinacci, F., Springel, V., Rodriguez-Gomez, V., Vogelsberger, M., and Hernquist, L. (2020). The fate of disc galaxies in IllustrisTNG clusters. *Mon. Not. R. Astron. Soc.*, 496(3):2673–2703.

Kauffmann, G., Heckman, T. M., White, S. D. M., Charlot, S., Tremonti, C., Brinchmann, J., Bruzual, G., Peng, E. W., Seibert, M., Bernardi, M., Blanton, M., Brinkmann, J., Castander, F., Csábai, I., Fukugita, M., Ivezic, Z., Munn, J. A., Nichol, R. C., Padmanabhan, N., Thakar, A. R., Weinberg, D. H., and York, D. (2003). Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.*, 341(1):33–53.

Kennicutt, Robert C., J. (1998). Star Formation in Galaxies Along the Hubble Sequence. *Annual Review of Astronomy & Astrophysics*, 36:189–232.

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., Grogin, N. A., Kocevski, D. D., Koo, D. C., Lai, K., Lotz, J. M., Lucas, R. A., McGrath, E. J., Ogaz, S., Rajan, A., Riess, A. G., Rodney, S. A., Strolger, L., Casertano, S., Castellano, M., Dahlen, T., Dickinson, M., Dolch, T., Fontana, A., Giavalisco, M., Grazian, A., Guo, Y., Hathi, N. P., Huang, K.-H., van der Wel, A., Yan, H.-J., Acquaviva, V., Alexander, D. M., Almaini, O., Ashby, M. L. N., Barden, M., Bell, E. F., Bournaud, F., Brown, T. M., Caputi, K. I., Cassata, P., Challis, P. J., Chary, R.-R., Cheung, E., Cirasuolo, M., Conselice, C. J., Roshan Cooray, A., Croton, D. J., Daddi, E., Davé, R., de Mello, D. F., de Ravel, L., Dekel, A., Donley, J. L., Dunlop, J. S., Dutton, A. A., Elbaz, D., Fazio, G. G., Filippenko, A. V., Finkelstein, S. L., Frazer, C., Gardner, J. P., Garnavich, P. M., Gawiser, E., Gruetzbauch, R., Hartley, W. G., Häussler, B., Herrington, J., Hopkins, P. F., Huang, J.-S., Jha, S. W., Johnson, A., Kartaltepe, J. S., Khostovan, A. A., Kirshner, R. P., Lani, C., Lee, K.-S., Li, W., Madau, P., McCarthy, P. J., McIntosh, D. H., McLure, R. J., McPartland, C., Mobasher, B., Moreira, H., Mortlock, A., Moustakas, L. A., Mozena, M., Nandra, K., Newman, J. A., Nielsen, J. L., Niemi, S., Noeske, K. G., Papovich, C. J., Pentericci, L., Pope, A., Primack, J. R., Ravindranath, S., Reddy, N. A., Renzini, A., Rix, H.-W., Robaina, A. R., Rosario, D. J., Rosati, P., Salimbeni, S., Scarlata, C., Siana, B., Simard, L., Smidt, J., Snyder, D., Somerville, R. S., Spinrad, H., Straughn, A. N., Telford, O., Teplitz, H. I., Trump, J. R., Vargas, C., Villforth, C., Wagner, C. R., Wandro, P., Wechsler, R. H., Weiner, B. J., Wiklind, T., Wild, V., Wilson, G., Wuyts, S., and Yun, M. S. (2011). CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey—The Hubble Space Telescope Observations, Imaging Data Products, and Mosaics. *Astrophys. J. Suppl.*, 197(2):36.

Kornreich, D. A., Haynes, M. P., and Lovelace, R. V. E. (1998). A photometric method for quantifying asymmetries in disk galaxies. *The Astronomical Journal*, 116(5):2154.

Kruk, S. J., Lintott, C. J., Simmons, B. D., Bamford, S. P., Cardamone, C. N., Fortson, L., Hart, R. E., Häußler, B., Masters, K. L., Nichol, R. C., Schawinski,

K., and Smethurst, R. J. (2017). Galaxy Zoo: finding offset discs and bars in SDSS galaxies. *Mon. Not. R. Astron. Soc.*, 469(3):3363–3373.

Lagos, C. d. P., Theuns, T., Stevens, A. R. H., Cortese, L., Padilla, N. D., Davis, T. A., Contreras, S., and Croton, D. (2017). Angular momentum evolution of galaxies in EAGLE. *Mon. Not. R. Astron. Soc.*, 464(4):3850–3870.

Laine, S., Knapen, J. H., Muñoz-Mateos, J.-C., Kim, T., Comerón, S., Martig, M., Holwerda, B. W., Athanassoula, E., Bosma, A., Johansson, P. H., Erroz-Ferrer, S., Gadotti, D. A., de Paz, A. G., Hinz, J., Laine, J., Laurikainen, E., Menéndez-Delmestre, K., Mizusawa, T., Regan, M. W., Salo, H., Sheth, K., Seibert, M., Buta, R. J., Cisternas, M., Elmegreen, B. G., Elmegreen, D. M., Ho, L. C., Madore, B. F., and Zaritsky, D. (2014). Spitzer/Infrared Array Camera near-infrared features in the outer parts of S^4G galaxies. *Mon. Not. R. Astron. Soc.*, 444(4):3015–3039.

Lanzetta, K. M., Yahil, A., and Fernández-Soto, A. (1998). An Empirical Limit on Extremely High Redshift Galaxies. *Astron. J.*, 116(3):1066–1073.

Laporte, C. F. P., Gómez, F. A., Besla, G., Johnston, K. V., and Garavito-Camargo, N. (2018). Response of the Milky Way's disc to the Large Magellanic Cloud in a first infall scenario. *Mon. Not. R. Astron. Soc.*, 473(1):1218–1230.

Łokas, E. L. (2022). Lopsided galactic disks in IllustrisTNG. *Astron. Astrph.*, 662:A53.

Mahor, A., Reddy, J., Singh, A., and Singh, S. (2023). Estimating dynamical parameters of two interacting galaxies using deep learning. *Mon. Not. R. Astron. Soc.*, 521(3):3441–3450.

Masood, A., Al-Jumaily, A., and Anam, K. (2014). Texture analysis based automated decision support system for classification of skin cancer using sa-svm. In Loo, C. K., Yap, K. S., Wong, K. W., Teoh, A., and Huang, K., editors, *Neural Information Processing*, pages 101–109, Cham. Springer International Publishing.

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., Kanaan, A., Overzier, R. A., Molino, A., Sampedro, L., Coelho, P., Barbosa, C. E., Cortesi, A., Costa-Duarte, M. V., Herpich, F. R., Hernandez-Jimenez, J. A., Placco, V. M., Xavier,

H. S., Abramo, L. R., Saito, R. K., Chies-Santos, A. L., Ederoclite, A., Lopes de Oliveira, R., Gonçalves, D. R., Akras, S., Almeida, L. A., Almeida-Fernandes, F., Beers, T. C., Bonatto, C., Bonoli, S., Cypriano, E. S., Vinicius-Lima, E., de Souza, R. S., Fabiano de Souza, G., Ferrari, F., Gonçalves, T. S., Gonzalez, A. H., Gutiérrez-Soto, L. A., Hartmann, E. A., Jaffe, Y., Kerber, L. O., Lima-Dias, C., Lopes, P. A. A., Menendez-Delmestre, K., Nakazono, L. M. I., Novais, P. M., Ortega-Minakata, R. A., Pereira, E. S., Perottoni, H. D., Queiroz, C., Reis, R. R. R., Santos, W. A., Santos-Silva, T., Santucci, R. M., Barbosa, C. L., Siffert, B. B., Sodré, L., Torres-Flores, S., Westera, P., Whitten, D. D., Alcaniz, J. S., Alonso-García, J., Alencar, S., Alvarez-Candal, A., Amram, P., Azanha, L., Barbá, R. H., Bernardinelli, P. H., Borges Fernandes, M., Branco, V., Brito-Silva, D., Buzzo, M. L., Caffer, J., Campillay, A., Cano, Z., Carvano, J. M., Castejon, M., Cid Fernandes, R., Dantas, M. L. L., Daflon, S., Damke, G., de la Reza, R., de Melo de Azevedo, L. J., De Paula, D. F., Diem, K. G., Donnerstein, R., Dors, O. L., Dupke, R., Eikenberry, S., Escudero, C. G., Faifer, F. R., Farías, H., Fernandes, B., Fernandes, C., Fontes, S., Galarza, A., Hirata, N. S. T., Katena, L., Gregorio-Hetem, J., Hernández-Fernández, J. D., Izzo, L., Jaque Arancibia, M., Jatenco-Pereira, V., Jiménez-Teja, Y., Kann, D. A., Krabbe, A. C., Labayru, C., Lazzaro, D., Lima Neto, G. B., Lopes, A. R., Magalhães, R., Makler, M., de Menezes, R., Miralda-Escudé, J., Monteiro-Oliveira, R., Montero-Dorta, A. D., Muñoz-Elgueta, N., Nemmen, R. S., Nilo Castellón, J. L., Oliveira, A. S., Ortíz, D., Pattaro, E., Pereira, C. B., Quint, B., Riguccini, L., Rocha Pinto, H. J., Rodrigues, I., Roig, F., Rossi, S., Saha, K., Santos, R., Schnorr Müller, A., Sesto, L. A., Silva, R., Smith Castelli, A. V., Teixeira, R., Telles, E., Thom de Souza, R. C., Thöne, C., Trevisan, M., de Ugarte Postigo, A., Urrutia-Viscarra, F., Veiga, C. H., Vika, M., Vitorelli, A. Z., Werle, A., Werner, S. V., and Zaritsky, D. (2019). The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters. *Mon. Not. R. Astron. Soc.*, 489(1):241–267.

Mucesh, S., Hartley, W. G., Palmese, A., Lahav, O., Whiteway, L., Bluck, A. F. L., Alarcon, A., Amon, A., Bechtol, K., Bernstein, G. M., Carnero Rosell, A., Carrasco Kind, M., Choi, A., Eckert, K., Everett, S., Gruen, D., Gruendl, R. A., Harrison, I., Huff, E. M., Kuropatkin, N., Sevilla-Noarbe, I., Sheldon, E.,

Yanny, B., Aguena, M., Allam, S., Bacon, D., Bertin, E., Bhargava, S., Brooks, D., Carretero, J., Castander, F. J., Conselice, C., Costanzi, M., Crocce, M., da Costa, L. N., Pereira, M. E. S., De Vicente, J., Desai, S., Diehl, H. T., Drlica-Wagner, A., Evrard, A. E., Ferrero, I., Flaugher, B., Fosalba, P., Frieman, J., García-Bellido, J., Gaztanaga, E., Gerdes, D. W., Gschwend, J., Gutierrez, G., Hinton, S. R., Hollowood, D. L., Honscheid, K., James, D. J., Kuehn, K., Lima, M., Lin, H., Maia, M. A. G., Melchior, P., Menanteau, F., Miquel, R., Morgan, R., Paz-Chinchón, F., Plazas, A. A., Sanchez, E., Scarpine, V., Schubnell, M., Serrano, S., Smith, M., Suchyta, E., Tarle, G., Thomas, D., To, C., Varga, T. N., Wilkinson, R. D., and DES Collaboration (2021). A machine learning approach to galaxy properties: joint redshift-stellar mass probability distributions with Random Forest. *Mon. Not. R. Astron. Soc.*, 502(2):2770–2786.

Nandapala, E. and Jayasena, K. (2020). The practical approach in customers segmentation by using the k-means algorithm. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 344–349.

Nelson, D., Pillepich, A., Genel, S., Vogelsberger, M., Springel, V., Torrey, P., Rodriguez-Gomez, V., Sijacki, D., Snyder, G. F., Griffen, B., Marinacci, F., Blecha, L., Sales, L., Xu, D., and Hernquist, L. (2015). The illustris simulation: Public data release. *Astronomy and Computing*, 13:12–37.

Nelson, D., Pillepich, A., Springel, V., Pakmor, R., Weinberger, R., Genel, S., Torrey, P., Vogelsberger, M., Marinacci, F., and Hernquist, L. (2019a). First results from the TNG50 simulation: galactic outflows driven by supernovae and black hole feedback. *Mon. Not. R. Astron. Soc.*, 490(3):3234–3261.

Nelson, D., Pillepich, A., Springel, V., Weinberger, R., Hernquist, L., Pakmor, R., Genel, S., Torrey, P., Vogelsberger, M., Kauffmann, G., Marinacci, F., and Naiman, J. (2018). First results from the IllustrisTNG simulations: the galaxy colour bimodality. *Mon. Not. R. Astron. Soc.*, 475(1):624–647.

Nelson, D., Springel, V., Pillepich, A., Rodriguez-Gomez, V., Torrey, P., Genel, S., Vogelsberger, M., Pakmor, R., Marinacci, F., Weinberger, R., Kelley, L., Lovell, M., Diemer, B., and Hernquist, L. (2019b). The IllustrisTNG simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1):2.

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., and Zumach, W. A. (1992). Automated Star/Galaxy Discrimination With Neural Networks. *Astron. J.*, 103:318.

O'Shea, K. and Nash, R. (2015). An Introduction to Convolutional Neural Networks. *arXiv e-prints*, page arXiv:1511.08458.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How many trees in a random forest? In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 154–168, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pardy, S. A., D'Onghia, E., Athanassoula, E., Wilcots, E. M., and Sheth, K. (2016). Tidally Induced Offset Disks in Magellanic Spiral Galaxies. *Astrophys. J.*, 827(2):149.

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., and Fouchez, D. (2019). Photometric redshifts from SDSS images using a convolutional neural network. *Astron. Astrph.*, 621:A26.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Peebles, P. J. E. (1998). The Standard Cosmological Model. *arXiv e-prints*, pages astro–ph/9806201.

Pérez-Montaño, L. E., Rodriguez-Gomez, V., Cervantes Sodi, B., Zhu, Q., Pillepich, A., Vogelsberger, M., and Hernquist, L. (2022). The formation of low surface brightness galaxies in the IllustrisTNG simulation. *Mon. Not. R. Astron. Soc.*, 514(4):5840–5852.

Phookun, B., Vogel, S. N., and Mundy, L. G. (1993). NGC 4254: A Spiral Galaxy with an M = 1 Mode and Infalling Gas. *Astrophys. J.*, 418:113.

Pillepich, A., Nelson, D., Springel, V., Pakmor, R., Torrey, P., Weinberger, R., Vogelsberger, M., Marinacci, F., Genel, S., van der Wel, A., and Hernquist, L. (2019). First results from the TNG50 simulation: the evolution of stellar and gaseous discs across cosmic time. *Mon. Not. R. Astron. Soc.*, 490(3):3196–3233.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., and Marinacci, F. (2018). Simulating galaxy formation with the IllustrisTNG model. *Mon. Not. R. Astron. Soc.*, 473(3):4077–4106.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., Bartolo, N., Battaner, E., Battye, R., Benabed, K., Benoît, A., Benoit-Lévy, A., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bonaldi, A., Bonavera, L., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Catalano, A., Challinor, A., Chamballu, A., Chary, R. R., Chiang, H. C., Chluba, J., Christensen, P. R., Church, S., Clements, D. L., Colombi, S., Colombo, L. P. L., Combet, C., Coulais, A., Crill, B. P., Curto, A., Cuttaia, F., Danese, L., Davies, R. D., Davis, R. J., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Désert, F. X., Di Valentino, E., Dickinson, C., Diego, J. M., Dolag, K., Dole, H., Donzelli, S., Doré, O., Douspis, M., Ducout, A., Dunkley, J., Dupac, X., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Farhang, M., Fergusson, J., Finelli, F., Forni, O., Frailis, M., Fraisse, A. A., Franceschi, E., Frejsel, A., Galeotta, S., Galli, S., Ganga, K., Gauthier, C., Gerbino, M., Ghosh, T., Giard, M., Giraud-Héraud, Y., Giusarma, E., Gjerløw, E., González-Nuevo, J., Górski, K. M., Gratton, S., Gregorio, A., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Hansen, F. K., Hanson, D., Harrison, D. L., Helou, G., Henrot-Versillé, S., Hernández-Monteagudo, C., Herranz, D., Hildebrandt, S. R., Hivon, E., Hobson, M., Holmes, W. A., Hornstrup, A., Hovest, W., Huang, Z., Huffenberger, K. M., Hurier, G., Jaffe, A. H., Jaffe, T. R., Jones, W. C., Juvela, M., Keihänen, E., Keskitalo, R., Kisner, T. S., Kneissl, R., Knoche, J., Knox, L., Kunz, M., Kurki-Suonio, H., Lagache, G., Lähteenmäki, A., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Leahy, J. P., Leonardi, R., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Linden-Vørnle, M., López-Caniego, M., Lubin, P. M., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marchini, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Masi, S., Matarrese, S., McGehee, P., Meinhold, P. R., Melchiorri, A., Melin, J. B., Mendes, L., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Moneti, A., Montier, L., Mor-

gante, G., Mortlock, D., Moss, A., Munshi, D., Murphy, J. A., Naselsky, P., Nati, F., Natoli, P., Netterfield, C. B., Nørgaard-Nielsen, H. U., Noviello, F., Novikov, D., Novikov, I., Oxborrow, C. A., Paci, F., Pagano, L., Pajot, F., Paladini, R., Paoletti, D., Partridge, B., Pasian, F., Patanchon, G., Pearson, T. J., Perdereau, O., Perotto, L., Perrotta, F., Pettorino, V., Piacentini, F., Piat, M., Pierpaoli, E., Pietrobon, D., Plaszczynski, S., Pointecouteau, E., Polenta, G., Popa, L., Pratt, G. W., Prézeau, G., Prunet, S., Puget, J. L., Rachen, J. P., Reach, W. T., Rebolo, R., Reinecke, M., Remazeilles, M., Renault, C., Renzi, A., Ristorcelli, I., Rocha, G., Rosset, C., Rossetti, M., Roudier, G., Rouillé d'Orfeuil, B., Rowan-Robinson, M., Rubiño-Martín, J. A., Rusholme, B., Said, N., Salvatelli, V., Salvati, L., Sandri, M., Santos, D., Savelainen, M., Savini, G., Scott, D., Seiffert, M. D., Serra, P., Shellard, E. P. S., Spencer, L. D., Spinelli, M., Stolyarov, V., Stompor, R., Sudiwala, R., Sunyaev, R., Sutton, D., Suur-Uski, A. S., Sygnet, J. F., Tauber, J. A., Terenzi, L., Toffolatti, L., Tomasi, M., Tristram, M., Trombetti, T., Tucci, M., Tuovinen, J., Türler, M., Umana, G., Valenziano, L., Valiviita, J., Van Tent, F., Vielva, P., Villa, F., Wade, L. A., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Wilkinson, A., Yvon, D., Zacchei, A., and Zonca, A. (2016). Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrph.*, 594:A13.

Rajendra Acharya, U., Vinitha Sree, S., Alvin, A. P. C., and Suri, J. S. (2012). Use of principal component analysis for automatic classification of epileptic eeg activities in wavelet framework. *Expert Systems with Applications*, 39(10):9072–9078.

Ratra, B. and Vogeley, M. S. (2008). The beginning and evolution of the universe. *Publications of the Astronomical Society of the Pacific*, 120(865):235–265.

Reichard, T. A., Heckman, T. M., Rudnick, G., Brinchmann, J., and Kauffmann, G. (2008). The Lopsidedness of Present-Day Galaxies: Results from the Sloan Digital Sky Survey. *Astrophys. J.*, 677(1):186–200.

Reichard, T. A., Heckman, T. M., Rudnick, G., Brinchmann, J., Kauffmann, G., and Wild, V. (2009). The Lopsidedness of Present-Day Galaxies: Connections to the Formation of Stars, the Chemical Evolution of Galaxies, and the Growth of Black Holes. *Astrophys. J.*, 691(2):1005–1020.

Richter, O. and Sacisi, R. (1994). Asymmetries in disk galaxies - how often - how strong. *Astronomy & Astrophysics*, 290(1):L9–L12.

Rix, H.-W. and Zaritsky, D. (1995). Nonaxisymmetric Structures in the Stellar Disks of Galaxies. *Astrophys. J.*, 447:82.

Rogstad, D. H. (1971). Aperture synthesis study of neutral hydrogen in the galaxy M101: II. Discussion. *Astron. Astrph.*, 13:108–115.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Rudnick, G., Rix, H.-W., and Kennicutt, Robert C., J. (2000). Lopsided Galaxies, Weak Interactions, and Boosting the Star Formation Rate. *Astrophys. J.*, 538(2):569–580.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Salvato, M., Ilbert, O., Hasinger, G., Rau, A., Civano, F., Zamorani, G., Brusa, M., Elvis, M., Vignali, C., Aussel, H., Comastri, A., Fiore, F., Le Floc'h, E., Mainieri, V., Bardelli, S., Bolzonella, M., Bongiorno, A., Capak, P., Caputi, K., Cappelluti, N., Carollo, C. M., Contini, T., Garilli, B., Iovino, A., Fotopoulou, S., Fruscione, A., Gilli, R., Halliday, C., Kneib, J. P., Kakazu, Y., Kartaltepe, J. S., Koekemoer, A. M., Kovac, K., Ideue, Y., Ikeda, H., Impey, C. D., Le Fevre, O., Lamareille, F., Lanzuisi, G., Le Borgne, J. F., Le Brun, V., Lilly, S., Maier, C., Manohar, S., Masters, D., McCracken, H., Messias, H., Mignoli, M., Mobasher, B., Nagao, T., Pello, R., Puccetti, S., Perez-Montero, E., Renzini, A., Sargent, M., Sanders, D. B., Scodeggio, M., Scoville, N., Shopbell, P., Silvermann, J., Taniguchi, Y., Tasca, L., Tresse, L., Trump, J. R., and Zucca, E. (2011). Dissecting Photometric Redshift for Active Galactic Nucleus Using XMM- and Chandra-COSMOS Samples. *Astrophys. J.*, 742(2):61.

Sánchez Almeida, J. and Allende Prieto, C. (2013). Automated Unsupervised Classification of the Sloan Digital Sky Survey Stellar Spectra using k-means Clustering. *Astrophys. J.*, 763(1):50.

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., Förster, F., Eyheramendy, S., Elorrieta, F., Bauer, F. E., Cabrera-Vives, G., Estévez, P. A., Catelan, M., Pignata, G., Huijse, P., De Cicco, D., Arévalo, P., Carrasco-Davis, R., Abril, J., Kurtev, R., Borissova, J., Arredondo, J., Castillo-Navarrete, E., Rodriguez, D., Ruz-Mieres, D., Moya, A., Sabatini-Gacitúa, L., Sepúlveda-Cobo, C., and Camacho-Iñiguez, E. (2021). Alert Classification for the ALeRCE Broker System: The Light Curve Classifier. *Astron. J.*, 161(3):141.

Sandage, A. (1961). *The Hubble Atlas of Galaxies*.

Schade, D., Lilly, S. J., Crampton, D., Hammer, F., Le Fevre, O., and Tresse, L. (1995). Canada-France Redshift Survey: Hubble Space Telescope Imaging of High-Redshift Field Galaxies. *Astrophys. J. Let.*, 451:L1.

Schmidt, M. and Green, R. F. (1983). Quasar evolution derived from the Palomar bright quasar survey and other complete quasar surveys. *Astrophys. J.*, 269:352–374.

Schweizer, F. (1976). Photometric studies of spiral structure. I. The disks and arms of six Sb I and Sc I galaxies. *Astrophys. J. Suppl.*, 31:313–332.

Sellwood, J. A. (2013). Dynamics of Disks and Warps. In Oswalt, T. D. and Gilmore, G., editors, *Planets, Stars and Stellar Systems. Volume 5: Galactic Structure and Stellar Populations*, volume 5, page 923.

Sheth, K., Hinz, J., Gil de Paz, A., Regan, M., Menendez-Delmestre, K., Schinnerer, E., Elmegreen, B., Elmegreen, D., Athanassoula, L., Buta, R., Bosma, A., Jarrett, T., Ho, L., Armus, L., Madore, B., Zaritsky, D., Munoz-Mateos, J. C., Helou, G., Gaditto, D., Peng, C., Surace, J., Masters, K., Ogle, P., Mobasher, B., Seibert, M., Koda, J., Capak, P., Laurikainen, E., Salo, H., and Knapen, J. (2008). The Spitzer Survey of Stellar Structure in Galaxies (S4G). Spitzer Proposal ID #60007.

Shlens, J. (2014). A Tutorial on Principal Component Analysis. *arXiv e-prints*, page arXiv:1404.1100.

Smith, M. J. and Geach, J. E. (2023). Astronomia ex machina: a history, primer and outlook on neural networks in astronomy. *Royal Society Open Science*, 10(5):221454.

Snyder, G. F., Peña, T., Yung, L. Y. A., Rose, C., Kartaltepe, J., and Ferguson, H. (2022). Mock galaxy surveys for hst and jwst from the illustristng simulations. *Monthly Notices of the Royal Astronomical Society*, 518(4):6318–6324.

Springel, V. (2010). E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh. *Mon. Not. R. Astron. Soc.*, 401(2):791–851.

Stoughton, C., Lupton, R. H., Bernardi, M., Blanton, M. R., Burles, S., Castander, F. J., Connolly, A. J., Eisenstein, D. J., Frieman, J. A., Hennessy, G. S., Hindsley, R. B., Ivezić, Ž., Kent, S., Kunszt, P. Z., Lee, B. C., Meiksin, A., Munn, J. A., Newberg, H. J., Nichol, R. C., Nicinski, T., Pier, J. R., Richards, G. T., Richmond, M. W., Schlegel, D. J., Smith, J. A., Strauss, M. A., SubbaRao, M., Szalay, A. S., Thakar, A. R., Tucker, D. L., Vanden Berk, D. E., Yanny, B., Adelman, J. K., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Bartelmann, M., Bastian, S., Bauer, A., Berman, E., Böhringer, H., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Carey, L., Carr, M. A., Chen, B., Christian, D., Colestock, P. L., Crocker, J. H., Csabai, I., Czarapata, P. C., Dalcanton, J., Davidsen, A. F., Davis, J. E., Dehnen, W., Dodelson, S., Doi, M., Dombeck, T., Donahue, M., Ellman, N., Elms, B. R., Evans, M. L., Eyer, L., Fan, X., Federwitz, G. R., Friedman, S., Fukugita, M., Gal, R., Gillespie, B., Glazebrook, K., Gray, J., Grebel, E. K., Greenawalt, B., Greene, G., Gunn, J. E., de Haas, E., Haiman, Z., Haldeman, M., Hall, P. B., Hamabe, M., Hansen, B., Harris, F. H., Harris, H., Harvanek, M., Hawley, S. L., Hayes, J. J. E., Heckman, T. M., Helmi, A., Henden, A., Hogan, C. J., Hogg, D. W., Holmgren, D. J., Holtzman, J., Huang, C.-H., Hull, C., Ichikawa, S.-I., Ichikawa, T., Johnston, D. E., Kauffmann, G., Kim, R. S. J., Kimball, T., Kinney, E., Klaene, M., Kleinman, S. J., Klypin, A., Knapp, G. R., Korienek, J., Krolik, J., Kron, R. G., Krzesiński, J., Lamb, D. Q., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., McLean, B., Menou, K., Merelli, A., Mo, H. J., Monet, D. G., Nakamura, O., Narayanan, V. K., Nash, T., Neilsen, Jr., E. H., Newman, P. R., Nitta, A., Odenkirchen, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. S., Petravick, D., Pope, A., Pordes, R., Postman, M., Prosapio, A., Quinn, T. R., Rechenmacher, R., Rivetta, C. H., Rix, H.-W., Rockosi, C. M., Rosner, R., Ruthmansdorfer,

K., Sandford, D., Schneider, D. P., Scranton, R., Sekiguchi, M., Sergey, G., Sheth, R., Shimasaku, K., Smee, S., Snedden, S. A., Stebbins, A., Stubbs, C., Szapudi, I., Szkody, P., Szokoly, G. P., Tabachnik, S., Tsvetanov, Z., Uomoto, A., Vogeley, M. S., Voges, W., Waddell, P., Walterbos, R., Wang, S.-i., Watanabe, M., Weinberg, D. H., White, R. L., White, S. D. M., Wilhite, B., Wolfe, D., Yasuda, N., York, D. G., Zehavi, I., and Zheng, W. (2002). Sloan Digital Sky Survey: Early Data Release. *Astron. J.*, 123(1):485–548.

Strauss, M. A., Weinberg, D. H., Lupton, R. H., Narayanan, V. K., Annis, J., Bernardi, M., Blanton, M., Burles, S., Connolly, A. J., Dalcanton, J., Doi, M., Eisenstein, D., Frieman, J. A., Fukugita, M., Gunn, J. E., Ivezić, Ž., Kent, S., Kim, R. S. J., Knapp, G. R., Kron, R. G., Munn, J. A., Newberg, H. J., Nichol, R. C., Okamura, S., Quinn, T. R., Richmond, M. W., Schlegel, D. J., Shimasaku, K., SubbaRao, M., Szalay, A. S., Vanden Berk, D., Vogeley, M. S., Yanny, B., Yasuda, N., York, D. G., and Zehavi, I. (2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. *Astron. J.*, 124(3):1810–1824.

Tully, R. B. and Fisher, J. R. (1977). A new method of determining distances to galaxies. *Astron. Astrph.*, 54:661–673.

van Eymeren, J., Jütte, E., Jog, C. J., Stein, Y., and Dettmar, R. J. (2011). Lopsidedness in WHISP galaxies. II. Morphological lopsidedness. *Astron. Astrph.*, 530:A30.

Vanzella, E., Cristiani, S., Fontana, A., Nonino, M., Arnouts, S., Giallongo, E., Grazian, A., Fasano, G., Popesso, P., Saracco, P., and Zaggia, S. (2004). Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS. *Astron. Astrph.*, 423:761–776.

Varela-Lavin, S., Gómez, F. A., Tissera, P. B., Besla, G., Garavito-Camargo, N., Marinacci, F., and Laporte, C. F. P. (2023). Lopsided galaxies in a cosmological context: a new galaxy-halo connection. *Mon. Not. R. Astron. Soc.*, 523(4):5853–5868.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Bird, S., Nelson, D., and Hernquist, L. (2014). Properties of galaxies reproduced by a hydrodynamic simulation. *Nature*, 509(7499):177–182.

Walker, I. R., Mihos, J. C., and Hernquist, L. (1996). Quantifying the Fragility of Galactic Disks in Minor Mergers. *Astrophys. J.*, 460:121.

Watts, A. B., Power, C., Catinella, B., Cortese, L., and Stevens, A. R. H. (2020). Global H I asymmetries in IllustrisTNG: a diversity of physical processes disturb the cold gas in galaxies. *Mon. Not. R. Astron. Soc.*, 499(4):5205–5219.

Weinberg, M. D. (1998). Dynamics of an interacting luminous disc, dark halo and satellite companion. *Mon. Not. R. Astron. Soc.*, 299(2):499–514.

White, S. D. M. and Rees, M. J. (1978). Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183(3):341–358.

Wilcots, E. M. and Prescott, M. K. M. (2004). H i observations of barred magellanic spirals. ii. the frequency and impact of companions. *The Astronomical Journal*, 127(4):1900.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., Melvin, T., Nichol, R. C., Raddick, M. J., Schawinski, K., Simpson, R. J., Skibba, R. A., Smith, A. M., and Thomas, D. (2013). Galaxy zoo 2: detailed morphological classifications for 304,122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860.

York, D. G., Adelman, J., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., Castander, F. J., Chen, B., Colestock, P. L., Connolly, A. J., Crocker, J. H., Csabai, I., Czarapata, P. C., Davis, J. E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B. R., Evans, M. L., Fan, X., Federwitz, G. R., Fiscelli, L., Friedman, S., Frieman, J. A., Fukugita, M., Gillespie, B., Gunn, J. E., Gurbani, V. K., de Haas, E., Haldeman, M., Harris, F. H., Hayes, J., Heckman, T. M., Hennessy, G. S., Hindsley, R. B., Holm, S., Holmgren, D. J., Huang, C.-h., Hull, C., Husby, D., Ichikawa, S.-I., Ichikawa, T., Ivezić, Ž., Kent, S., Kim, R. S. J., Kinney, E., Klaene, M., Kleinman, A. N., Kleinman, S., Knapp, G. R., Korienek, J., Kron, R. G., Kunszt, P. Z., Lamb,

D. Q., Lee, B., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R. H., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., Meiksin, A., Merelli, A., Monet, D. G., Munn, J. A., Narayanan, V. K., Nash, T., Neilsen, E., Neswold, R., Newberg, H. J., Nichol, R. C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. L., Petravick, D., Pier, J. R., Pope, A., Pordes, R., Prosapio, A., Rechenmacher, R., Quinn, T. R., Richards, G. T., Richmond, M. W., Rivetta, C. H., Rockosi, C. M., Ruthmansdorfer, K., Sandford, D., Schlegel, D. J., Schneider, D. P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W. A., Smee, S., Smith, J. A., Snedden, S., Stone, R., Stoughton, C., Strauss, M. A., Stubbs, C., SubbaRao, M., Szalay, A. S., Szapudi, I., Szokoly, G. P., Thakar, A. R., Tremonti, C., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Wang, S.-i., Watanabe, M., Weinberg, D. H., Yanny, B., Yasuda, N., and SDSS Collaboration (2000). The Sloan Digital Sky Survey: Technical Summary. *Astron. J.*, 120(3):1579–1587.

Zaritsky, D. and Rix, H.-W. (1997). Lopsided Spiral Galaxies and a Limit on the Galaxy Accretion Rate. *Astrophys. J.*, 477(1):118–127.

Zaritsky, D., Salo, H., Laurikainen, E., Elmegreen, D., Athanassoula, E., Bosma, A., Comerón, S., Erroz-Ferrer, S., Elmegreen, B., Gadotti, D. A., Gil de Paz, A., Hinz, J. L., Ho, L. C., Holwerda, B. W., Kim, T., Knapen, J. H., Laine, J., Laine, S., Madore, B. F., Meidt, S., Menendez-Delmestre, K., Mizusawa, T., Muñoz-Mateos, J. C., Regan, M. W., Seibert, M., and Sheth, K. (2013). On the Origin of Lopsidedness in Galaxies as Determined from the Spitzer Survey of Stellar Structure in Galaxies (S^4G). *Astrophys. J.*, 772(2):135.

Zhang, Y., Ma, H., Peng, N., Zhao, Y., and Wu, X.-b. (2013). Estimating Photometric Redshifts of Quasars via the k-nearest Neighbor Approach Based on Large Survey Databases. *Astron. J.*, 146(2):22.

Zhong, X. and Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, 67:126–139.