UNIVERSIDAD DE GRANADA

Instituto de Astrofísca de Andalucía,
Consejo Superior de Investigaciones Científicas
(IAA-CSIC)

Identification and characterization of emission line objects in J-PAS using artificial neural network

Thesis submitted by Ginés Martínez Solaeche
for the degree of Doctor of Philosophy

Supervisors: Dr. Rosa María González Delgado
Dr. Ruben García Benito

Programa de Doctorado
en Física y Ciencias del Espacio Granada, Agosto 2022

Editor: Universidad de Granada. Tesis Doctorales
Autor: Ginés Martínez Solaeche
ISBN: 978-84-1117-571-5
URI: https://hdl.handle.net/10481/77691

Agradecimientos

Más allá del relato fantástico que asocia el existo profesional o personal a las sola conducta de los individuos, la realidad muestra que la vida es más bien un conjunto fortuito de causalidades continuas donde el buen o mal juicio de nuestras decisiones operan en un margen muy estrecho y que es solo mediante la cooperación y la ayuda de los otros que uno puede llegar a buen puerto. La ciencia, en este sentido, es un claro ejemplo. Por tanto, agradecer la suerte de nuestros éxitos no es solamente una obligación moral sino también un reconocimiento a nuestra naturaleza dependiente.

Muchas han sido las personas que a través de su acción directa o indirecta han hecho posible las tesis que se defienden en este manuscrito, la lista es larga y seguramente me deje a alguien por mencionar en el camino. Pido disculpas de antemano. Sin duda esto no sería posible sin haber tenido la inmensa suerte de nacer en el seno de una familia que me la ha dado todo: las condiciones materiales necesarias para poder estudiar y formarme, su apoyo moral, su ejemplo, pero sobre todo su amor infinito. Gracias Papá. Gracias Mamá. Quiero agradecer también a mis profesores y profesoras, desde las que me enseñaron a leer y a escribir hasta los que me enseñaron las leyes de la física cuántica, sin ellos yo no habría llegado hasta aquí. Gracias. En especial, me gustaría recordar a Jesus Carnicer quien me enseño que la ciencia tiene mucho más valor cuando se comparte. Gracias profesor. Gracias también a Jacques que confió en mí y con él que comencé mi andadura en el mundo de la astronomía.

A Granada llegué por casualidad, no desvelo nada si digo que conseguir financiación para hacer una tesis es una tarea ardua, la escasa inversión junto con la alta competitividad dificultan sustancialmente el asunto. Sea como fuere, encontré en Granada una ciudad abierta, acogedora y con una arraigada tradición científica. En el IAA he encontrado grandes profesionales con una gran disposición a ayudar en cualquiera que sea el problema. En primer lugar quiero agradecer a mis directores de tesis por su apoyo continuo y su gran calidad humana. Gracias Rosa por

ser la luz de la experiencia, por enseñarme parte de tu gran intuición científica, por escuchar y apoyar mis propuestas pese a lo muy disparatadas que fueran. Gracias Rubén, por tu paciencia infinita, por tu ayuda meticulosa, por ser foro de debate donde las ideas se fraguan. Gracias a todas las personas de J-PAS, indispensables para que este trabajo saliera adelante. Gracias a Luis y Roberto por meter el dedo en la llaga si los argumentos no era suficientemente sólidos. Gracias a Enrique por ayudarme a dar mis primeros pasos en el IAA y por escucharme. Gracias a toda la gente del grupo de quásares con los que también he aprendido a trabajar en equipo. Gracias a Julio por su disposición a echar una mano con lo que hiciera falta. Gracias a Iris por darme el impuso final que necesitaba para poder escribir esta tesis.

A Granada no llegue solo, viene con una gran amiga, este camino lo hemos hecho juntos. Gracias Arezu por ser bálsamo en la tempestad. Gracias a Janlys que vino después pero que apoyo mis empresas en la trinchera de la entropía. Allí seguiremos cavando mientras nos dejen. Gracias a Ali que nos enseñó el mejor Ghormeh sabzi de Teheran. Gracias a Ricardo, que no entiende el lenguaje de las palabras, pero sabe lo que es el amor. Gracias a *Refuerzo Positivo* que le puso música a la primavera. Gracias a mis amigos de toda la vida, a Miguel, a David, a Fran. Seguiremos arreglando el mundo desde cualquier banco. Gracias a mi tía Mercedes que ya no está, pero estaría muy orgullosa de su sobrino. Gracias a mi hermana, siento no haber descubierto ninguna galaxia a la que ponerle tu nombre. Gracias a las gentes de Granada por hacer de esta ciudad un lugar mágico. Al camarero de las cervezas de tapa, a la panadera del queso de tarta, a los gitanos de música ardiente, al árabe del té verde. Y finalmente y no por ello menos importante, extiendo estos agradecimientos a todos los ciudadanos españoles que con sus impuestos han financiado esta investigación sin las que yo, naturalmente, no podría haberle dedicado cuatro años de mi vida.

Resumen

En los años venideros el sondeo J-PAS cartografiará $\sim$ 8000 grados^2 del hemisferio norte con 56 colores proporcionando así una cantidad de imágenes de objetos astronómicos sin precedente. Antes de la llegada de la cámara JPCam al OAJ, la colaboración J-PAS ha observado 1 grado^2 del campo de AEGIS con el mismo sistema fotométrico que J-PAS. Más de 60 000 objetos fueron detectados y compartidos con la comunidad científica en lo que se conoce como el cartografiado miniJPAS.

El objetivo principal de esta tesis es identificar y caracterizar objetos con lineas de emisión en J-PAS. En particular estudiamos las galaxias que presentan lineas de emisión y las propiedades que se pueden obtener tanto del análisis de estas lineas como de las poblaciones estelares. Además, dedicamos un capítulo a la detección de quásares. A diferencia de otros sondeos que usan filtros estrechos para detectar lineas de emisión, las características únicas de J-PAS nos permiten estudiar estos objetos en un rango continuo de redshift. Por ejemplo, podremos detectar las líneas de emisión de Hα o [N$_{\rm II}$] en galaxias desde 0 hasta $z \sim 0.35$, y $z \sim 1$, respectivamente. Del mismo modo, la línea de emisión Lyα de los cuásares se detectará desde redshift 2.1 hasta redshift 4.

Los métodos tradicionales que miden la anchura equivalente (EW) de una línea de emisión se basan generalmente en el contraste fotométrico. Aunque este método puede dar buenos resultados, tiene grandes limitaciones. En primer lugar, hay líneas de emisión como Hα y [O II] que están muy próximas entre sí en el espectro. Por lo tanto, ambas contribuyen al flujo total observado en el filtro, lo que hace difícil desentrañar la contribución individual de cada una de las líneas de emisión. Esto es particularmente relevante para estimar la relación [N$_{\rm II}$]/Hα y determinar los principales mecanismos de ionización de las galaxias. Además, en al menos la mitad de las galaxias observadas con J-PAS las líneas de emisión caerán en el centro de dos filtros adyacentes. Por lo tanto, la medición del la anchura equivalente ya no será posible mediante este método.

En esta tesis hemos desarrollado nuevas técnicas basadas en la inteligencia artificial para superar las limitaciones previamente expuestas. A diferencia de los métodos tradicionales, los algoritmos de aprendizaje automático son capaces de encontrar patrones en los datos sin necesidad de hacer ninguna suposición empírica o teórica. Sin embargo, se necesitan grandes cantidades de estos para poder entrenarlos de manera eficiente. Es por ello que hemos necesitado simular datos de J-PAS a partir de una colección de espectros de otros sondeos como CALIFA, MaNGA y SDSS. En concreto, en el capítulo 3 presentamos un tipo de algoritmos de aprendizaje automático llamado redes neuronales artificiales (RNA). Con RNA mostramos que podemos predecir a partir de los colores de J-PAS la anchura equivalente de las líneas de emisión de Hα, Hβ, [O III] y [N$_{II}$]. Para cada espectro, se disponía previamente de las mediciones de estas líneas en los catálogos. Con este método hemos demostramos que la señal-ruido mínima que necesitamos en la fotometría para medir una línea con una anchura equivalente de 10 Å en Hα, Hβ, [N$_{II}$], y [O III] es de 5, 1.5, 3.5, y 10 respectivamente. En cambio, los métodos basados en el contraste fotométrico necesitan para la misma anchura equivalente una señal ruido en la fotometría de al menos 15.5. Con un conjunto de entrenamiento compuesto por galaxias de CALIFA y MaNGA hemos logrado alcanzar una precisión de 0.092 y 0.078 dex en los ratios de [N$_{II}$]/Hα y [O III]/Hβ. Sin embargo, hemos encontrado más dificultades para determinar los ratios de estas lineas en galaxias que albergan un agujero negro activo.

También hemos usado RNA para distinguir entre las galaxias con líneas de emisión intensas y débiles. De hecho, probamos que el régimen de baja emisión ($\sim$ 3 Å) puede ser explorado en J-PAS. Esto se debe a que los algoritmo inteligentes son capaces de encontrar relaciones mucho más complejas en los datos, por lo que aunque no tengamos suficiente sensibilidad en el seudo-espectro de J-PAS para distinguir las galaxias con líneas de emisión muy débiles, los algoritmos son capaces de encontrar otros patrones en los datos.

Con el objetivo de validar la capacidad de las RNA para predecir sobre datos reales hemos estudiado una muestra de galaxias observadas por miniJPAS con redshift $0 < z < 0.35$. Esto lo desarrolamos en el capítulo 4. Los resultados muestras que podemos identificar galaxias con líneas de emisión y distinguir además aquellas cuyo mecanismo principal de ionización proviene de la formación estelar de aquellas en las que el gas se ioniza por la emisión de un agujero negro activo. Podemos también estimar la tasa de formación estelar a través del flujo de Hα, situar estas galaxias en la secuencia principal de formación estelar o calcular la evolución de la densidad de formación estelar cósmica hasta redshift 0.35. Además, nuestros resultados derivados de las propiedades de las líneas de emisión

concuerdan con los resultados obtenidos mediante el análisis de las poblaciones estelares. Por ejemplo, demostramos que las galaxias que son azules (rojas) en miniJPAS están compuestas por una población estelar más joven (más vieja) y presentan líneas de emisión más fuertes (más débiles).

Por último, en el capítulo 5 hemos abordado el problema de la clasificación de objetos astronómicos con el objetivo de distinguir entre cuásares de bajo redshift, cuásares de alto redshift, galaxias y estrellas. Los resultados indican que la principal fuente de confusión aparece entre los cuásares de bajo redshift y las galaxias. Esto se debe a que la galaxias que albergan cuásares a veces son lo suficientemente brillantes que su luz contribuye al espectro observado. De este modo, estos objetos presentan características en el espectro que provienen tanto de la luz de las estrellas en la galaxia como de la luz del núcleo activo y en consecuencia son más difíciles de clasificar. Hemos prestado especial atención a la fiabilidad de las ”probabilidades” que estiman los algoritmos, algo que a menudo no se investiga en detalle. En particular hemos estudiado el efecto de aumentar el volumen del conjunto de entrenamiento a través de la hibridación. Esta técnica consiste en mezclar los espectros de galaxias, cuásares y estrellas para generar objetos híbridos con probabilidades mixtas. Desafortunadamente, no observamos una mejora global en el rendimiento de los algoritmos. De hecho, observamos que las probabilidades RNA se suelen subestimar cuando se aplica la hibridación. Creemos que esto se debe probablemente a la propia naturaleza de las observaciones astronómicas. A diferencia de otros campos de investigación dónde la hibridación sí se ha aplicado con éxito para clasificar objetos, en astronomía los errores son inseparables de las observaciones, por lo que la ‘hibridación’ aparece de manera natural a medida que disminuye la señal ruido de las fuentes observadas.

Aunque los métodos y técnicas desarrollados en esta tesis tienen algunas limitaciones, este trabajo sienta las bases para estudiar mejor las propiedades de los objetos de líneas de emisión en J-PAS. Tan pronto como J-PAS comience a observar el cielo, nuestros métodos se pondrán a prueba en grandes muestras de galaxias, por lo que será posible refinarlos y mejorarlos.

Abstract

In the years to come the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) will map $\sim$ 8000 deg^2 of the northern sky in 56 colours (J-spectrum), providing an unprecedented amount of images of astronomical objects. Before arrival of JPCam to the Observatorio Astrofísico de Javalambre (OAJ), the J-PAS collaboration observed 1 deg^2 of the AEGIS field with the J-PAS-*Pathfinder* camera, using the same photometric system of J-PAS. More than 60 000 objects were detected in what is known as the miniJPAS survey.

The main goal of this thesis is to identify and characterize emission line objects with J-PAS. In particular, we study emission line galaxies (ELG) and the properties that can be derived from the analysis of both the emission lines and the stellar populations. Furthermore, we dedicate one chapter to the detection of quasars. Unlike others photometric surveys that use few narrow band filters, the unique characteristics of J-PAS allows us to study these objects in a continuous range of redsfhit. For instance, we will be able to detect the emission lines of Hα or [O II] in galaxies from 0 to $z \sim 0.35$, and $z \sim 1$, respectively. Similarly, the Lyα emission line of quasars will be detected from redshift 2.1 up to redshift 4.

Traditional methods that measure the equivalent width (EW) of an emission line are generally based on the photometry contrast. Although, this methods gives a very good first approximation, it is limited in many ways. Firstly, there are emission lines such as Hα and [N II] which are very close to each other in the spectrum. Therefore, they both contribute to the total observed flux in the filter, making difficult to disentangle the individual contribution of each emission line. This is particularly relevant in order to estimate the [N II]/Hα ratio and determine the main ionization mechanisms of galaxies. What is more, in at least half of the observed galaxies by J-PAS the emission lines will fall in the middle of two adjacent filters. Consequently, measuring the EW is no longer feasible with the photometry contrast approach.

In this thesis we developed new techniques based on machine learning (ML) in

order to overcome these limitations. Unlike traditional methods, ML algorithms are able to find patterns in the data without making any empirical or theoretical assumptions. Nevertheless, large data sets are needed to train them efficiently. For this purpose, we generated mock J-PAS data, which are based on a collection of spectra from CALIFA, MaNGA, and SDSS. In chapter 3 we trained artificial neural networks (ANN) in order to predict from the generated syntethic J-PAS colors the EW of Hα, Hβ, [O III], and [N II] emission lines. Direct measurements of these lines were available in the catalogues for each spectrum. We showed that the minimum S/N that we need in the photometry to measure a line with an EW of 10 Å in Hα, Hβ, [N II], and [O III] is 5, 1.5, 3.5, and, 10 respectively. Instead, methods based on the photometry contrast need for the same EW a S/N in the photometry of at least 15.5. With a training set composed of CALIFA and MaNGA galaxies, we reached a precision of 0.092 and 0.078 dex in the [N II]/Hα and [O III]/Hβ ratios. Nevertheless, we found that there ratios are more difficult to constrain in galaxies hosting an active galactic nuclei (AGN).

We also trained an ANN to distinguish between strong and week emission line galaxies (ELG). We proved that the regime of low emission ($\sim$ 3 Å) can be explored in J-PAS. This is because ML algorithms are able to find much more complex relations between features, so even though we do not have enough sensitivity in the J-spectrum to distinguish galaxies with very low emission lines, the algorithms are able to find other patterns in the data to make this possible.

As a proof of concept we applied our techniques to a sample of galaxies observed by miniJPAS in the redshift range $0 < z < 0.35$. This is done in chapter 4. We showed that we are able to make a selection of emission line galaxies (ELG), distinguish AGNs from star forming galaxies based on the [N II]/Hα and [O III]/Hβ ratios, estimate the star formation rate (SFR) in galaxies throughout the flux of Hα, recover the star formation main sequence of galaxies or constrain the evolution of the cosmic star formation density up to redshift 0.35. Furthermore, our results derived from the properties of the emission lines are in agreement with the products obtained through the analysis of the stellar populations. For instance, we showed that blue (red) galaxies in miniJPAS are composed of a younger (older) stellar population and present stronger (weaker) emission lines.

Finally, in chapter 5 we addressed the problem of source classification in order to distinguish between low redshift quasars, high redshift quasars, galaxies, and stars. We found that the main source of confusion appears between low redshift quasars and galaxies. This is because the host galaxy of low redshift quasars is sometimes bright enough to contribute to the observed spectrum. Thus, these objects present mixing features and consequently they are more difficult to classify.

We paid special attention to the reliability of the 'probabilities' yield by the algorithms, something that is very often neglected in the community. In particular we investigated the effect of data augmentation via hybridisation. This technique consists in mixing the spectra from galaxies, quasars, and stars so as to generate hybrid objects with mixing probabilities. Unfortunately, we do not observe a global improvement in the performance of the algorithms. As a matter of fact, we observed that the ANN becomes under-confidence in their prediction. We believe this is likely due to the intrinsic nature of astronomical observations where errors are attached to observations, thus 'hybridisation' turns out to be a natural outcome as the S/N of the sources decreases.

Although, the methods and techniques developed in this thesis are limited in some aspects, this work lays the foundations on which to study better the properties of emission line objects in J-PAS. As as soon as J-PAS begins to observe the sky, our methods will be tested in large sample of galaxies, thus it will be possible to improve them even further.

Contents

Chapter 1

Introduction

Due to the huge timescale of galaxy evolution, we cannot study the transformation of individual objects during our lifetime. Instead, astronomers attempt to reconstruct the scene by observing galaxies at different evolutionary epochs. Almost one century ago the notion of galaxy was about to emerge. In 1925 Edwin Hubble inferred for the first time the distance to several nebulae, as they were called at the time, including the Andromeda Galaxy and the Triangulum Galaxy. The measured distances were too large compared to any other detected object within our Milky Way, thus these new entities were more likely to be independent systems. Latter on, the very same Hubble published the so-called *Hubble sequence* (Hubble 1926) and showed that galaxy morphology exhibits a wide variety of shapes and forms. The stars that make up galaxies turned up to be arranged following a spiral structure (S or SB), a elliptical shape (E) or even spread irregularly (Irr) across the galaxy. In the end, Immanuel Kant's idea postulated back in the XVII century of nebulae being 'island universes' was not far from real.

During the decade of 1960 and 1970, astronomer begun to develop the first models of stellar populations (Tinsley 1968; Searle et al. 1973; Tinsley & Gunn 1976). The idea was simple but revolutionary, the relation between the stellar masses and its metal content with their luminosity across the spectrum can be used to infer the global properties of galaxies. Certainly, galaxies composed of very massive and young stars with very short lifetimes ($\sim$ 10 Myr) would show prominent emission in the blue part of the spectrum and therefore it would indicate that the star formation is taking place at high rate. In the same vein, galaxies where the star formation have been quenched should be redder because they would be mainly populated by old low masses stars. Of course, the picture is much more complex than that. The spectrum of galaxy might look red simply due to the pres-

ence of interstellar dust around stars. Dusts grains are typically of the size of $\sim 0.3\ \mu m$ and absorb preferentially the bluest wavelength in the spectrum. Furthermore, more metal rich stars emit less radiation in the blue part of the spectrum as compared with their metal poor counterparts.

The stellar evolutionary synthesis models for galaxies have been developed substantially over the subsequent decades (see Conroy 2013, and references therein). Although there is still room for improvement, our ability to recover the properties of galaxies depends not as much in the model but mainly on how well we can observe their spectral energy distribution (SED). With high resolutions optical spectrographs one can know almost everything about a galaxy: the chemical abundance, the electron density, the amount of interstellar dust, the age of the stellar population, the pressure of the interstellar medium, the star formation rate, whether or not there is an actively feeding supermassive black hole in the centre, etc. Nevertheless, obtaining high resolution spectra from galaxies is very time consuming. Consequently, it lowers the sample size under study and restrict the observations to a small patch of the sky. The alternative comes at the cost of the spectral resolution. With photometric filters huge and deep images of the cosmos can be taken at a sets of particular wavelengths in the spectrum. But before going any further into that, what are actually the key questions regarding the evolution of galaxies than any astronomer would like to answer? Well, we know that galaxies exhibit different morphologies and their properties have changed along the cosmic history. What make them to look so different? Is there a relation between morphology and the youth of a galaxy? What are the responsible mechanisms that quench the star formation? Is the environment surrounding galaxies playing a major role?

In this chapter, I present the astrophysical context in which this PhD work is immersed. In section 1.1 I give a brief summary of the most important observational evidence of galaxy proprieties, what they can tell us about galaxy evolution, and I discuss the role of large galaxy survey to explore the cosmos. Then, in section 1.2 I present the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS, Benitez et al. 2014) and I discuss its particularities and potentiality to study galaxy properties. In section 1.3 I review the use of the machine learning in astronom since some of these techniques have been used during this thesis. Finally, in section 1.4 I briefly define the goals of this thesis and I explain the contents of each chapter.

1.1 Galaxies

Galaxies are gravitationally bound systems composed of stars, interstellar dust, gas, and dark matter. According to the most accepted cosmological framework, the so-called Lambda cold dark matter (Λ−CDM) model, the primordial density fluctuation in the early universe are the building block of galaxies. In an expanding universe, such over-dense regions grows with time by the successive fusion of dark matter halos, increasing as well the baryonic matter over-densities. The first generation of stars are thought to be form in these halos. Although they have not been discovered yet, they are predicted to have masses in the range $\sim 10 - 1000$ $M_\odot$ (Hosokawa et al. 2016; Hirano et al. 2018). Their very short lifetime caused by such huge masses make them to explode rapidly in very powerful supernova spreading their metals through the Universe. The next generations of stars are formed from the ashes of the first ones, and they would be more metal rich and less massive. They would not have enough power to destroy their surrounding dark matters halos during the explosions as supernova. Hence, it would make possible for larger gravitational structures to form. Indeed, the variety of morphological galaxy shapes that we observe in the Universe today are linked to different evolutionary stages.

1.1.1 Morphological classification

Galaxies were first classified according to their morphological structures by Hubble. In his *Real of the Nebulae* (Hubble 1936) Hubble proposed the tuning-fork diagram (Fig. 1.1) where manly two big families of galaxies can be distinguished: spiral and elliptical galaxies. Spiral galaxies are composed of a central bulge and a thin disk with spiral arms attached to the bulge. The rotation of the stars in the disk prevent the gravitational collapse of the galaxy. These galaxies can be further divide into subclasses (Sa to Sd) where Sd (Sa) are more disk-dominated (bulge-dominated) with more (less) separated spiral arms. The surface-brightness profile of such galaxies is well described with a Sérsic profile (Sérsic 1963):

$$I(r) = I_0 \exp\left[-b_n\left(\left(\frac{r}{r_e}\right)^{1/n} - 1\right)\right] \tag{1.1}$$

where I_0 is the central intensity, and r is a scale radius. The quantity b_n is a function of the Sérsic index n, and is chosen so that the effective radius (r_e) capture half of the total luminosity of the galaxy. The surface brightness profile of a disk

is thought to have an exponential profile (n=1) although this assumption might be inappropriate at the galactic center (Breda et al. 2020). The bulges on the other hand are better describe with a Vauconleurs profile (n=4, de Vaucouleurs 1948).

Elliptical galaxies do not appear to have a disk and the galaxy is supported against gravitational collapse by the random motion of the stars. These galaxies are well characterized by their ellipticity. The subfamily of elliptical (E0 to E6) is the result of projection effect due to relative angle of the observer. The light profile of elliptical galaxies resemble that of the bulge of spiral galaxies with a similar Sérsic index ($n = 4$). Finally, irregular galaxies do not follow any clear pattern or structure. At the time where the Hubble's diagram was published it was proposed that elliptical galaxies (also called early-type galaxies) would eventually evolve to become spiral ones (late-type galaxies). Nowadays, although we still do not fully understand which are the formation and/or evolution processes that give rise to variety of spiral galaxy that we observe in the Universe, we do know they are not the result of the evolution of elliptical galaxies. The important of the Hubble sequences is the fact that morphology can actually be related to the main physical properties of galaxies such as the colour, the amount of gas and dust or the kinematic properties. Therefore, the physical processes that govern the evolution of galaxies should intimately be related to their morphological type.

1.1.2 Color-mass diagram

Generally, spiral galaxies are blue, gas-rich, low-mass, metal-poor and present high star formation activity. On the other hand, elliptical galaxies exhibit red intrinsic colours, are more massive, metal-rich and contain very old population of stars with quenched star formation (Strateva et al. 2001; Baldry et al. 2004; Schawinski et al. 2014; Díaz-García et al. 2019a). They are part of the so-called blue cloud and the red sequence, respectively. Strong evidences support the existence of these two populations beyond the nearby Universe suggesting that they might already be in place even at $z = 4$ (Muzzin et al. 2013; Ilbert et al. 2013; Tomczak et al. 2014). In Fig. 1.2 we observe such bi-modal distribution for a sample of galaxies observed by SDSS (Schawinski et al. 2014) in the nearby universe. While the number density of red sequence galaxies has increased by a factor of two from redshift one to the present day, the number density of blue galaxies has remain roughly constant (Faber et al. 2007; Ilbert et al. 2010; Pozzetti et al. 2010). This fact has strong implications because it is suggesting that blue galaxies has been migrating from the blue cloud to the red sequence at least in the last 8 Gyr by process in which the star formation (SF) is suppressed (see eg. Peng et al.

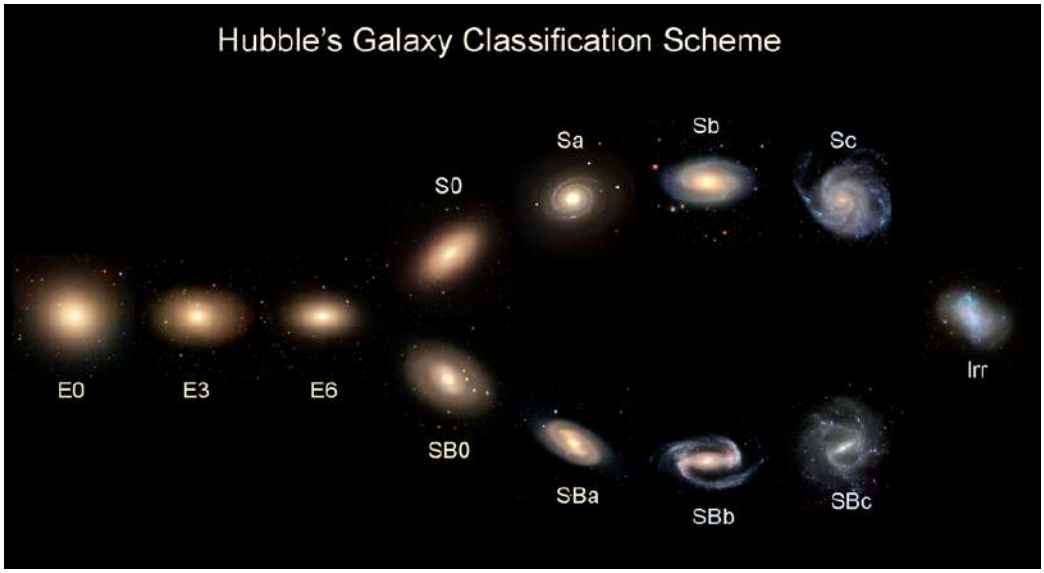


Figure 1.1: Hubble's morphological galaxy classification diagram. Elliptical galaxies are shown on the left (early-type). Spiral galaxies are placed on the right (late-type). An example of an irregular galaxy is also shown on the far right.

2010; Darvish et al. 2016).

Between the red sequence and the blue cloud there is a third population of galaxies with intermediate colors and masses, the so-called Green valley. Galaxies within this group are often interpret as galaxies in transition. Nevertheless, this group is more heterogeneous that it might seem in the first place. The distribution of late-type galaxies in a color-mass diagram is more extend and disperse with respect to early-type galaxies. In other words, we find many more late-type galaxies with red and green intrinsic colors than early-type galaxies with blue and green colors. The quenching mechanism that are taken place are operating at different timescales and therefore they should be different. For instance, Schawinski et al. (2014) studied a sample of galaxies from the Sloan Digital Sky Survey (SDSS, York et al. 2000) in the nearby Universe and estimate that late-type galaxies takes several Gyr to quench the star formation while early-type galaxies should exhaust their gas reservoir almost instantaneously ($\sim$ 100 Myr). Similarly, Noirot et al. (2022) investigated a sample of galaxies at intermediate redshift ($1 < z < 1.8$) and found different evolutionary tracks from the blue cloud to the red sequence: one population in a fast mode with a star formation rate (SFR) e-folding time (τ) lower that 0.5 Gyr, and a slow mode with $\tau > 1.5$ Gyr. Which are the mechanisms

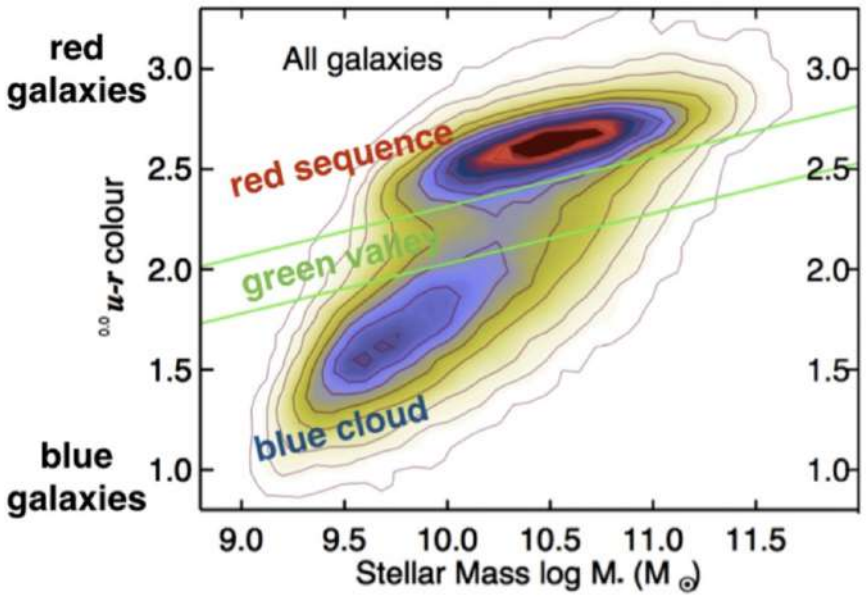


Figure 1.2: Colour-mass diagram for a sample of galaxies observed by SDSS. In the top left, all galaxies are shown, whereas on the right, only early-type (top) and late-type galaxies (bottom) are shown; green lines show the green valley defined by the all-galaxy diagram. The figure is taken from Schawinski et al. (2014).

that are turning off the stellar formation in galaxies? Beside, why some galaxies underwent such a sharp transitions and other evolve slowly across cosmic time? We will discuss some of the proposed quenching mechanism in section 1.1.6 but firstly, how can we actually measure the star formation in galaxies?

1.1.3 Tracing the star formation rate

Hydrogen is the most abundant elements in the Universe. It is found in the interstellar medium (ISM) in forms of giant molecular clouds (H_2) with masses $\sim 10^{5-6}$ $M_\odot$, diameters $\sim$ 50 pc, average densities of $< n_{H_2} > \sim 10^2$ cm^{-3}, and very low temperature (T = 10 K Williams et al. 2000). They are in hydrostatic equilibrium so the gas pressure is in balance with the gravitational pull. However, disruptive events such as the explosion of supernovae or the collapse of two molecular clouds can provoke instabilities which trigger the gravitational collapse of certain regions within the cloud and therefore the birth of new stars takes places. In theory, the initial mass function (IMF), i.e. the relative number of stars at different masses that are created in a SF event, depends on the initial conditions of the molecular clouds. For high temperature or low metallicities, the gas pressure increases, thus more massive gas clouds are required to defeat the radiation pressure which leads eventually to the creation of more massive stars. This is particularly important in the formation of the first stars because the temperature was higher and metals were not presence yet. However, studies of local young and old clusters and associations suggest that the vast majority of the SF events were drawn from a 'universal' IMF even though the exact shape is yet to be known (Bastian et al. 2010). The estimation of the SFR in galaxies lies on the previous statement and strong variations of the IMF from galaxy to galaxy or within different regions in a galaxies might affect significantly our predictions.

The most massive stars (10-100 $M_\odot$) in a galaxy are also the ones with the shortest lifetimes ($\sim$ 10 Myr) and the highest luminosity ($\sim 10^{6-7}$ $L_\odot$). Thus, in order to measure the 'present' SFR of a galaxy one only needs to count how many of those stars are shining, the IMF will tell us the rest. The ultraviolet (UV) emission in a galaxy spectrum is fully dominated by the presence of young stars. Hence, the SFR can be calibrated by measuring the direct or indirect effect of such radiation (Kennicutt 1998). In the UV, the wavelength window between 1250 and 2500 Å is optimal because is sufficiently far away form Lyα forest but close enough to minimize the contamination from older stellar population, so it is sensitive to the light of stars younger than 100 Myr. Fortunately for us, the earth atmosphere absorb the UV radiation. Consequently, direct observations of the UV

emission are only available from the space such as those performed by the Galaxy Evolution Explorer (*GALEX*, Martin et al. 2005). Furthermore, the interstellar dust grains surrounding stars absorb preferentially the shortest wavelength in the spectrum and re-radiate this energy to the infrared (IR, $\lambda \sim 10 - 300\,\mu$m). Thus, the UV SFR calibrator need to be corrected form the dust extinction which can reach up to $\sim$ 2.5 magnitudes (Burgarella et al. 2013; Cucciati et al. 2012). To alleviate this problem some authors have proposed hybrid or composite calibrators (Hao et al. 2011; Catalán-Torrecilla et al. 2015; Boquien et al. 2016). The SFR is then a function of the observed luminosities in the UV and the IR. For instance, Catalán-Torrecilla et al. (2015) found that SFR can be derived as:

$$\mathrm{SFR}(M_{\odot}yr^{-1}) = 4.6 \times 10^{-44}[L(FUV_{obs}) + 4.08 \times L(22\mu m)] \tag{1.2}$$

where $L(FUV_{obs})$, and $L(22\mu m)$ are the observed luminosity measured by *GALEX*, and the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010) at 1516 Å, and 22 μm, respectively.

Another approach is to look at nebular emission lines. Photons with energies above 13.6 eV are able to ionize the neutral hydrogen embedded within molecular clouds. Then, the hydrogen atoms recombine to excited levels, and the new excited atoms decay to lower and lower levels by radiative transitions, eventually reaching the ground level. During this process, line photons are emitted and originate the Balmer lines observed in all gaseous nebulae. The transitions between the $n = 3$ and $n = 2$ quantum states create the Hα emission line, i.e. photons with $\lambda = 6562.8$ Å. Consequently, the total flux of Hα line is directly related to the presence of massive young stars (< 20 Myr) and therefore with the ongoing SFR (Kennicutt 1998):

$$\mathrm{SFR}(M_{\odot}yr^{-1}) = 5.5 \times 10^{-42} L(H\alpha_{corr}) \tag{1.3}$$

where a Kroupa (2001) IMF, and solar metallicity are assumed. Although the flux of Hα is less affected by dust extinction ($A(H\alpha) \sim 0.2 - 1.5$ mag, Garn & Best 2010; Sobral et al. 2016; Duarte Puertas et al. 2017), it is necessary to take it into account in order to obtain reliable SFR. If the Hβ emission line is also measured, the dust extinction can be estimated from the Balmer decrement (Domínguez et al. 2013). Those are some of the most important direct methods to derive the present SFR in star-forming galaxies. They are only valid as as long as the star formation remain constant during the time where calibrators are sensitive.

Finally, indirect methods include fitting the SED assuming either a parametric or a non-parametric star formation history (SFH) (López Fernández et al. 2018;

Asari et al. 2007). In principle, this is the most robust approach since it combines the information across the electromagnetic spectrum to retrieve the galaxy properties. However, the derived SFH depends on the reliability of models and it can be degenerated with other model outputs. If possible, one should employ different SFR calibrators and check they are consistent between each other.

1.1.4 Optical emission line diagnostic

The conversion from the Hα to SFR is only valid as long as the main ionization mechanism is dominated by the radiation of young stars. Nevertheless, the presence of active galaxy nucleus (AGN) or shocks waves as a results of massive stellar winds, gas collisions due to mergers, jets, etc. are also able to ionise the interstellar gas. Therefore, determining the main ionization mechanism of galaxies is essential not only to recover SFRs but also to understand the processes that are taking place within galaxies and the role they play in their evolution. In addition to the Balmer series, collisionally excited emission lines (CEL) are very frequent in the spectra of the ionized gas under astrophysical conditions. Collision between thermal electrons and ions such as O^+, O^{++}, and N^+ are able to excite their low-lying energy levels even though they are much less abundant than H or He. This is because their excitation potential are of the order of the kinetic energy of the electrons in the nebula. At low densities ($N_e < 10^4$ cm^{-3}) collisional de-excitation is much less probable than the transition probabilities to the ground state via spontaneous radioactive decay, thus every excitation leads to emission of a photon and consequently CEL such as [O II] $\lambda\lambda$3727, 3729, [O III] $\lambda\lambda$4959, 5007, [N II] $\lambda\lambda$6548, 6584 or the [S II] $\lambda\lambda$6717, 6731 doublets among others are present in the spectrum of a nebulae.

Line optical ratios depend on physical quantities such as the metallicity of the ionized gas, the ISM pressure, the hardness of the ionizing radiation field or the ionization parameter. Baldwin et al. (1981) proposed to use the ratios of [O III] λ5007/Hβ, [N II] λ6584/Hα, and [O II] λ6300/Hα in order to distinguish among normal H II regions, planetary nebulae or objects photoionized by a harder radiation field. The BPT diagrams, as they are known today, define a star-forming sequence where galaxies evolve from low mass and low metallicity to high mass and high metallicity (Left wing on Fig. 1.3). Kewley et al. (2001, Ke01) used a combination of stellar population synthesis, photoionization, and shock models to derive an upper limit for a maximum star-formation in this diagram. Galaxies above this line cannot be ionized by only star formation events and must contain ionization from AGN or shock excitation. Latter on, Kauffmann et al. (2003a, Ka03) studied

a complete sample of galaxies from the SDSS survery and shifted the Ke01 line to provide a more accurate demarcation. Objects between these two curves might have contribution from several ionization mechanisms, they are often called composite objects. The AGN branch in the BPT has also been dived empirically by Schawinski et al. (2007) to define a region populated by Low Ionization Nuclear Emission Regions (LINER, Heckman 1987) and pure AGN. LINER emission was associated with Low Luminosity AGN with a harder ionizing radiation field and a lower ionization parameter (Kewley et al. 2006). Nevertheless, shocks by jets or other outflows may be needed to power the emission of LINERs (Molina et al. 2018). Furthermore, galaxies with very weak emission lines that are in the process of quenching might cross the Ka03 line and finally end up in the LINER region. The ionizaiton mechanism of these galaxies might be caused by post-AGB stars (Cid Fernandes et al. 2011; Hsieh et al. 2017). On the top of that, the position of AGN in the BPT diagram might change at high redshift where metals were even less abundant (Kewley et al. 2013). In essence, the BPT diagram in its most popular representation is a very useful tool to select a sample of star-forming galaxies in the local Universe but it might not be the ideal diagnostic to differentiate between other ionization mechanisms. This is reason why combinations of other line ratios that includes the UV have been proposed (Kewley et al. 2019).

1.1.5 The main sequence of star formation

The SFR for star-forming galaxies correlates almost linearly with their stellar mass (SFR $\sim$ M$^{\alpha}$). This relation, which is referred to as the star formation main sequence (SFMS), has been proven to be true regardless the star formation tracer used (see e.g. Oliver et al. 2010; Boogaard et al. 2018; Shin et al. 2021). Although the exact value of the slope of the SFMS (α) is already under debate, most of the works agree that a sub-linear slope ($0.6 < \alpha < 1$) is more likely to govern the relation between the SFR and the stellar mass (see the compilation of works in Speagle et al. 2014). This implies that the SFR efficiency or the specific star formation rate (sSFR = SFR/M$_*$) decreases as galaxies grow in mass. Furthermore, some authors have found that the relation between the SFR and the stellar mass turns over at mass of 10^{10} M$_\odot$ from which the slope becomes flatter ($\alpha \sim 0.3$, Whitaker et al. 2014; Lee et al. 2015; Schreiber et al. 2015; Tomczak et al. 2016). Nevertheless, such flattening might be the result of different effects, the star-forming selection criteria, the accuracy to determine the SFR, the selection bias of the survey, etc. Determining the exact shape of the SFMS is crucial to understand the mass assembly history of galaxies. In fact, the relatively small scatter found in

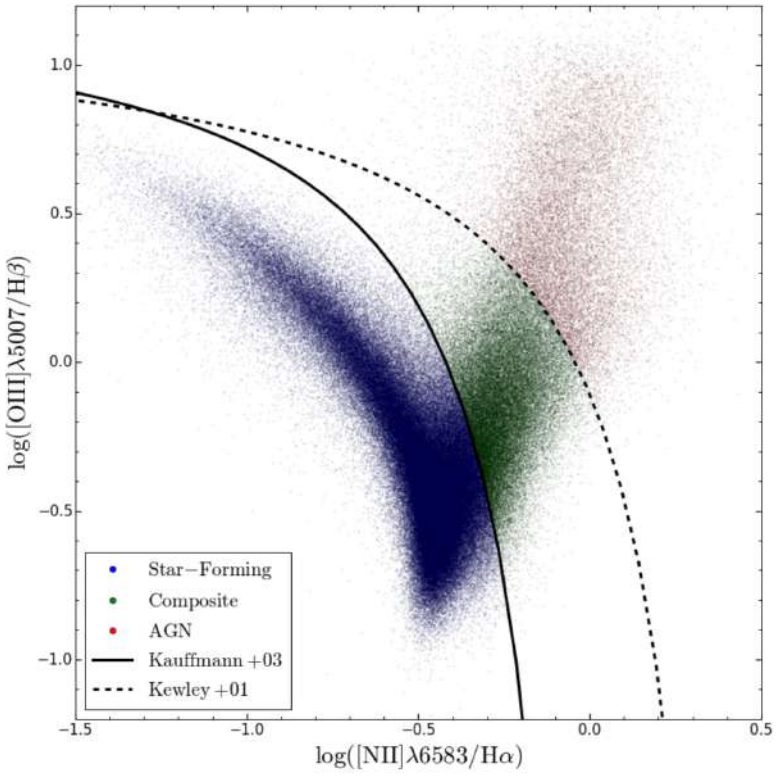


Figure 1.3: BPT diagram for a sample of SDSS galaxies. Figure taken from Duarte Puertas et al. (2017).

star-forming galaxies around the SFMS ($\sigma \sim 0.15 - 0.5$ dex, Whitaker et al. 2012; Salmi et al. 2012; Speagle et al. 2014; Schreiber et al. 2015; Ilbert et al. 2015) has been interpreted as a sign that galaxies undergo preferentially a secular evolution rather than violent episodes of star-formation. We would expect strong variations in the SFR for star-forming galaxies if mergers were the main driver of galaxy evolution.

Observations of galaxies at high redshifts prove that the SFMS relation holds true at early epochs with a global increase of the SFR (Oliver et al. 2010; Rodighiero et al. 2011; Karim et al. 2011; Whitaker et al. 2012; Ilbert et al. 2015; Schreiber et al. 2015). Therefore, the SFR can be parameterized as function of the mass and the redshift (e.g. SFR $\sim M^{\alpha}(1+z)^{\beta}$) which allow us to set constrains on the average SFHs of galaxies. For instance Ciesla et al. (2017) used the best-fit found by Schreiber et al. (2015) to recover the evolutionary path in the SFMS for different mass seeds at redshift 5 and estimated the most realistic parametrization of the SFH. In terms of the cosmic star formation rate density (ρ_{SFR}), the Universe reached a peak at approximately 3.5 Gyr after the Big Bang ($z \sim 2$) and it has been decreasing exponentially since then, with an e-folding timescale of 3.9 Gyr (Madau & Dickinson 2014). Unfortunately, only the brightest and more massive galaxies can be observed at high redshift (see Fig. 1.4) which reduce our ability to recover the SFR-M relation at early time. Alternatively, reconstruction of the SFHs can be done thorough the records of present stars in nearby galaxies, the so-called fossil record method. Studies of well resolved galaxies in the local Universe with surveys such as the Calar Alto Legacy Integral Field Area (CALIFA, Sánchez et al. 2012) or the Mapping Nearby Galaxies at Apache Point Observatory (MaNGA, Bundy 2015) have confirmed the shape of the ρ_{SFR} (López Fernández et al. 2018; Sánchez et al. 2019). The comparison between these two method can be seen in Fig. 1.5.

1.1.6 Galaxy quenching

Although the mechanisms that quench the star formation in galaxies are not yet fully understood, two main scenarios are invoked: *mass quenching* and *environmental quenching*. Both galaxy mass and the environment in which galaxies reside have been proven to play a key role in the cessation of star formation activity (Peng et al. 2010; Kovač et al. 2010; Paccagnella et al. 2016; Gu et al. 2021). Figure 1.6 is quite illustrative in this regard. Peng et al. (2010) who studied the relation between mass, SFR, and environment in a sample of SDSS and zCOSMOS (Lilly et al. 2007) galaxies up to redhsift one, showed that the fraction of

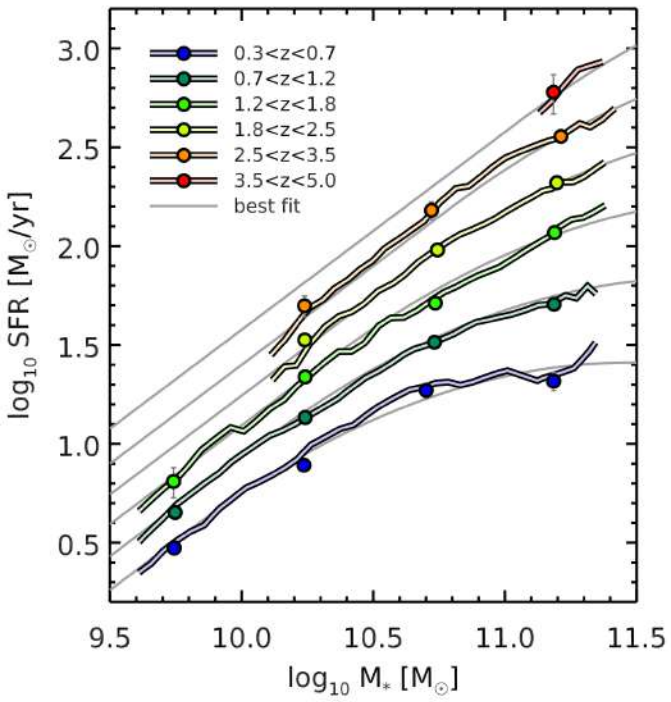


Figure 1.4: Star formation main sequence as a function of redshift derived from deepest *Herschel* images. Light gray curves show the best-fit relation to the main sequence. Figure taken from Schreiber et al. (2015).

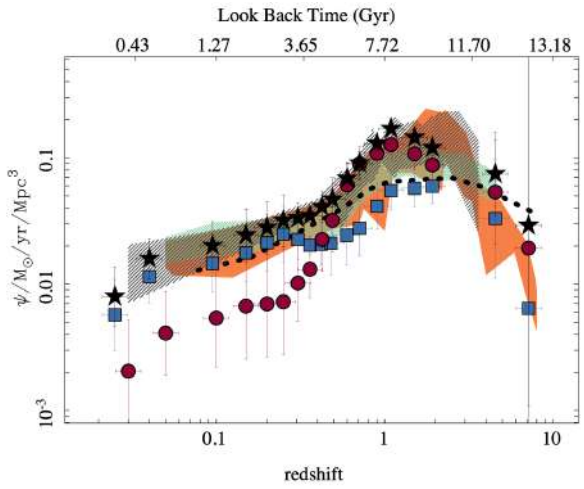


Figure 1.5: Cosmic evolution of the SFR density (ρ_{SFR}) obtained with a sample of MaNGA (black solid stars, Sánchez et al. 2019) and CALIFA (black dotted points, López Fernández et al. 2018) galaxies using the fossil record method. The ρ_{SFR} is broken into star-forming (blue solid squares) and quiescent galaxies (red solid squares) for the MaNGA sample. The shadowed regions correspond to the star-formation rate densities derived from direct observations based on cosmological surveys compiled by Madau & Dickinson (2014). Figure taken from Sánchez et al. (2019).

red galaxies increases either with galaxy mass or the environmental density. In other words, galaxies with masses lower than 10^{10} $M_\odot$ are actively forming stars unless they are found in high density environment. On the other hand, the most massive galaxies ($\geq 10^{11}$ $M_\odot$) are always quenched no matter they are found in the field or within a galaxy cluster.

Any processes that aim to characterize the quenching of star formation should propose an explanation of why the cold gas reservoir is depleted. Locally, at galaxy level, there are evidences that suggest that AGN might be responsible of heating the gas and eventually shutting down the star formation. For instance, Penny et al. (2018) studied a sample of low mass galaxies ($\leq 10^9$ $M_\odot$) with the MaNGA survey and found that among those galaxies where the line ratios at the galactic center are compatible with the presence of an AGN, the gas was not in dynamical equilibrium. Besides, Cicone et al. (2014) studied the massive molecular outflows traced by the carbon monoxide emission lines in a sample of nearby galaxies and found that galaxies hosting an AGN are indeed able to increase the outflow rate and reduce the depletion time scale of gas consumption. From a theoretical point of view, in order to reproduce the observed fractions of quenched

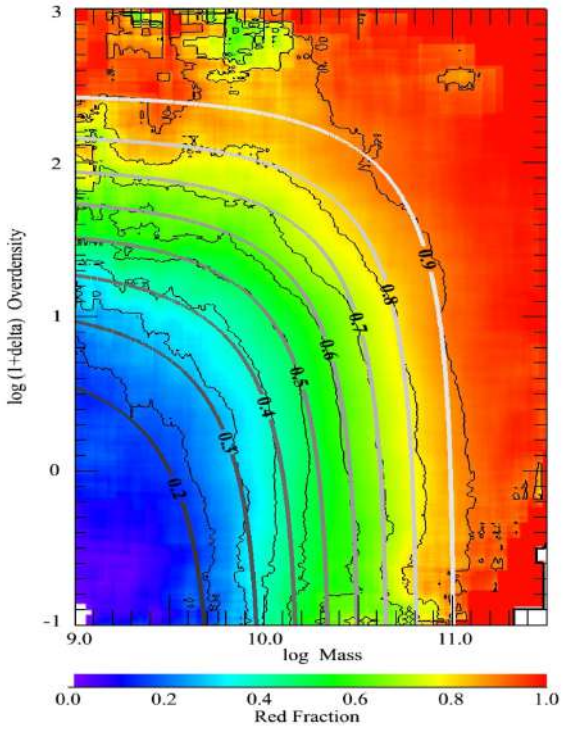


Figure 1.6: Red fraction of a sample of SDSS and zCOSMOS galaxies between $0 < z < 1$ as functions of stellar mass and environment. Figure taken from Peng et al. (2010).

galaxies and color distributions some form of AGN feedback is required (Weinberger et al. 2018; Donnari et al. 2021). Rather than the AGN feedback, other mechanism have been proposed, that includes the supernova feedback and stellar winds (Cantalupo 2010; Chevance et al. 2022). Most likely, *mass quenching* is a process driven by several quenching mechanism whose relative importance might even vary as galaxies grow in mass.

In very dense environment, such as those found in groups or clusters, galaxies moving through the intercluster medium experience a pressure that depends on the intracluster gas density and the speed of the galaxy. Such pressure can eventually strip out part of the gas that was gravitationally bound to the galaxy. This phenomenon is know as ram pressure stripping (Boselli et al. 2022) and it has been observed in many galaxies which exhibit a 'jellyfish' shape with tentacles of gas that appear to be stripped from the main body of the galaxy (Ebeling et al. 2014; Poggianti et al. 2016, 2017). Furthermore, it has also been proposed that the gas supply might be suppressed if galaxies become satellites of a larger

halo, thus the formation of new stars is stopped as the available gas is consumed through a process called 'starvation' or 'strangulation'. For instance, Peng et al. (2015) measure the stellar metallicity in quiescent and star-forming galaxies observed by SDSS. The authors argue that 'starvation' is the dominant quenching mechanism since stellar metallicity in quiescent galaxies increased significantly respect to star-forming ones at a given mass. Such metal enrichment would not be expected if the quenching were driven by processes where the gas is instead removed abruptly from the galaxy. However, the recovered distribution of stellar metallicity for both star-forming and quiescent galaxies reveal a strong variation from the mean values which suggest that actually other processes might also be in place. In addition to ramp pressure stripping and starvation, other processes occurring in dense environment include tidal interactions (Chung et al. 2007) and galaxy harassment (Moore et al. 1996). Finally, galaxy merger, since it is a very disruptive event, is able to stop the formation of new stars even though galaxies might undergo a star-burst phase ($t \sim 10-100$ Myr) in the fist stages of the merger process (see e.g. Cortijo-Ferrero et al. 2017).

In conclusion, galaxy quenching is a very complex process produced by many different phenomena that can act even simultaneously. Certainly, one of the major challenges in astronomy in the upcoming years will be to determine at which conditions the mentioned mechanisms are more relevant to explain the cessation of the star formation in galaxies. In this regard, analyzing large and unbias samples of galaxies will be crucial to push the envelope.

1.1.7 Large galaxy surveys

Large galaxy surveys are conducted by telescopes either from the ground (e.g. GTC, VLT, VISTA) or from the space (e.g. HST, JWST). Certainly, the starting point to design the strategy of any galaxy survey is to have a clear idea of what are the scientific questions that want to be addressed, the measurements that will be performed, and the desired accuracy. Ideally, one would like to map the whole sky with the highest possible precision and reach the fainter galaxies in the Universe. Unfortunately, the finite amount of funding assigned to scientific projects together with our current technological limitations make this dream impossible. Therefore, galaxy surveys follow a wedding-cake approach, where either we observe huge but shallow areas of the sky, large semi-deep fields or small very deep fields. For instance, the Two Micron All-Sky Survey (2MASS, Skrutskie et al. 2006) mapped the whole sky in three IR phtometric bands, the J, H, and Ks bands (1.25, 1.65, and 2.17 μm) detecting 471 million sources. However, the depth of

the survey in J band reaches only $\sim$ 15 mag. On the other hand, the VISTA Deep Extragalactic Observations survey (VIDEO, Jarvis et al. 2013) was able to detect galaxies at redshift 4 with a magnitude depth of 24.4 magnitude in the same band. Nevertheless, only 12 deg^2 of the southern sky were mapped. While 2MASS provided large samples of galaxies in the IR, the VIDEO survey detected objects at very high redshift. The science goals of these surveys are different but they complement each other. Definitely, the depth and the collected area of the survey is one the most important aspect but the wedding-cake comes in a variety of flavours as observations are carried out at different wavelengths. That includes observations in Gamma rays, X-ray, UV, optical, IR, microwaves or radio. Depending on the physical phenomenon we are interested in, some portion of the electromagnetic spectrum might be more desirable. For example, the light emitted by the stellar population within galaxies is usually analyzed in the optical and near-infrared wavelengths while the structure of the neutral hydrogen is observed at 21 cm. Nevertheless, other astrophysical phenomena, for example the AGN activity, might need a multi-wavelength analysis in order to trace the diversity of physical processes that take place around SMBH.

Moreover, astronomical surveys can be divided in two big categories depending on whether spectroscopy or photometry is used. On the one hand, photometric surveys observe the sky with a set of filters obtaining high resolution images at specific wavelengths. For instance, in Fig. 1.7 we show an example of how M101 looks like with photometric bands that go from the UV to the NIR. On the other hand, spectroscopic surveys are able to observe the SED with much more detail as they can measure the energy of photons in a continuous range. Certainly, SDSS is one of most successful spectroscopic galaxy surveys up-to-date. It observed millions of galaxy spectra in the nearby universe ($z \sim 0.1$), but also collected the biggest quasar census with more 700 000 objects. Alternatively, other spectroscopy surveys performed deeper observations such as the DEEP2 Galaxy Redshift Survey (Newman et al. 2013) that obtained spectra for nearly 53,000 galaxies being complete up to redshift 1 in 2.8 deg^2 or the VIMOS Ultra Deep Survey (Le Fèvre et al. 2005) that observed the COSMOS field and was able to take galaxy spectra in the redshift range $2 < z < 4$ with the HST. One of the drawback of these surveys is that they are limited to one spectrum per galaxy, and very often they lose a significant fraction of the light, specially for nearby resolved galaxies. For example, SDSS used a fiber of 3 arcsec leading to aperture effects that might bias our scientific analysis (Duarte Puertas et al. 2017). This problem is alleviated with the use of Integral Field Spectroscopy (IFS), a technique that uses integral field unit (IFU) to take spectra from individual regions within the

galaxy. Past and on-going IFU surveys includes Atlas3D (Cappellari et al. 2011), CALIFA (Sánchez et al. 2012), MaNGA (Bundy 2015), MUSE-WISE (Urrutia et al. 2019) or SAMI (Croom et al. 2021) to name but a few. Without going into greater detail, the IFUs utilized by these surveys differ in their wavelength range, their spectral and spacial resolution and their field of view (FoV) coverage. This fact together with the galaxy sample selection of each survey allow us to study different aspects of well-resolved galaxies in the nearby universe. However, they still have important limitations. For instance, they cannot trace the environment of nearby galaxies because they observe segregate areas of the sky. Furthermore, they suffer from aperture selection effects either because they are not able to capture the outskirt regions of galaxies or either because they exclude galaxies with angular sizes that do not fit within the FoV of the IFUs. In contrast, narrowband photometric surveys such as SHARDS (Pérez-González et al. 2013; Lumbreras-Calle et al. 2019), ALHAMBRA (Moles et al. 2008; Molino et al. 2014) or the Javalambre Photometric Local Universe Survey (J-PLUS, Cenarro et al. 2019) do not experience these effects and they are able to detect fainter objects than their spectroscopic counterparts under the same observational conditions. In this regard, Fig. 1.8 is very informative. The FoV of the T80Cam (Marin-Franch et al. 2015) used in the observations of J-PLUS is compared with several IFUs.

Upcoming surveys will observe the cosmos even with more filters improving their ability to describe the SED of galaxies. This is the case of J-PAS, which has the most powerful photo-metric system in terms of wavelength coverage up-to-date. Let us review now the details of J-PAS.

1.2 Javalambre Physics of the Accelerating Universe Astrophysical Survey

In this section, I describe the particularities of J-PAS data. For that, I begin in section 1.2.1 by introducing the facilities where J-PAS will be carried out. Then, in section 1.2.2 I explain the characteristics of J-PAS, i.e. the JPCam and the J-PAS photo-metric filter system. Finally, in section 1.2.3 I discuss the scientific potential of J-PAS, specially for galaxy evolution studies.

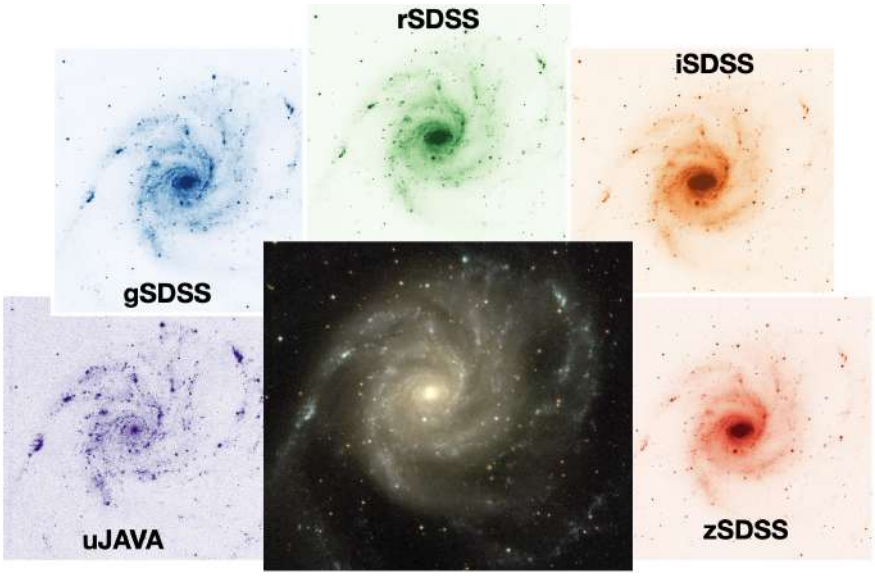


Figure 1.7: Multi-wavelength view of M101. Each image represents the flux measured by J-PLUS with broad band filters, from UV to NIR. A composite image of the galaxy is shown in the center.

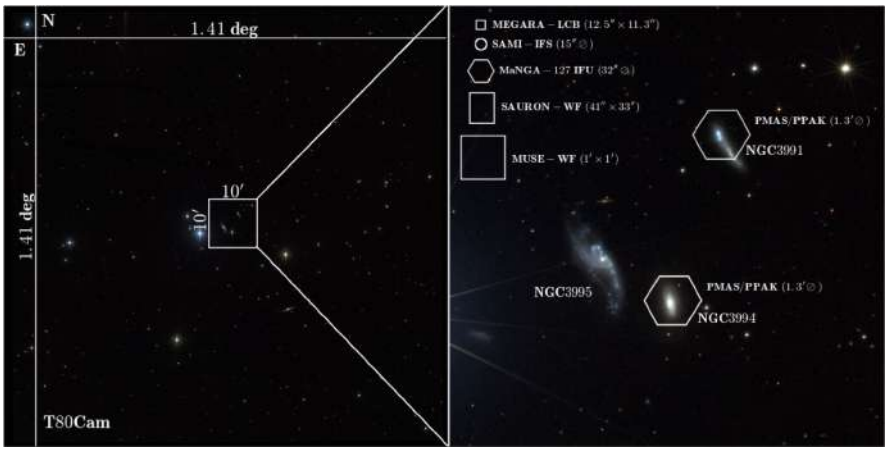


Figure 1.8: Left panel: colour composite image of the SVD pointing '1500041-Arp313', illustrating the 2 deg^2 FoV of T80Cam. Right panel: $10' \times 10'$ zoom covering the Arp313 triplet, where the galaxies NGC3991, NGC3994, and NGC3995 are visible. The FoV of several IFUs is displayed: MEGARA (Gil de Paz et al. 2016), SAMI (Croom et al. 2021), MaNGA (Bundy 2015), SAURON (Bacon et al. 2001), MUSE (Bacon et al. 2010) and PMAS/PPAK (Roth et al. 2005; Sánchez et al. 2012). Figure taken from Logroño-García et al. (2019).

Figure 1.9: View of the Observatorio Astrofísico de Javalambre in the Pico del Buitre, in the Sierra de Javalambre, Teruel (Spain). The image of the JST/T250 telescope is attached on the top left-hand side of the image. Image provided by CEFCA.

1.2.1 The Observatorio Astrofísico de Javalambre

The Observatorio Astrofísico de Javalambre (OAJ) is a Spanish astronomical facility designed to perform large sky photometric surveys. Two professional telescopes with large FoV are the main infrastructures at the OAJ, the 2.5m Javalambre Survey Telescope (JST250, Cenarro et al. 2019) and the 80cm Javalambre Auxiliary Survey Telescope (JAST/T80, Marin-Franch et al. 2015). The OAJ was developed by the Centro de Estudios de Física del Cosmos de Aragón (CEFCA) who is responsible of its management and the exploitation of the data products obtained from the surveys conducted at the telescopes. The observatory is located at the Pico del Buitre (see Fig. 1.9), in the Sierra de Javalambre, Teruel, (Spain), 1957 m above the sea level. The site meets exceptional conditions in terms of the night sky surface brightness (V $\sim$ 22.1 mag arcsec $^{-2}$), the number of clear nights, the median seeing (0.71 arcsec in V band) the transparency or the photometric stability (Moles et al. 2010).

1.2.2 The JPCam and the J-PAS filter system

The Javalambre Panoramic Camera (JPcam, Cenarro et al. 2019; Taylor et al. 2014) is the main scientific instrument of the OAJ installed in the JST/T250 telescope in order to conduct the J-PAS survey. The JPCam is a camera of 1.2 Gpixel

with a plate scale of 0.46 arcsec pix^{-1} and it will cover an area of 4.6 deg^2 with a 14-CCD mosaic. That means that the JPCam can observe in one single pointing the whole extension of Andromeda galaxy as it shown in Fig. 1.10.

In Fig. 1.11 we show J-PAS fitler system, which is composed of 56 bands of which 54 are narrowband (NB) filters in the optical range, and two are medium-band filters, one in the near-UV and another in the near-infrared. NB filters are separated by 100 Å and have a width of $\sim$ 145 Å, which provides a pseudo-spectrum of an equivalent resolving power of $R \sim 60$. However, a J-sepctrum, as we often called it, differs for a low-resolution spectrum in two senses. Firstly, there is overlapping between filters since their widths are greater than their wavelength separation. Therefore, photons with the same energy can contribute to total flux of two continuous filters. Consequently, a narrow emission line might look broader in the J-spectrum than what it would be appear from the observations taken with an actual $R \sim 60$ spectrometer. Secondly, the sky images are not collected with all filters at the same time, which imply having different observational conditions for each tray of filters. Concretely, each tray contains 14 NB filters (see Fig. 1.10). These are important details to keep in mind in order to understand the nature of J-PAS data.

1.2.3 Galaxy evolution studies with J-PAS

The special design of the J-PAS filter system together with its large area coverage ($\sim$ 8000 deg^2) makes J-PAS survey unique and very versatile. Certainly, J-PAS will be competitive across many different astrophysical fields, from the study of very metal-pool stars, globular clusters, galactic stellar streams, withe drafts to the evolution of galaxies or large scale structure through the measurement the of Baryonic acoustic oscillations (Benitez et al. 2014; Bonoli et al. 2021).

When it comes to galaxies, J-PAS will be in a extraordinary position to study them, both in the nearby universe and since $z \sim 1$. As a low spectral resolution IFU, J-PAS will observe thousands of spatially resolved galaxies covering a large range of masses and morphological types. We will be able to analyse them in diverse environments, from galaxies in voids to galaxies in groups and clusters. Definitely, the unprecedented spatial resolution of its NB filters ($\sim$ 0.46 arcsec/pixel[1]) make possible to resolve the stellar populations and derive the age-metallicity gradients at different scales. For instance, galaxies in the Local Volume such as M101

[1]The actual pixel size of JPCam is 0.23 arcsec/pixel, but observations will be sample in pair of pixels for NB filters

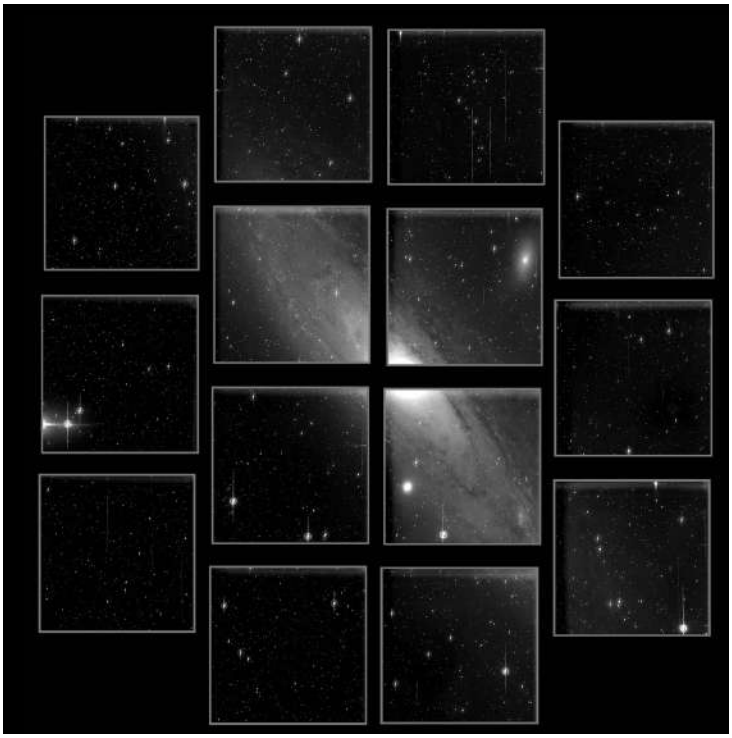


Figure 1.10: Andromeda Galaxy (M31). Technical First Light Image taken by the JPCam in the JST/T250 on June 29, 2020. 14 CCD detectors are arranged in the filter tray. Image provided by CEFCA.

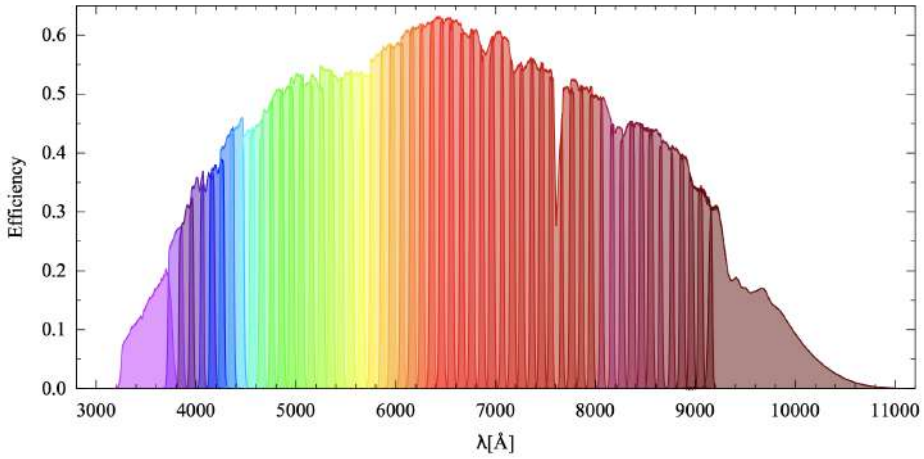


Figure 1.11: Transmission curves of the J-PAS filter system. Figure taken from (Bonoli et al. 2021).

can be resolved at $\sim$ 18 pc, while more distance galaxies ($z < 0.1$) a scale of 2 kpc can be reached. With such resolution different components and subcomponents, the disk, the spiral arms, bars, rings, or bulges can be treated separately, shedding light to the build-up history of galaxy sub-structures. Furthermore, the radial migrations of gas and stars across the disk might be identified systematically by looking at their footprints in the age-metallicity relation. In contrast with modern IFU surveys such as CALIFA or MaNGA, J-PAS will be able to reach the low surface brightness external parts of disks beyond what the FoV of the IFU can capture. What is more, it will provide a much larger sample of galaxies.

With the J-PAS photometric system the main spectral features in the optical range of galaxies, quasars and stars can be captured with enough precision that it allows us to identify and characterize them as function of cosmic time. For instance, nebular emission lines such as Hα or [O II] will be observed in continuous range of redshift, all the way from 0 to $z \sim 0.35$, and $z \sim 1$, respectively. Similarly, the Lyα emission line of quasars will be detected from redshift 2.1 up to redshift 4. This is an important difference compared to other surveys using narrow band filters such as HiZELS (Matthee et al. 2017) or LAGER (Khostovan et al. 2020) where emission lines can only be detected in a set of redshift windows.

Although J-PAS has not started observations yet, the first scientific operations at the JST250 were conducted with JPAS-*Pathfinder* camera, an instruments designed to test the telescope performance and demonstrate to the scientific community the capability of J-PAS to give insights in a variety of astronomical fields. A detail description of this survey, named the miniJPAS survey, together with some of the most relevant results regarding galaxy evolution are described in chapter 2.

1.3 Machine learning in astronomy

Over the last two decades astronomical data have underwent an enormous growth both in size and complexity. From the success of past surveys such as SDSS, MaNGA, CALIFA or GALEX, the new generation of surveys such as the Dark Energy Spectroscopic Instrument (DESI, Levi et al. 2013), the Large Synoptic Survey Telescope (LSST, Ivezić et al. 2019) the Square Kilometer Array (SKA, Dewdney et al. 2009), J-PAS or the James Webb Space Telescope (Gardner et al. 2006) among others, will observe of the order of millions or even billions of objects. Such unprecedented amount of data will offer astronomers the opportunity and the need to apply the most sophisticated machine learning (ML) algorithms in order to fully exploit its content scientifically.

ML is a brunch of artificial intelligence where models or algorithms 'learn' to perform particular set of tasks on sample data often called the trained set. Then, they are tested on an unseen data refereed to as the test set. The power of ML algorithms is precisely the ability to find patterns in the data without making any empirical or theoretical assumptions on how different observable, or *features* as they are called within the ML community, are related. Broadly speaking, ML algorithms can be divided in three main categories: *supervised learning*, *unsupervised learning*, and *reinforcement learning*. In this section I will only discuss supervised and unsupervised ML learning algorithms which have been used more frequently within the astronomical field. I will pay more attention to their use rather than the mathematical framework behind the algorithms. For a more comprehensive and detail review I recommend to read Baron (2019).

1.3.1 Supervised ML learning

Supervised machine learning algorithms are able to find the relation between *features* in an arbitrary high dimensional space to a given set of outputs which have been previously labeled by humans. For instance, let us suppose we would like to distinguish between images of different species of birds. Bird images need to be classified or labeled in the first place by humans so the algorithm can be trained in a supervised manner. In such a case, the inputs or *features* are represented by matrices (images), perhaps with different colors or channels. Then, after seeing many examples, the algorithm has to find the map that connect a set of matrices to a set of integers which represent each one of the bird species. For simple problems, such function might even be analytical so one does not need ML algorithms to perform a classification task. However, most of the time the mapping function is so complex that only via ML we are able to address the problem. Instead of classifying bird images we might also want to determine the sizes of their wings where the outputs are no longer discrete but continuous. In this case we refer to it as a regression task.

There is a zoo of ML algorithms that can actually success in doing these tasks, the literature is extended (see e.g. Géron 2019, for a comprehensive overview) and depending on the problem one should consider to employ different ones. Nonetheless, it is worthy to briefly describe two of the big families of ML algorithms: *artificial neural networks (ANN)* and *decision tree (DT)* based algorithms. ANN are based on a collection of neurons arranged in a set of layers. Each neuron is connected to all the neurons in the continuous layers by a matrix of weights and bias. During the process of training neurons are switched on and off according

to an activation function. This is key ingredient of ANN that allows to find an approximation of the mapping function, which is often highly non-linear (Hornik et al. 1989). In order to find the best combinations of weights and bias, the ANNs minimize a certain loss function that compare the predicted values with the actual ones. Some ANNs include convolutional layers which employ convolutional kernels to extract meaningful features that are afterwards fed into the fully connected layers. They are called convolutional neural networks (CNN) and they have become the gold standard for classifying images in large datasets.

DT algorithms makes predictions by splitting the data set in the feature space in an iterative process until data cannot longer be split. Each decision boundary is encoded in a node of the tree. In order to find the optimal strategy that lead to the best classification, the algorithm evaluate the information gain of splitting the data using each one of the features. *Random forest (RF)* and *Gradient Boosting* are some of the most used DT algorithms by the ML community.

Certainly, one of most common used of supervised ML algorithm in astronomy is the estimation of photometric redshifts. In the pioneering work of Collister & Lahav (2004) the authors trained an ANN to predict the redshift of galaxies with SDSS phtometric bands. They proved that an ANN was competitive respect to traditional template fitting methods. Nowadays, in almost all photometric surveys, ML has been employed to estimate photometric redshift. More sophisticated algorithms are also able to predict the probability density function (PDF, Graham et al. 2018; Ramachandra et al. 2022) which is particular important given the intrinsic degeneracies of the mapping function. Other works employed CNNs trained on photometric images, rather than integrated fluxes or magnitudes, which are specially powerful to extract the spatially resolved information of galaxies (Pasquet et al. 2019; Zhou et al. 2022).

Source classification is another important application of supervised ML algorithm in astronomy, specially in the context of large astronomical surveys where we do not only need accurate identification of astronomical objects but also methods that are computationally faster. Some works explored the classification between point-like and extended sources (see e.g. Burke et al. 2019; Baqui et al. 2021) or between galaxies, stars and quasars (Logan & Fotopoulou 2020; Xiao-Qing & Jin-Meng 2021; He et al. 2021). Others focus on the classification of galaxies according to their morphological types (Domínguez Sánchez et al. 2018; Cheng et al. 2021a), the spectral classification of stars (Sharma et al. 2020) or even combine source detection and classification (González et al. 2018; He et al. 2021).

Besides photometric redshift, other regression-like problems have been ad-

dressed with supervised ML to predict quantities of physical interest. For instance, Bonjean et al. (2019) trained a RF to estimate the SFR and the stellar masses of galaxies from WISE observations at near and mid-infrared wavelengths. The SFR was previously determined by measuring the Hα emission line in SDSS spectra for the same objects. Furthermore, with ML one can investigate the importance of each feature to predict the target variables. The authors demonstrated that luminosity at 3.4 μm is the most important quantity to derived the SFR, while the color between the 4.6 μm and 12 μm bands is also relevant to determine the total stellar mass. Dust particles heated by the radiation field of newborn stars re-emit preferentially at 3.4 μm which explain the correlation between the SFR and the luminosity at this band. The stellar mass also correlates with the 3.4 μm luminosity but the color is needed to differentiate between different stellar populations. We might argue that we do not need ML algorithm to confirm what our current physical knowledge can already tells us. Nevertheless, observations are becoming more and more complex every time, thus the feature space is more difficult to explore and hence correlations between physical quantities might not be straightforward. In Dobbels & Baes (2021) the authors succeed in predicting maps of specific dust luminosity, specific dust mass, and dust temperature by training a RF with a set of surface brightness images of galaxies from UV to mid-infrared wavelengths. They studied the relation between stellar light and dust properties at different pixel scales in order to test the limitation of the energy balance approximation at resolved scales. They concluded that at 400 pc scales the energy balance approximation holds true although they admit spacial correlations of nearby pixels might affect their concussions since the algorithm was trained pixel by pixel. This is an example of how ML can help us to give insight in the physical processes that take place within galaxies.

Another important application of the RF algorithm has been found by Bluck et al. (2022) who proposed a way to distinguish correlation from causality in astronomical data. They trained a RF classifier to differentiate between quiescent and star-forming galaxies in SDSS, MaNGA and the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy survey (CANDELS Grogin et al. 2011). They found that at all redhisft the mass of the bulge is the most predictive parameter of quenching. However, if the central velocity dispersion of stars was also used to train the RF, then it became even more predictive than bulge mass. Both quantities are related but the central velocity dispersion might have a more fundamental connection with cessation of star formation in galaxies.

Supervised ML have also been used to reduce the computational cost of SED fitting by order of magnitudes. The idea is not to retrieve the model parameters

from a set of observation, which would eliminate the information of the posterior probability, but to approximate with ANNs the theoretical SED function given the model parameters. With this approach one can still use Bayesian inference and quantify the uncertainties and degeneracies among parameters. Emulators or speculators, as they are called, have been developed to generate stellar population synthesis models and predict the photometry or the spectra of galaxies (Alsing et al. 2020), the supernova spectra (Kerzendorf et al. 2021) or to estimate the 21 cm signal from the epoch of reionization parameter space (Kern et al. 2017). Hahn & Melchior (2022) went an step further and proposed to employ Amortized Neural Posterior Estimation to analyse the SED of galaxies and estimate the posterior probability distribution over the full range of observations. The authors tested their method in a sample of galaxies from the SDSS survey taking only 1 second per galaxy to obtain the posterior for 12 model parameters.

1.3.2 Unsupervised ML learning

Unlike supervised ML algorithms, unsupervised learning do not require the data to be tagged by an expert. Let us go back to our previous example of bird images and assume there are three different species. For instance, a goldfinch, a nightingale and a swallow. This time, we would not provide to the algorithms which image correspond to each species. With sufficient amount of data the algorithm would be able to identify three different clusters by looking at properties that each image shares with the others. This is one of the application of unsupervised learning called *clustering*. Some of the most popular clustering algorithms are *K-means*, *Hierarchical clustering*, or *Gaussian mixture model* (see chapter 9 for details on how these algorithms work, Géron 2019). For example, Turner et al. (2019) studied a sample of galaxies from the Galaxy And Mass Assembly (GAMA, Driver et al. 2011) survey in order to find sub-populations of galaxies that go beyond the well known bi-modal distribution. The authors trained a *K-means* algorithm using as features the total stellar mass, the sSFR, the Half-light radius (HLR), the u-r colour and the Sérsic index of each galaxy. They found that 2,3,5, and 6 clusters can explain the structure of the data, suggesting the existence of different sub-populations of galaxies. In the same vein, Tammour et al. (2016) studied a sample of quasars from SDSS and trained the *K-means* algorithm with the properties of some of the UV emission lines present in the spectra of quasars. They were able to recover clusters that follow a smooth distribution of physical properties such as the hardness of the SED which have not seen by the algorithm.

Clustering algorithms have the power of unveiling the existence of different

populations of objects or different physical regimes with only a data-driven approach. Nevertheless, there is one critical assumption that is behind that: the distance between objects in the features space is assumed to be caused by their intrinsic physical difference. Therefore, features need to be chosen carefully, in this case by an expert astronomer, in order to attribute a physical meaning to the clusters found by the algorithm. In Turner et al. (2019) some of the features are observational properties (e.g. the HLR a galaxy) and others are the results of a model output such as the mass or sSFR. Only with the previous knowledge that this set of features is indeed meaningful to characterize galaxy populations one can classify them in different groups.

Sometimes, we might not have a clear insight on which features are the most relevant to explain the physical nature of the problem or we do not simply want to restrict ourselves to the product of physical models which have their own biases. This is the reason why *Dimensionality Reduction Algorithms* are often used. The idea is to find a representation of data in a lower dimensional space that encodes the most relevant information. In order to do that, the algorithms are forced to reconstruct the original dimensions of the input data from this compact representation. For instance, Portillo et al. (2020) proved that using variational autoencoders, i.e. a type of ANN that performs a non-linear dimensionality reduction, the spectra of different galaxies such as extreme emission line galaxies or galaxies hosting an AGN appear naturally in different locations in the reduced dimensional space called the latent space. What is more, one can define tracks within the latent space that yields sequences of realistic spectra that interpolate between different types of galaxies even though the algorithm was not trained with physical information. Clustering algorithms have been applied after reducing the dimension of the feature space (see e.g. Jiménez et al. 2020) but some recent works have shown that combining both task in one single algorithm outperform the previous approach (Cheng et al. 2021b; Teimoorinia et al. 2022). This is because the algorithms is force to find a meaningful representation of the data at the same time the clustering performance is optimized.

Anomaly detection is another important application of unsupervised ML algorithms in astronomy. Some works relied on the reconstruction error to detect outliers (Ichinohe & Yamada 2019; Portillo et al. 2020). Since the algorithms only learn to reduce the dimensions of objects that follow the general distribution, extreme cases will not be reconstructed properly. For instance, Ichinohe & Yamada (2019) simulated X-ray spectra and they were able to differentiate between the 'normal' single-temperature electron plasma from the two-temperature plasma and electron plasma out of the collisional ionization equilibrium which

they consider to be anomalous. Another interesting approach was proposed by Baron & Poznanski (2017) who searched for peculiar objects classified as galaxies by SDSS. The authors used as features the original fluxes at each wavelength in the spectrum. Then, they generated synthetic spectra by sampling from the marginalized distribution of each feature so the covariance between features is lost in the synthetic sample. A RF classifier was trained to distinguish between real and synthetic spectra so the algorithm was able to identify the covariance in the data. Finally, they used the decisions taken by all trees within the forest to define a similarity distance. The weirdest objects are those which are far away from the whole sample. This method proved to be successful in identifying peculiar object within the training set such as galaxies with high ionisation lines, galaxies which host supernovae or galaxies with unusual gas kinematics. Certainly, the most challenging part concerning anomaly detection is to distinguish between 'interesting' anomalies which might be an indication of new physics from those caused by observations issues either due to instrumental effects or low signal to noise-ratio. A possible solution to deal with this problem was proposed by Sarmiento et al. (2021) who used *contrastive learning* on MaNGA galaxies. Briefly, the authors applied mathematical transformations to a set of maps containing information of the stellar population and kinematics of galaxies that leave unchanged their physical properties. For example, adding noise to a map or rotate a galaxy certain angle does not modify their physical information. Then, they use a CNN to project the maps to a representation space where they maximized the agreement between the original data object and its transformed pair. With this approach they were able to identify differences between galaxies caused only by real physical dissimilarity.

One of the major limitation of ML algorithms is that they need enormous amount of data to work properly. Very frequently, we have access to sufficiently large observational data but only a small fraction of the data are labelled. For example, as we will see in next chapter, the miniJPAS survey detected plenty of galaxies in the AEGIS field. However, spectroscopic redshift is only available for a fraction of them. In this case, one possible approach to estimate photometric redhisft for miniJPAS with ML is to use *semi-supervised learning*. In a first step the ML algorithm is trained in an unsupervised manner so as to learn the most important features of the data, e.g. the uncertainty in the photometry, the depth of the survey, the colors of galaxies etc. Subsequently, with the fraction of galaxies that have spectroscopic redshift the algorithm is retrained to predict them. This is called *transfer learning* since the knowledge from the first phase of the training is transferred to the second one. A similar approach has been used by Eriksen et al. (2020) in order to predict photometric redshifts for the observations per-

formed by the Physics of the Accelerating Universe Survey (PAU, Serrano et al. 2022). However, the authors employed simulated data to train their networks in the first phase. Transfer learning can also be used in a purely supervised manner. For example, Domínguez Sánchez et al. (2019) transferred the knowledge of a CNN trained to classified galaxies in SDSS to perform the same task in the Dark Energy Survey (DES). The authors proved that the size of the training size to adapt the classification from one survey to another can be reduced by one order of magnitude.

All in all, the use of ML in astronomy is diverse and very promising given the fact that the era of big data is upon us. As a final remark, I would like to stress some of the limitation that are intrinsic to the field. As we show, ANN and RF algorithms are excellent approximators of high dimensional mapping functions. Nevertheless, the ability to make successful predictions on unseen data relies on how representative the training set is respect to the target population. If ML algorithms are trained on simulations, they should mimic observations in the best possible way. If instead observation are used for the training, special care should be paid to the amount of training examples of different physical properties as very few examples might not be enough to capture the entire richness of the physical problem. Another important limitation, specially for ANN, is the lack of interpretability. Sometimes it is difficult to understand how the algorithms came to a decision. Consequently, we might be unable to improve its performance further. Finally, it is important to bear in mind what is the physical problem we would like to address and whether or not the use of ML can actually be helpful. Simple problems require simple solutions, the laws of physics have been developed for centuries while only in last two decades we begun to have access to large data sets and being in a position to apply ML to our science. These two worlds are nothing but independent, so the great challenge lies on our capability to find the right meeting point.

1.4 Scope of the thesis

The aim of this thesis is to understand better the properties of emission line objects in J-PAS. Thanks to the J-PAS photometric system we will be able identify and characterize galaxies and quasars with no selection bias rather than the depth of the survey. Certainly, the main spectral features of galaxies and quasars will be capture with enough precision that it can act, in practice, as a low resolution spectrograph. Therefore, it will be possible to study the properties of emission line

galaxies covering a large range of masses and morphological types. The emission lines of Hα or [O II] will be observed in continuous range of redshift, from 0 to $z \sim 0.35$, and $z \sim 1$, respectively. In the same vein, the Lyα emission line of quasars will be detected from redshift 2.1 up to redshift 4. This is a fundamental differences respect to other photometric surveys where NB filters can only select emission line objects in a set of redsfhit windows.

We begin in chapter 2 with an overview of the miniJPAS survey giving details to the nature and properties of the data. This data has been used as a proof of concept for the techniques developed in this thesis. We discuss how the observations were carried out, the photometric system used, the data calibration process, the methods employed to derive photometric redshift, etc. Furthermore, we make a brief summary of the results related to galaxy evolution studies. In chapter 3 we present an algorithm based on ANN to first detect emission line galaxies (ELGs) and second predict the equivalent width of the main emission lines in the optical spectrum. In chapter 4 we made used of this algorithm to understand the properties of ELGs in miniJPAS. We retrieve the main ionization mechanism of ELG, the star formation main sequence, the cosmic evolution of the star formation rate density and the relation between the properties of the gas and the properties of the stellar populations. Subsequently, in chapter 5 we address the problem of source classification in order to distinguish between low redshift quasars, high redshift quasars, galaxies, and stars in miniJPAS and we investigate the effect of augmenting the data through hybridisation. Finally, we summarized our results and conclusions in chapter 6. We also discuss some of the works that might be done in the future.

Chapter 2

The miniJPAS survey

2.1 Observations

Before the arrival of the JPCam at the OAJ, the J-PAS collaboration carried out the miniJPAS survey and observed 1 deg^2 along the Extended Groth Strip (AEGIS, Davis et al. 2007). For that, the J-PAS-*Pathfinder* camera, a single CCD direct imager (9.2k $\times$ 9.2k, 10 μ m pixel) with a pixel scale of 0.23 arcsec pix $^{-1}$, was installed at the center of the JST/T250 FoV. Observations of the AEGIS field were conducted with the photometric filter system detailed in section 1.2.2 plus four SDSS-like BB filters, u_{uJAVA}, g_{SDSS}, r_{SDSS}, and i_{SDSS}.

The wealth of multi-wavelength observations available at the AEGIS location makes this field optimal for a science verification program. In Fig. 2.1 we show the footprint of miniJPAS field together with the footprint of other projects such as SDSS, ALHAMBRA or the HSC-SSP (Aihara et al. 2018) wide field. With four pointings of 0.27 deg^2, i.e. the FoV of the J-PAS-*Pathfinder* camera, the AEGIS field was covered almost entirely. After taking the mask into account, the effective area in the four tiles amount to 0.895 deg^2.

Unlike the JPCam that can observe the sky with 14 CCD at the same time, the observations with the J-PAS-*Pathfinder* were performed in groups of six filters between May and September 2018. A minimum of 4 exposures, with a dithering of 10 arcsec along the horizontal and vertical direction of the CCD were carried out to generate each tile image. The exposure time for the broad bands g_{SDSS}, r_{SDSS}, and i_{SDSS} was set to 30 s, while for the 56 J-PAS filters and the u_{JAVA} filter was 120 s. The depths of both NB filters and BB filters reached by the miniJPAS survey are shown in Fig. 2.3. Differences in the depth for adjacent filters are due to dif-

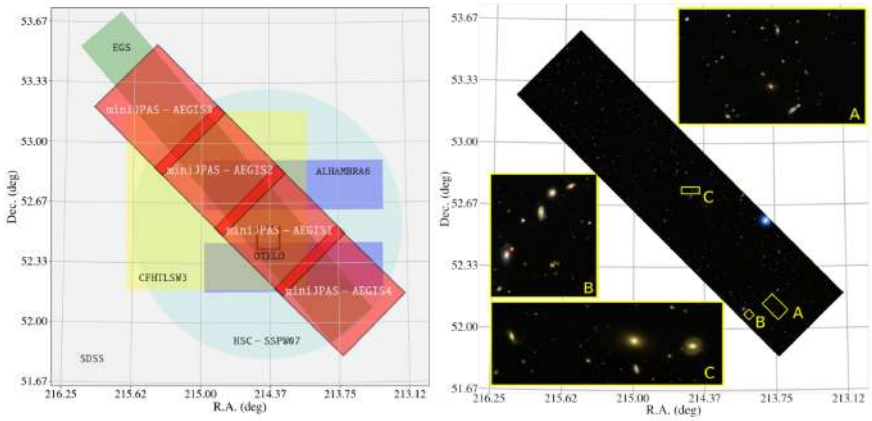


Figure 2.1: Footprint of the miniJPAS field (red squares), the Extended Groth Strip (in green), pointing #6 of the ALHAMBRA Survey (in violet), the W07 wide field of the HSC-SSP (large circle in pale blue), field W3 of the CFHTLS (in yellow), OSIRIS Tunable Emission Line Object survey (OTELO) (small square close to the center of the figure) and SDSS (in light gray occupying the whole area). Right: g_{SDSS}, r_{SDSS}, and i_{SDSS} composite image of miniJPAS with zoom in three selected areas. Figure taken from Bonoli et al. (2021).

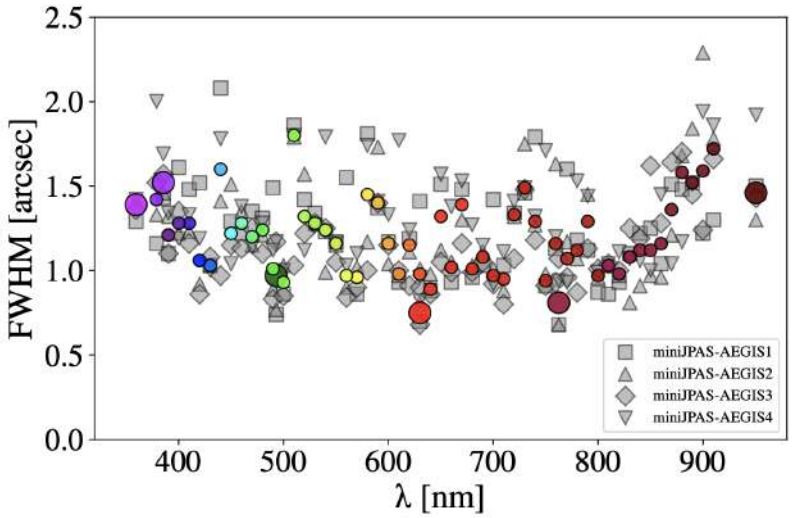


Figure 2.2: Average FWHM of the PSF. The coloured symbols represent the average values for each filter, while the gray ones are the value for each pointing. The larger symbols indicate the FWHM of the the broad bands. Figure taken from Bonoli et al. (2021).

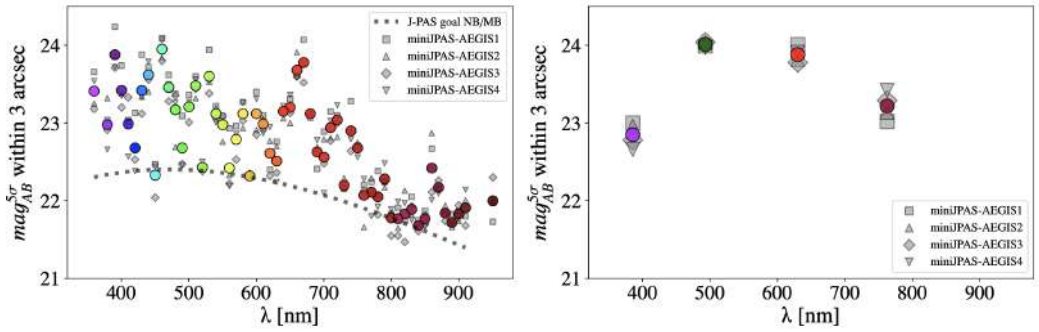


Figure 2.3: Estimated depths (5σ at 3 arcsec aperture), computed from the noise in each tile, for the NB (left) and BB (right). The coloured symbols show the average values for each filter, while the gray ones are the values for the co-added images of each pointing. For the NB, the dashed gray line indicates the approximate targeted minimum depth, as defined in Benitez et al. (2014). Figure taken from Bonoli et al. (2021).

ferent observational conditions and on the final number of combined images. The average full-width-half-maximum (FWHM) of the point spread function (PSF) as a function of the tile and the filter are shown in Fig. 2.2. Due to the fact that the reddest filters were scheduled to be observed at the end of the observing campaign, the AEGIS field reached the lowest elevation and therefore the FWHMs are slightly larger for these filters despite the sky conditions did not change. It is expected that observations with the JPCam will improve this situation as the FHWM will be homogenized among filters with a target value around 1 arcsec.

The data Processing and Archiving Unit (UPAD) is the group at CEFCA in charge to process all data collected by the OAJ. As far as image reduction is concerned, the UPAD team needed to account for several issues such as the subtraction of background patterns, removal of strong vignetting in the outer parts of the CCD or fringing effects in the reddest filters. Using the software Scamp (Bertin 2010) and the Gaia DR2 (Gaia Collaboration et al. 2018) as reference catalogue, astrometric calibration of the images was performed.

The software `SExtractor` was used to detect sources within miniJPAS field and estimate their fluxes and AB magnitudes in a set of apertures. `SExtractor` was run in two complementary modes. In the *dual-mode* catalogue, `SExtractor` was first run in the rSDSS band, the detection band, from which the rest of the photometry is computed. If an object is not detected in particular band the fluxes might be negatives since the sum of the light only accounts for the sky background. In *single-mode*, however, source detection is performed individually in each band. Thus, every detection leads to a reliable measurement of the flux within an aper-

ture but the position of the objects might vary from filter to filter. One of the main advantage of this mode is the possibility to identify emission line objects that are too faint to be detected in the reference band. Total fluxes and AB magnitudes for each object using different methods implemented in `SExtractor` are provided in the photometric catalogues, that includes the `FLUX_AUTO`, `FLUX_ISO`, and `FLUX_PETRO` photometry. Furthermore, the integrated flux within a set of elliptical apertures (`FLUX_APER_...`) that ranges from 0.8 to 6 arcsec are available. Additionally, in order to take into account the variations of the PSF from filter to filter, the users can obtain some of the previous photometry measured after degrading each image to the worst PSF (`FLUX_ISO_WORSTPSF` and `FLUX_APER3_WORSTPSF`) or work with the `FLUX_PSFCOR` photometry that corrects the differences in the PSF among different bands with a Kron aperture (Molino et al. 2017).

The miniJPAS survey contains more than 64 000 sources in the dual-mode catalogue. For point-like sources the miniJPAS catalogue is 99 % complete up to $r_{\rm SDSS} \sim 23.6$ (`MAG_AUTO`), while for extended sources this limit is constrained at $r_{\rm SDSS} \sim 22.7$. The procedure to compute the completeness of the survey has been incorporated as a part of the J-PAS pipeline. It is based on the synthetic injection of sources and the fraction of successful detection under the observational conditions of the field. In Fig. 2.4 we show some examples of Luminous Red galaxies (LRGs), Emission Line galaxies (ELGs), and quasi-stellar objects (quasars) at different redshift and magnitudes that are presented both with in miniJPAS and SDSS. The main features in the spectrum such as the 4000 Å break, the emission lines or the absorption lines are clearly captured thanks to the J-PAS filter system. All data is publicly available and can be accessed from the J-PAS website[1]. Moreover, CEFCA developed a Science Web Portal[2]in order to make the access to the data easy and user friendly. In Fig. 2.5 we show a screenshot of the webpage, and the sky navigator. Each object can be visually inspected by clicking on any marker. Additional information to fully explore the Science Web Portals is given in the user's manual[3].

[1] https://j-pas.org/datareleases/minijpas_public_data_release_pdr201912

[2] http://archive.cefca.es/catalogues

[3] http://archive.cefca.es/catalogues/static/manuals/science_archive_users_manual_v1_18.pdf

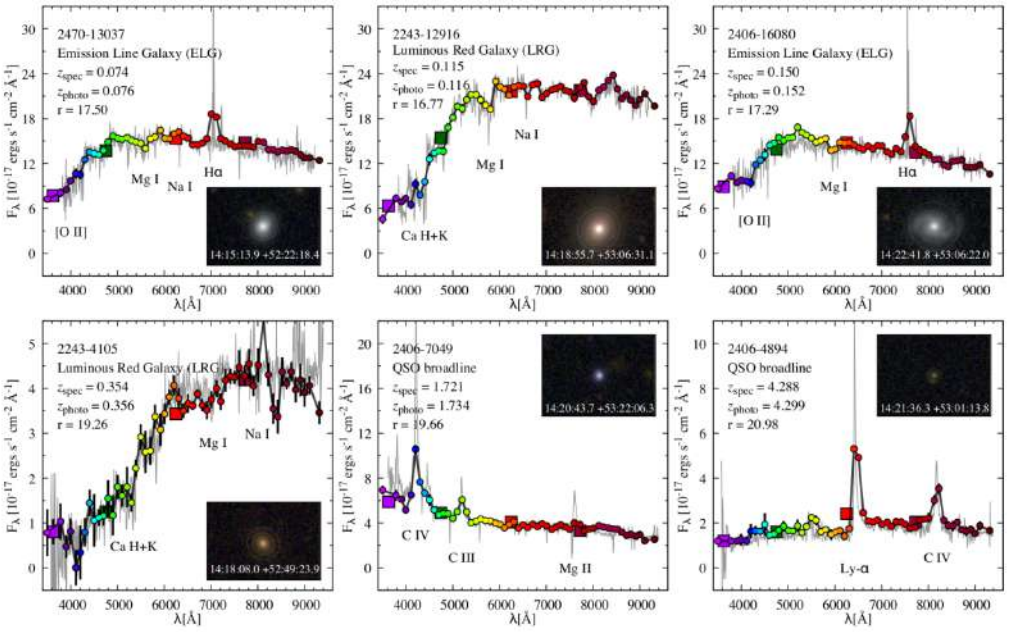


Figure 2.4: Comparison of the J-sepctra (coloured dots) with the SDSS spectra (grey lines) for galaxies and quasars in the miniJPAS field. The miniJPAS object ID, the r magnitude, the spectroscopic redshift, and the photometric redshift are provided in the legend. A multi-colour RGB image centred on the object covering 30 arcsec across is included for each object. Figure taken from Bonoli et al. (2021).

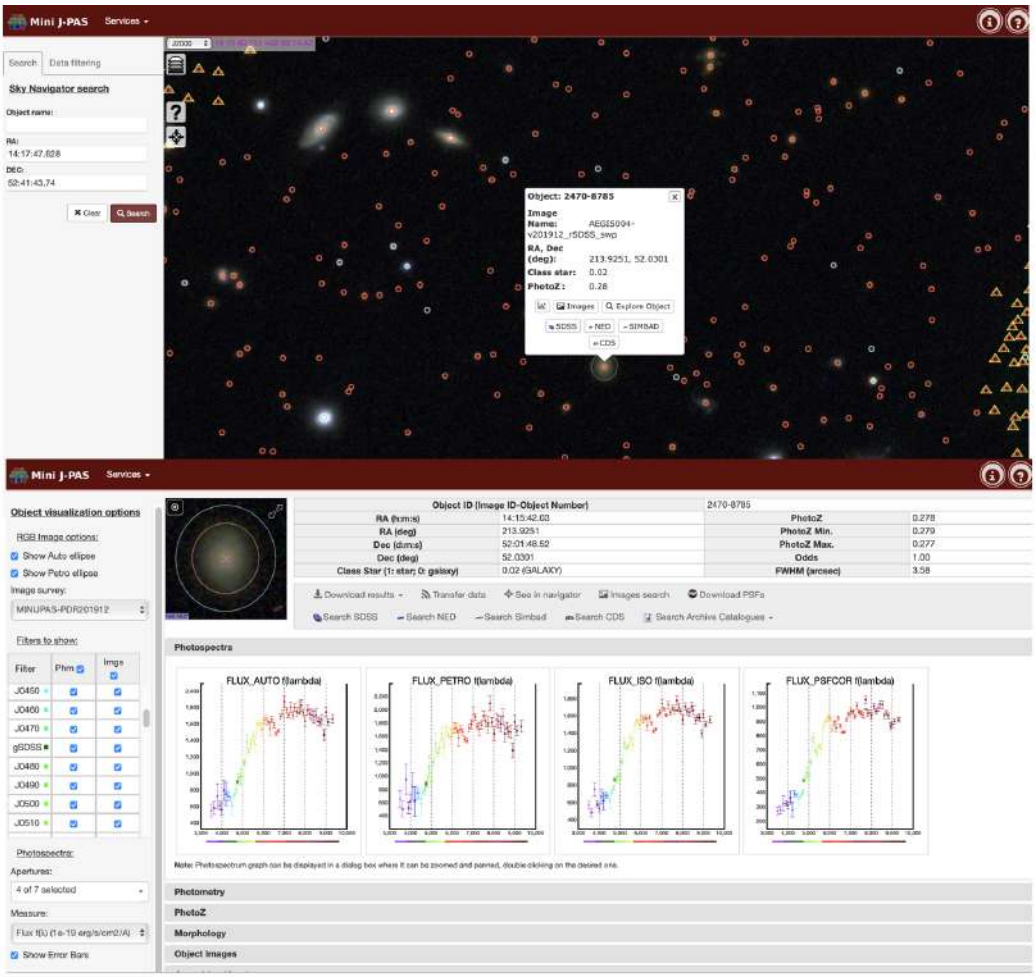


Figure 2.5: Screenshots of the Science Web Portal (see links on the footnotes). The sky navigator shows a region of the AEGIS field observed by miniJPAS. The photometry of one galaxy is visually inspected.

2.2 Photometric redhisft

One of the primary goals of J-PAS is to provide accurate photometric redshift (photo-z) for extragalactic objects so as to conduct cosmological experiments. As it was described in Benitez et al. (2014), it will be possible to measure baryon acoustic oscillations (BAO) not only for LRGs down to $z \sim 1$ but also for ELGs, and quasars up to redshift $z \sim 1.5$, and $z \sim 3$, respectively. Naturally, the photo-z precision decreases as objects are detected with lower S/N. In order to estimate photo-z for galaxies, Hernán-Caballero et al. (2021) customised a version of `LePhare` (Arnouts & Ilbert 2011) using a set of templates optimised for the J-PAS filter system, and studied a sample of 5 266 galaxies spcectocopically confirmed by SDSS and DEEP in the miniJPAS field. This sample is representative of the main observed properties of galaxies within miniJPAS. Thus, a certain desired photo-z accuracy, either in terms of a target $\sigma_{\rm NMAD}$ or a maximum outlier rate (η), can be achieved by setting constrain on the model's outputs, for instance the photometric redshift error or the *odds* (see Hernán-Caballero et al. (2021) for details). In miniJPAS there are $\sim$ 17 500 galaxies with $r_{\rm SDSS} < 23$ that present valid photo-z estimates with an average $\sigma_{\rm NMAD} = 0.013$ and outlier rate of $\eta = 0.39$. Around $\sim$ 4 200 of them are expected to have $|\Delta z| < 0.003$. In Fig. 2.6 we compare the estimated photo-z with the spectroscopic redshifts for the whole sample (left) and for *odds* > 0.61 (right). In these samples, 64 %, and 87 % of the objects have $|\Delta z| < 0.03$, respectively.

Photometric redshift can also be predicted for quasars (Bonoli et al. 2021). Queiroz et al., (in prep.) developed a method that estimate the photo-z of quasars through a PCA modeling of the spectral variations and a reddening law that takes into account the change in the slope of dusty-rich quasars. The eigenspectra of the PCA capture the most relevant features of a selection of broad-line quasars from SDSS. For the subsample of 97 SDSS quasars in the miniJPAS footprint with $z_{\rm spec} < 3.5$, and $r_{\rm SDSS} < 22$, the method reaches an uncertainty of $\sigma_{\rm NMAD} = 0.0059$ with an outlier fraction of 4.1% (Fig. 2.7). In the future, other methods, not only based on template fitting but also on ML, will be available in the J-PAS collaboration to predict phot$-z$ for quasars and galaxies. Thus, we will be able to improve further the current photo-z accuracy.

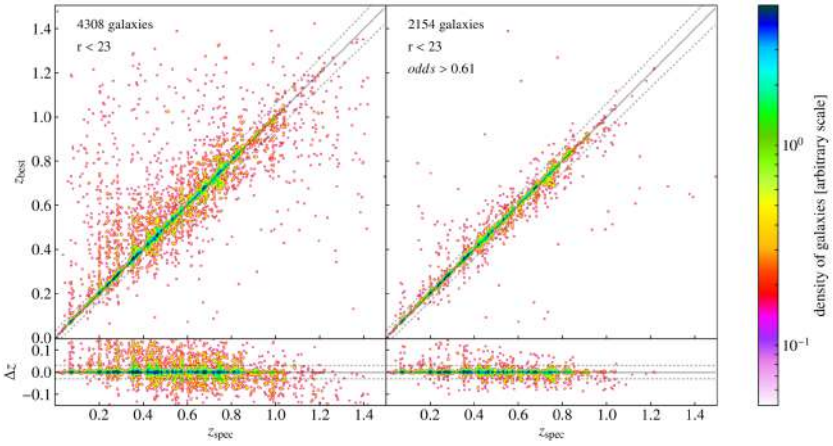


Figure 2.6: Comparison between photometric and spectroscopic redshifts for individual miniJPAS galaxies in the spectroscopic sample. The left panel includes all galaxies with $r_{SDSS} < 23$ and valid photo-z estimates, while the right one contains only half the sample (those with higher odds). The bottom panels show the redshift errors, $|\Delta z|$. A 2-D Gaussian smoothing is applied to the data to improve the visualisation of the density of points. The solid line marks the 1:1 relation, while the dotted lines indicate the $|\Delta|z = 0.03$ threshold used to define outliers. Figure taken from Hernán-Caballero et al. (2021).

2.3 Galaxy evolution studies with miniJPAS

Since the publication of miniJPAS data in December 2019 many works have proven the capability of J-PAS to conduct galaxy evolution studies. González Delgado et al. (2021) characterized the populations of unresolved galaxies in miniJPAS down to $z \sim 1$ by means of fitting the SED with parametric and non-parametric stellar population codes. The J-PAS filter system allows for the determination of the main properties of galaxies such as the total stellar mass, rest frame colours, the stellar extinction or the average stellar age. Some of the most relevant results of this work are shown in Fig. 2.8 and Fig. 2.9. Overall the population of red galaxies decreases at high redshift while the amount of blue galaxies increases. Although some discrepancies are found between codes at high redshift, those are most likely due to the drop in the number of galaxies per bin and the lower median S/N of such population that reaches values below 5. The average mass-weighted age of the stellar population decreases as a function of the redshift for both galaxy types regardless the SED fitting routine employed. Finally, a sub-population of galaxies in the nearby universe ($0.05 \leq z \leq 0.15$) was used to derive the cosmic evolution of the star formation rate density obtaining a great

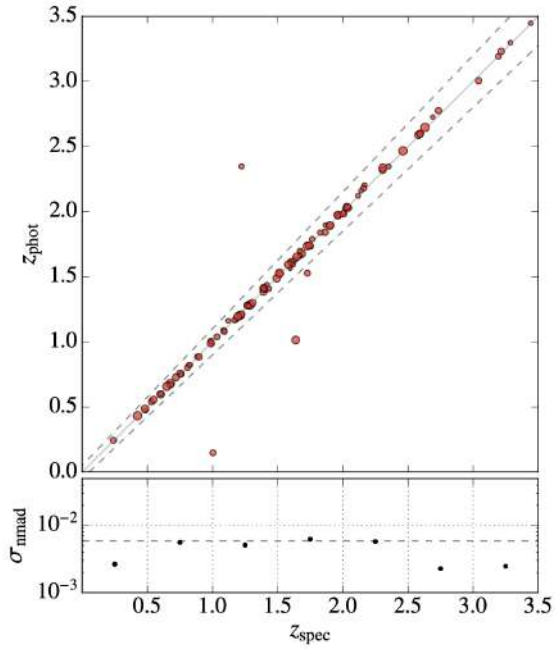


Figure 2.7: Upper panel: photometric versus spectroscopic redshifts for a sample of 97 DR14 quasars with $z_{spec} < 3.5$, and $r_{SDSS} < 22$. Larger symbols represent higher median S/N. The solid diagonal line indicates the 1:1 relation and the dashed lines correspond to $z_{phot} = z_{spec} \pm 0.05(1 + z_{spec})$. Bottom panel: photo-$z$ precision as a function of redshift, with the horizontal dashed gray line indicating the average photo-z uncertainty. Figure taken from Bonoli et al. (2021).

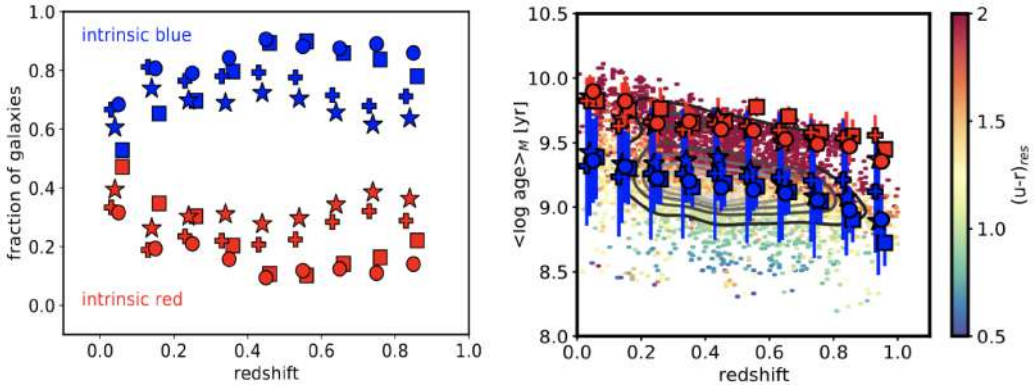


Figure 2.8: Evolution in the fraction of red and blue galaxies (right panel) and its age as a function of redshift obtained by `BaySeAGal` with a delayed-τ SFH (circles), `MUFFIT` (squares), `AlStar` (stars), and `TGASPEX` (crosses). The color-code on the right panel indicates the intrinsic ($u_{SDSS} - r_{SDSS}$) colour. Figures taken from González Delgado et al. (2021). More details of how these SED codes work can be found there.

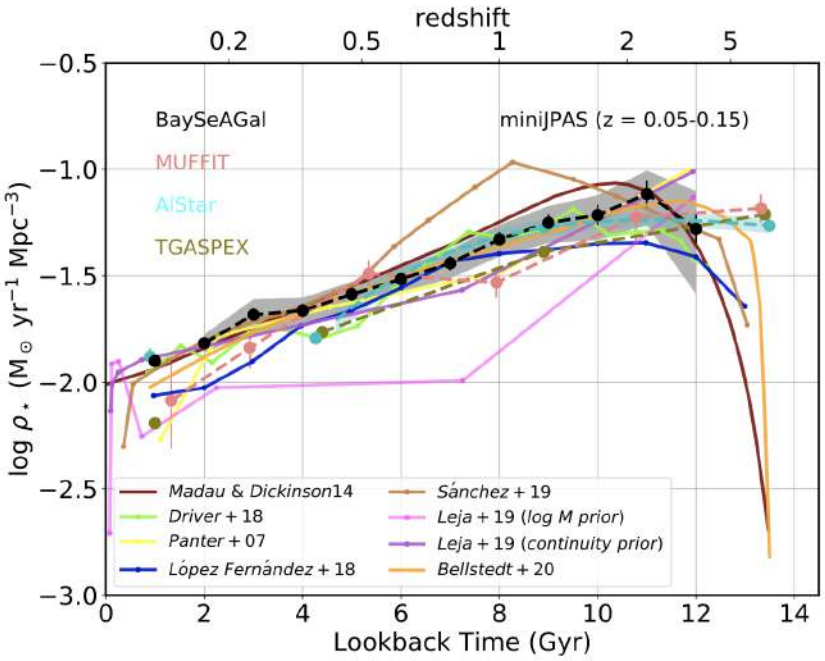


Figure 2.9: Cosmic evolution of the SFRD (ρ_*) obtained from the SED-fitting results of `BaySeAGal` (black dots), `MUFFIT` (coral dots), `AlStar` (cyan dots), and `TGASPEX` (olive dots) with nearby galaxies ($0.05 \leq z \leq 0.15$). The different lines represent the ρ_* obtained in other works. Figure taken from González Delgado et al. (2021).

agreement with other works based either on the fossil record method with galaxies from CALIFA (López Fernández et al. 2018) and MaNGA Sánchez et al. (2019) or the results derived from cosmological surveys (Sobral et al. 2013; Madau & Dickinson 2014; Driver et al. 2018).

The properties of galaxies as a function of the environment has also been studied very recently by González Delgado et al. (2022). In this work, the role that groups and clusters plays in quenching the star formation is investigated by comparing how their properties changes respect to galaxies that are found in the field. As expected, the fraction of red galaxies increases with the galaxy mass but it is always higher in groups than in the field. Similarly, the fraction of quenched galaxies ($sSFR < 0.1$ Gry^{-1}) is much higher in groups than it is in the field. Furthermore, Rodríguez Martín et al. (2022) prove that the same tools can be applied to analyze individual clusters such as mJPC2470-1771, the most massive cluster found in miniJPAS, and study the properties of each galaxy as a function of their location respect to the cluster center.

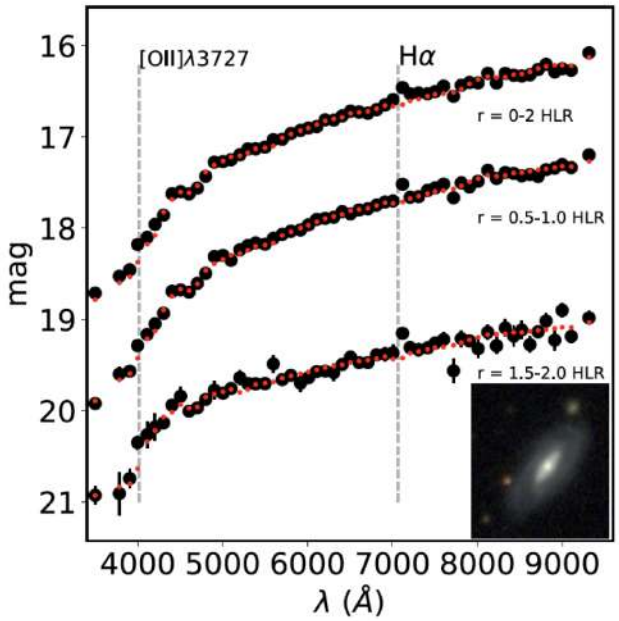


Figure 2.10: J-spectra of an ELG from miniJPAS at redshift $z = 0.077$. We show the total integrated J-spectrum (top) and the J-spectra within two elliptical rings (see text on the image). Note the variations in the intensity of the emission lines or the slope of the spectrum as a function of the radial distance.

Although very few galaxies are bright enough to be resolved spatially in the miniJPAS footprint, preliminary tests have shown the potential of J-PAS to work as a low resolution IFU. In Fig. 2.10 we show an emission line galaxy at redshift $z = 0.077$ with a HLR of ~ 7.5 kpc. The photometry has been extracted assuming elliptical rings apertures and shows variations of the J-spectrum from the center to the periphery, where the star formation is taking place at a higher rate. Another example is shown in Fig. 2.11, i.e. a quiescent galaxy at redshift $z = 0.075$ with a HLR of ~ 13.9 kpc that has also been observed by the MaNGA survey. The SED in the inner part of the galaxy ($r < 0.5$ HLR) obtained by the miniJPAS photometry shows a great agreement with the integrated spectrum of MaNGA over the same region. Furthermore, the average luminosity age of the stellar population derived by fitting the spectrum with `STARLIGHT` (Cid Fernandes et al. 2005) coincides with the results of `BaySeAGal` and `AlStar` that fit the miniJPAS photometry with different assumption on the SFH. While the FoV of MaNGA only reaches 1 HLR, miniJPAS images can go all the way up to 2 HLR. As we pointed out before, this is one of the major advantage of J-PAS compared to IFU-like surveys.

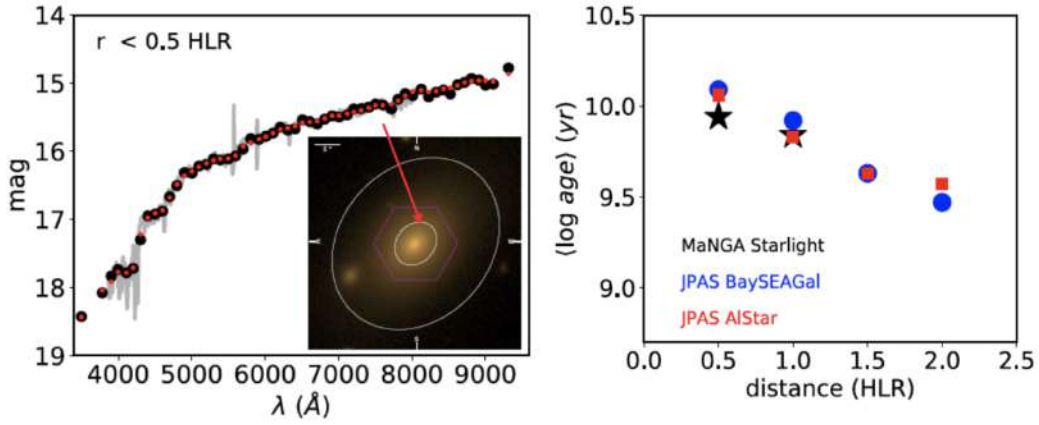


Figure 2.11: Left panel: The MaNGA spectrum of the central ring of 0a.5 HLR (grey line) of a galaxy at $z = 0.075$ is compared with miniJPAS data (black dots). The result of the best fit to miniJPAS data is also plotted (red dots). The inset shows an image of the galaxy in the r_{SDSS} band where two ellipses are overlaid at 0.5 HLR and 2 HLR. The FoV of the MaNGA survey is over-plotted as a red hexagon. Right panel: Comparison of the radial variation of the average luminosity age $< logt >_L$ derived from miniJPAS data, with the non-parametric code `AlStar` (red dots) and the parametric code `BaySeAGal` (blue dots), and from the MaNGA data analysed with the `STARLIGHT` code (black stars). Figure taken from Bonoli et al. (2021).

Chapter 3

Predicting emission lines with artificial neural networks in J-PAS

This chapter is based on the publication:

"J-PAS: Measuring emission lines with artificial neural networks"
by G. Martínez Solaeche, R. M. González Delgado, R. García-Benito et. al

Published in A&A, 647, A158 (2021)
https://doi.org/10.1051/0004-6361/202039146

3.1 Introduction

The study of the formation and evolution of galaxies through cosmic time has been addressed in recent decades through an understanding of how their physical properties leave footprints in the spectral energy distribution (see e.g., Díaz-García et al. 2019b, and references therein). Both the analysis of the light coming from stars and the ionized interstellar gas can be converted, via well-known recipes, to physical quantities such as the stellar mass, star formation rate (SFR), dust attenuation, luminosity-age, and gas-phase metallicity. In addition, they may unveil the main ionization mechanism responsible for the optical emission lines that we can observe in the spectrum (for some of the most recent reviews on these topics, see Conroy 2013; Madau & Dickinson 2014; Kewley et al. 2019).

The most massive and youngest stars that lie within galaxies are responsible for the ultraviolet emission in the spectrum, but very often, the presence of dust grains does not allow ultraviolet photons to travel freely through the interstellar medium; consequently, this makes it difficult to constrain the SFR from the blue part of the spectrum alone. However, those stars can actually ionize the surrounding interstellar gas. Very rapidly, hydrogen atoms recombine, leaving tracks the in form of emission lines at a particular wavelength in the spectrum. The Hα emission line at 6562.8 Å is less affected by dust extinction, thus it serves as an excellent tracer for measuring SFRs up to $z \sim 0.4$ in the optical range (Catalán-Torrecilla et al. 2015).

Other lines, such as the forbidden [O III] $\lambda\lambda$4959, 5007 Å and [N II] $\lambda\lambda$6548, 6584 Å doublets[1], are sensitive to the gas-phase metallicity, which is ideal for investigating the metal enrichment of gas throughout cosmic time (Maiolino & Mannucci 2019). The [N II]/Hα and [O III]/Hβ ratios, among others, were used to construct the so-called BPT diagrams (Baldwin et al. 1981), which distinguish galaxies where the gas is preferentially ionized by the presence of an active galactic nucleus (AGN) from those where the main ionization mechanism comes from high rates of star formation in the galaxy or shock-ionized gas regions.

Even though spectroscopic surveys have revolutionized astronomy across a number of fields, they provide a limited picture of the universe in many senses. Both multi-object spectroscopy (MOS) and integral field units (IFUs) surveys are partially biased due to the pre-selection of samples, where some properties such as fluxes, redshift, or galaxy-size are limited to a certain range. Some of these

[1] In the remaining of this chapter, [O III] λ5007 and [N II] λ6584 will be denoted [O III] and [N II] , respectively.

issues can partially be solved with narrow band photometric surveys. Although they have been historically limited to few filters, they can act as low-resolution spectrographs and they are able to map the sky quickly and deeply – therefore they offer a more comprehensive snapshot of the universe. Needless to say, some astrophysical analyses will always require the highest possible spectral resolution to fully exploit all the information encoded in the spectrum.

One of the most competitive astrophysical surveys designed to overcome the weaknesses of both photometry and spectrography, functioning halfway between them, could well be the Javalambre-Physics of the Accelerating Universe (J-PAS, Benitez et al. 2014). It will sample the optical spectrum with 54 narrow-band filters for hundreds of millions of galaxies and stars over $\sim$ 8000 deg^2. This is equivalent to a resolving power of $R \sim 60$ (J-spectrum hereafter). Initially thought to explore the origin and nature of the dark energy in the universe, J-PAS is also ideal for galaxy evolution studies and to detect emission line objects (Bonoli et al. 2021). However, the large number of galaxies peaking over a wide range of redshift makes it difficult to employ traditional methods, such as subtracting from the emission line flux the image of the stellar continuum (Vilella-Rojo et al. 2015). Furthermore, line fluxes will contribute to several J-PAS filters, which also vary with the redshift of the object. Consequently, it is necessary for new techniques and algorithms to be developed in order to completely leverage the capability of J-PAS.

Machine learning (ML) techniques have effectively become a powerful tool across many fields where large quantities of data are available. The capability of these algorithms to find patterns in the data without making any empirical or theoretical assumptions has turned out to be their main advantage. In recent decades, astrophysical surveys are increasingly releasing vast amounts of data, which brings the opportunity of employing the most sophisticated up-to-date algorithms in order to analyze them faster and more efficiently. The applications range from the estimation of photometric redshifts (Pasquet et al. 2019; Cavuoti et al. 2017) and the identification of stars (Whitten et al. 2019) up through the classification of galaxies (Domínguez Sánchez et al. 2018) and the separation between galaxies and stars (Baqui et al. 2021) up to the determination of the SFR (Delli Veneri et al. 2019; Bonjean et al. 2019) – to cite some of the most recent research. In this chapter, we developed a new method based on artificial neural networks (ANN) to detect and measure some of the main emission lines in the optical range of the spectrum: Hα, Hβ, [N II], and [O III].

This chapter is organized as follows. In section 3.2, we present the J-PAS data together with data from other surveys that have been used to train and test

the ANNs. In section 3.3, we describe in detail the main characteristics of the ANNs along with a discussion of how they can be trained and tested to deal with the uncertainties associated to the data. In section 3.4, we show the performance of ANNs in SDSS simulated data sets and discuss its main weaknesses. In section 3.5, we test our method in galaxies that have been observed both in miniJPAS and SDSS. Finally, we present a summary in section 3.6 and point out the steps needed to improve and extend the performance of the ANN in detecting and predicting emission lines.

3.2 J-PAS and spectroscopic data

In this section, we present J-PAS and the spectroscopic data used throughout this chapter for training and testing our ML codes. The reader can skip the following section 3.2.1 if she/he went thought section 1.2.

3.2.1 J-PAS

J-PAS is an astrophysical survey (Benitez et al. 2014) that is aimed at mapping out close to 8000 deg^2 of the northern sky with 56 bands, namely, 54 narrow-band filters in the optical range plus 2 medium-band – one in the near-UV (uJAVA band) and another in the NIR (J1007 band). With a separation of 100 Å, each narrow-band filter has a full width at half maximum (FWHM) of $\sim$ 145 Å, whereas the FWHM of the uJAVA band is 495 Å and the J1007 is a high-pass filter. The observations will be carried out with the 2.55 m telescope (T250) at the Observatorio Astrofísico de Javalambre, a facility developed and operated by CEFCA, in Teruel (Spain) using the JPCam, a wide-field 14 CCD-mosaic camera with a pixel scale of 0.46 arcsec and an effective field of view of $\sim$ 4.7 deg^2 (see Cenarro et al. 2019; Taylor et al. 2014; Marin-Franch et al. 2015). The survey is expected to detect objects with an apparent magnitude equivalent to $i_{AB} < 22.5$, up to $z \sim 1$ and with a photo-z precision of $\delta z \leq 0.003(1+z)$ for luminous red galaxies.

The J-PAS project started its observations taking data with the Pathfinder camera observing four AEGIS fields with 60 optical bands amounting to 1deg^2. These data allow us to build a complete sample of galaxies up to $r_{SDSS} \leq 22.5$ mag (Bonoli et al. 2021). More than 60.000 objects have been detected and can be downloaded from the website of the survey[2]. The survey, referred as to miniJPAS,

[2] http://www.j-spas.org/

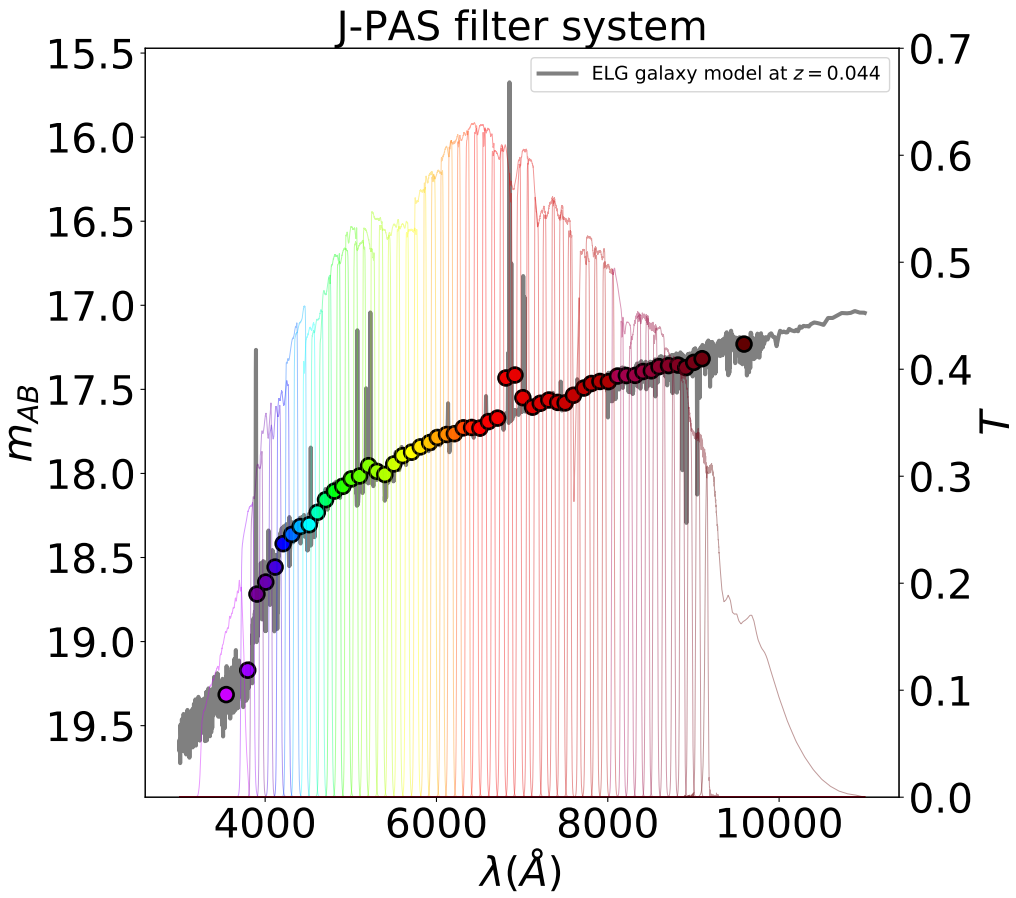


Figure 3.1: Synthetic photometry (colored dots) of an emission line galaxy model (gray line) at $z = 0.044$ in the J-PAS photometric system.

was described in detail in section 2 but we provide a short description in section 3.5.1. One example of how a nearby star-forming galaxy would be observe with J-PAS is shown in Fig. 3.1. We also show in the same figure the transmission curves of the J-PAS filter system.

3.2.2 CALIFA survey

The Calar Alto Legacy Integral Field Area (CALIFA, Sánchez et al. 2012; García-Benito et al. 2015) is an integral field spectroscopy survey which observed 600 spatially resolved galaxies in the local universe ($0.005 < z < 0.03$). The obser-

vations were taken with the 3.5 m telescope at the Calar Alto observatory with the Postdam Multi Aperture Spectrograph (PMAS, Roth et al. 2005) in the PPaK mode (Kelz et al. 2006), which contains 331 fibers of 2.7" in diameter. With a field of view of $71'' \times 64''$ and a spatial sampling of 1 arcsec/spaxel, CALIFA observed each galaxy in the wavelength range of 3700 – 7300 Å with two different overlapping setups. Here, we use the spectra taken in the low-resolution setup (V500) that provides spectra from 3745 to 7500 Å with a spectral resolution of 6 Å to generate J-PAS synthetic photometry.

There are measurements of the emission lines available for a total of 275787 spectra corresponding to 466 galaxies processed through the reduction pipeline of García-Benito et al. (2015). These spectra include emission patterns of many different zones within the galaxy. Therefore, even though the integrated spectra of CALIFA galaxies might not be heterogeneous enough to build a training set, the individual zones cover plenty of diverse physical states. The properties of the stellar populations and the state of the ionized interstellar gas change from one region to another in each individual galaxy. Hence, with the amount of galaxies observed with CALIFA, we can expect to see a rich representation of the most likely physical scenarios. The emission lines in each spaxel were measured from the residuals spectra obtained after subtracting the stellar continuum with STARLIGHT (Cid Fernandes et al. 2005).

3.2.3 MaNGA survey

The Mapping Nearby Galaxies at Apache Point Observatory (MaNGA, Bundy 2015) is an ongoing integral field spectroscopic survey that plans to observe spatially resolved spectra for ten thousand galaxies in the nearby universe ($z < 0.15$). With a wavelength coverage of 3600 – 10300 Å at a resolution of $R \sim 2000$, MaNGA is equipped with an IFU, in total 19 fibers of 12″ and 127 of 32″. In this chapter, we use the catalog available in [3] and processed by PIPE3D pipeline in MaNGA SDSS-IV datacubes Sánchez et al. (2016b,c). The analysis of the stellar populations and ionized gas provides spatially-resolved information of the strongest emission lines in the optical range for a total of 4670507 spaxels from 2755 galaxies.

[3] https://www.sdss.org/dr14/manga/manga-data/manga-pipe3d-value-added-catalog

3.2.4 SDSS survey

The Sloan Digital Sky Survey (SDSS, York et al. 2000) contains spectroscopic measurements for more than three million astronomical objects and deep images of one third of the sky in five optical bands. The spectra were taken with a fiber of 3″ in diameter and a spectral coverage of 3800 – 9200 Å at a resolution of $R \sim 2000$. Here, we use the publicly available MPA-JHU DR8 catalog from the Max Planck Institute for Astrophysics and Johns Hopkins University (Kauffmann et al. 2003b; Brinchmann et al. 2004). All the information regarding the catalog and the fitting procedure of the galaxy physical properties can be consulted online [4]. The catalog provides a total of 818333 galaxies with redshift up to $z \sim 0.35$. We only consider galaxies with reliable emission line measurements. Thus, we exclude the objects with `RELIABLE` = 0 and/or `ZWARNING` > 0 from the sample. We also discard galaxies where J-PAS synthetic magnitudes can not be calculated due to the lack of data in certain wavelength range of SDSS spectra. Finally, we ended up with spectra from 701975 galaxies.

3.3 Method of analysis.

In this section, we describe the architecture of the network in section 3.3.1 and the strategies used for training and testing the model in section 3.3.2. We also explain how to deal with photo-redshift uncertainty in section 3.3.3, how errors can be estimated in section 3.3.4, and how to treat missing data in section 3.3.5.

3.3.1 Architecture of the Network

In this chapter we use a class of ANN that is referred to as a fully connected neural network. The implementation was carried out with `Tensorflow` (Abadi et al. 2015) and `Keras` libraries (Chollet et al. 2015) in `Python`. It is composed of a set of layers which have a specific number of neurons. The first layer contains the inputs (features) of the network. In this chapter, the inputs are the colors of J-PAS measured with respect to the filter corresponding to Hα for each spectrum. For instance, in nearby galaxies (z< 0.015), the Hα emission line will be captured by the J0660 band. Then the color in the filter J_i is defined as the difference respect to the magnitude measured in the J0660 band ($C_i = m_{AB}(J0660) - m_{AB}(J_i)$). The final layer contains the output of the network, sometimes also named targets in

[4] www.sdss3.org/dr10/spectro/galaxy_mpajhu.php

the machine learning argot. Our targets are the equivalent width (EW) of Hα, Hβ, [N II], and [O III]. We built two different ANNs: one performs a regression task and obtains the values of these EWs, this network will be referred to as ANN$_R$. The other, ANN$_C$, carries out a classification between galaxies without emission lines (below a given threshold) and emission line galaxies (ELG) by imposing cuts in the EWs of the mentioned lines. We could have performed this classification based on the values yielded by the ANN$_R$ but an algorithm specifically constructed for a given task gives better results.

As we mentioned earlier, emission line fluxes have contribution to different bands according to the redshift of the source and the width of the emission line. The redshift might be treated as an input in the model but that would imply to train the ANN with a uniform distribution in this parameter, otherwise the ANN would not be able to make predictions equally at all redshifts. Furthermore, this approach would reduce our sample size and limit our range of predictability due to the different redshift coverage of CALIFA, MaNGA, and SDSS. For these reasons, we trained a different ANN for each redshift, going from 0 to 0.35 with a step of 0.001. We shifted all the spectra of the training set in wavelength at the same redshift and we computed the colors within the common wavelength range between J-PAS and the spectroscopic surveys described in Sect 3.2. This range depends on the redshift and, consequently, the number of inputs vary between 28 and 39 colors.

Between the input and the output layers, the ANN can hold inner layers, commonly called 'hidden' layers, with no restrictions to the number of layers and neurons in it. There is no standard recipe to find the optimal architecture of a network. Theoretically, with only one hidden layer and sufficient amount of neurons is possible to model the most complex function. However, deep ANN, i.e. those with mores hidden layers, have a much higher parameter efficiency, thus they are able to model complex functions by using much fewer neurons (Géron 2019). Few hidden layers are normally sufficient if the relation between input and output is not very complex. Certainly, this is our case since emission lines are clearly visible in the J-spectra. In addition, other features, such as the color of the J-spectra can help to estimate the emission line patterns because they linearly connected to the inputs.

The amount of neurons in the hidden layer varies between the size of the input and the size of the output layers. Our ANNs have 2 hidden layers with 20 neurons each, which is in between the number of inputs (34 colors in average) and the number of outputs (four EWs for the ANN$_R$ and two classes in the case of the ANN$_C$). A schematic view of the ANN$_R$ used in this chapter can be seen in

Fig. 3.2.

All the neurons in a given layer are connected to the neurons in the contiguous layer by a matrix of weights, **W,** and a bias, **B**:

$$\mathbf{L_n} = g(\mathbf{W_n} \cdot \mathbf{L_{n-1}} + \mathbf{B_n}), \tag{3.1}$$

where $\mathbf{L_n}$ refers to layer n. Also, L_0 are the inputs of the ANN and g is the activation function of neurons. It worth mentioning the importance of such a function, as it is responsible for the non-linear behavior in the network. Otherwise, the outputs would be simply a linear combination of the inputs, which would not be sufficient to address non-linar problems. We use the so-called Rectified Linear Unit (ReLU) activation function (Nair & Hinton 2010), which has become the default activation function in recent years due to its advantages (Glorot et al. 2011).

Typically, ANN are trained using an algorithm commonly referred to as back-propagation. Adjusting the set of weights and bias that minimizes a certain loss-function is the actual process of training. For regression-like problems the most common loss-function is usually a mean square error, while for binomial classification the binary cross entropy is frequently employed. We make use of these functions in our models.

One important aspect to take heed of when when we are training an ANN is to avoid overfitting. Improving the loss-function indefinitely would lead to the algorithm fitting features of certain data that do not represent the general trend. Consequently, the predictability of the network would be compromised. One way to avoid that is to impose a maximum value over the weights that each neuron can carry.

Optimising the architecture of the network is a process that requires tweaking many hyper-parameters. As part of these efforts, we tested different architectures, increasing and decreasing the number of neurons or hidden layers or by using alternative loss functions such as the mean absolute error or the mean relative error for regression. Sometimes even different architectures can obtain very similar results. The model that we describe in this chapter is among the ones we tested that better perform.

3.3.2 Training strategy

We generate synthetic J-PAS data by convolving the spectra presented in section 3.2 with the J-PAS filter system. Since the wavelength coverage of CALIFA,

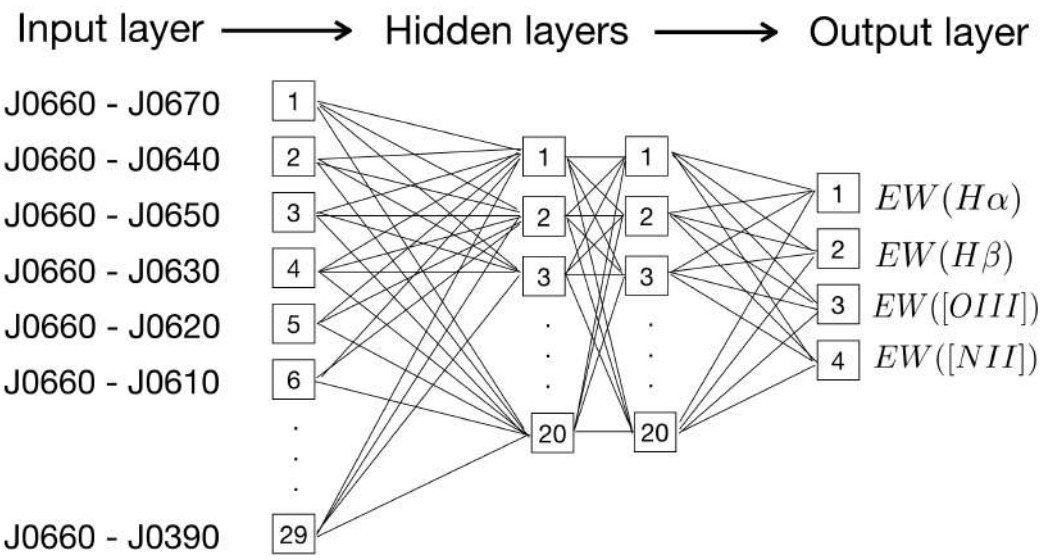


Figure 3.2: Schematic diagram of the ANN_R used for predicting lines emission at rest frame. The J0660 filter is our reference band for colors.

MaNGA, SSDS, and J-PAS are different, in our model, we only use the common wavelength range of the four instruments at $z = 0$, which is 3810 – 6850 Å.

The training sample is built differently, depending on whether we are dealing with a classification or a regression task. In a classification problem, an unbalanced number of classes in the training sample might under-predict the minor class (see e.g., Ali et al. 2015, for a review in the topic). Therefore, when is possible, a balance training set is more desirable. In regression-like problems the optimal training set is the one that better covers the parameter space of the target variables. For instance, a training set built for classifying galaxies above and below 3 Å in the EW of Hα will be different from one that aims to compute the same EW in the range between 0 and 20 Å. Simply because we would need many more galaxies in the interval from 3 to 20 Å than would be needed below 3 Å.

Considering the data that we have at hand, there are other aspects that need to be taken into account to build the training sample. First, in order to ensure the algorithm receives the most reliable information, we would wish to select only the spectra where emission lines have been measured with high signal-to-noise ratio (S/N). However, being too strict in the selection criterium induces a bias towards line-emmiting galaxies and reduces significantly the size of the sample. Second, while CALIFA and MaNGA have observed the nearby universe spatially resolving the physical properties of the interstellar medium within galaxies, SDSS can

only see the inner parts of nearby galaxies but with the advantage of covering distances further away in the universe. It has been shown how spatial resolution affects the location of points (spaxels) in the BPT, possibly altering AGN classification or simulating it via mixed spectral featured (Gomes et al. 2016). Finally, the emission line catalogs obtained from these surveys have been derived with different fitting tools, which makes it difficult to compare them in equal terms.

In essence, there is not a simple and unique way of putting together all these data and build the training set that better represents the universe as J-PAS will look at it. Instead, we propose to train the ANN with different training sets in order to understand the source of errors and inaccuracies of the model.

Training and testing sets in the ANN for classification

With the aim of identifying galaxies with low and high emission lines, we train a ANN classifier to perform a binary classification based on the EW of Hα, Hβ, [N II] or [O III]. This type of classification might allows us to disentangle the structure of the bimodal distribution found in the EW of Hα in CALIFA and SDSS galaxies (Bamford et al. 2008; Lacerda et al. 2018). In these works the authors found that the mentioned bimodal distribution has its minimum around 3 Å. In the regime of low emission the J-PAS filter system is not sensitive enough to detect emission lines and hence, it is only via machine learning, which can extract features from the J-spectra much more complex, it is possible to address this problem.

Galaxies are considered emitting-line galaxies or *Class* 1 according to the following criteria:

$$\begin{aligned} EW(\mathrm{H}\alpha) > EW_{min} \,||\, EW(\mathrm{H}\beta) > EW_{min} \,||\, \\ EW([\mathrm{O\,III}]) > \mathrm{EW_{min}} \,||\, \mathrm{EW}([\mathrm{O\,III}]) > \mathrm{EW_{min}}, \end{aligned} \tag{3.2}$$

and *Class* 2 in the rest of our cases. We trained several classifiers where EW_{min} takes the following values: 3, 5, 8, 11, and 14 Å. In short, if a galaxy has an EW greater than the EW_{min} in any of these lines, it will be considered as *Class* 1. If all the EWs in a galaxy are below the threshold then it will be tagged as *Class* 2.

In most of the cases, Hα is the most powerful emission line and, consequently, it determines whether galaxies belong to one class or other. There is nothing special in the values chosen for EW_{min} except that they are in the regime of low emission. With the ANN classifier we prove that this regime can be explored in J-PAS. In addition, any other EW_{min} around these values could be implemented in the future.

The combination of data from different surveys used in this chapter does not improve or worsen the performance of the ANN classifier. Consequently, for the sake of simplicity, we train only with CALIFA synthetic J-spectra and we test with SDSS galaxies. We do not impose any cut in the errors of the EWs, but we ensure to have the same amount of J-spectra in both classes in the training set. We end up with 200000 synthetic J-spectra to perform the training.

Training and testing sets in the ANN for regression

For the purpose of obtaining the values of the EWs of galaxies in J-PAS, we propose two training sets. The first one, which we call the CALMa set, is only composed of CALIFA and MaNGA synthetic J-spectra, while the second one, the SDSS set, includes only SDSS galaxies.

We test the performance of the model by randomly removing 15000 synthetic J-spectra from the training samples: 5000 from CALIFA, 5000 from MaNGA and 5000 from SDSS. Those synthetic J-spectra are considered as validation or test samples depending on the training sample. For instance, if we train with the CALMa set, we use MaNGA and CALIFA samples to tune the hyper-parameters of the model (validation samples) and SDSS galaxies to actually evaluate the model; and the other way around: if we train with the SDSS sample, SDSS galaxies plays the role of the validation sample and CALIFA and MaNGA synthetic J-spectra are used for testing purpose. In this way, we ensure that the color terms that might appear as a result of fitting tools used to derive the emission lines or the instruments that obtained the spectra are not playing a major role in the prediction made by the ANN. If that were the case, building samples with different surveys in the training and testing sets would allow us to identify any potential bias arising from such circumstances.

We add only those synthetic J-spectra to the training set that have emission lines with an error below a certain threshold. In the case of MaNGA galaxies, spaxels with a S/N below 10 in the flux of Hα, Hβ, [N II] or [O III] were discarded. However, we were more flexible with spaxels in CALIFA and SDSS galaxies, going down to a S/N of 2.5. Such flexibility allows us to increase the amount of low-emitting galaxies in the samples. In addition, when it comes to the CALMa set, we achieved a more equilibrated weight between the prominence of CALIFA and MaNGA in the training sample. We also excluded from the training set those spectra where the EWs are greater than 600 Å (these are very rare cases, 10 in total). Since the loss function is quadratic in the EWs, this type of spectra force the ANN$_R$ to fit, at the same time, two antagonistic regimes: low-emitting and

extreme emission line galaxies. Consequently, it would worsen the performance of the ANN_R in the range of interest. Finally, we ended up with a training set of 134000 synthetic J-spectra from CALIFA, 280270 from MaNGA that altogether make up the CALMa set, as well as 135300 galaxies in the SDSS set.

3.3.3 Photo-redshift uncertainty

Even though J-PAS will provide redshifts with a high precision (Benitez et al. 2014, $\delta z \leq 0.3$ % [5] for luminous red galaxies), the performance of the ANN could be compromised in certain cases. Let us assume, for example, that we aim to predict the EWs of a galaxy at redshift 0.3 with $\Delta z = 0.003$. In the best-case scenario, the galaxy redshift would be between 0.296 and 0.304. According to our redshift bin, we have eight possible ANNs to test with. While in the vicinity of the true redshift the ANN can do a reasonably good job, in the extremes, the EWs would dramatically be underestimated. Since colors are computed with respect to a filter far away from the one corresponding to Hα, the ANN will interpret as an absorption line what indeed is an emission line. Although the probability density functions (PDFs) of the photo-z can help to improve the predictability in assigning weights to each redshift; whenever we found a non-gaussian PDF with, for instance, an asymmetric distributions with two peaks, it would be difficult for the ANN to make reasonable predictions.

One way to obtain better results in galaxies where the uncertainty in the redshift is high is to consider only the configurations (redshifts) that maximize a certain function. Certainly, for emission line galaxies, the redshift where the sum of all EWs reaches the highest value is close to the true redshift. However, this redshift overestimates the EWs in galaxies with low emission. In order to minimize such an effect, we average over the five configurations (redshifts) that maximize the sum of all EWs within the photo-redshift uncertainty (Δz). The fact that these configurations might be found in non-contiguous redshift bins can help in those cases where there are asymmetric PDF distributions of photo-redshifts.

As we go on to discuss in section 3.4.4, this method is capable of somehow recomputing the redshift of the galaxy, correcting a possible deviation from the spectroscopic redshift in galaxies where $\sum EW_i > 20$ Å. Therefore, the method of the five maximum, hereafter *5max*, can certainly help the ANN_R to improve its performance but cannot be used with the ANN_C. Most probably, it would increase the amount of false positives as the redshift uncertainty increases. In section 3.4,

[5] Throughout this chapter we use the convention $\Delta z = (1 + z)\delta z$, where $\Delta z = z - z_{photo}$.

we quantify how the error in the redshift can impact the predictions of the ANN_C and the ANN_R. Fortunately, the ANN_C is less sensitive to that effect (see Fig. 3.3 and Table 3.1).

3.3.4 Estimation of errors

The uncertainty of the ANN method can be estimated by considering three sources of error: the error of the photometry, the error in the photometric redshift, and the intrinsic error of the ANN training. Before the training actually starts, weights and biases in ANN can be set to a certain value by initialising randomly according to any distribution function. Generally, each initialization state will converge to different local minimum of the loss-function. Even though it is possible to find the state that leads to the best score over the validation sample, it is usually a Monte Carlo approach called the committee (i.e. the mean of the individual predictions of a set of ANN) that will be a more robust and accurate estimate of the targets. Thus, the variations of the outputs in each individual member of the committee with respect to the mean provide an estimation of the uncertainty in the predictions intrinsically associated to the training procedure. The paragraphs bellow details the steps to follow in order to account for the contribution of each uncertainty to the errors budget.

Photometric error: we input the ANN with $N + 1$ different values of the magnitude, where one corresponds to the nominal value and the other N are randomly drawn from a gaussian distribution centred on the nominal value and with standard deviation equal to the photometric error. The median (M) and the median absolute deviation (MAD) of N+1 predictions give us the prediction and the weight of one member in one committe:

$$P_{iz_j} = \mathrm{M}[p_0^{iz_j}, p_1^{iz_j}, ..., p_{N+1}^{iz_j}],$$
$$W_{iz_j} = 1/\mathrm{MAD}[p_0^{iz_j}, p_1^{iz_j}, ..., p_{N+1}^{iz_j}],$$

where i stands for the committe member and z_j for the redshift.

ANN intrinsic error: the prediction of the committe in a given redshift can be estimated by computing the average (AVG) of all members in the committe with the weights obtained above. The error of the committe is simply the MAD of m(N+1) prediction, where m refers to the number of members in the committe. We found that averaging over five members is enough to obtain reliable results:

$$P_{z_j} = AVG[P_{0z_j}, P_{1z_j}, ..., P_{mz_j}; W_{0z_j}, W_{1z_j}, ..., W_{mz_j}],$$
$$\epsilon_{z_j}^{ANN} = \mathrm{MAD}[p_0^{1z_j}, ..., p_{N+1}^{1z_j}, p_0^{2z_j}, ..., p_{N+1}^{2z_j}, ..., p_0^{mz_j}, ..., p_{N+1}^{mz_j}],$$

Photo-redshift uncertainty: we compute the median value of n committes, one for each redshift. In the case of the ANN_R we select the five maximum setting (see section 3.3.3) and for the ANN_C, we consider all the redshift within the error range:

$$P_{ANN_R} = \mathrm{M}[P_{z_0}(max_0), P_{z_1}(max_1), ..., P_{z_4}(max_4)],$$
$$P_{ANN_C} = \mathrm{M}[P_{z_0}, P_{z_1}, ..., P_{z_n}],$$

Finally, the error is the quadratic sum of the median error of all committees plus the dispersion of these committees respect to the median, which gives us the contribution of the redshifts uncertainty.

$$\epsilon_{ANN} = \sqrt{\mathrm{M}[\epsilon_{z_0}^{ANN}, \epsilon_{z_1}^{ANN}, ..., \epsilon_{z_n}^{ANN}]^2 + \mathrm{MAD}[P_{z_0}, P_{z_1}, ..., P_{z_n}]^2},$$

If the spectroscopic redshift of the object were known, the expression above would be simply: $\epsilon_{ANN} = \epsilon_{z_{spec}}^{ANN}$.

3.3.5 Missing data

There are a number of problems, both related to the data reduction or the observation, that could lead to incomplete or missing data. Consequently, a fraction of our sample will lack photometric measurements in some of the filters used by the ANN. Certainly, many such objects would have to be rejected automatically if the photometry is not reliable in the bands capturing the emission lines. However, there will be galaxies where the photometry might be problematic only in some of the bands dominated by the stellar continuum. For instance, in the miniJPAS area, among the galaxies that are below 0.35 in redshift and 22.7 magnitudes in the rSDSS band (2291), 30 % of them have at least one band where the photometry is not reliable. Most of the galaxies in this sample (70 %) have a median S/N below 10. Naturally, this fraction will decrease as the median S/N of the sample increases.

One solution to address the problem of missing date requires training several ANN and considering different configurations where part of the data is accessible. Nevertheless, this would imply testing the performance of the ANN in many scenarios and would be computationally very expensive. The other solution is to replace the missing data in the corresponding filter with the fluxes obtained from the spectral fitting of the stellar continuum. Several spectral fitting codes can be used, such as `MUFFIT` (Díaz-García et al. 2015) or `BaySeAGal` (Amorim

et al. in prep.). This analysis provides reliable photometric predictions for the missing data, as well as information regarding their stellar population properties (e.g., stellar mass, age, and extinction, which is always necessary for a more comprehensive picture). Furthermore, the stellar continuum is needed for obtaining absolute emission line fluxes. We follow this technique to treat the missing data in J-PAS.

3.4 Validation of the method.

In this section we perform several tests to study the predictability and limitations of the model. First, we evaluate the capability of the ANN_C in section 3.4.1. Second, in section 3.4.2, we compare the predictions of the EWs obtained by the ANN_R and trained with the CALMa set with the SDSS testing sample. In section 3.4.3, we compare the performance of the different training sets proposed in section 3.3.2. In section 3.4.4, we test the $5max$ method and we study the impact of the redshift uncertainty on the ANN_R predictions as a function of the EW in section 3.4.5. Finally, in section 3.4.6 we estimate the minimum EW measurable in function of the S/N of the photometry for each of the emission lines predicted by the ANN.

3.4.1 Classifying galaxies

The ANN_C is trained with the CALIFA training sample. To evaluate its efficiency, we explicitly selected a subset of 10000 galaxies from the SDSS catalog with of which 5000 galaxies belongs to $Class1$ and 5000 to $Class2$. (see section 3.3.2). Galaxies in each class are picked at random from the catalog. For each galaxy, the ANN_C yields a number between 0 and 1 indicating the probability of being one of the two classes. As we discuss in section 3.4.4, the $5max$ method (section 3.3.3) is not suitable for galaxies without emission lines. Most probably, it would increase the amount of false positives as the redshift uncertainty increases. Since we have noticed that the ANN_C is less sensitive to redshift and is able to classify galaxies even when the uncertainty is high, we simply compute the average of each one of the predictions within the redshift interval defined by δz.

We show in Fig. 3.3 the receiver operating characteristic (ROC) curve, which represents the true positive rate (TPR) versus the false positive rate (FPR) for $EW_{min} = 3$ Å. We also show how the ROC curve varies as a function of the redshift uncertainty. The ANN_C scores very high even when $\delta z = 0.01$ and loses

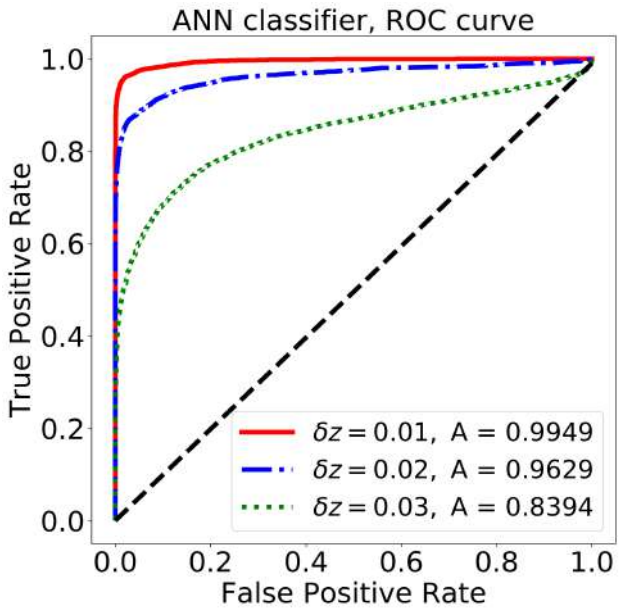


Figure 3.3: ROC curve of the ANN$_C$ for EW_{min} = 3 Å as a function of the redshift uncertainty for 10000 SDSS galaxies. The legend shows the areas under the ROC curves for each Δz. In Table 3.1 we show these values for other EW_{min} settings. Blue dashed line shows the performance of a random classifier.

efficiency gradually as the uncertainty in the redshift increases. We summarize in Table 3.1 the area under the ROC curves for others EW_{min}. The ROC curves do not show remarkable changes in function of the EW_{min} used in the classification.

3.4.2 ELG: EWs, line ratios and BPT diagram

In this section, we discuss how the CALMa training set (see section 3.3.2) scores in the SDSS testing sample. We use the spectroscopic redshift provided in the catalog without considering any error so as to separate the uncertainties intrinsically associated to the model from those related to redshift. We do not consider the errors of SDSS spectra; rather, we add Gaussian noise to each magnitude 100 times, assuming an average S/N of 10.

The testing set from CALIFA, MaNGA, and SDSS are composed of 5000 synthetic J-spectra with S/N in the EWs above 10. This criterion excludes preferentially galaxies with low emission. We also exclude the spectra where the EWs are greater than 600 Å to test the model in the range of which we trained the

EW_{min}	Area ($\Delta z = 0.01$)	Area ($\Delta z = 0.02$)	Area ($\Delta z = 0.03$)
3 Å	0.9949	0.9629	0.8394
5 Å	0.9948	0.9507	0.8160
8 Å	0.9938	0.9604	0.8407
11 Å	0.9915	0.9594	0.8547
14 Å	0.9894	0.9600	0.8614

Table 3.1: Area under the ROC curve as a function of the redshift uncertainty and the EW_{min} used in the classification.

ANN_R. Hence, even though we are able to identify strong and weak emission lines galaxies, their EWs might not be accurate due to these selection criteria on the training sample.

Equivalent widths

Fig. 3.4 compares the EWs predicted by the ANN_R and those in the SDSS testing sample (extracted from the MPA-JHU DR8 catalog). We do not plot the errors yielded by the ANN_R for visual reasons. A complete analysis of the errors estimated by the ANN_R, as discussed in section 3.3.4, is performed in section 3.4.6. The plots on the left are color-coded with the density of points and the ones in the middle with the redshift of the galaxy. The histograms on the right represents the relative difference between the ANN_R predictions and the SDSS testing set. We constrain better the EW of Hα followed by Hβ, [O III] and [N II] (see median and median absolute deviation in Fig. 3.4). The Hα line, which is the most powerful one, presents less dispersion and bias. Hβ and [O III] lines are recovered with similar precision and [N II] line shows more dispersion and bias. We observe that [N II] line saturates at high values, that is to say, the EWs tend to be underestimated as the strength of the line increases. The same effect occurs in the [O III] line in form of a second branch. We analyze this effect in section 3.4.2. We do not observe strong color gradients in the middle panels, indicating we are not biased with regard to the redshift of the galaxy.

In summary, the EWs of Hα, Hβ, [N II], and [O III] can be predicted with a relative standard deviation of 8.4 %, 13.7%, 14.8 %, and 15.7 % respectively. Hα, Hβ, [N II], and [O III] lines presents a relative bias of 0.03%, 5.0 %, 4.8 %, and -6.4 % respectively.

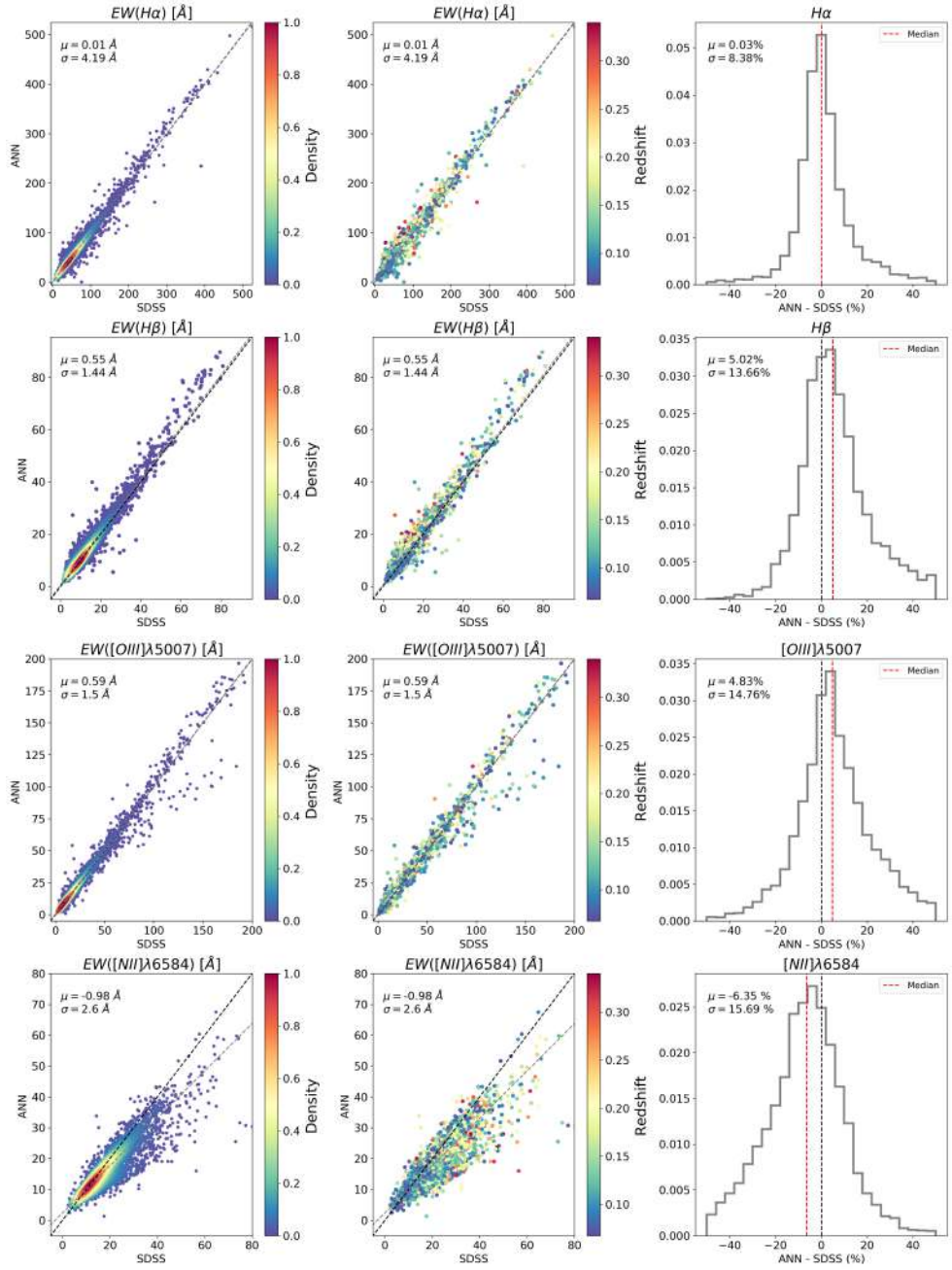


Figure 3.4: EWs of Hα, Hβ, [N II] and [O III] predicted by the ANN$_R$ compared to SDSS testing sample. The ANN$_R$ is trained with the CALMa set. The color-code represents the density in arbitrary units (right panel) and the redshift (left panel). The normalized histograms show the relative difference between both values. Black and blue numbers are the median and the MAD of the difference. Black line is the 1:1 relation and grey dashed lines represents the best linear fit. The red dashed line represents the median.

Ratios between emission lines

Based on the EWs, we can easily obtain the ratios of [N II]/Hα and [O III]/Hβ under the approximation that each couple has the same stellar continuum. From that, we also obtain the metallicity indicator O3N2 $\equiv$ log{([O III]/Hβ)/([N II]/Hα)} (Pettini & Pagel 2004). Fig. 3.5 shows the comparison between the logarithmic ratios obtained with ANN$_R$ and the SDSS testing sample. As in Fig. 3.4, the plots are color-coded with the density of points (left column) and the redshift of the galaxy (middle panel). The histograms on the right show the logarithmic difference between the ANN$_R$ predictions and the SDSS testing set.

The [N II]/Hα ratio is predicted within 0.092 dex and a bias of -0.02 dex. The [O III]/Hβ ratio is slightly better constrained, with no bias and a dispersion of 0.078 dex. Finally, the O3N2 is recovered within 0.108 dex and a bias of 0.04 dex. The saturation of the [N II] line at high values is responsible of the same effect observed in the [N II]/Hα ratio. Since MaNGA and CALIFA surveys observed galaxies spatially resolved, the number of star-forming spaxel is much more numerous in the training sample and consequently the ANN$_R$ has few spectra to constrain the ratio of [N II]/Hα in galaxies hosting an AGN. To a lesser extent, that also occurs in the [O III]/Hβ ratio for galaxies with values higher than 3.2 and in the form of a second branch in the [O III] line.

BPT diagram

In Fig. 3.6, we compare the BPT diagram recovered by the ANN$_R$ (left plot) and the one obtained from the SDSS testing sample (right plot). Galaxies are color-coded with the density of points and are grouped into four classes by three dividing lines: star-forming, composite, Seyfert, and LINER. The solid curve is derived empirically using the SDSS galaxies (Kauffmann et al. 2003a, hereafter ka03). The dashed curve is determined by using both stellar population synthesis models and photoionization (Kewley et al. 2001, hereafter Ke01). The dotted line is a empirical division between Seyfert and LINER found by (Schawinski et al. 2007, hereafter S07). The sequence of metal enrichment experienced by star-forming galaxies from high to low values of the [O III]/Hβ ratio is clearly visible and well reproduced in the diagram. We will refer to that as the SF-wing. However, the saturation of the [N II]/Hα and [O III]/Hβ ratios produces the migration of galaxies from right to left and from top to bottom lowering the percentage of Seyferts (from 10.04 % to 6.78 %), composite (from 15.4 % to 10.33 %) and LINERS galaxies (from 1.7 % to 0.21 %) and increasing the percentage of star-forming galaxies

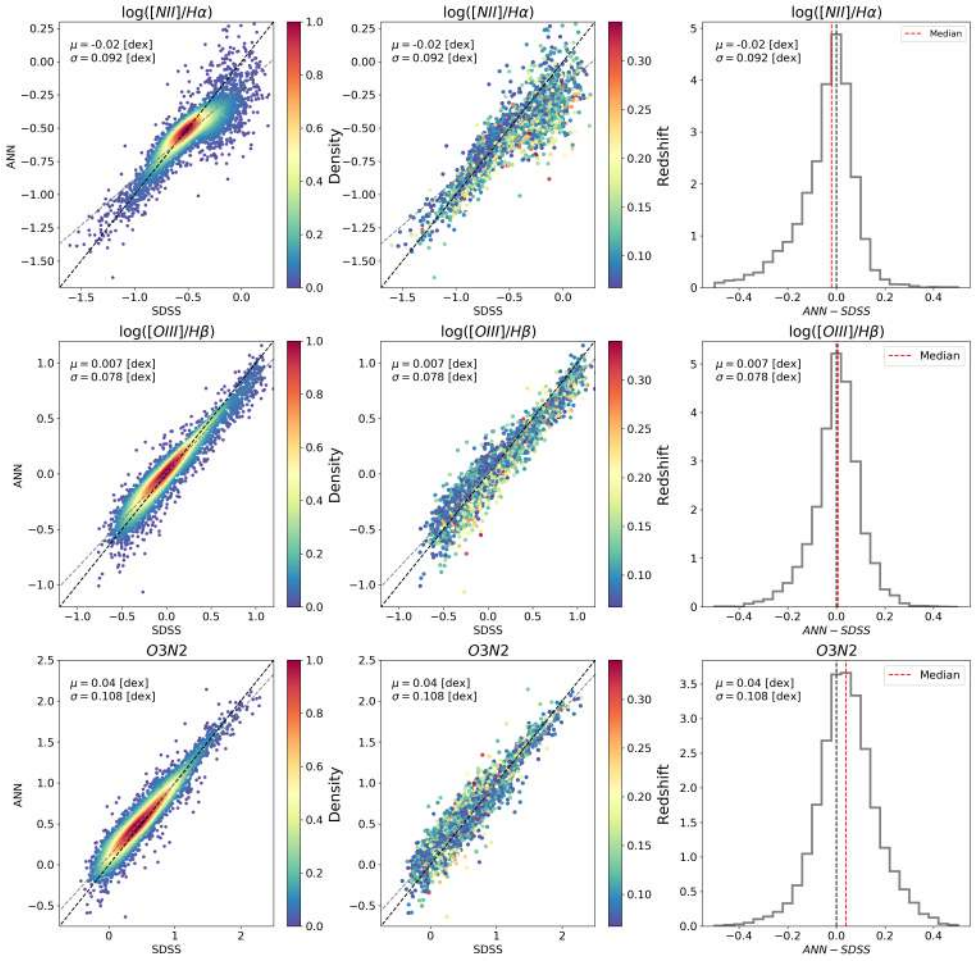


Figure 3.5: Comparison between [N II]/Hα, [O III]/Hβ and O 3N 2 ratios estimated by the ANN$_R$ and SDSS testing sample. Same scheme of Fig. 3.4. The ANN$_R$ is trained with the CALMa set.

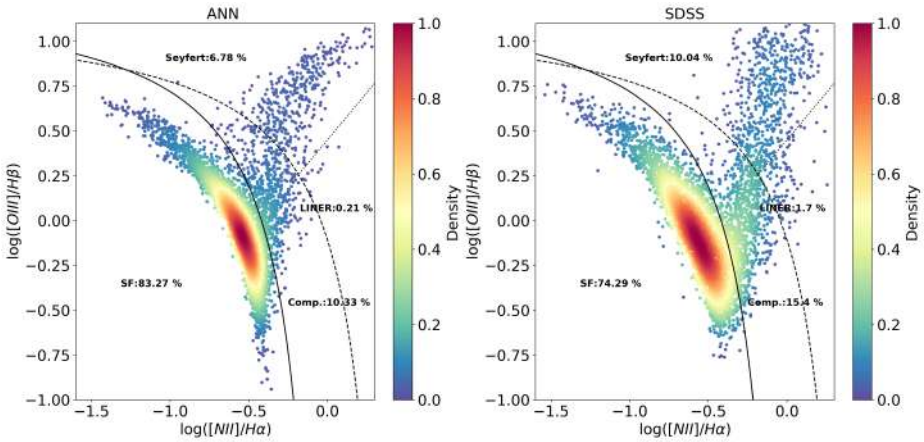


Figure 3.6: BPT diagram obtained with the ANN$_R$ and SDSS testing sample from the MPA-JHU DR8 catalog. The ANN$_R$ is trained with the CALMa set. The color-code indicates the density of points. The solid (ka03), dashed (Ke01) and dotted lines (S07) define the regions for the four main ionization mechanism of galaxies. The percentage for each group is shown in black.

(from 74.29 % to 83.27 %).

Another way to look at this is Fig. 3.7. We show the direction towards the location which galaxies should be placed in the BPT according to the SDSS MPA-JHU DR8 catalog. The vectors are color-coded with the distance of each galaxy between the two BPT diagrams and those at a greater distance are plotted last. On average, star-forming galaxies deviate 0.10 dex while Seyfert and composite galaxies do 0.12 dex. In the right panel of Fig. 3.7, we plot the angular distribution of star-forming, Seyfert, and composite galaxies. The angle is defined as a clockwise rotation towards the x axis. While star-forming galaxies do not show any preferential direction, Seyfert and composite galaxies point with an average angle of 45^o in the diagram. The CALMa set is very good at predicting the SF-wing because the main ionization mechanism in most of the regions in CALIFA and MaNGA galaxies is dominated by star-formation process. However, galaxies with a high [N II]/Hα ratio are more difficult to constrain.

3.4.3 Comparison between different ANN$_R$ training sets

As we pointed out in the section 3.3.2 we trained the ANN$_R$ with two different training samples. In Appendix A, we show the results obtained with the SDSS training set in the SDSS testing sample. A quick look at these plots (Appendix A.1, A.2, and A.3) proves the importance of testing the model on data with a

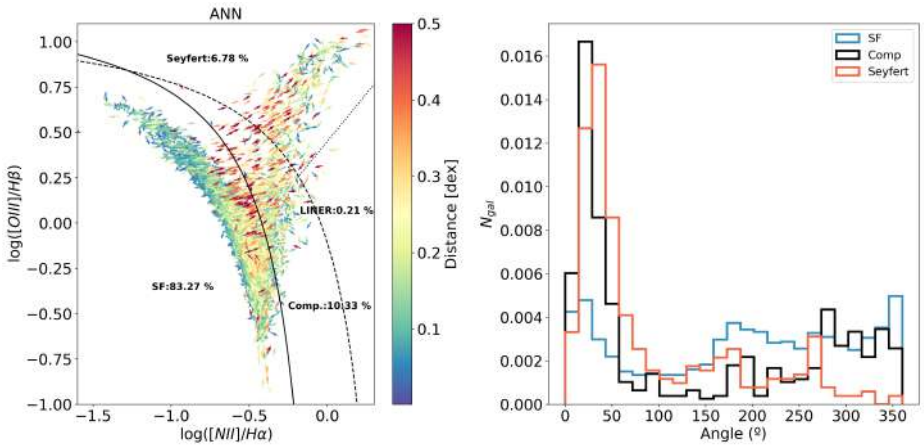


Figure 3.7: BPT diagram obtained by the ANN_R trained with the CALMa set. Arrows point in the direction towards the location where galaxies should be placed according to their position in the SDSS MPA-JHU DR8 catalog. The color represents the distance for each point between the two BPT diagrams. The solid (ka03), dashed (Ke01) and dotted lines (S07) define the regions for the four main ionization mechanisms of galaxies. The percentage for each group is shown in black. The histograms on the rights represent the angular distribution of the arrows for Star forming, Seyfert and composite galaxies. The angle is defined as a clockwise rotation towards the x axis.

different observational setup and calibration. Considering the fact that the EWs are estimated from a pseudo-spectrum (J-spectrum) with a much lower resolving power, the performance of the SDSS training set in SDSS testing sample is outstanding. Nevertheless, it would not be realistic to deduce from that the actual capability of this method to predict in J-PAS data. Testing the CALMa training set with SDSS galaxies or vice versa gave us a better picture of the weakness and inaccuracies of the model. For instance, the predictions made by ANN_R that were trained with SDSS set on the [N II]/Hα and [O III]/Hβ ratios of MaNGA and CALIFA spaxels tend to be overestimated. This is the opposite effect observed when the ANN_R is trained with CALMa training set and tested on SDSS galaxies. The performance on the validation samples, i.e. the data that belongs to the same survey, is generally better.

In Table 3.4.3 and 3.4.3 we show the performance of both training sample (SDSS test and CALMa set) in each one of the testing sets (CALIFA, MaNGA and SDSS). There is always an emission line that is better recovered in one particular simulation, for example, Hα in CALMa versus SDSS, however, the overall performance of the ANN_R is generally more accurate using data from the same

survey.

Training vs Test	Hα (%)	Hβ (%)	[O III] (%)	[N II] (%)
SDSS vs SDSS	-0.4 ± 8.1	-2.1 ± 12.4	1.9 ± 16.0	2.7 ± 16.4
SDSS vs CALIFA	-6.3 ± 10.7	-12.5 ± 13.5	-5.3 ± 21.1	-2.3 ± 21.4
SDSS vs MaNGA	-2.4 ± 11.1	-8.1 ± 13.7	-3.5 ± 19.9	9.8 ± 22.1
CALMa vs CALIFA	-4.4 ± 8.1	-4.9 ± 12.2	1.5 ± 19.2	-3.8 ± 15.3
CALMa vs MaNGA	-2.3 ± 8.6	-1.7 ± 12.2	0.4 ± 17.4	8.4 ± 18.2
CALMa vs SDSS	0.03 ± 8.4	5.0 ± 13.7	4.8 ± 14.8	-6.4 ± 15.7

Table 3.2: Relative difference between the EWs (in percentage) predicted by ANN_R and the true values. Two training sample are used for training: CaLMA and SDSS, and three for testing: SDSS, CALIFA and MaNGA.

Training vs Test	[N II]/Hα [dex]	[O III]/Hβ [dex]	O 3N 2 [dex]
SDSS vs SDSS	0.019 ± 0.089	0.027 ± 0.080	0.014 ± 0.12
SDSS vs CALIFA	0.018 ± 0.122	0.04 ± 0.102	0.023 ± 0.159
SDSS vs MaNGA	0.06 ± 0.105	0.033 ± 0.096	-0.031 ± 0.148
CALMa vs CALIFA	0.003 ± 0.088	0.035 ± 0.089	0.037 ± 0.131
CALMa vs MaNGA	0.051 ± 0.083	0.019 ± 0.077	-0.03 ± 0.125
CALMa vs SDSS	-0.020 ± 0.092	0.007 ± 0.078	0.04 ± 0.108

Table 3.3: Relative difference between the EWs ratios (in dex) predicted by ANN_R and the true values. Two training sample are used for training: CaLMA and SDSS, and three for testing: SDSS, CALIFA and MaNGA.

3.4.4 The 5max method in practice

A simple test to confirm the capability okf the *5max* method for retrieving the redshift of the object is to verify whether the average redshift over the five configuration is far from the true redshift. Normally, we would predict the EWs only in the redshift within the PDF of photo-z before applying the *5max*, but let us assume we do not have any information regarding the redshift of the object. Then, we have to calculate the EWs in all the redshift from 0 to 0.35 inside the grid and pick only the five redshifts that maximize their sum. Fig. 3.8 shows this scenario where points are color-coded with the spectroscopic redshift. For emission line galaxies ($\sum EW_i > 20$ Å), this method is able to obtain the redshift of the object with high precision; what is more, the redshift is not needed as an input. Nevertheless, the *5max* is not able to retrieve the redshift of the object when galaxies have low emission. The set of redshifts that maximizes the sum of the EWs is

largely uncertain and consequently we do need the PDFs to constrain the redshift value.

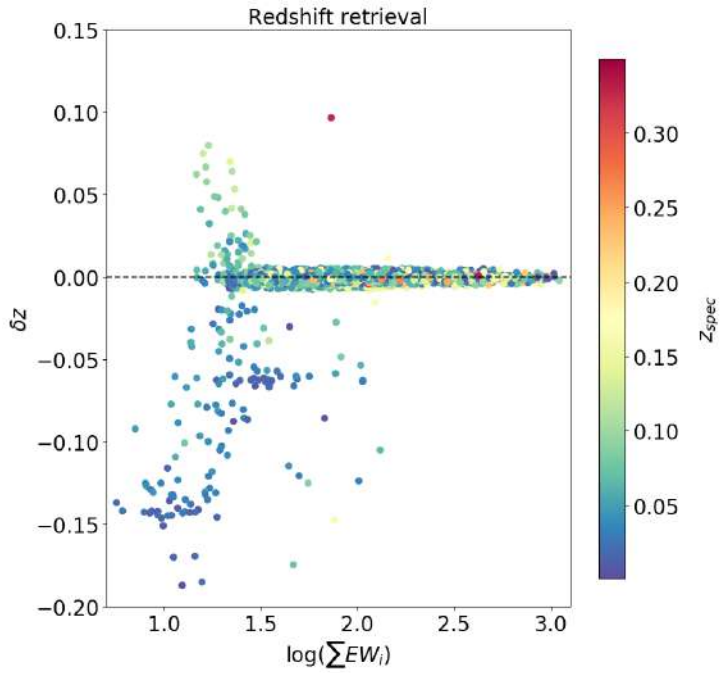


Figure 3.8: δz obtained from the difference between the spectroscopic redshift and the median redshift in the *5max* setting in function of the sum of the EWs provided in the SDSS catalog for a total of 10000 galaxies. Points are color-coded with the spectroscopic redshift.

3.4.5 Dependency on the EW and redshift uncertainty

In order to explore the limitation of the model as a function of the redshift uncertainty and the EW of each one of the emission lines, we assembled galaxies in bins by the EW provided in the SDSS catalog and computed the ratio (R) between the predicted and observed EW. Each bin contains 500 galaxies in the interval $10^{\gamma} < EW_{SDSS} < 10^{\gamma+0.1}$ with γ ranging from 0.8 to 2.5 for Hα, from 0.8 to 2.2 for [O III], from 0.8 to 1.8 for Hβ and from 0.8 to 1.8 for [N II]. As we observe in Fig. 3.9, Hα is clearly more affected by the *5max* strategy when $EW(\mathrm{H}\alpha) \leq 10^{1.2}$ Å. Independently of the redshift uncertainty, the ANN$_R$ trained with the CALMa set has more difficulties to constrain the [N II] line underestimating its value as the EW increases. It also presents more dispersion, which is an indication that high

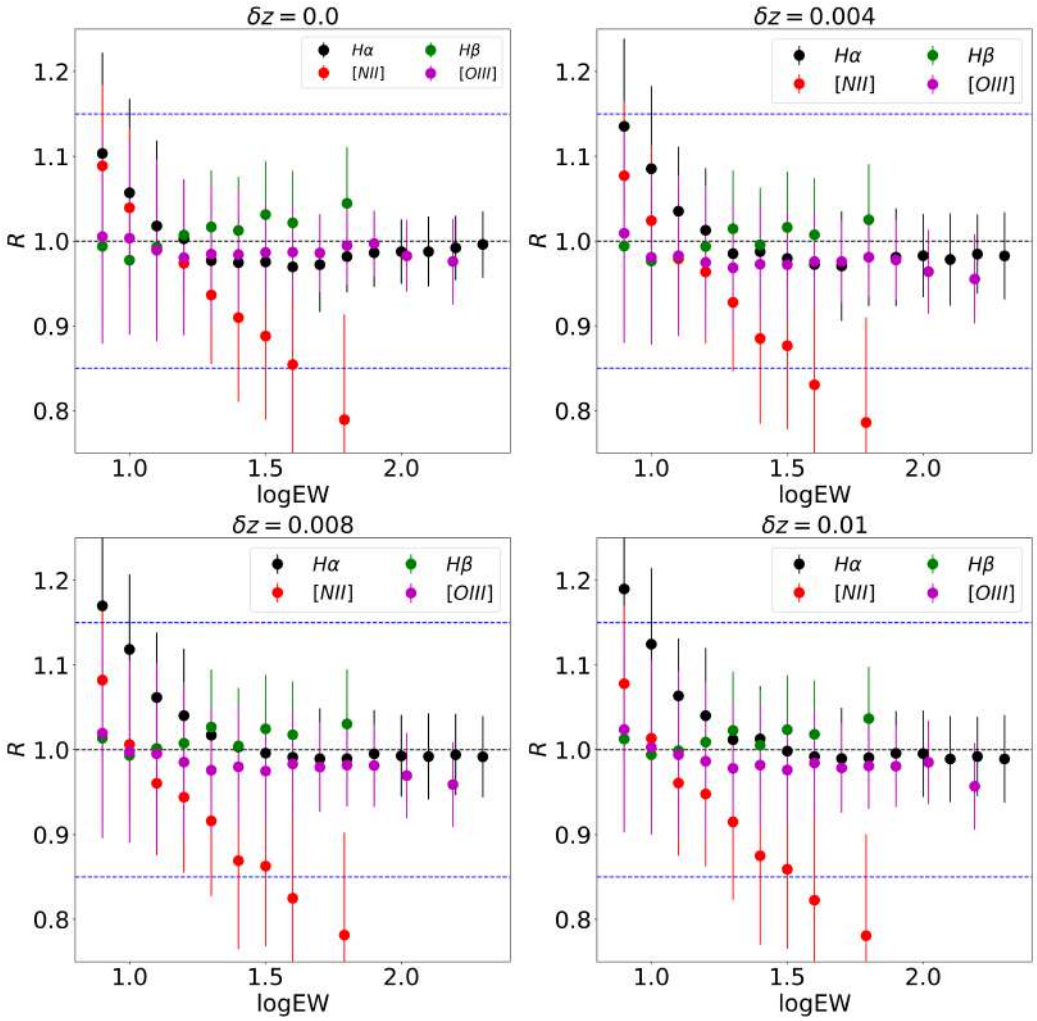


Figure 3.9: Each point represents the median ratio between the predicted and the observed SDSS EWs and bars indicate the mean absolute deviation. Each bin contains 500 galaxies in the interval $10^{\gamma} < EW_{SDSS} < 10^{\gamma+0.1}$ with γ ranging from 0.8 to 2.5 for Hα, from 0.8 to 2.2 for [O III], from 0.8 to 1.8 for Hβ and from 0.8 to 1.8 for [N II]. From left to right and top to bottom we increase the uncertainty in the redshift. Dashed blue lines point to a ratio of 1.15 and 0.85 respectively. Dash black line represent zero bias between the predicted and observed EWs.

values of the [N II] line implies a higher number of galaxies hosting an AGN. Nonetheless, we are able to constrain the EW of galaxies with a bias less than 10 % for most of the lines – even with high uncertainty in the redshift.

3.4.6 EW limit

The minimum EW measurable in a photometry system using a traditional method depends only on the S/N of the photometry and the effective width of filters in the system. Let us assume that an emission line falls within one filter (f_i) and we know with high precision the redshift of the object. The EW of an emission line can be computed assuming the line is infinitely thin, as:

$$EW = \Delta'(\lambda_z)(Q - 1), \tag{3.3}$$

where Δ' is the effective width of filter f_i and Q is the ratio between the flux with and without emission line see (see Pascual et al. 2007, for details) or simply:

$$Q = 10^{-(m_{AB}^{obs} - m_{AB}^{cont})/2.5}, \tag{3.4}$$

in AB magnitudes. Then, if we are able to estimate the flux of the stellar continuum in the filter tracing the emission line, obtaining the EW is straightforward. The S/N of such line can be expressed in terms of Q and the S/N of the photometry in the filter f_i through the following equation:

$$S/N_{EW} = \frac{Q - 1}{Q} S/N_{phot}. \tag{3.5}$$

The minimum EW measurable can be written as:

$$EW_{min} = \frac{\Delta'}{S/N_{phot} - 1}. \tag{3.6}$$

For $S/N_{phot} = 10$ only lines with EW greater that 16.1 Å can be measured in a filter width of 145 Å.

In Fig. 3.10, we determine the relation between the S/N of each line obtained with the ANN in function of the S/N of the photometry. As before, we assume no errors in the redshift of the objects. We analyze here the same galaxies used in the previous section in order to study the dependence with the EW. Each color represents the average S/N obtained in the line for 500 SDSS galaxies with the same EW. The red dashed line follows Eq. 3.5 for $EW = 10$ Å, which is the

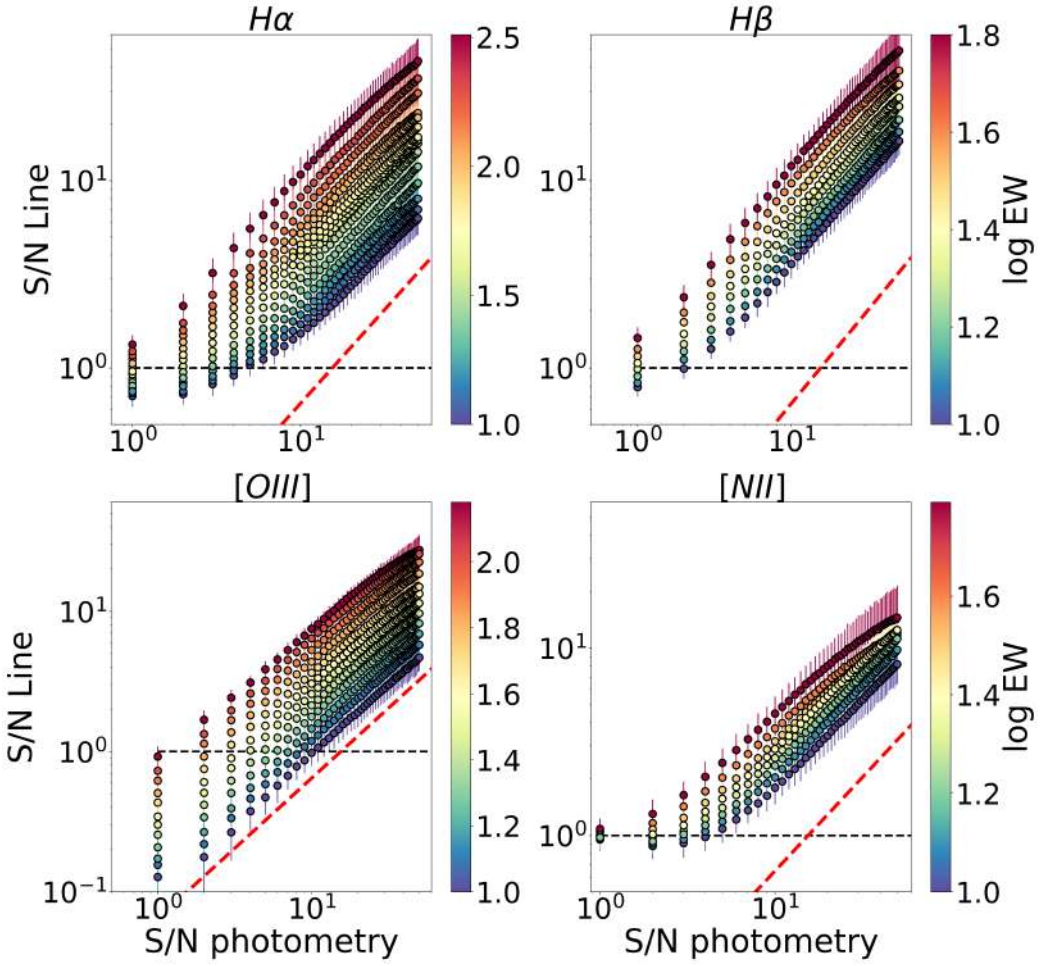


Figure 3.10: Predicted S/N of Hα, Hβ, [O III] and [N II] lines in function of the S/N in the photometry. For a given S/N in the photometry, each point represent the mean S/N obtained in the line for 500 SDSS galaxies in the interval (color-coded) $\gamma < \log EW_{SDSS} < \gamma + 0.1$ with γ ranging from 0.8 to 2.5 for Hα, from 0.8 to 2.2 for [O III], from 0.8 to 1.8 for Hβ and from 0.8 to 1.8 for [N II]. Errors bars indicate the mean absolute deviation. Dashed red line represents Eq. 3.5 for EW = 10 Å.

lowest EW bin considered in the simulations. All the lines estimated with the ANN can be measured with a precision higher than a method based on the contrast between the emission line flux and the stellar continuum.

Hβ is the line that can be predicted with the highest S/N for the same EW, with even better precision than Hα. This is not surprising since the algorithm has found the implicit relation between Hα and Hβ constrained by the Balmer series and the amount of interstellar dust. Therefore, an EW in Hβ of 10 Å, which corresponds on average to an EW in Hα of about 30 Å, is measured with the same S/N. More complex relations, such as the one between Hα and [N II] has also been found, but we observe a flattening of the S/N of the [N II] line for the highest EW with an increase in the scatter. This regime is populated with more AGN-like galaxies and consequently it is more difficult to constrain it with the CALMa set. This finding agrees with the behaviour observed in Fig. 3.9, where higher values of [N II] are systematically underestimated. Finally, the [O III] line is generally more difficult to constrain as we obtain lower S/N. Nevertheless, it can be recovered with better precision than a method based only on the photometry contrast.

To sum up, with an ANN one can measure a EW of 10 Å in Hα, Hβ, [N II], and [O III] lines with a S/N in the photometry of 5, 1.5, 3.5, and, 10 respectively. However, methods based on the photometry contrast need for the same EW a S/N in the photometry of at least 15.5. These facts illustrate once again the capability of machine learning algorithms to go beyond in precision and accuracy respect to traditional methods when large amount of data sets are available.

3.5 Comparison between miniJPAS and SDSS

In this section, we analyze and compare the data from the SDSS survey that has also been observed with miniJPAS in the AEGIS field. First, we describe the miniJPAS survey in section 3.5.1. We analyze and compare the properties of galaxies in terms of their emission lines in section 3.5.2.

3.5.1 The miniJPAS survey

The miniJPAS survey (Bonoli et al. 2021) is the result of the J-PAS-Pathfinder observation phase carried out with the 2.55 m telescope (T250) at the Observatorio Astrofísico de Javalambre in Teruel (Spain). the miniJPAS survey was conducted with the Pathfinder camera, the first instrument installed in the T250 before the arrival of the Javalambre Panoramic Camera (JPCam, Cenarro et al. 2019; Taylor

et al. 2014; Marin-Franch et al. 2015). The JPAS-Pathfinder instrument is a single CCD direct imager ($9.2k \times 9.2k$, $10\mu m$ pixel) located at the center of the T250 FoV with a pixel scale of 0.23 arcsec pix^{-1}, that is vignetted on its periphery, providing an effective FoV of 0.27 deg^2. The miniJPAS data includes four pointings of 1 deg^2 along the Extended Groth Strip (called the AEGIS field). We use the same photometric system of J-PAS. Thus, AEGIS was observed with 56 narrow band filters covering from $\sim$ 3400 to $\sim$ 9400 Å. Observations in the four broad bands (u_{JPAS}, and SDSS g, r, and i) were also taken. More than 60000 objects were detected in the r band, allowing to build a complete sample of extended sources up to $r \leq 22.7$ (AB). A detailed description of the survey is in Bonoli et al. (2021). The data is accessible and open to the community through the web page of the survey[6].

3.5.2 The miniJPAS versus SDSS

For this comparison, we selected all galaxies observed with SDSS and miniJPAS with redshift below $z \leq 0.35$ and a minimum average S/N of 20 in J-PAS narrow band filters. By a visual inspection we get rid of all quasars in the sample. We ended up with a total of 89 objects. Whenever photometry measurements are lacking or the S/N in a particular filter is below 2.5, we replace it with the best-fit obtained from the stellar population analysis of the galaxy, as we discuss in section 3.3.5. For this comparison, we employ `BaySeAGal` (Amorim in prep), a Bayesian parametric approach which assumes a tau-delayed star formation model for the star-formation history.

Generally, galaxy properties vary within the galaxy: the distribution of the gas, its temperature and its density, the distribution of interstellar dust or the stellar populations change as a function of the position in the galaxy (González Delgado et al. 2015). Consequently, if the SFR of a galaxy were higher in the outer parts, the galaxy would look younger in the integrated spectrum than in the central part. Similarly, the AGN of a galaxy would not leave the same imprint in the spectrum if the integrated areas covered regions dominated by other ionization mechanisms. Therefore, ideally, it would be optimal to analyse the same region in both surveys, which implies integrating over the same area. However, the aperture corresponding to the 3 arcsec fiber of SDSS is not sufficiently large to ensure that the point spread function (PSF) in miniJPAS observations is not affecting the photometry in the filters where the seeing is worse. For this reason, we make use of

[6] `http://www.j-pas.org`

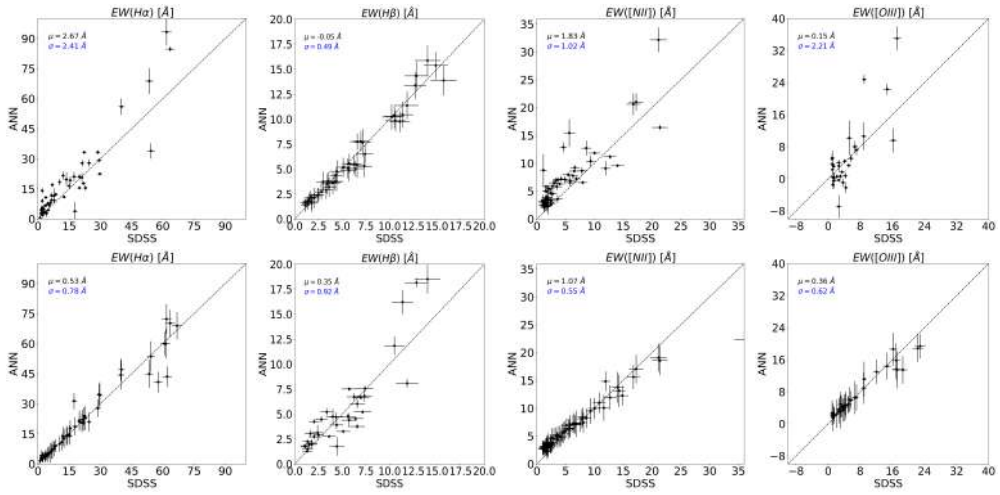


Figure 3.11: Comparison between the EWs of Hα, [N II], Hβ and [O III] measured in the SDSS spectra and the predictions made by the ANN on miniJPAS data using the `MAG PSFCOR` (top panel) and synthetic J-PAS magnitudes obtained from the SDSS spectra (bottom panel). Black and blue numbers are the median and the median absolute deviation of the difference. Dashed black line is line with slope one.

the `MAG_PSFCOR` photometry which corrects each magnitude individually by considering the light profile of the galaxy and the PSF for each filter (Molino et al. 2014, 2019). As a consequence, the integrated area varies from galaxy to galaxy, going from 2 to 7 arcsec, and should be taken into account to interpret fairly this comparison. Although the ANN$_R$ only use colors as inputs, we scale the SDSS spectrum to match the rSDSS miniJPAS magnitude in each galaxy for a visual inspection.

Figure 3.11 shows the EWs obtained by the ANN$_R$ on miniJPAS photometric data (column 1) and on the synthetic J-PAS magnitudes obtained after convolving SDSS spectra with J-PAS filters (column 2) and assuming an average S/N of 20. We compare those values with the EWs derived as a result of fitting a Gaussian function to each one of the emission lines in the spectrum (x-axis). We do not include in this comparison the emission lines where EWs are below 1 Å, which indeed are compatible with zero. The number of galaxies in each row are from top to bottom 57, 37, 64, and 31. We find an excellent agreement when it comes to SDSS synthetic magnitudes, which is in line with the simulations performed with the SDSS dataset. We also find a remarkable correlation in Hα, Hβ and [N II] with J-PAS magnitudes, but we obtain in most of the cases higher values with an increase in the dispersion (see median and MAD in Fig. 3.11). The agreement

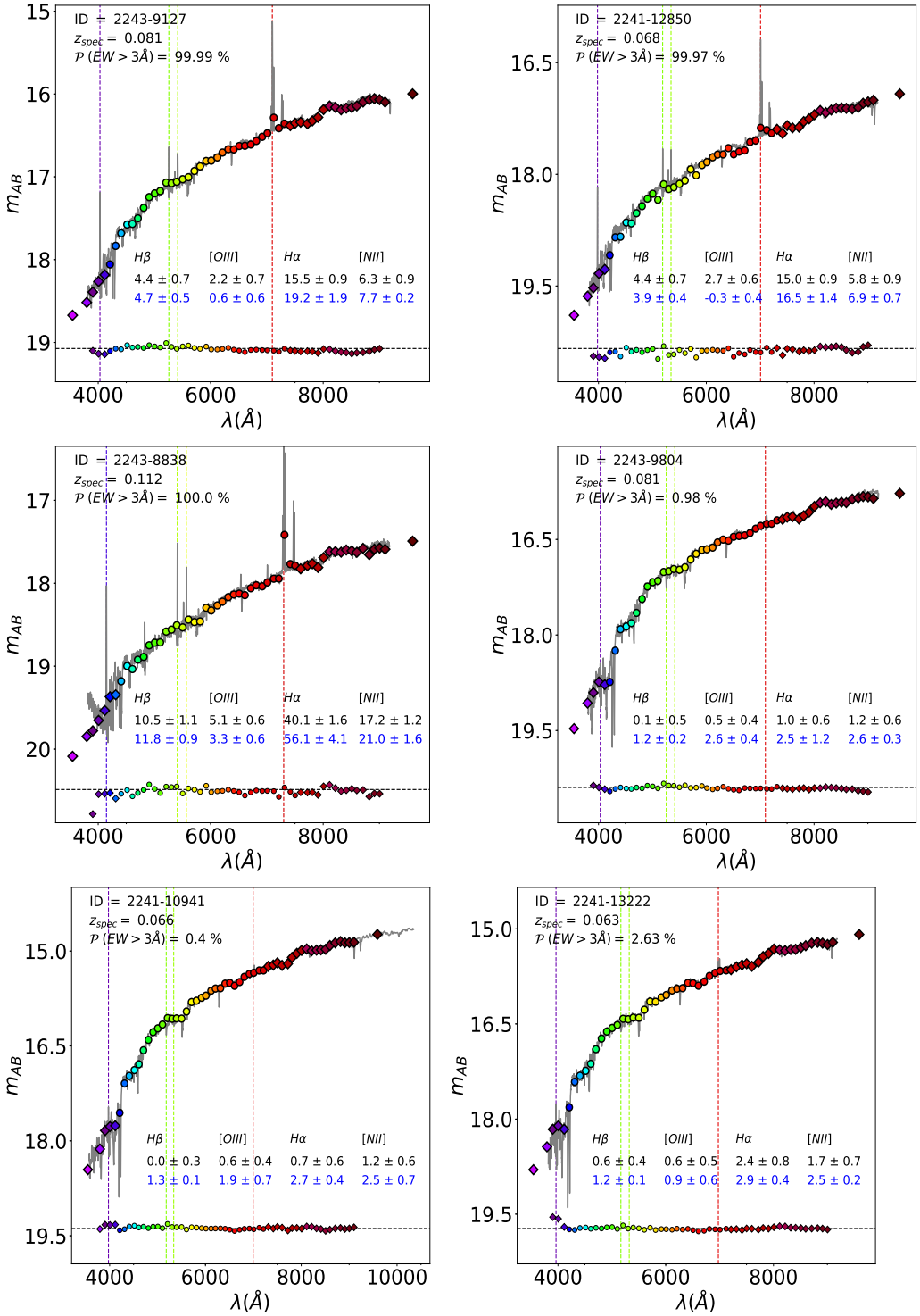
is less favourable for the [O III] line. Nevertheless, we should bear in mind the limited number of galaxies used here in order to avoid drawing any conclusion that may not be supported from a statistical point of view. Instead, we consider more appropriate to analyze the origin of these discrepancies by visually examining each object.

In Fig. 3.12, we show several galaxies analyzed in this comparison. We re-scaled the SDSS spectrum to match the rSDSS J-PAS magnitude. We compare the values of the EWs measured in the SDSS spectrum (black) with the values predicted by the ANN_R (blue) for each one of these galaxies. On the bottom part, we show in each filter the difference between miniJPAS data and SDSS synthetic photometry, which can certainly help to shed light on the origin of the discrepancies.

The third first images in Fig. 3.12, are emission line galaxies where the agreement in most of the EWs is remarkable. Although ANNs are often difficult to interpret, it is evident after a visual inspection that the filters capturing the fluxes of the emission lines are the most relevant in determining the values of the EWs. The excess in the flux of Hα in galaxy 2243-8838 explains the increase in its EW respect to what it is obtained from a direct measurement in the spectrum or with the synthetic fluxes by means of the ANN_R. In the same vein, the drop in the flux observed in the [O III] line in galaxy 2241-12850 clarifies the differences found in the EW. Second-order terms include the relation between emission lines (Balmer decrement or recombination lines) and the colors of galaxies. Certainly, the excess in the flux of Hβ in galaxy 2243-9127 does not only increase the value of such line, but it also contributes to the enlargement of the EW of Hα.

2243-9804, 2241-10941, and 2241-13222 are early-type galaxies (ETGs) where the differences between miniJPAS data and SDSS synthetic fluxes are negligible. The ANN_C estimates very low probability for these galaxies to have any emission line with a EW greater than 3 Å, which is in agreement with the measurements performed in SDSS spectra. As we discussed in section 3.4.5 the ANN_R tends to overestimate the EWs in the regime of low emission and consequently a zero level bias appears in these galaxies. Nonetheless, for many of these lines the values are compatible with the uncertainty and never overcome the 3 Å limit.

Finally, the fluxes observed by miniJPAS and SDSS present evident differences in the blue part of the spectrum in the last three galaxies of Fig. 3.12. Most probably, the integrated areas in miniJPAS for 2243-9209 and 2406-4867 galaxies are capturing regions with greater populations of young stars. Such populations raise the number of ionizing photons being responsible for the increase in the EWs of emission lines that we observe. The opposite effect occurs in galaxy


ID = 2243-9127
z_{spec} = 0.081
$\mathcal{P}$ ($EW > 3Å$) = 99.99 %
Hβ [OIII] Hα [NII]
4.4 ± 0.7 2.2 ± 0.7 15.5 ± 0.9 6.3 ± 0.9
4.7 ± 0.5 0.6 ± 0.6 19.2 ± 1.9 7.7 ± 0.2
ID = 2241-12850
z_{spec} = 0.068
$\mathcal{P}$ ($EW > 3Å$) = 99.97 %
Hβ [OIII] Hα [NII]
4.4 ± 0.7 2.7 ± 0.6 15.0 ± 0.9 5.8 ± 0.9
3.9 ± 0.4 -0.3 ± 0.4 16.5 ± 1.4 6.9 ± 0.7
ID = 2243-8838
z_{spec} = 0.112
$\mathcal{P}$ ($EW > 3Å$) = 100.0 %
Hβ [OIII] Hα [NII]
10.5 ± 1.1 5.1 ± 0.6 40.1 ± 1.6 17.2 ± 1.2
11.8 ± 0.9 3.3 ± 0.6 56.1 ± 4.1 21.0 ± 1.6
ID = 2243-9804
z_{spec} = 0.081
$\mathcal{P}$ ($EW > 3Å$) = 0.98 %
Hβ [OIII] Hα [NII]
0.1 ± 0.5 0.5 ± 0.4 1.0 ± 0.6 1.2 ± 0.6
1.2 ± 0.2 2.6 ± 0.4 2.5 ± 1.2 2.6 ± 0.3
ID = 2241-10941
z_{spec} = 0.066
$\mathcal{P}$ ($EW > 3Å$) = 0.4 %
Hβ [OIII] Hα [NII]
0.0 ± 0.3 0.6 ± 0.4 0.7 ± 0.6 1.2 ± 0.6
1.3 ± 0.1 1.9 ± 0.7 2.7 ± 0.4 2.5 ± 0.7
ID = 2241-13222
z_{spec} = 0.063
$\mathcal{P}$ ($EW > 3Å$) = 2.63 %
Hβ [OIII] Hα [NII]
0.6 ± 0.4 0.6 ± 0.5 2.4 ± 0.8 1.7 ± 0.7
1.2 ± 0.1 0.9 ± 0.6 2.9 ± 0.4 2.5 ± 0.2
m_{AB}
$\lambda(Å)$

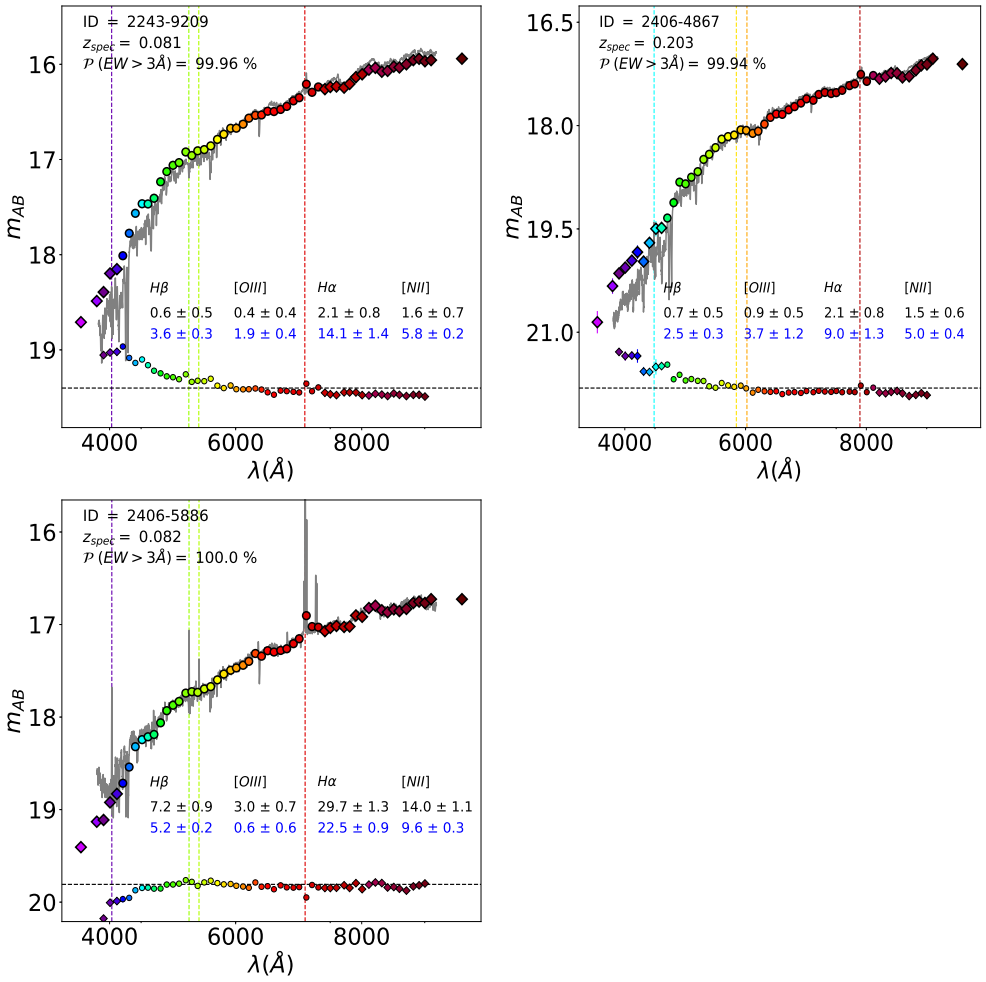


Figure 3.12: Examples of J-PAS galaxies in the AEGIS field with SDSS spectrum. The SDSS spectrum is re-scaled to match the rSDSS J-PAS magnitude. Diamonds correspond to the filters not used by the ANN. Blue and black numbers show, respectively, the predictions made by the ANN_R on the EWs and the values measured in the SDSS spectrum. On the top-left part of the plot, we indicate the J-PAS ID of the object, its redshift and the prediction of the ANN_C for EW_{min} = 3 Å. At the bottom, we show the difference in magnitude between the synthetic fluxes obtained from SDSS spectra and miniJPAS data. Dashed lines mark from left to right the position of [O II], Hβ, [O III], and Hα emission lines.

2406-5886, the galaxy looks redder with miniJPAS data and the flux in Hα is less intense. Therefore, the predictions of the ANN$_R$ in the EWs are below the values measured in the SDSS spectrum.

To sum up, despite of the fact that this comparison suffer from several difficulties and it would need many more galaxies to be statistically robust, results are coherent with the simulations presented in section 3.4 and lay the foundations to better understand and interpret the whole sample of galaxies observed in the AEGIS field, which we will analyze in the following chapter.

3.6 Summary and conclusions

We have developed a new method based on ANNs to measure and detect emission lines in J-PAS up to $z = 0.35$. We can classify galaxies according to the EWs of the emission lines, even with high uncertainty in the redshift. This will allow us to better study the density function of emitting-line galaxies in J-PAS.

Using the synthetic photometry of CALIFA, MaNGA or SDSS spectra, we trained an ANN$_R$ to estimate the EWs of Hα, Hβ, [N II], and [O III] lines. We present two training samples to undertake this task.

First, we trained the ANN$_R$ with only synthetic J-spectra from MaNGA and CALIFA surveys and we used SDSS to evaluate the performance of the model. The lack of a large enough number of AGN-like synthetic J-spectra leads to a saturation of [N II]/Hα and [O III]/Hβ ratios at high values, which compromises the ability of the model to deal with galaxies where the main ionization mechanism is not dominated by star-formation processes. Nevertheless, we are able to constrain those ratios within 0.078 and 0.092 dex. Furthermore, we are able to reach 0.070 and 0.087 dex, respectively, if one considers only star-forming galaxies. While a method based on the photometry contrast need for an EW of 10 Å a S/N in the photometry of at least 15.5, the ANN can measure the same EW in Hα, Hβ, [N II], and [O III] lines with a S/N in the photometry of 5, 1.5, 3.5, and, 10, respectively.

Second, we trained the ANN$_R$ with SDSS galaxies and we revealed the importance of testing the model with data coming from different surveys. Otherwise, the performance of the model may be overestimated. While the SDSS training set scores very high with SDSS testing sample, the performance worsens when we compare it with the MaNGA or CALIFA test sample.

Finally, we estimate the EWs of a set of galaxies observed both in SDSS and miniJPAS. We compare the performance of ANN$_R$ in the synthetic SDSS fluxes with the performance in the fluxes measured by miniJPAS. Despite the difficulty

of comparing data from different surveys in equal terms, we reached an overall agreement. We argue that the origin of the discrepancies might be attributed to differences between the integration areas in miniJPAS and SDSS and/or photometry artefacts that appear as a result of the PSF. Many more data would be needed to be conclusive.

In this chapter, our model is limited to redshift below $z = 0.35$ in order to ensure Hα line is measurable with the J-PAS filter system. However, J-PAS will be able to detect galaxies up to $z \sim 1$. Other emission lines, such as the $[OII]\lambda\lambda$ 3726,3729 doublet, are visible in the optical range up to redshift $z < 1.6$ and has been used to trace the star formation (Kewley et al. 2004; Sobral et al. 2012). Such line might be include in a future version of the model.

Another important limitation of our work lies on the unavoidable gap between simulations and observations. The data reduction process is full of assumptions and limitations than can impact the photometry and the error estimates provided in the final catalogues. For instance, the JPCam will not take pictures of the sky with its 56 filters simultaneously. Instead, the observations will be carried out in trays of 14 CCDs, thus different observational conditions will be present in the SED of individual galaxies. One way to fill this gap might be to used transfers learning. As as soon as J-PAS begins to observe the sky, we will have J-PAS data for galaxies already observed by spectroscopic surveys. Therefore, the model might be retrain with actual observations. An ultimate version of our models should take into account those facts and build a more sophisticated and complete training sample so as to be able to overcome the limitations and inaccuracies mentioned and fully exploit the potentiality of J-PAS. Our main conclusions are summarized below:

- The ANN_C can classify galaxies according to the EWs of the emission lines beyond the contrast that can directly be measured with sufficient significance in J-PAS ($\sim$ 16 Å) and in the case of high uncertainty in the redshift as well.

- The ANN_R trained with the CALMa set can estimate the EWs of Hα, Hβ, [N II], and [O III] in SDSS galaxies with a relative standard deviation of 8.4 %, 13.7 %, 14.8 %, and 15.7 %, respectively. The Hα, Hβ, [N II], and [O III] lines present a relative bias of 0.03 %, 5.0 %, 4.8 %, and −6.4 % respectively. For a S/N of 3, the minimum EW measurable in Hα, Hβ, [O III] and [N II] lines is 18, 6, 40, and, 13 Å, respectively.

- The [N II]/Hα is constrained within 0.092 dex and a bias of −0.02 dex and the [O III] Hβ ratio with no bias and a dispersion of 0.078 dex in SDSS

galaxies. The O 3N 2 is recovered within 0.108 dex and a bias of 0.04 dex.

- We found an overall correlation between miniJPAS and SDSS galaxies in the EW of Hα, Hβ and, [N II] lines. The correlation in the EW of [O III] is less strong. More data will be needed to unveil the origin of such discrepancy. Certainly, the problems associated with the integrated areas play an important role.

Chapter 4

Identification and characterization of the emission line galaxies down to $z < 0.35$ in the AEGIS field

This chapter is based on the publication:

"The miniJPAS survey: Identification and characterization of the emission line galaxies down to $z < 0.35$ in the AEGIS field"
by G. Martínez Solaeche, R. M. González Delgado, R. García-Benito et. al

Published in A&A, 661, A99 (2022)
https://doi.org/10.1051/0004-6361/202142812

4.1 Introduction

The Hα emission line is an excellent tracer for estimating the current star formation rate (SFR) in galaxies because it is less affected by dust extinction than UV light (Kennicutt 1998; Garn et al. 2010; Oteo et al. 2015; Catalán-Torrecilla et al. 2015). The Hα line can be observed in the optical range up to $z \sim 0.4$. Thus, it is very useful for the identification of emission line galaxies (ELGs) in spectroscopic and photometric surveys. The detection of other emission lines, such as [O III] $\lambda\lambda$4959, 5007 Å and the [N II] $\lambda\lambda$6548, 6584 Å doublets[1] , is crucial to determine the main ionization mechanism of ELGs (see, e.g., Cid Fernandes et al. 2011; Belfiore et al. 2016; Sánchez et al. 2018; Lacerda et al. 2020; Kalinova et al. 2021). Diagrams such as the WHAN (EW(Hα) vs. [NII]/Hα) (Cid Fernandes et al. 2011) or the BPT (Baldwin et al. 1981) (e.g., [OIII]/Hβ vs. [NII]/Hα) can differentiate galaxies in which the gas is ionized by young stars or by an active galactic nucleus (AGN), from low ionization nuclear emission regions (LINERs, Heckman 1980), or extended low-ionization emission lines (see, e.g., Lacerda et al. 2018), in which the ionization might be attributed to old and hot stars. Furthermore, the characterization of the galaxy populations through the SFR and its correlation with other galaxy properties, such as stellar mass, colors, ages, metallicity, and neutral gas content (Kewley et al. 2019; Förster Schreiber & Wuyts 2020), is essential to obtain insight into the formation and evolution of galaxies.

Galaxies grow in mass mainly through star formation, which is fed by gas accretion from the cosmic web. While massive galaxies undergo a larger fraction of their star formation at early times, less massive galaxies are still forming stars at a high rate today. The star formation main sequence (SFMS), a tight quasi-linear relation between stellar mass, ($M_\star$), and the SFR in log scale (Zahid et al. 2012; Renzini & Peng 2015; Cano-Díaz et al. 2016; Duarte Puertas et al. 2017; Belfiore et al. 2018; Boogaard et al. 2018; Sánchez et al. 2019; Cano-Díaz et al. 2019; Shin et al. 2021; Vilella-Rojo et al. 2021), can reveal indications how this process takes place. Galaxies that are undergoing a starburst, for instance, lie above the SFMS, while galaxies that have already quenched their star formation lie below this relation.

The SFMS and its evolution with redshift are expected outcomes of hydrodynamical models. The currently best cosmological hydrodynamical simulations of galaxy formation such as Illustris (Sparre et al. 2015) or EAGLE (Furlong et al.

[1] In the remaining of the chapter, [O III] λ5007 and [N II] λ6584 are denoted [O III] and [N II], respectively.

2015) predict a slope near unity. Semi-analytical models favor a sublinear slope that is generally higher than 0.8. For instance, Dutton et al. (2010) predicted a slope of 0.96 for galaxies with stellar masses between 10^9 and $10^{11} M_\odot$. However, Mitchell et al. (2014) used GALFORM and retrieved a slope of 0.87 at $z = 0.1$.

The slope of the SFMS in observations ranges from 0.6 to 1, depending on the data, the SFR tracer, and method used (see, e.g., the study of Speagle et al. 2014, and references therein). The discrepancies found by different studies are expected. On the one hand, spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) have aperture effects that can cause an underestimation of the total SFR within the galaxy (Duarte Puertas et al. 2017). On the other hand, the SFR derived from photometric surveys throughout Hα measurements needs to be corrected for the [N II] and dust extinction, which become the main sources of uncertainty.

The definition of the SFMS itself might also lead to significant differences between different works, even though they all trace the SFR through the Hα line. Some authors (e.g., Vilella-Rojo et al. (2021, $z \leq 0.017$) or Shin et al. (2021, $z \sim 0.07 - 0.5$)) relied on color-color diagrams. Others selected star-forming (SF) galaxies based on the BPT diagrams with a cut in the equivalent width (EW) of Hα or Hβ. For example, Cano-Díaz et al. (2016, $0.005 \leq z \leq 0.03$) imposed a minimum EW in Hα of 6 Å while Duarte Puertas et al. (2017, $0.005 \leq z \leq 0.22$) used instead 3 Å and Zahid et al. (2012, $z = 0.07, 0.8$ and 2.26) adopted a EW of 4 Å in Hβ. In addition, the SFMS has also been defined as the ridge line in the M$\star$-N-SFR- plane where N account for the number of galaxies in every M$\star$-SFR bin ($0.02 \leq z \leq 0.085$, Renzini & Peng 2015).

In essence, there is no unique and homogeneous definition of the galaxies that belong to the SFMS. Furthermore, any dividing line between star-forming and quiescent galaxies affects the analysis of the SFMS because it includes or excludes some of the galaxies in the the so-called 'green valley' (GV), that is, galaxies that are in transition and are interpreted as a crossroads in galaxy evolution (see, e.g., Mendez et al. 2011; Gonçalves et al. 2012; Schawinski et al. 2014; Díaz-García et al. 2019a). Sánchez et al. ($0.03 \leq z \leq 0.2$, 2019) attributed the constancy of the SFMS slope across galaxy mass to the selection criterion (based on sSFR cut). There is no drop in the SFR at high masses. In the same vein, Belfiore et al. ($0.03 \leq z \leq 0.15$, 2018), who also used the Hα line as an SFR tracer, found that the flattening in the slope of the SFMS only occurs if galaxies with quiescent central regions (cLIERs) are included in the fit.

In addition, the detection limit and particularities of each study might lead to a specific bias in the selection criteria. For instance, a photometric survey that

selects ELGs based on a minimum contrast would be limited to the minimum EW that can be measured and would therefore be biased toward highly actively SF galaxies. As a consequence, it produces an increase in normalization constant and a shallower slope (Khostovan et al. 2021). Finally, the minimization method employed in the fitting takes the uncertainties into account in different ways. It might therefore also have an impact on the shape of the SFMS.

Another important aspect that helps to understand how galaxies assemble their mass throughout cosmic time is estimating the intrinsic scatter of galaxies in the SFMS. It is expected that low-mass galaxies are more sensitive to stochastic events such as starbursts or feedback from supernovae. Theoretical simulations (Hopkins et al. 2014; Domínguez et al. 2015; Matthee & Schaye 2019) and observations (Salim et al. 2007; Emami et al. 2019; Boogaard et al. 2018; Santos et al. 2020) have both found an increase in scatter for low-mass galaxies ($< 10^9 M_\odot$).

Other studies (Willett et al. 2015; Davies et al. 2019) found that the dispersion along the SFMS follows a U-shaped distribution, meaning that galaxies with high and low stellar masses scatter more from the SFMS. Interestingly, the U-shape depends on the way the SFMS is defined. While selecting SF galaxies based on $u - r$ colors or morphology causes the SFMS to have higher scatter for galaxies at high mass, a selection based on a minimum sSFR, which is equivalent to a minimum EW in Hα, produces a decrease in scatter as the mass of the galaxy increases (see, e.g., Davies et al. 2019).

It has been proven by the analysis of stellar populations within galaxies through stellar continuum spectral energy distribution (SED) fitting that the SFMS holds true at high redshift with an increase in the global SFRs of galaxies (Daddi et al. 2004; Oliver et al. 2010; Karim et al. 2011; Ilbert et al. 2015; Schreiber et al. 2015; Tasca et al. 2015; Rodighiero et al. 2011). In terms of the SFR density ($\rho_{\rm SFR}$), the Universe reached a peak at $\sim$ 3 Gyr after the Big Bang, and it has been decreasing ever since (Madau & Dickinson 2014; Driver et al. 2018; López Fernández et al. 2018; Sánchez et al. 2019; Leja et al. 2019; Bellstedt et al. 2020). Through Hα measurements, astronomers are also able to measure $\rho_{\rm SFR}$ both in the nearby Universe and at intermediate redshift, which has confirmed this trend (Gallego et al. 1995; Ly et al. 2007; Shioya et al. 2008; Dale et al. 2010; Westra et al. 2010; Drake et al. 2013; Sobral et al. 2013; Gunawardhana et al. 2013; Sobral et al. 2015; Stroe & Sobral 2015; Van Sistine et al. 2016; Khostovan et al. 2020; Vilella-Rojo et al. 2021).

The incredible progress achieved in the past decades would not have been possible without the construction of large galaxy surveys. Multi-object spectroscopy (MOS) surveys such as the SDSS and the the Galaxy And Mass As-

sembly (GAMA; Driver et al. 2011) or integral field unit (IFU) surveys such as the Calar Alto Legacy Integral Field Area (CALIFA; Sánchez et al. 2012; García-Benito et al. 2015; Sánchez et al. 2016a) and the survey Mapping Nearby Galaxies at the Apache Point Observatory (MaNGA; Bundy et al. 2015; Law et al. 2015) provide a very detailed description of the optical SED of galaxies. However, they are partially biased through their preselection of samples, which is driven by some properties such as redshift, fluxes, or a galaxy size that is constrained to a particular range.

In contrast, narrowband photometric surveys such as HiZELS (Best et al. 2013; Sobral et al. 2013; Matthee et al. 2017), ALHAMBRA (Moles et al. 2008; Molino et al. 2014), DAWN (Coughlin et al. 2018), J-PLUS (Cenarro et al. 2019), S-PLUS (Mendes de Oliveira et al. 2019), the Deep and UDeep layers driven by the Subaru Strategic Program with the Hyper Suprime-Cam (HSC-SSP) (Hayashi et al. 2018, 2020), LAGER (Khostovan et al. 2020), or SHARDS (Pérez-González et al. 2013; Lumbreras-Calle et al. 2019), experience these effects to a lesser degree. In particular, narrowband photometric surveys are able to detect fainter objects than their spectroscopic counterpart at a fixed exposure time. Furthermore, they can fully observe galaxies whose light cannot be captured entirely by IFU-like surveys (see, e.g., Fig. 19 in Bonoli et al. 2021). However, their SED in the optical, infrared, or UV is limited by the number of filters and their width. More importantly, ELGs can only be detected in certain redshift intervals, which makes contamination from other sources more likely because the emission lines may be confused; for example, [O III] emitters may be detected as Hα emission line objects.

The special design of the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS, Benitez et al. 2014) enables overcoming some of the caveats for spectroscopic and traditional photometric surveys. J-PAS will play a crucial role in the upcoming years, which will be very competitive compared to the new generations of spectroscopic surveys such as DESI (DESI Collaboration et al. 2016), *Euclid* (Laureijs et al. 2011), or the WHT Enhanced Area Velocity Explorer-Stellar Population at intermediate redshift Survey (WEAVE-StePS; Costantin et al. 2019).

The unprecedented area that J-PAS will cover ($\sim$ 8000 deg^2 of the northern sky) is perhaps one of the main advantages compared to previous and current surveys. J-PAS will observe the sky with 56 bands: 54 narrowband filters in the optical range, plus two medium-band filters, one in the UV and another in the near-infrared. Separated by 100 Å, each narrowband filter has a width of $\sim$ 145 Å, which provides a resolving power of $R \sim 60$ (J-spectrum hereafter). These unique

characteristics make J-PAS an ideal survey for galaxy evolution studies (Bonoli et al. 2021), superseding the scientific impact achieved by other previous medium-band imaging surveys, such as ALHAMBRA ($R \sim 20$). The narrowband setup of J-PAS allows the detection and measurement of galaxies with emission lines in a continuous range in redshift within a nonsegregated area (Martínez-Solaeche et al. 2021). J-PAS observations will be carried out with the 2.55 m telescope (T250) at the Observatorio Astrofísico de Javalambre, a facility developed and operated by the Centro the Estudios de Física del Cosmos de Aragón (CEFCA, in Teruel, Spain) using JPCam, a wide-field 14 CCD-mosaic camera with a pixel scale of 0.2267 arcsec/px and an effective field of view (FoV) of ~ 4.6 deg^2 (Taylor et al. 2014; Marin-Franch et al. 2015; Bonoli et al. 2021).

The pathfinder camera of J-PAS started its observations using 60 optical bands in four fields of the sky that overlap with the All-wavelength Extended Groth Strip International survey (AEGIS; Davis et al. 2007), amounting to 1deg^2 with more than 60 000 objects[2]; hereafter, this is referred to as the miniJPAS survey (Bonoli et al. 2021). The pathfinder instrument used by the J-PAS collaboration is a single CCD direct imager ($9.2k \times 9.2k$, $10\mu m$ pixel) located at the center of the T250 FoV with a pixel scale of 0.23 arcsec pix^{-1}, vignetted on its periphery. This provides an effective FoV of 0.27 deg^2.

The goal of this chapter is to identify the ELG population in the AEGIS field and characterize them through their SFR and the stellar population properties. This work shows the potential of J-PAS data in this regard. We apply a method based on artificial neural networks (ANN) described in chapter 3 to obtain the EW of the main emission lines in the optical range: Hα, Hβ, [O III], and [N II]. Afterward, we analyze the main ionization mechanisms in galaxies through WHAN and BPT diagrams, and we compare the nebular properties of the gas with the properties of the stellar populations of their host galaxies derived in González Delgado et al. (2021). We characterize the SFR-M$_*$ relation derived from the flux of Hα, and we compute the cosmic evolution of $\rho_{\rm SFR}$ up to $z = 0.35$.

This chapter is organized as follows. In section 4.2 we present the galaxy sample taken from miniJPAS, which is the subject of this study. In section 4.3 we summarize the method we employed, which is based on chapter 3 and González Delgado et al. (2021). In section 4.4 we identify the ELG population by means of the EWs of the emission lines and their relations with the stellar population properties: stellar mass, intrinsic colors, luminosity-age, and so on. We derive the fraction of AGN, quiescent, and star-forming galaxies in miniJPAS. In section 4.5

[2]http://www.j-pas.org/

we characterize the star-forming galaxy population. We derive their SFR through Hα emission, and we fit the SFMS. In section 4.6 we discuss the implications of our results in detail and compare them with previous works. We derive the $\rho_{\rm SFR}$ up to $z = 0.35$. Finally, we provide the outlook for J-PAS in section 4.7, and we summarize in section 4.8. Throughout this chapter, we adopt a ΛCDM cosmology with $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_{\rm M} = 0.3$, and $\Omega_\Lambda = 0.7$. All magnitudes are presented in the AB system (Oke & Gunn 1983), and a Chabrier (2003) initial mass function (IMF) was employed.

4.2 Sample and data

The galaxy sample studied in this chapter is a subsample of the galaxies analyzed in González Delgado et al. (2021, see section 2.3). We selected all the objects detected in miniJPAS with a photometric redshift (photo-z) lower than 0.35, which is the highest redshift at which Hα can be observed in miniJPAS. The photo-z was estimated with the `JPHOTOZ` package developed by the photo-z team at CEFCA. This package is a customized version of the `LePhare` code (Arnouts & Ilbert 2011), which has a new set of stellar population synthesis galaxy templates that were optimized for the miniJPAS filter system (Hernán-Caballero et al. 2021). At the depth of miniJPAS (5σ limits between $\sim$ 21.5 and 22.5 mag for the narrowband filters and $\sim$ 24 mag for the broadband filters in a 3″ aperture), there are 17 500 galaxies per deg^2 with valid photo-z estimates (rSDSS $<$ 23), of which $\sim$ 4 200 have $|\Delta z| < 0.003$. The typical error for rSDSS $<$ 23 galaxies is $\sigma_{\rm NMAD} = 0.013$ with an outlier rate of $\eta = 0.39$. The target photo-z accuracy $\sigma_{\rm NMAD} = 0.003$ is achieved after imposing $odds > 0.82$ (see Hernán-Caballero et al. 2021, for details).

We imposed a maximum `CLASS_STAR` probability of 0.1, as defined in `SExtractor`, in order to select only extended sources. We discarded galaxies with an S/N lower than 1.8 in the filters to capture the flux of the emission lines. The estimates of the EWs with the ANN for galaxies with a very low S/N yield large errors. Therefore, these errors indicate the limit to which galaxies can be analyzed. For this reason, we favor a more conservative approach by setting a very low constraint on the S/N of the filters with which the flux of the emission lines is captured. Thus, we can exclude galaxies a posteriori if their EW predictions are not reliable. The magnitude limit cut of the sources was set at 22.5 mag in the rSDSS band. This is near the completeness limit for miniJPAS extended sources (Bonoli et al. 2021). Finally, the sample is composed of 2154 galaxies in total.

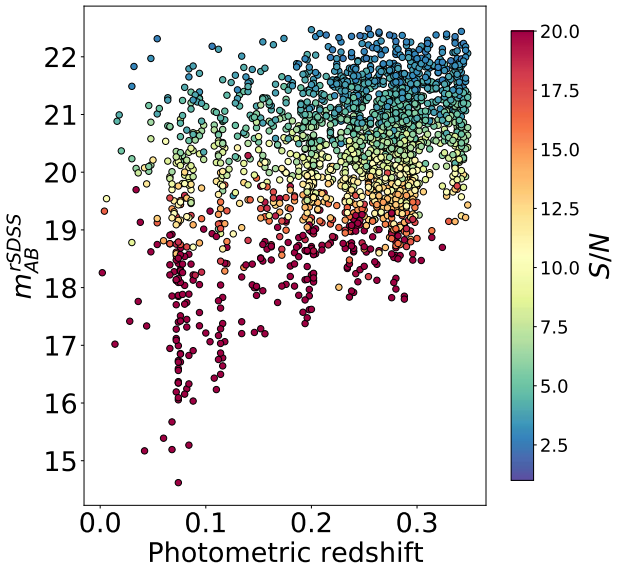


Figure 4.1: Relation between the apparent magnitude in the rSDSS band and redshift for all galaxies in the parent sample. We used the `MAG_AUTO` photometry. Dots are color-coded according to the median S/N of the J-PAS narrowband filters.

In Fig. 4.1 we show the relation between the apparent magnitude in the rSDSS band and the redshift for the galaxies in the parent sample. The color bar indicates the median S/N measured in the J-PAS narrowband filters. In this chapter, we made use of the `MAG_AUTO` photometry from the miniJPAS dual-mode catalog because it captures the entire light from the galaxy. Most of the galaxies in this sample ($\sim$ 68 %) are higher than 0.205 in redshift and have an S/N lower than 10.

In Fig. 4.2 we show some examples of galaxies in this sample at different redshift and magnitude bins. Emission lines such as Hα or [O III] are clearly visible in most of them. Some lines are captured by more than one filter (see, e.g., 2241-6186). This is caused by the overlapping adjacent filters, whose separation (100 Å) is smaller that their width ($\sim$ 145 Å).

4.3 Method

4.3.1 Artificial neural networks

This section is a summary of the previous chapter. Thus, the reader can skip it if he/she already read it.

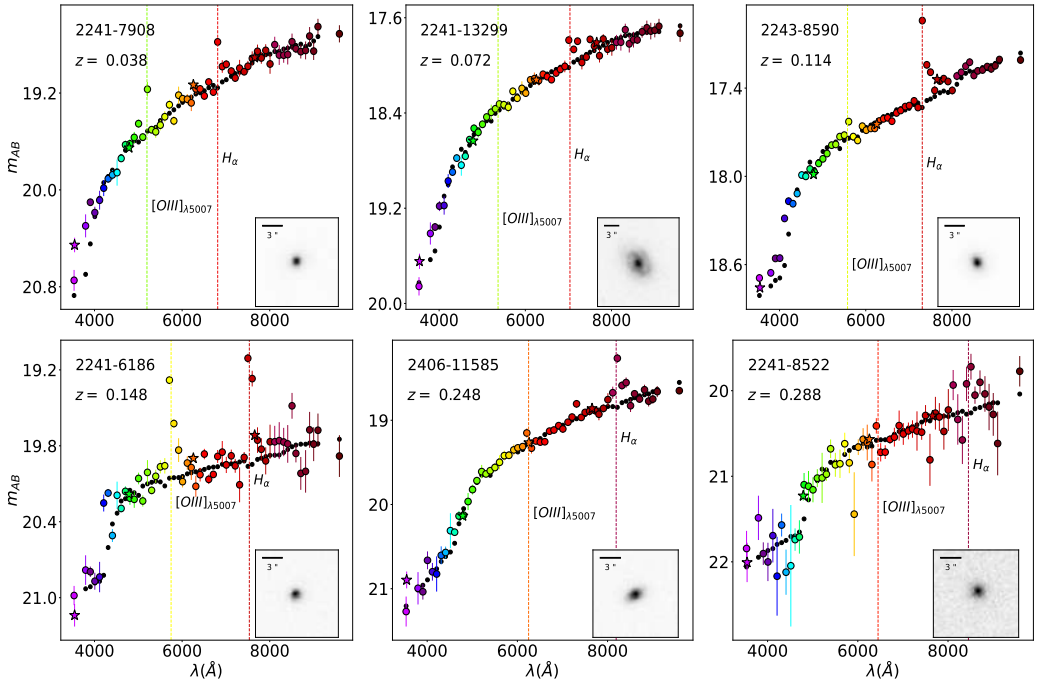


Figure 4.2: J-spectra in magnitudes (`MAG_AUTO` photometry) for a set of galaxies within the AEGIS field observed by miniJPAS. Stars correspond to broadband filters (u_{JPAS}, and SDSS g, r, and i). Black dots are the best fit obtained with `BaySeAGal` to the stellar continuum. Filters including the wavelength of Hα and [O III] emission lines within their bandpass are marked with dashed vertical lines. The images of these galaxies in the rSDSS band are attached in the lower left inset. The miniJPAS ID and the photo-z are shown in black in the left corner of each figure.

The analysis of the emission lines was carried out with a machine-learning code based on ANN. Different ANNs were trained with the J-PAS synthetic photometry extracted from CALIFA, MaNGA, and SDSS galaxies after convolving the spectra with the J-PAS photometric system. The ANNs learned to perform different tasks. First, an ANN was trained to estimate the EW values for the main emission lines in the optical range: Hα, Hβ, [O III], and [N II]. This ANN is referred to as $\mathrm{ANN_R}$. As inputs, the ANNs used photometry colors measured with respect to the J-PAS filter, in which the Hα flux dominates. As outputs, the ANNs received the values of the EWs that were measured directly in the spectrum. We estimated the uncertainty in the EWs with a Monte Carlo approach. We considered the error in the photo-z and the error in photometric fluxes. Second, another ANN was trained to distinguish galaxies with emission lines from those without them. This classifier ($\mathrm{ANN_C}$) relies on the EWs, but it is independent of the prediction from the $\mathrm{ANN_R}$. Galaxies were previously classified as class 1 or class 2 depending on whether they exceeded a preselected EW threshold in any of the emission lines. Several $\mathrm{ANN_C}$ with different thresholds ($\mathrm{EW_{min}} = 3, 5, 8, 11, \mathrm{and}\ 14$ Å) were trained in order to better study the regime of low emission, in which the $\mathrm{ANN_R}$ is less sensitive.

As we discussed in chapter 3, there are many ways of combining the CALIFA, MaNGA, and SDSS surveys to build up a training set. Each survey has its own observational biases, and the emission lines were measured with different approaches. In this chapter, we made use of the CALMa training set for the $\mathrm{ANN_R}$, which performs better in unseen data (SDSS test sample). The CALMa training set employs both CALIFA and MaNGA spectra from spatially resolved regions over many diverse physical states, including AGN emission and SF regions. With the CALMa training set, we are able to fully reproduce the position of SF galaxies in the BPT diagram. We reached a precision of 0.092 and 0.078 dex for log ([N II]/Hα) and log ([O III]/Hβ), respectively, assuming an average S/N in the photometry of 10. We can predict an EW of 10 Å in the Hα, Hβ, [N II], and [O III] lines with a median S/N of 5, 1.5, 3.5, and 10, respectively.

For the the $\mathrm{ANN_C}$ classifier, we employed the CALIFA set, which is a subset of the CALMa set, but only includes CALIFA galaxies. The two training sets performed very similarly in the SDSS test sample. For the sake of simplicity, we therefore employed the CALIFA set.

4.3.2 Stellar population analysis

The stellar population properties of the galaxies in this sample were analyzed with `BaySeAGal` (Amorim et al. in prep., González Delgado et al. 2021). This is a Bayesian parametric code that fits stellar metallicity ($Z_\star$), dust attenuation (τ_V), and the parameters related to the star formation history of galaxies. We assumed a delayed-τ model of the form

$$\Psi(t) = \phi \frac{t_0 - t}{\tau} \exp\left[-(t_0 - t)/\tau\right], \tag{4.1}$$

where t is the lookback-time, t_0 is the starting point of star formation in lookback-time, τ is the SFR e-folding time, and ϕ is the normalization constant related to the total mass formed in stars. t_0 and τ are sampled uniformly in logarithmic scale, which can vary between 1.4 and the maximum age at the redshift of the galaxy (13.7 Gyr at $z = 0$), and between 0.1 and 10 Gyr, respectively. For the present chapter, we chose the attenuation law proposed by Calzetti et al. (2000), which adds a unique foreground screen with a fixed ratio of $R_V = 4.05$ (the average value for the Milky Way).

The code used the 2017 version of the Bruzual & Charlot (2003) stellar population (SSP) synthesis models (hereafter CB17). The SSP covers the metallicity range $\log Z_\star/Z_\odot$ = -2.3, -1.7, -0.7, -0.4, 0, and +0.4, and the ages span from 0 to 14 Gyr. The CB17 models follow the PARSEC evolutionary tracks (Marigo et al. 2013; Chen et al. 2015) and use the Miles (Sánchez-Blázquez et al. 2006; Falcón-Barroso et al. 2011; Prugniel et al. 2011) and IndoUS (Valdes et al. 2004; Sharma et al. 2016) stellar libraries in the spectral range observed by J-PAS.

It is important to emphasize that filters capturing the nebular emission lines are masked and were not used in the SED fitting. Therefore, the galaxy properties are only based on the stellar continuum, and it does not include the emission of nebular regions or the result of the AGN activity. The stellar continuum is derived from the ensemble of best fits and allows us to determine stellar masses (M_*) , metallicities (Z_*), the amount of dust attenuation (A_V), or the luminosity-weighted age ($< \log t >_L$) of galaxies. Furthermore, it is also used to extrapolate the photometry in the filters that lack a measurement or have a very low S/N (lower than 1.8). Because the ANN (as we designed it) cannot work with missing data, these extrapolations allow the ANN to access all the inputs needed (photometric fluxes). This does not apply to the filters containing emission lines at each redshift and the filters that are immediately next to them. For instance, the Hα emission line is captured by the J0660 filter for a galaxy in the local Universe ($z = 0$).

Therefore, the fluxes in filters J0650, J0660, and J0670 are never extrapolated. When problems in the photometry with these filters occurred, we did not include the corresponding galaxies in our sample.

The use of alternative SED fitting codes to derive stellar population properties of miniJPAS galaxies does not affect the main results in this chapter. González Delgado et al. (2021) analyzed in detail how the main properties derived for galaxies might change with different SED fitting approaches. The results are consistent between each other: nonparametric codes such as `MUFFIT` (Díaz-García et al. 2015), `Alstar` (the algebraic version of starlight Cid Fernandes et al. 2005), or `TGASPEX` (Magris C. et al. 2015) and `BaySeAGal` all obtained similar distributions of rest-frame $(u-r)$ color, stellar mass, age, and metallicity up to $z = 1$.

A summary of the stellar population properties of the galaxies we analyzed is shown in Fig. 4.3. The distributions of the galaxy ages and the τ/t_0 ratio are bimodal. `BaySeAGal` provides rest-frame colors and extinction-corrected colors. In particular, $(u-r)_{int}$ is very useful for distinguishing between red and blue galaxies. We followed the criterion of Díaz-García et al. (2019a, hereafter the color criterion) in order to distinguish them. This criterion was adapted to match the miniJPAS photometric system. For a galaxy to be part of the red sequence, this criterion establishes a limit in $(u-r)_{int}$ from the galaxy stellar mass and redshift,

$$(u-r)^{\rm lim}_{\rm int} = 0.16 \times (\log M_\star - 10) - 0.3 \times (z - 0.1) + 1.7. \quad (4.2)$$

Galaxies with $(u-r)_{\rm int}$ above $(u-r)^{\rm lim}_{\rm int}$ are classified as red galaxies, otherwise, they are considered to be blue. Furthermore, `BaySeAGal` provides the probability distribution function (PDF) for the model parameters. The uncertainty on the derived stellar population properties is defined as the standard deviation. As expected, the uncertainty depends on the S/N of the photometry. The median errors are lower in the red sequence than in the blue cloud. That is, $\langle\sigma(\log M_\star)\rangle = 0.16 \pm 0.03$ dex, $\langle\sigma(\langle\log\, t\rangle_{\rm L})\rangle = 0.19 \pm 0.05$ dex, $\langle\sigma(A_V)\rangle = 0.19 \pm 0.07$ mag, and $\langle\sigma(\tau/t_0)\rangle = 0.10 \pm 0.04$ for galaxies in the red sequence, and $\langle\sigma(\log M_\star)\rangle = 0.28 \pm 0.04$ dex, $\langle\sigma(\langle\log\, t\rangle_{\rm L})\rangle = 0.25 \pm 0.05$ dex, $\langle\sigma(A_V)\rangle = 0.33 \pm 0.05$ mag, and $\langle\sigma(\tau/t_0)\rangle = 0.5 \pm 0.19$ for those in the blue cloud.

4.4 Identification of ELGs

In this section, we show the potential of our methods to identify ELG in the AEGIS field and determine their main ionization mechanism. The EW of Hα, Hβ, [O III], and [N II] and their relative strengths allow us to distinguish between

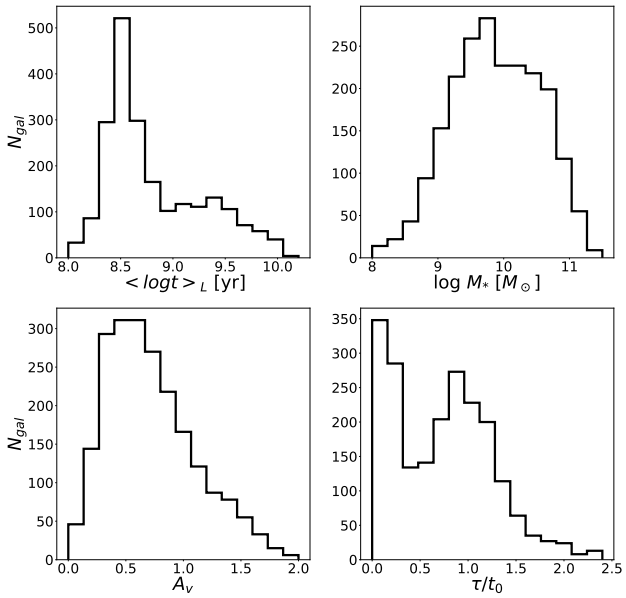


Figure 4.3: Distributions of mean stellar luminosity-weighted age (top left panel), galaxy stellar mass (lower right panel), extinction (lower left panel), and τ/t_0 ratio (bottom right panel) obtained by `BaySeAGal` for the sample of galaxies described in section 4.2.

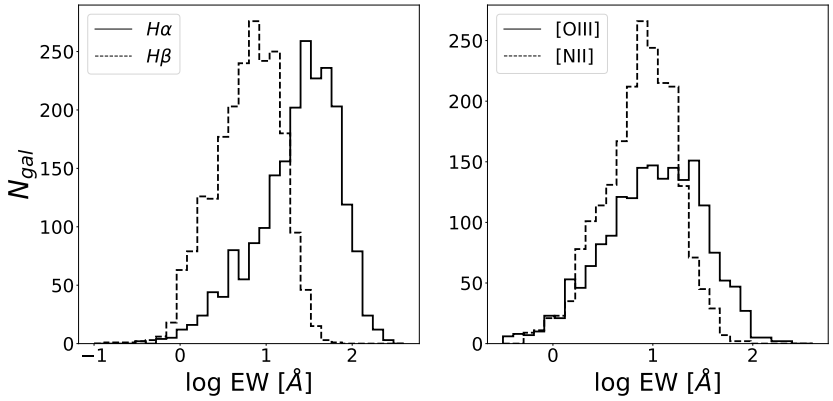


Figure 4.4: Distribution of the EW of Hα and Hβ (left), [O III], and [N II] (right) in log scale as obtained with the ANN$_R$.

different types of ELGs and derive the fraction of star-forming, Seyfert, and quiescent galaxies in miniJPAS.

4.4.1 Identification with ANN$_R$: EW distributions

First, we show the EW distribution of the Hα, Hβ, [O III], and [N II] lines in Fig. 4.4 derived with the ANN$_{\rm R}$. We excluded from the histograms galaxies where the EWs are below zero. Even though the ANN$_{\rm R}$ was not trained with absorption lines, certain configurations can indeed lead to negative values of the EWs. If the fluxes in the filters in which the emission lines are expected to appear are suppressed or are highly uncertain, or if they mimic the shape of an absorption line, the ANN$_{\rm R}$ might predict EWs that are below zero. We find 20, 2, 299, and 23 galaxies with negative EWs in Hα, Hβ, [O III], and [N II], respectively. The median S/N in the EWs for these galaxies is below one, which indicates that these values are compatible with positive and null values.

Generally, blue galaxies are star-forming galaxies, while red galaxies are quiescent. However, a galaxy might appear to be part of the red sequence due to the presence of dust, which absorbs a fraction of the total radiation more efficiently on the blue side of the spectrum. Therefore, it is important to correct for dust extinction in order to distinguish between red and dust-reddened star-forming galaxies.

Figure 4.5 shows as expected that blue galaxies contain young populations of stars with high values of EW(Hα), while red galaxies are older and lack Hα emission or have very low values of EW(Hα). Between the red sequence and the blue cloud, we observe galaxies in the GV with intermediate ages and moderate values in the EWs of Hα.

4.4.2 Identification with the ANN$_C$: Strong and weak ELGs

In addition to the color-criterion, we can also make use of the predictions of the ANN$_{\rm C}$ to distinguish between galaxies above and below a certain threshold limit in the EW. The EW of Hα quantifies the relative intensity of the emission line flux with respect to the stellar continuum, and therefore it is a good indicator of the sSFR in the galaxy (Mármol-Queraltó et al. 2016; Khostovan et al. 2021). In Fig. 4.6 we plot the log EW(Hα) as a function of the stellar mass. In the left panel, we indicate in blue (red) the galaxies that belong to the blue cloud (red sequence) following the color criterion. On the right panel we show a similar scheme but galaxies are separated according to the class defined by the ANN$_{\rm C}$ with EW$_{\rm min}$ = 3 Å. In other words, galaxies are considered strong ELs if any

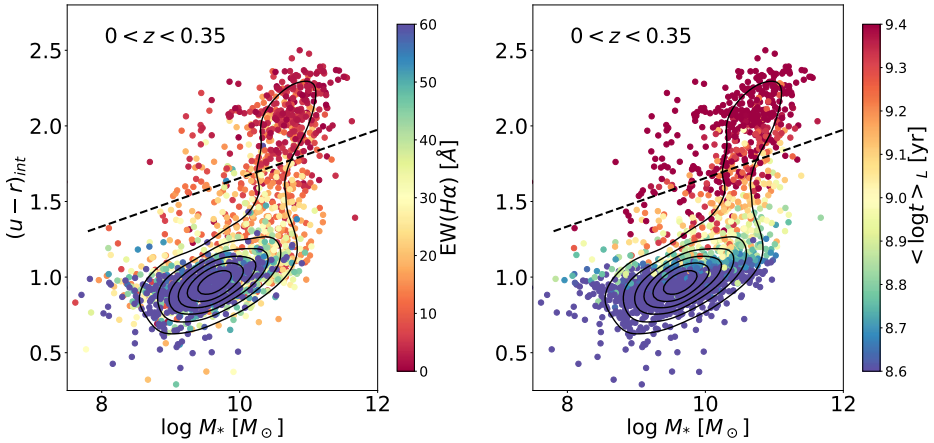


Figure 4.5: Color–mass diagram for our sample of galaxies. The (u – r) color-corrected for dust extintion vs. stellar mass. Galaxies are color-coded with the EW of Hα (the luminosity-weighted stellar age) on the left side (right side). The intrinsic color, stellar mass, and luminosity-weighted age are obtained via `BaySeAGal`. Dashed black lines separate blue and red galaxies following Eq. 4.2, where we considered the median redshift of the sample (z = 0.25). Density contours are drawn in black at the top.

of the emission lines present an EW greater than 3 Å and weak ELs if all lines are below this limit. For a threshold of 0.1 in the $\mathrm{ANN_C}$ probability, strong ELs represent 83 % of the sample, while weak ELs are the remaining 17 %. With the color criterion, 82 % of the galaxies in the parent sample are classified as red and the remaining 18 % are blue.

The dashed line in Fig. 4.6 illustrates the EW(Hα) = 3 Å limit. As expected, most of the galaxies below this limit are classified as weak ELs. However, we detect a non-negligible number of weak ELs or red galaxies above this limit in both panels. We have to take into account that the $\mathrm{ANN_R}$ is less accurate at low EWs and has a tendency to overestimate their values. Moreover, the relative errors in this regime are higher. Therefore, it is not surprising to find a fraction of weak EL galaxies above this limit. Moreover, although Hα leads the $\mathrm{ANN_C}$ classification, the algorithm includes other emission lines in addition to Hα, which might occasionally overcome this limit. At high EWs, the number of weak EL galaxies decreases significantly, and the discrepancy between the $\mathrm{ANN_R}$ and the $\mathrm{ANN_C}$ can be explained by the high uncertainty found in the photo-z or a low S/N in the photometric fluxes.

In two panels in Fig. 4.6, the two methods of classifying galaxies present a

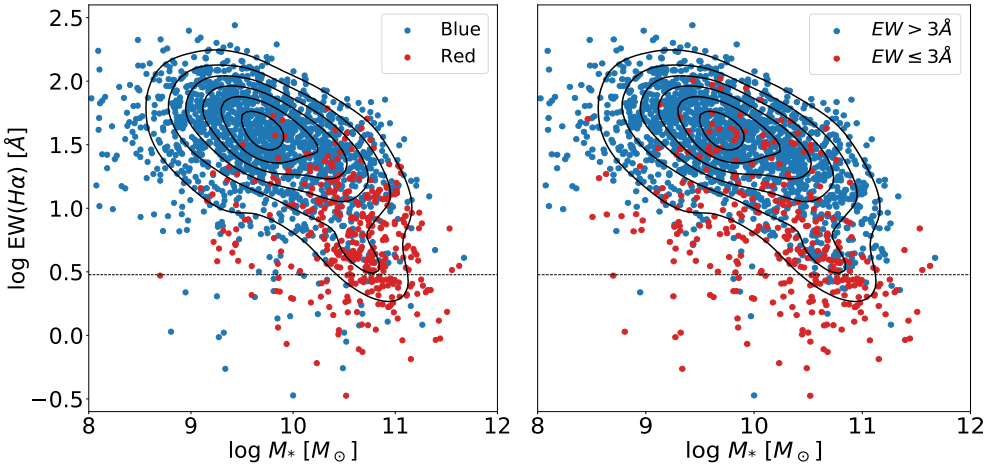


Figure 4.6: Equivalent width of Hα as a function of the stellar mass of the galaxy. In the left panel, we used Eq. 4.2 to distinguish between red and blue galaxies. In the right panel, we relied on the classification performed with a machine-learning code trained with strong EL and weak EL galaxies. Strong ELs were defined as those with EWs greater than 3 Å in any of the following emission lines: Hα, Hβ, [O III], or [N II], and weak ELs are all others. The dashed horizontal lines mark the 3 Å limit in the EW(Hα). Density contours are drawn in black at the top.

consistent picture. Most of the blue galaxies are strong ELs, and red galaxies are weak ELs. Nevertheless, we found some disagreement between the last two populations. While the ANN_C is trained to separate galaxies as a function of the EW, Eq. 4.2 depends mainly on the global color and the mass of the galaxy. Thus, it is expected to find some galaxies with red intrinsic colors and a low level of star formation reflected on the nebular emission with EWs greater than 3 Å.

Finally, it is clear from these diagrams that galaxies are less efficient at forming new stars as the mass of the galaxy increases at $z < 0.35$. At some point around $M_* = 10^{11} M_\odot$, the EW of Hα falls sharply, with most galaxies above this mass showing red colors and low values in the EW(Hα), suggesting that the main sequence of star-forming galaxies has already ended.

4.4.3 Identification of star-forming galaxies and AGNs: BPT and WHAN diagrams

The BPT diagram (log[O III]/Hβ versus log[N II]/Hα) provides a means to unveil the main ionization mechanism of galaxies. It involves four emission lines, and galaxies are classified into four groups by three dividing lines: star-forming, com-

posite, Seyfert, and LINERs. The Kauffmann et al. (2003a, hereafter Ka03,) curve is derived empirically using the SDSS galaxies and defines the region populated by SF galaxies. Usually referred to as the SF wing, galaxies evolve from high (low) to low (high) [O III]/Hβ ([N II]/Hα) ratios, increasing their mass (Maiolino & Mannucci 2019). The Kewley et al. (2001, hereafter Ke01,) curve is determined using both stellar population synthesis models and photoionization. It defines the AGN wing that is dominated by AGN (including LINER or LINER-like emission, and shocks). Between these two lines lies the composite region, which might be populated by galaxies with a composite spectrum, that is, the ionization mechanism is a mix of star-formation processes and AGN activity or galaxies with very weak emission lines that are leaving the SFMS. Finally, the Schawinski et al. (2007, hereafter S07,) line is an empirical division that distinguishes between Seyfert and LINER galaxies.

We show the BPT diagram for the galaxies in the parent sample with error lower than 0.2 dex in [O III]/Hβ and [N II]/Hα in the left panel of Fig. 4.7. In the right panel, we relax this threshold to 0.5 dex. These thresholds are arbitrary, and they have been chosen to show how the BPT diagram changes when galaxies with a high uncertainty in the predicted emission lines are included. However, they are not used for the final selection of SF galaxy sample. The stellar mass distribution of galaxies in the BPT is consistent with expectations: galaxies grow in mass while they evolve through the SF wing. However, as the error increases (right panel), some galaxies populate regions that are less likely to be occupied (the narrowest wedge at the top left within the composite region).

Galaxies with very faint emission lines may be misclassified as LINERs from a BPT diagnostic. Sometimes called fake AGN (Cid Fernandes et al. 2011), one of the advantages of the WHAN (log EW(Hα) versus log ([N II])/Hα)) diagram is that it can identify these galaxies. Even more important is the fact that the WHAN diagram provides a simpler way of determining the main ionization mechanism of galaxies.

Fig. 4.8 we show the WHAN diagram for the galaxies in the parent sample. The solid and dashed vertical lines represent the optimal projection of Ka03 and Ke01 onto the log EW(Hα) versus log ([N II])/Hα) space, that is, the dividing lines that better distinguish galaxy types in the WHAN diagram as they are defined in the BPT (Cid Fernandes et al. 2010, 2011). Similarly, the division between Seyferts and LINERs at EW(Hα) = 6 Å corresponds to the optimal projection of S07. Finally, the area below the dashed horizontal line at EW(Hα) = 3.16 Å is composed of galaxies with highly uncertain line predictions that are therefore compatible with quiescent galaxies. We did not distinguish between retired and

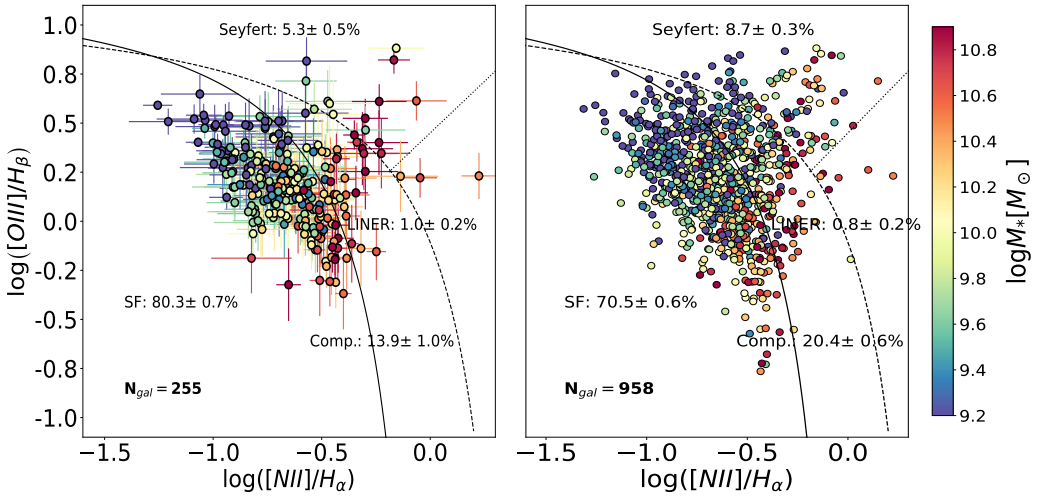


Figure 4.7: BPT diagram for the galaxies in the sample with an error of 0.2 dex (0.5 dex) in the [O III]/Hβ and [N II]/Hα ratios in the left (right) panel. The errors are not plotted in the right panel for clarity. The color bar indicates the stellar mass of the galaxy. The solid (Ka03), dashed (Ke01), and dotted lines (S07) define the regions for the four main spectral classes. The relative percentage of each galaxy type in each subsample is indicated in the figure. In each panel, the number of galaxies is specified in the lower left corner. The parent sample contains 2154 galaxies.

passive galaxies as in Cid Fernandes et al. (2011) because our precision is not high enough to predict values of the EWs in the range of a few Å.

In the left panel of Fig. 4.8 we show galaxies with an error smaller than 0.2 dex in both the EW(Hα) and the [N II]/ Hα ratio, while in the right panel, we relax this requirement to 0.5 dex. The percentage of each galaxy type is indicated in the legend. Galaxies with lower EW(Hα) have higher relative errors. Furthermore, many red galaxies do not appear in this diagram.

The color gradient in Fig. 4.8 indicates that galaxies are more massive as the EW(Hα) decreases and the [N II]/Hα ratio increases. Therefore, star-forming galaxies are on average less massive than Seyferts, while LINERs and passive galaxies are the most massive galaxies.

By comparing the position of each galaxy (i.e., their values with errors) in both diagrams, it is noticeable that the values of a given galaxy in the BPT convey more uncertainties than in their counterpart spot in WHAN. The reason is that the error in the y-axis of the BPT diagram stems from two sources: the error in the [O III] and Hβ emission lines. However, in the WHAN diagram, the only error source is the Hα emission line. As a consequence, with a maximum error of 0.2 dex, we

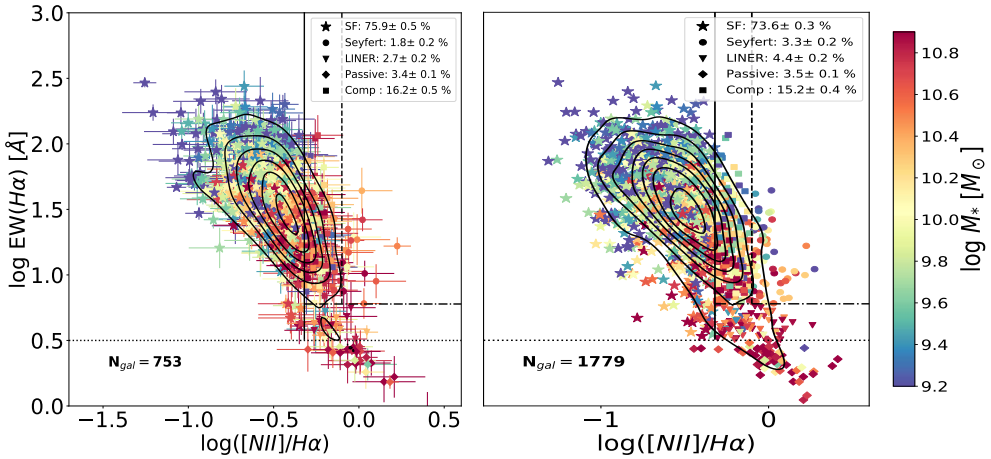


Figure 4.8: WHAN diagram for galaxies with an error smaller than 0.2 dex (0.5 dex) in both the EW(Hα) and the [N II]/ Hα ratio in the left (right) panel. The errors are not shown in the right panel for clarity. The color bar indicates the stellar mass of the galaxy. The inset shows the relative percentage of each galaxy type in each subsample. Dashed and solid vertical lines define the optimal projections of the Ke01 and the Ka03 lines in the WHAN diagram (Cid Fernandes et al. 2010, 2011). Similarly, the dash-dotted horizontal line at EW(Hα)= 6 Å is the optimal transposition of the S07, and the dotted line at log EW(Hα)= 0.5 Å defines the limit of ELGs. In each panel, the galaxy counts are specified in the lower left corner. The parent sample contains 2154 galaxies. Density contours are drawn in black at the top.

can estimate the position of only 255 galaxies of the sample in the BPT and 753 galaxies in the WHAN. The median S/Ns of these subsamples in the narrowband filters are 10.7 and 11.4.

4.4.4 Fraction of galaxy types in miniJPAS

We identify 83 % of the galaxies (1787) from the parent sample (2154 galaxies) in the AEGIS field as strong ELGs, and the remaining 17% (367 galaxies) are weak ELGs. In Table 4.1 we show the percentages of each galaxy type according to the WHAN diagram for all galaxies with an error smaller than 1 dex in the EW(Hα) and [N II]/ Hα ratio. This criterion is fulfilled by 2000 galaxies, which leaves 154 galaxies from the parent sample unclassified. We eliminate the composite population, but we indicate the percentage of SF and Seyfert galaxies in the different separation curves: Ka03, Ke01, or Stasińska et al. (2008, hereafter S08). Although we showed in Fig. 4.8 the percentages for LINERs and passive galaxies, we grouped both classes together in this table. The emission lines for

LINER galaxies are at the limit of what we can detect with the ANN given the S/N in the photometry. Hence, it is more challenging to distinguish them in the low S/N regime. We estimated the percentages and the errors of each galaxy type with a Monte Carlo (MC) method using the position of each galaxy in the diagram and its errors. Then, we computed the median and the standard deviation.

[N II]/ Hα	Star-forming [%]	Seyfert [%]	Quiescent [%]
$\leq$ 0.79 (S08)	89.8 $\pm$ 0.2	3.5 $\pm$ 0.2	6.7 $\pm$ 0.2
$\leq$ 0.48 (Ke01)	72.8 $\pm$ 0.4	17.7 $\pm$ 0.4	9.4 $\pm$ 0.2
$\leq$ 0.40 (Ka03)	62.4 $\pm$ 0.3	27.5 $\pm$ 0.4	10.1 $\pm$ 0.2

Table 4.1: Percentage of each galaxy type according to the WHAN diagram. Quiescent galaxies include LINERs and passives.

Finally, we studied how the fractions of SF, Seyfert, and quiescent (passive or LINER) galaxies varied when we imposed brighter flux limit constraints. For this purpose, we generated new samples of galaxies that are below 20.5, 21.5, and 22.5 mag in the rSDSS band and computed the fraction of each galaxy type. The results are shown in Fig. 4.9. We do not observe a strong correlation with the rSDSS apparent magnitude. The fraction of each galaxy type is more uncertain when one or another of the separation curves is chosen.

4.5 Characterization of star-forming galaxies

In this section, we characterize the star-forming galaxy population in miniJPAS. We traced the SFR through the Hα emission line. First, we selected a suitable sample of star-forming galaxies with the identification tools we presented in the previous section. Then, we corrected the Hα flux from nebular extinction and derived the position of SF galaxies in the SFMS. We also analyzed the correlation between nebular and stellar extinction and the relation between the star formation history (SFH) of galaxies obtained with the SED fitting and their position in the SFMS.

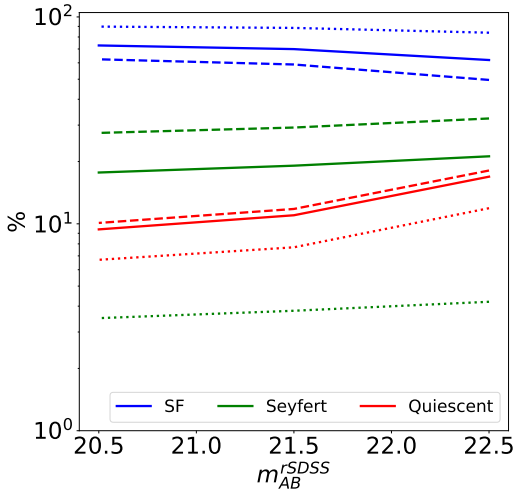


Figure 4.9: Fraction of SF, Seyfert, and quiescent (passive or LINER) galaxies as a function of the maximum rSDSS apparent magnitude of each subsample. Solid, dashed, and dotted lines represent the fraction of each galaxy type according to the Ka03, Ke01, and S08 curves, respectively.

4.5.1 Selection of star-forming galaxies

Our sample of star-forming galaxies was obtained from the parent sample (section 4.2) by imposing different constraints. We relied on the WHAN diagram to exclude the galaxies in which the main ionization mechanism is not driven by star formation (AGN-like galaxies). We chose the Ka03 curve. In order to consider a galaxy as a member of the main sequence, we therefore imposed a maximum [N II]/Hα of 0.48. We also discarded galaxies with very low emission in the diagram (LINER and passive galaxies). Finally, galaxies must be classified as blue with the color criterion and the $\mathrm{ANN_C}$ to be part of our sample. We found 1178 galaxies in total (SF sample hereafter).

In Fig. 4.10 we show the relation between the total stellar mass and the redshift for all galaxies in the parent sample. The solid black line indicates the limit at which galaxies cannot be observed in our flux-limited sample (see section 4.2). In order to be complete in mass, we would need to discard a large fraction of galaxies and risk to loose statistical reliability. Furthermore, the mass dynamical range would be significantly reduced at high redshift. Therefore, we fit the SFMS in two cases: using the whole SF sample, or using only galaxies in the SF sample that are above the stellar mass detection limit (see section 4.5.3). We will also study how stronger flux limit constraints affect the shape of the SFMS. As soon as

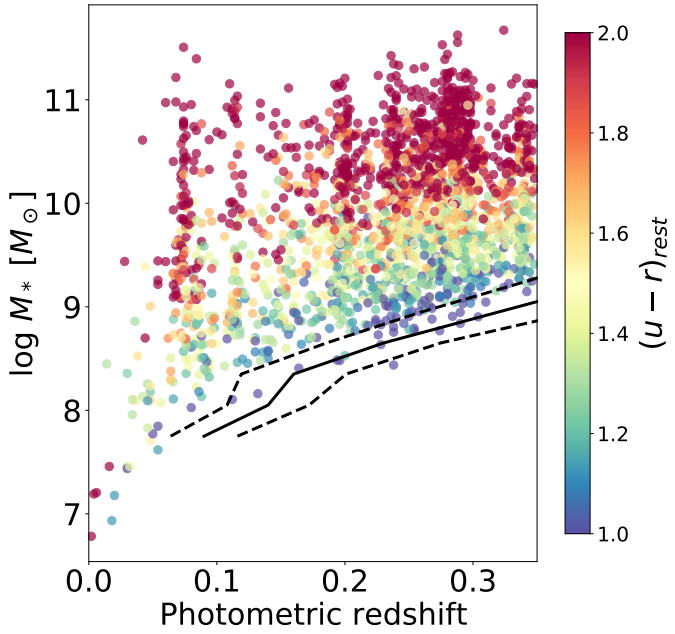


Figure 4.10: Relation between galaxy stellar mass and redshift for all galaxies in the parent sample. The solid black line is the limit at which galaxies can no longer be observed with the criteria we used to select the sample (see section 4.2). Dashed black lines represent the uncertainty limit ($\pm\sigma$.) Galaxies are color-coded according to their (u–r) rest-frame color.

J-PAS observes larger areas of the sky, we will be able to be more conservative in the mass limit of the selected sample.

4.5.2 Dust correction

In order to account for the extinction of dust, we followed the empirical extinction relation described in Calzetti et al. (1994). The intrinsic luminosity of galaxies ($L_{\rm int}$) is attenuated by interstellar dust through the following equation:

$$L_{\rm int}(\lambda) = L_{\rm obs}(\lambda)10^{0.4A_\lambda} = L_{\rm obs}(\lambda)10^{0.4k(\lambda)E(B-V)}, \tag{4.3}$$

where $L_{\rm obs}$ is the observed luminosity, A_λ is the extinction at wavelength λ, and $k(\lambda)$ is the reddening curve. We considered the reddening curve of Calzetti et al. (2000) with $R_{\rm V} = 4.05$. The nebular color excess $E(B-V)$ can be obtained from the Balmer decrement assuming regular gas conditions in star-forming galaxies (for a detailed description, see, e.g., Domínguez et al. 2013, and references

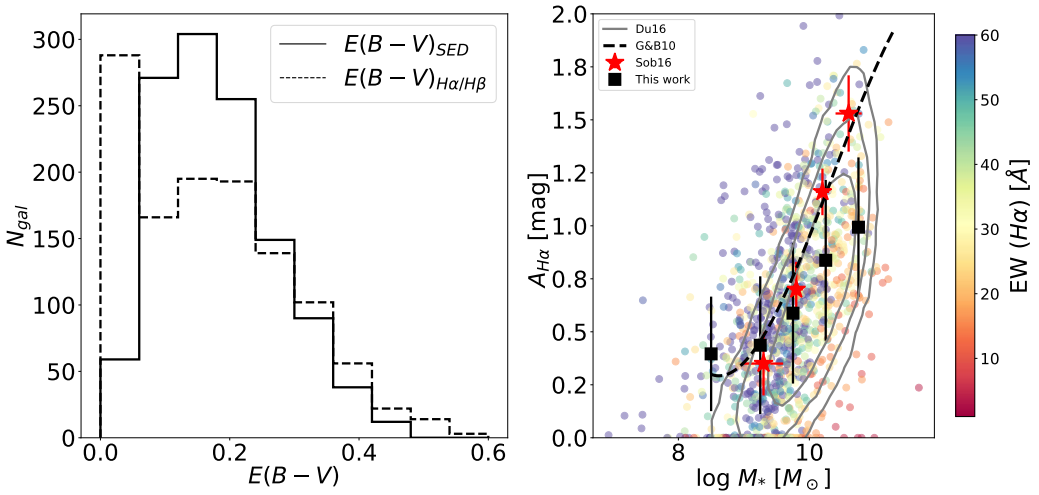


Figure 4.11: Distribution of the nebular ($E(B-V)_{H\alpha/H\beta}$) and stellar ($E(B-V)_{SED}$) color excess (left). Nebular extinction at the Hα wavelength as a function of stellar mass (right). Galaxies are color-coded with the EW of Hα and belong the SF sample described in section 4.5.1. Black squares are the median obtained in the following stellar mass bins: $8 < \log M_* \leq 9$, $9 < \log M_* \leq 9.5$, $9.5 < \log M_* \leq 10$, $10 < \log M_* \leq 10.5$, and $10.5 < \log M_* \leq 11$. The error bars on the y-axis represent the standard deviation, gray contours represent the density of sources for 1 σ, 2 σ, and 3 σ derived from SDSS galaxies in Duarte Puertas et al. (2017). Red stars are the values obtained by Sobral et al. (2016) by means of spectroscopy measurements in SF galaxies within the cluster Cl0939+4713 at $z = 0.41$. The dashed blakc line is the best polynomial fit obtained by Garn & Best (2010) in a sample of SDSS galaxies.

therein) as follows:

$$E(B-V) = 1.97 \log_{10}\left[\frac{(\mathrm{H}\alpha/\mathrm{H}\beta)_{\mathrm{obs}}}{2.86}\right], \tag{4.4}$$

where Hα and Hβ stand for the emission line fluxes. As the $\mathrm{ANN_R}$ provides the values of the EWs, we used the stellar continuum derived from `BaySeAGal` at the Hα and Hβ wavelengths to compute the total flux of the emission lines.

In the left panel of Fig. 4.11 we show the distribution of the nebular ($E(B-V)_{H\alpha/H\beta}$) and stellar ($E(B-V)_{SED}$) color excess. A fraction of galaxies in the SF sample ($\sim$ 15 %) have a Balmer decrement below the theoretical value (2.86), but very close to it. Furthermore, its errors indicate that nebular extinction for these galaxies is compatible with null or very low values. Either way, we set the $E(B-V)_{H\alpha/H\beta}$ to zero for these galaxies. $E(B-V)_{SED}$ is 0.017 mag higher on average than $E(B-V)_{H\alpha/H\beta}$ with a dispersion of 0.072 mag. The median error

on the $E(B-V)_{H\alpha/H\beta}$ and $E(B-V)_{SED}$ is 0.089 and 0.015, respectively. Some authors reported that $E(B-V)_{H\alpha/H\beta}$ is twice $E(B-V)_{SED}$ on average (Calzetti et al. 2000; Qin et al. 2019; Koyama et al. 2019). However, other studies found similar levels of nebular and stellar extinction (Kashino et al. 2013; Puglisi et al. 2016). In particular, we found agreement with the results of Kouroumpatzakis et al. (2021, see Fig. 8 and Table 1,) who argued that nebular extinction is much more pronounced in the nuclear regions, affecting the relations found by single spectroscopic surveys such as the SDSS, which cannot capture the whole light produced in galaxies.

The right panel of Fig. 4.11 shows the nebular extinction at the Hα wavelength ($A_{H\alpha}$) as a function of the galaxy stellar mass. We found a similar trend as in other studies. Red stars are the values obtained by Sobral et al. (2016) by means of spectroscopy measurements in SF galaxies within the cluster Cl0939+4713 at $z = 0.41$. Gray contours represent the density of sources for 1 σ, 2 σ, and 3 σ derived from all SDSS SF galaxies in Duarte Puertas et al. (2017). Finally, the dashed black line is the best polynomial fit obtained by Garn & Best (2010) in a sample of SDSS galaxies. Applying aperture correction to the Hα/Hβ ratio as in Duarte Puertas et al. (2017) lowers the extinction 0.2 mag in average.

4.5.3 Fitting the star formation main sequence

The SFR was obtained from the Hα luminosity using the Kennicutt et al. (1994) relation converted to employ a Chabrier IMF (Chabrier 2003) and assuming case B recombination,

$$\mathrm{SFR}[M_\odot\ \mathrm{yr}^{-1}] = 4.9 \times 10^{-42} L_{\mathrm{H}\alpha}[\mathrm{erg/s}]. \tag{4.5}$$

We used this relation to derive the SFR from the corrected Hα luminosity. Then, we fit the SFMS for the galaxies in the SF sample assuming a power-law relation between the stellar mass (M_*) and the SFR,

$$\log \mathrm{SFR} = \alpha \times \log M_* + \beta. \tag{4.6}$$

We assumed that galaxies deviate from this relation with a scatter perpendicular to the line that we parameterized in terms of the scatter along the y-axis (σ_y), often called σ_{int}. We employed a Bayesian approach to derive the posterior distribution of σ_y, α, and β. We followed Robotham & Obreschkow (2015) in order to

construct the likelihood function,

$$\ln L = -\frac{1}{2}\sum_{i=0}^{N_{\rm gal}} \frac{(\log \mathrm{SFR}_i - \alpha \log M_{*,i} - \beta)^2}{\sigma_i^2} + \ln \sigma_i^2 - \ln(\alpha^2 + 1), \tag{4.7}$$

where σ_i^2 reads

$$\sigma_i^2 = \sigma_y^2 + \sigma^2_{\log \mathrm{SFR}_i} + \alpha^2 \sigma^2_{\log \mathrm{M}_i}. \tag{4.8}$$

We assumed that the errors in the SFR and stellar mass of the galaxies are not correlated. This hypothesis is justified because both quantities are derived independently from each other. Although the flux of stellar continuum at Hα wavelength is used to estimate the total Hα flux, its error is negligible compared to the error in the EW. The errors are considered Gaussian and heteroscedastic, that is, each data point is drawn from a different Gaussian distribution. The last term in Eq. 4.7 ensures that the data are rotationally invariant. In other words, data have no defined predictor or response variable, and therefore we can predict the SFR from the stellar mass of the galaxy and vice versa.

The posterior distribution was sampled with the Markov chain Monte Carlo (MCMC) method, using the `emcee` Python implementation (Foreman-Mackey et al. 2013), with 250 walkers and 5000 steps per walker. We used a burn-in phase of 3500 steps.

Figure 4.12 shows the SFMS for the galaxies in the SF sample; in black we plot the ensemble of best fits obtained with the Bayesian routine. Galaxies are color-coded with the τ/t_0 ratio, which is an indicator of the SFH (see Eq. 4.1). High values of τ/t_0 indicate an SFH with almost constant SFR throughout cosmic time, while low values are related to galaxies with a burst of star formation long ago with a decreasing SFR ever since.

On the one hand, the color gradient observed in Fig. 4.12 suggests that galaxies with higher values of τ/t_0 are more likely to be found above the SFMS and preferentially have stellar masses below $10^{10} M_\odot$. On the other hand, lower values of τ/t_0 are associated with massive galaxies that lie below the SFMS.

We investigated how the parameters of the SFMS are affected when we included only the galaxies in the SF sample that lie above a certain flux limit. Additionally, we generated a new sample of galaxies that were selected from the SF sample with stellar masses above $10^9 M_\odot$ (SF0 sample). This is the stellar mass detection limit for the redshift between 0 and 0.35 (black line in Fig. 4.10). Subsequently, we studied again how the flux limit cut affects the parameters of the

rSDSS	α	β	σ_y
≤ 22.5	$0.90^{+0.02}_{-0.02}$	$-8.85^{+0.19}_{-0.20}$	$0.20^{+0.01}_{-0.01}$
≤ 21.5	$0.93^{+0.02}_{-0.02}$	$-9.15^{+0.21}_{-0.21}$	$0.21^{+0.01}_{-0.01}$
≤ 20.5	$0.93^{+0.03}_{-0.03}$	$-9.27^{+0.26}_{-0.27}$	$0.22^{+0.01}_{-0.01}$
≤ 22.5	$0.93^{+0.03}_{-0.03}$	$-9.17^{+0.29}_{-0.29}$	$0.21^{+0.01}_{-0.01}$
≤ 21.5	$0.95^{+0.03}_{-0.03}$	$-9.37^{+0.30}_{-0.33}$	$0.21^{+0.01}_{-0.01}$
≤ 20.5	$0.97^{+0.04}_{-0.04}$	$-9.66^{+0.30}_{-0.30}$	$0.23^{+0.02}_{-0.01}$

Table 4.2: Parameters of the SFMS with different selection cuts in the rSDSS band for the SF (SF0) sample at the top (bottom).

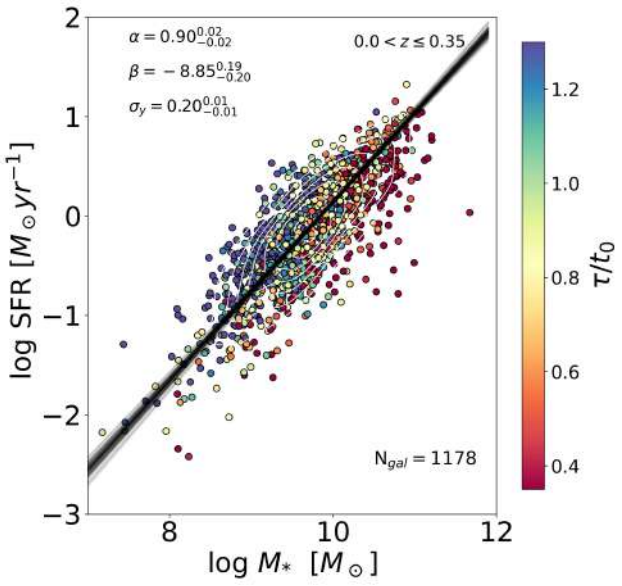


Figure 4.12: SFR vs. stellar mass for the galaxy sample described in section 4.5.1. Galaxies are color-coded with the τ/t_0 ratio (see section 4.3.2). Black lines are the best fits obtained with the Bayesian routine. The median posterior value and 1σ confidence interval are shown for each of the parameters.

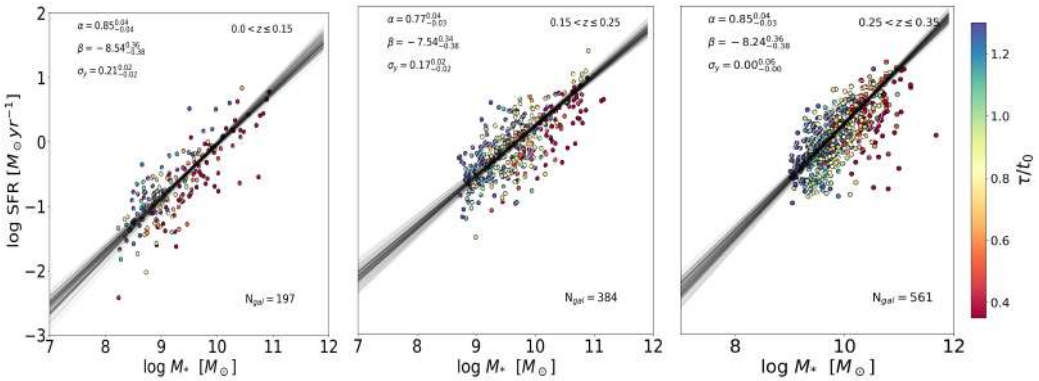


Figure 4.13: SFR vs. stellar mass for galaxies in different redshift bins color-coded with their the τ/t_0 ratio (see section 4.3.2.) Black lines are the best fits obtained with the Bayesian routine. The median posterior value and 1σ confidence interval are shown for each of the parameters. The number of galaxies within each redshift bin is also indicated.

SFMS. The results are summarized in Table 4.2. We conclude that the selection function that depopulates the SFMS below $m_{AB} = 22.5$ in the rSDSS band does not affect the shape of the SFMS. The results for the SF and SF0 sample are consistent (compatible within the errors).

4.5.4 SFR at different redshift

The relation of the SFR and the stellar mass is expected to change as a function of the redshift due to changes in the cosmic gas accretion rates and the gas depletion timescales. Some authors modeled this relation with a power law (SFR $\propto (1+z)^a$, Boogaard et al. 2018; Schreiber et al. 2015), others assumed that the evolution takes place in the zeropoint (log SFR $\propto \beta z$, Shin et al. 2021). Another common approach is to split the sample into redshift bins and fit them independently (e.g., Davies et al. 2016; Thorne et al. 2020). Because the redshfit range of the SF sample is limited, we decided to employ the latter approach and fit the SFMS in three different redshift bins: $0 < z \leq 0.15$, $0.15 < z \leq 0.25$, and $0.25 < z \leq 0.35$. We removed all galaxies in each sample that lay below the stellar mass limiting value (solid black line in Fig. 4.10).

We show the results in Fig. 4.13. A small flattening of the relation is seen at intermediate redshifts, but it may not be significant. As expected due to the anticorrelation between the slope and the zeropoint, the latter becomes higher in the $0.15 < z \leq 0.25$ bin. Most likely, these discrepancies are caused by the effect

of fitting the SFMS within a smaller dynamical range of mass and by the lower statistics. The intrinsic scatter of galaxies along the SFMS decreases at higher redshifts. This may be caused by a dependence on stellar mass rather than on redshift. Galaxies below $1.6 \times 10^8 M_\odot$, $5 \times 10^8 M_\odot$, and $10^9 M_\odot$ for $0 < z \leq 0.15$, $0.15 < z \leq 0.25$, and $0.25 < z \leq 0.35$, respectively, cannot be detected with fluxes brighter than 22.5 in the rSDSS band. We discuss the implication of this result in more detail in section 4.6.1.

4.5.5 Turnover mass hypothesis

Several studies have shown evidence that the relation between the SFR and the stellar mass turns over at a mass of $M_* \sim 10^{10}\ M_\odot$ (Whitaker et al. 2014; Lee et al. 2015; Schreiber et al. 2015; Tomczak et al. 2016). In this section, we investigate this scenario by fitting a quadratic power law (Eq. 4.9) and a broken power law (Eq.4.10) to the SF sample,

$$\log \mathrm{SFR} = \alpha \times \log M_* + \gamma \times (\log M_*)^2 + \beta \tag{4.9}$$

$$\log \mathrm{SFR} = \beta - \log\left[1 + (M_*/M_0)^{-\alpha}\right]. \tag{4.10}$$

We obtained a turnover mass ($\log M_0 = 10.93^{+0.22}_{-0.17}$) that is very close to the highest mass that we have in the SF sample ($\log M_*^{\mathrm{max}} = 11.2$). Furthermore, only 14 out of 1178 galaxies have a mass higher than M_0. For the quadratic model, we obtained a quadratic term near zero ($\gamma = -0.08^{+0.02}_{-0.02}$). In Table B.1 (see next section) we show the best-fitting parameters for different separation curves. We employed the Bayesian information criterion (BIC) to determine the model that better describes the observed SMFS. The BIC is defined as $\mathrm{BIC} = n_{\mathrm{param}} \ln N_{\mathrm{gal}} - 2 \ln L$, where n_{param} is the number of parameters in the model, N_{gal} is the number of galaxies, and L is the likelihood function. The linear model (Eq. 4.6) obtained the lowest value. Therefore, it is the most likely model.

4.5.6 AGN selection criteria

The exclusion of AGN-like galaxies from the SF sample is based on the [N II]/Hα ratio and the EW of Hα. We chose the curve of Ka03 to select SF galaxies, but we could have relied on other separation curves, such as Ke01 or S08. In this section we study how these choices can impact our result.

In Table B.1 we show the best-fit parameter values as a function of the separation curves, the redshift bin, and the fitting equation used to model the SFMS. The

results are marginally consistent, meaning that the retrieved parameter does not change the main conclusion of the previous sections. Nevertheless, we observed a trend in the slope, the quadratic term, and in the turnover mass as we relaxed the maximum [N II]/Hα ratio allowed to be part of the SFMS. Galaxies at the border of the dividing lines populate the high-mass end. As a consequence, the quadratic terms and the turnover mass increase as the slope of the SFMS flattens. Nonetheless, the intrinsic scatter exhibits little variation, except for the highest redshift bin, where higher-mass galaxies increase the scatter. This exercise demonstrates that the SFMS can be affected by AGN contamination, which is only one ingredient in the definition of the SFMS. Other criteria based on color cuts or sSFR thresholds are also important and can have a non-negligible impact on the derived parameters of the SFMS (Belfiore et al. 2018; Sánchez et al. 2019; Khostovan et al. 2021).

4.6 Discussion

In the following sections, we compare the results of the SFMS with the literature. We derive the cosmic evolution of the star formation rate density up to $z = 0.35$, and we discuss the differences we found with respect to other studies that did not trace the SFR with Hα emission line.

4.6.1 SFMS: Comparison with the literature

We have modeled the SFMS in the mass range from 10^8 up to $10^{11} M_\odot$ in the redshift range $0 < z < 0.35$. We employed a Bayesian approach (section 4.5.3) that considers the intrinsic scatter of the SFMS and the heteroscedastic errors on the stellar masses and the SFRs. We derived the SFRs from the Hα emission line, and we corrected for dust extinction through the Balmer decrement. We relied on the [N II]/Hα ratio to remove from the sample galaxies hosting an AGN. The linear model explains the relation between the log SFR and log M_* for the sample of SF galaxies better. Our selection criteria combine color-cut and emission line diagnostics and consequently favour a pure rather than a complete sample of SF galaxies. Most probably, we also excluded most of the GV population, and this might explain why the turnover-mass scenario is not compatible with our results. We compare our results with the literature below. We focus our attention on the slope of the SFMS and on the intrinsic scatter.

Slope

We find a result very similar to those of Sánchez et al. (2019) (MaNGA) and Cano-Díaz et al. (2016) (CALIFA), but our slope is steeper than those of Belfiore et al. (2018) and Cano-Díaz et al. (2019), who used MaNGA data. Our results are also consistent with the recent work of Vilella-Rojo et al. (2021), who studies the SFR of galaxies in the nearby Universe with J-PLUS data. SDSS galaxies have also been used to analyze the SFMS. The slopes found by Zahid et al. (2012) and Renzini & Peng (2015) are flatter than our results. Nevertheless, Duarte Puertas et al. (2017) applied aperture correction based on CALIFA data (Iglesias-Páramo et al. 2016) to recover the total flux from SDSS fiber spectroscopy and found a slope of 0.935, which is very close to our slope, which we obtained with the SF sample in the $0 < z \leq 0.35$ redshift range (see Fig. 4.14). Shin et al. (2021) obtained a flatter slope than we did based on galaxies from the Subaru Deep Field at intermediate redshift ($0.1 < z \leq 0.5$). However, we recovered a slope that is marginally consistent with the one found by Boogaard et al. (2018), who used data from the Multi Unit Spectroscopic Explorer (MUSE) and employed the same method as we used to fit the SFMS.

Intrinsic scatter

The amount of intrinsic scatter is hard to constrain because the scatter caused by the measurements errors in both the stellar masses and the SFRs needs to be accounted for. As pointed out by Boogaard et al. (2018), this is one of the advantages of using the fitting model of Robotham & Obreschkow (2015). We obtained an intrinsic scatter of 0.20 dex for the SF sample ($0 < z \leq 0.35$). This is consistent with previous works, which found values ranging from 0.15 up to 0.5 dex (see, e.g., Whitaker et al. 2012; Salmi et al. 2012; Speagle et al. 2014; Schreiber et al. 2015; Ilbert et al. 2015).

Many factors than can impact the amount of intrinsic scatter. First of all, different SFR indicators account for variations in the SFH on different timescales (see, e.g., Davies et al. 2016, and references therein). For instance, while Hα provides a direct measure of the current SFR in galaxies ($< 10-20$ Myr), UV-like tracers can detect changes in the SFH in only the last 100 Myr and are therefore less sensitive to recent episodes in the SFH that enhance or suppressed the star formation in the galaxy. Secondly, the selection criteria that defined the SFMS can boost or decrease artificially the scatter by excluding or including a fraction of galaxies that 'belong' or not to the SFMS.

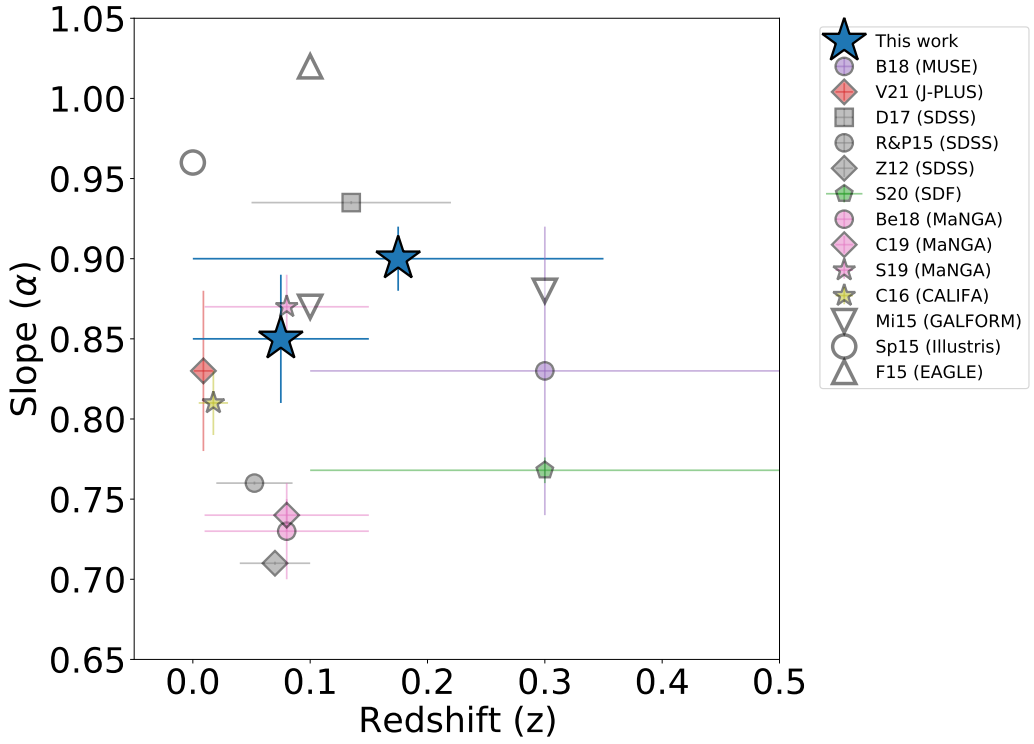


Figure 4.14: Slope of the SFMS derived from the Hα emission line by different works as a function of the redshift. The bars on the x-axis represent the redshift range of the galaxies involved in each study. Our best fit of the SFMS is shown with large blue stars for the lowest redshift range ($0 < z \leq 0.15$) and the SF sample ($0 < z \leq 0.35$). The results of the literature are from Boogaard et al. (2018) (B18), Vilella-Rojo et al. (2021) (V21), Duarte Puertas et al. (2017) (D17), Renzini & Peng (2015) (R&P15), Zahid et al. (2012) (Z12), Shin et al. (2021) (S20), Belfiore et al. (2018) (Be18), Cano-Díaz et al. (2019) (C19), Sánchez et al. (2019) (S19), and Cano-Díaz et al. (2016) (C16). We also include the results derived by GALFROM (a semianalytical model) (Mitchell et al. 2014) (Mi15), and from hydrodynamical simulations, Sparre et al. (2015) (Sp15) and Furlong et al. (2015) (F15).

The results obtained in each redshift bin show a decrease in intrinsic scatter for galaxies with higher redshift. The MC approach predicts σ_y to be compatible with zero in the last redshift bin. This might be the effect of the method. When we averaged over all galaxies in Eq. 4.8 and solved for σ_y , we found σ_y = 0.19, 0.09, and 0.17 dex for $0 < z \leq 0.15$, $0.15 < z \leq 0.25$, and $0.25 < z \leq 0.35$, respectively. However, we found a very similar value for the SF sample of galaxies (0.22 dex). As we pointed out in section 4.5.4, the selection function in the SF sample together with the low statistics in each redshift bin might affect the results.

SFMS with `BaySeAGal`

The SED fitting performed by `BaySeGal` yields the SFH of galaxies, and therefore we can estimate the current SFR in each galaxy by summing all the mass that formed stars in the last 30 Myr. Since tau-delayed models cannot account for a bursty SFH, any value between 10 to 200 Myr provides essentially the same SFR. A comparison of the results of the SFMS derived from the flux of Hα with a different and independent technique provides valuable information about the potential inaccuracies and strengths of our method.

In Fig. 4.15 we show the SFMS for the same sample of galaxies described in section 4.5.1 that is plotted in Fig. 4.12. The color code now represents the EW of Hα. As expected, galaxies with higher values in the EW of Hα are placed above the main sequence. This suggests that the two methods are consistent overall. Nevertheless, we obtained a zeropoint that is higher, meaning that the SFR derived from the analysis of the stellar populations gives higher values on average. This discrepancy later translates into the cosmic SFR density and the number of ionizing photons. In section 4.6.3, we discuss the possible origin of this difference in detail.

We obtain a slope that is slightly flatter, but still closer to what we retrieved with Hα. The different assumptions made by each method mean that this difference is expected. While the Hα flux is very sensitive to recent changes in the star formation activity of a galaxy, the SFR derived from the SED fitting traces the SFR on longer timescales. As a consequence, recent episodes that enhance or suppress the SFR might result in a global change in slope with respect to an SFMS derived from the average SFR over the last 200 Myr.

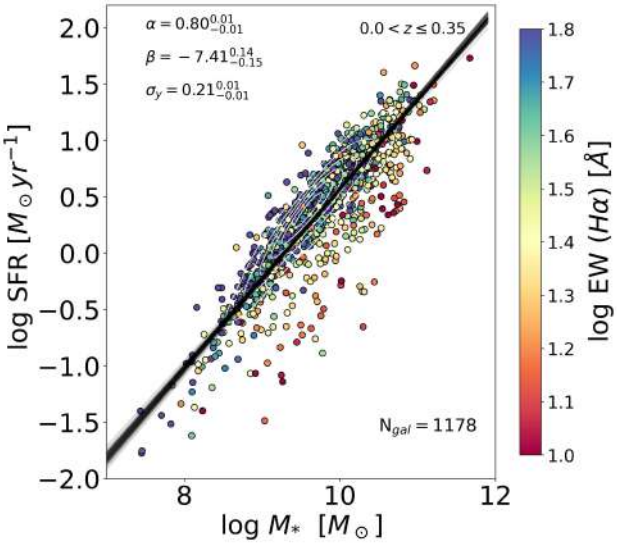


Figure 4.15: SFR vs. stellar mass for the galaxy sample described in section 4.5.1. SFRs are derived from `BaySeGal`. Galaxies are color-coded with the EW of Hα. Black lines are the best fits obtained with the Bayesian routine. The median posterior value and the 1σ confidence interval are shown for each of the parameters.

4.6.2 Cosmic evolution of the star formation rate density

The star formation rate density of the universe has been estimated by different means. Galaxy redshift surveys found that $\rho_{\rm SFR}$ peaks at $\sim$ 3.5 Gyr after the Big Bang ($z \sim 2$) and has decreased ever since (e.g., Gunawardhana et al. 2013; Sobral et al. 2013; Madau & Dickinson 2014; Driver et al. 2018). A similar trend was confirmed with galaxies in the nearby Universe using the so-called fossil record method (López Fernández et al. 2018; Sánchez et al. 2019; Bellstedt et al. 2020). Very recently, González Delgado et al. (2021) employed this method to derived the $\rho_{\rm SFR}$ from a subsample of galaxies in miniJPAS ($0.05 \leq z \leq 0.15$). The agreement with cosmological surveys is remarkable, even though different SED-fitting codes were used. In this section, we estimate the $\rho_{\rm SFR}$ from the SFR derived with the flux of Hα at the same redshift bins as described in section 4.5.4.

The miniJPAS area comprises only 0.895 deg^2 of the central regions of the AEGIS field. Therefore, our cosmological volume is somewhat limited, especially at low redshift. In this regard, a study of the $\rho_{\rm SFR}$ using miniJPAS data maybe affected by cosmic variance effects (Driver & Robotham 2010; Moster et al. 2011). The main source of uncertainty of $\rho_{\rm SFR}$ comes from this effect. We followed Eq. 4

in Driver & Robotham (2010) to quantify the cosmic variance of miniJPAS at different redshift bins,

$$\begin{aligned}\zeta_{\text{Cos. Var.}}(\text{per cent}) = [1.00 - 0.03\sqrt{A/B - 1}] \\ \times [219.7 - 52.4\log(AB \times 291.0)] \\ +3.21\log(AB \times 291.0)^2/\sqrt{NC/291.0},\end{aligned} \tag{4.11}$$

where N is the number of fields observed by miniJPAS (simply one), A and B are the median transverse lengths, and C is the radial depth. We obtained a cosmic variance for the comoving number density of galaxies of 37% (0.16 dex), 27% (0.12 dex), and 21% (0.09 dex) for the volumes within $0 < z \leq 0.15$, $0.15 < z \leq 0.25$, and $0.25 < z \leq 0.35$, respectively. In the future, J-PAS will scan $\sim$ 8000 deg^2 in the northen sky, and the effect of cosmic variance will be negligible (less than 1%).

In order to estimate ρ_{SFR} , we computed the total sum of the SFR for the galaxies in our sample and divided it by the volume contained in each redshift bin (V_{int}). We selected them from the parent sample with the same criteria as we used in section 4.5.1 to generate the SF sample. However, we relied on the Ke01 curve to exclude AGNs. We found a total of 1361 galaxies. In this way, we ensured that we did not underestimate ρ_{SFR} by excluding objects that lie between the Ke01 and Ka03 lines, which might contribute much to the flux of Hα through ionized interstellar gas. In any case, the difference between selecting SF galaxies with the Ka03 or the Ke01 line is only 0.05 dex in $\log \rho_{\text{SFR}}$.

The photometric depth of miniJPAS prevents us from detecting a fraction of galaxies below a certain mass limit. This effect becomes stronger for galaxies at higher redshift. Therefore, we have to apply volume corrections to reduce the impact of the lack of low-mass galaxies in the highest resdshift bins in this chapter. We used the classical V_{int}/V_{max} technique described originally in Schmidt (1968) and Huchra & Sargent (1973), (see Appendix C in Vilella-Rojo et al. 2021, for a detailed discussion of this correction). This is formally expressed as:

$$\rho^{int}_{\text{SFR}} = \sum_{i \in j} \frac{\text{SFR}_i}{V_{int}} w_i, \tag{4.12}$$

where $w_i = V_{int}/V^{max}_i$ is the weight that each galaxy has in the total ρ^{int}_{SFR} , and V_{max} is the maximum volume occupied by a galaxy assuming that it cannot be observed at a magnitude fainter than 22.7. For galaxies with $V_{int} \leq V^{max}_i$, the weight is simply one, but galaxies with $V_{int} > V^{max}_i$ will contribute more.

A direct comparison of $\rho_{\rm SFR}$ with the results obtained in González Delgado et al. (2021) also requires applying a correction to account for the galaxies that are detectable in the rSDSS band and are consequently fitted by the SED-fitting codes, but their emission lines cannot be measured because of the low S/N ratio. From the galaxies that belong to this group, we took those that were classified as blue by the color criterion and used their mass to place them in SFMS derived in section 4.5.3. In this way, we can estimate their SFR with Eq. 4.6 and add their contribution to $\rho_{\rm SFR}$. These corrections are indeed minor, as shown in Fig. 4.16 (red stars are the corrected values, and empty stars represent the uncorrected stars), but become slightly stronger at higher redshift.

In Fig. 4.16 we also show the values obtained by several studies that used the Hα flux to estimate the $\rho_{\rm SFR}$ at different redshift bins (squares, see references in Table 4.3). It is remarkable that most of them predict lower values of $\rho_{\rm SFR}$ than works that used the stellar continuum (solid line). Finally, black circles show the values obtained with the fossil record method by González Delgado et al. (2021) for miniJPAS galaxies in the range $0.05 < z \leq 0.15$.

Our results reproduce the $\rho_{\rm SFR}$ well that was found with other studies using Hα as a tracer to measure the SFR. Nevertheless, we found a non-negligible difference with respect to the results found by studies based on the stellar populations (Madau & Dickinson 2014; Driver et al. 2018; López Fernández et al. 2018; Sánchez et al. 2019; Leja et al. 2019; Bellstedt et al. 2020; González Delgado et al. 2021). Our estimation of $\rho_{\rm SFR}$ does not take the SFR into account that is ongoing in galaxies hosting an AGN.

4.6.3 Differences between the SFR derived through Hα and the SED fitting

The star formation rate density derived in this work is compatible with previous studies that used the Hα luminosity to determine its evolution with cosmic time in the nearby Universe. Nevertheless, our predictions are lower than those obtained with other methods based on the SED fitting of the stellar continuum. Even though $\rho_{\rm SFR}$ might be lower in the miniJPAS field, meaning we are affected by the large cosmic variance, our results differ from those derived with the analysis of the stellar populations in González Delgado et al. (2021).

In order to shed light on this difference, we compared the ionizing photon rates expected from Hα luminosity and from the SED fitting. When we assume that

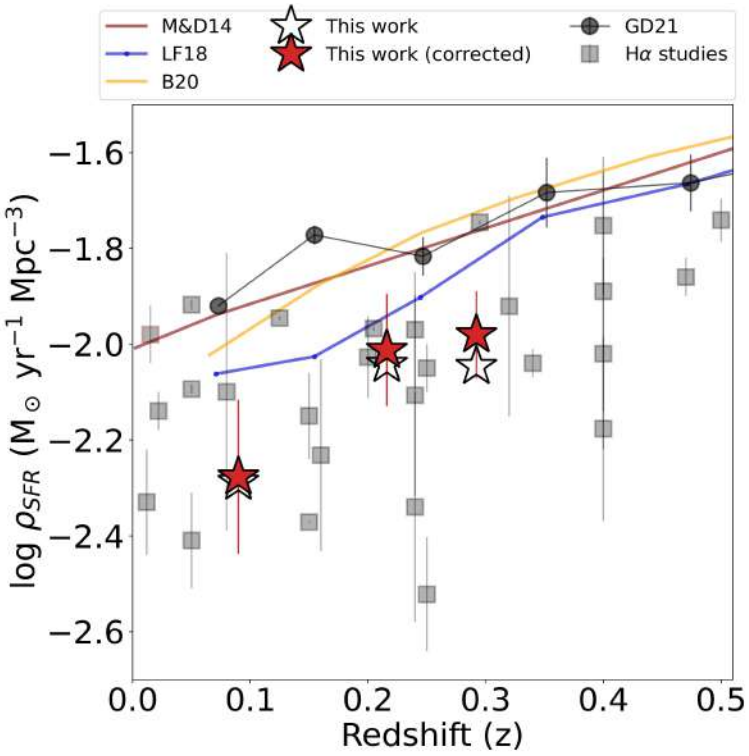


Figure 4.16: Star formation rate density at $z < 0.35$. Red stars show the values obtained in this chapter from the luminosity of Hα. Empty stars are uncorrected values that do not take galaxies with undetectable nebular emission lines or with very low S/N (see text in section 4.6.2) into account. Black circles are the values obtained by González Delgado et al. (2021) applying the fossil record method to a sample of miniJPAS galaxies in the range $0.05 < z \leq 0.15$. Squares are studies based on Hα (see references in Table 4.3). Solid lines represents the trends obtained by different studies based on the stellar continuum: Madau & Dickinson (2014, M&D14) , López Fernández et al. (2018, LF18), and Bellstedt et al. (2020, B20). All values are scaled to the Chabrier (2003) IMF.

References	**Redshift**	**log $\rho_\star$**
Gallego et al. (1995)	0.022	-2.14 ± 0.04
Ly et al. (2007)	0.08	-2.01 ± 0.29
	0.24	-2.34 ± 0.24
	0.4	-2.02 ± 0.20
Shioya et al. (2008)	0.24	-1.97 ± 0.12
Dale et al. (2010)	0.16	-2.23 ± 0.20
	0.24	-2.11 ± 0.21
	0.32	-1.92 ± 0.23
	0.40	-1.89 ± 0.25
Westra et al. (2010)	0.05	-2.41 ± 0.10
	0.15	-2.15 ± 0.09
	0.25	-2.05 ± 0.05
	0.34	-2.04 ± 0.03
Drake et al. (2013)	0.25	-2.52 ± 0.12
	0.4	-2.18 ± 0.19
	0.5	-1.74 ± 0.05
Sobral et al. (2013)	0.40	-1.75 ± 0.15
Gunawardhana et al. (2013)	0.05	-1.92 ± 0.06
(GAMA)	0.125	-1.95 ± 0.06
	0.205	-1.97 ± 0.09
	0.295	-1.75 ± 0.09
Gunawardhana et al. (2013)	0.05	-2.01 ± 0.06
(SDSS)	0.15	-2.37 ± 0.09

Stroe & Sobral (2015)	0.2	-2.03 ± 0.09
Van Sistine et al. (2016)	0.015	-1.98 ± 0.06
Khostovan et al. (2020)	0.47	-1.86 ± 0.04
Vilella-Rojo et al. (2021)	0.012	-2.34 ± 0.11
This work	0.09	-2.28 ± 0.16
	0.216	-2.02 ± 0.11
	0.292	-1.98 ± 0.09

Table 4.3: Compilation of star formation rate densities derived from Hα. All values are scaled to Chabrier (2003) IMF. $\log \rho_\star$ is in units of $M_\odot \mathrm{yr}^{-1}\mathrm{Mpc}^{-3}$.

no photons escape from H II regions, the relation between the dust-corrected luminosity of Hα and the ionizing photon rates is

$$Q_H^{H\alpha} = x_{H\alpha} \frac{L_{H\alpha}}{h\nu_{H\alpha}}, \tag{4.13}$$

where $x_{H\alpha} = 2.206$ for case B hydrogen recombination.

In the case of the SED fitting, `BaySeAGal` provides the mass fraction (μ_j) of each SSP that better describes the observed spectrum. In other words, for each galaxy, we can reproduce the SFH. Therefore, we can retrieve the ionizing photon rates by weighting the number of H ionizing photons emitted per unit time and initial mass for the jth SSP ($q_{H,j} = q_H(t_j, Z_j)$),

$$Q_H^{SFH} = M_\star \sum_{j=1}^{221} \mu_j q_{H,j}. \tag{4.14}$$

We compare the two quantifies in Fig. 4.17. Q_H^{SFH} is 0.54 dex higher than $Q_H^{H\alpha}$ on average. We observe a clear trend with the nebular extinction (color bar) and the EW of Hα. Galaxies where we estimated low values of the nebular extinction lie farther away from the 1:1 line. On the same line, the differences between Q_H^{SFH} and $Q_H^{H\alpha}$ become smaller as the EW of Hα increases.

Interestingly, Q_H^{SFH} and $Q_H^{H\alpha}$ are closer at higher values. This trend has also been found in comparisons between the SFR derived from Hα and from the UV

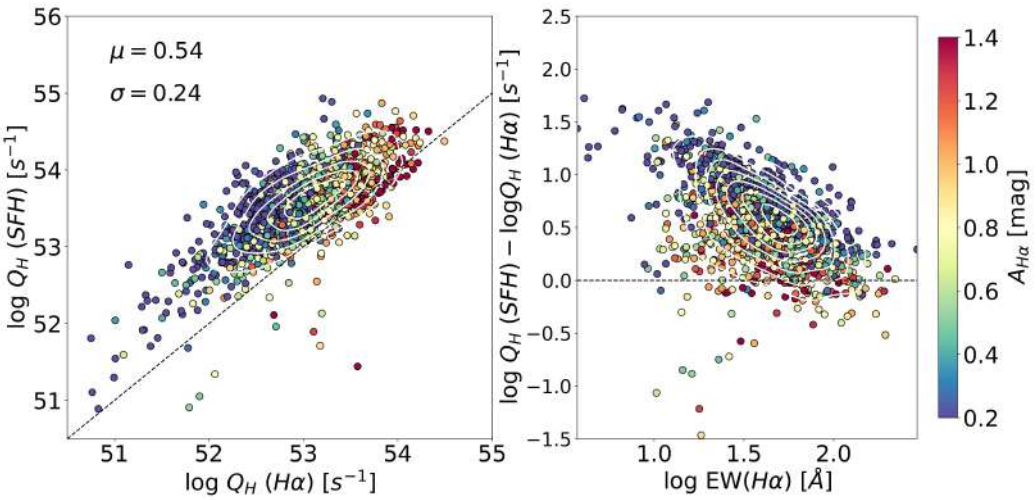


Figure 4.17: Comparison of the ionizing photon rates computed from Hα emission line and from the fit obtained with the analysis of the stellar populations with `BaySeAGal` (left; see text in section 4.6.3). The dashed black line represents the 1:1 relation. μ and σ are the bias and the standard deviation. The right panel shows the difference between these quantities as a function of the EW of Hα. Density contours are drawn in black. In both cases, the galaxies are color-coded with the extinction of the interstellar gas calculated from the Balmer decrement.

both in the integrated spectrum and in spatially resolved galaxies (Lee et al. 2009, 2016; Byun et al. 2021). Specifically, Byun et al. (2021) concluded that deficient Hα fluxes in the extended disks of galaxies are tightly correlated with recent starbursts, which are being rapidly suppressed over the last 10 Myr. This phenomenon can explain the difference found in the slope of the SFMS in section 4.6.1. Because galaxies with a low Hα luminosity have higher SFRs according to the SED fitting, the slope becomes flatter.

Q_H^{SFH} might also be overestimated if the mass fraction attributed to young stellar populations (YSP) were higher than it should be. This might happen if the SFH in the last 20 Myr were different from the global SFH that accounts for the formation and growth of mass in galaxies on scales of billion years and/or because our parametric code overestimated the fraction of mass that formed in recent epochs with respect to nonparametric codes that are more flexible to varying the fraction of the young stellar population on a shorter timescale. In order to determine how our result might be affected by different assumptions of the SFH, we used the SFH from ALSTAR and computed Q_H^{SFH}. We found that there is a bias of 0.81 dex, which is even higher than the results found with `BaySeAGal`.

Studies that retrieved the stellar population properties of a sample of galaxies based on optical spectra (either form SDSS or CALIFA) and based on photometry from the GALEX survey showed that when the UV part of the spectrum is not included in the SED fitting, a brighter YSP contribution is found (López Fernández et al. 2016; Werle et al. 2019). However, this excess of light in the UV does not have a strong impact on the mass content of YSP because the mass is dominated by older stars.

`BaySeAGal` does not yet include a model of nebular emission lines. Therefore, the SED fitting only accounts for the emission of the stellar continuum and masks the filters in which the emission lines peak. We do not know how this might affect the shape of the SFH and the mass fraction attributed to the young stellar population. Moreover, a delta-delayed model might not be sufficient to describe SFHs with a recent burst of SFR. In the future, we expect to explore this aspect further.

Furthermore, other hypotheses need to be taken into account to explain this discrepancy. First, we should consider whether we underestimate the nebular extinction. Certainly, we would expect that galaxies with very low S/N show this effect more. When we rebuild Fig. 4.17 and include only galaxies with an error in Hα luminosity smaller than 0.25 dex, the bias decreases by 0.17 dex. Additionally, when we assume for the SF sample that the nebular extinction is underestimated by a factor of two, which would mean $E(B-V)_{H\alpha/H\beta} \sim 2E(B-V)_{SED}$, as some studies reported (Qin et al. 2019; Koyama et al. 2019), the difference would only be reduced by 0.22 dex. In other words, it is plausible that we did not properly estimate the nebular extinction for a fraction of galaxies in the SF sample, but in the worst scenario, this effect alone cannot explain the difference between Q_H^{SFH} and $Q_H^{H\alpha}$.

Another effect that might also contribute to this difference is the ionizing radiation that leaks from the H II regions. In this case, Eq. 4.13 would underestimate the Hα ionizing photon rates. Several studies have shown precisely that there is a fraction of ionizing photons that escapes, and they are therefore unable to ionize the interstellar gas (Giammanco et al. 2005; Otí-Floranes & Mas-Hesse 2010; Pellegrini et al. 2012; Anderson et al. 2015). Nevertheless, the average fraction is still debated and can vary from galaxy to galaxy and from region to region within the same galaxy. Unfortunately, there is no means to quantify this effect with the data employed in this work. Nonetheless, when we assume that 30 % of the ionizing radiation leaks from H II regions, the difference could be reduced 0.16 dex.

Most probably, the difference that we observe between Q_H^{SFH} and $Q_H^{H\alpha}$ is a combination of all these factors. Certainly, fitting the SED of miniJPAS galaxies

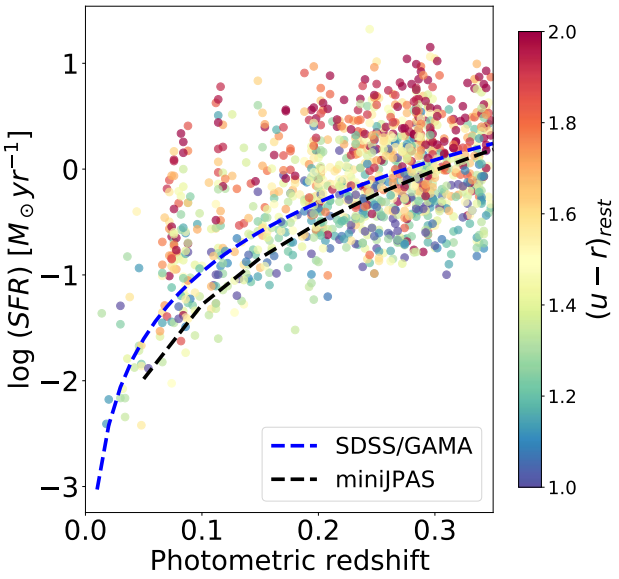


Figure 4.18: Relation between the SFR derived from Hα and redshift for the galaxy sample described in section 4.5.1. The blue dotted line is the approximate SFR completeness limit for GAMA and SDSS galaxies (Gunawardhana et al. 2013), and the dotted black line is the 95 % completeness limit from blue galaxies in miniJPAS. Galaxies are color-coded with their (u-r) rest-frame color.

with information from the UV from GALEX or HST-UV observations and/or the IR from *SPITZER* would be very useful to unveil the origin of the discrepancy and test some of the previous hypotheses. However, this analysis is not the main goal of this chapter.

4.7 Outlook for J-PAS

The results presented in this chapter prove that the main properties of ELGs can be studied with J-PAS data. The miniJPAS Pathfinder instrument allowed us to test and combine different methods of analysis to fully exploit the scientific potential of the data and draw the baseline for the prospect of J-PAS.

The vast amount of data to be collected by J-PAS will allow us to perform a more comprehensive research, exploring other aspects that remained elusive or were limited within the area covered by miniJPAS. For instance, we will be able to derive the properties of blue and SF galaxies in groups and clusters, the fraction of AGN, and their role in the quench of SF galaxies within dense and very low

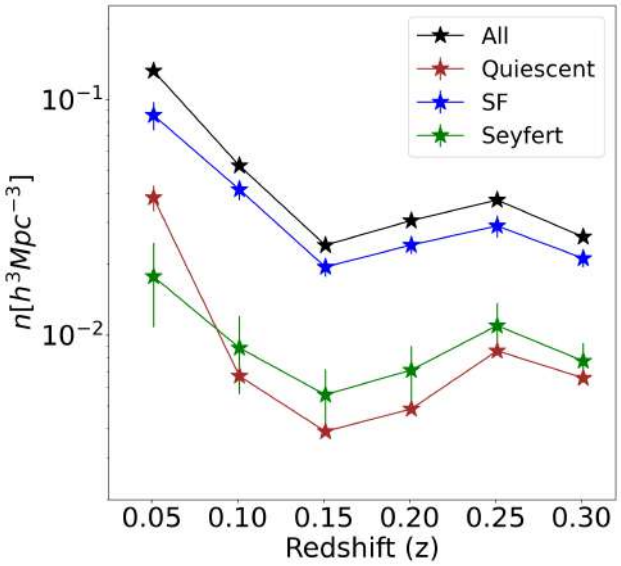


Figure 4.19: Comoving number density of galaxies in miniJPAS as a function of redshift. The total galaxy population (black star) is broken into star-forming (blue stars), AGN-like (green stars), and quiescent galaxies (red stars). We used the WHAN diagram with the Ka03 dividing line to separate AGN and SF galaxies. Quiescent galaxies include LINERs and passive galaxies. The uncertainty due to the cosmic variance is not included in the error budget.

density environments.

For instance, if in 1 deg^2 we were able to estimate the position of 255 galaxies in the BPT with an error smaller than 0.15 dex, the ionization mechanism of about two million galaxies in the Universe ($z < 0.35$) could be studied at the end of the J-PAS survey. With this amount of data, we will be able to determine the SFMS parameters better and place constraints on the evolution of $\rho_{\rm SFR}$ at least up to 0.35 in redshift. Thus, it will be possible to further explore the discrepancies found in section 4.6.2.

The SFR coverage of J-PAS will be at least as competitive as that of the SDSS or GAMA surveys. In Fig. 4.18 we show the SFR as a function of the resdshift for our SF galaxy sample. The dotted blue line is the approximate SFR completeness limit assuming a flux limit of $F_{H\alpha} = 10^{-18} Wm^{-2}$ for GAMA and SDSS galaxies (Gunawardhana et al. 2013). The dotted black line represents the 95 % completeness limit of miniJPAS for blue galaxies (Díaz-García et. al. in prep). We used the best fit obtained in section 4.5.3 to transform the completeness limit in mass into SFR.

Finally, in Fig. 4.19 we show the comoving number density of galaxies in miniJPAS as a function of redshift for the total galaxy population (black stars) for the star-forming galaxies (blue stars), for AGN-like galaxies (green stars), and for quiescent galaxies (red stars). Error bars represent the variation in the number density when a different division line in the WHAN diagram is considered, for example, k03, Ke01, or S08.

4.8 Summary and conclusion

We analyzed a subsample of galaxies (a total of 2154) from the AEGIS field observed by miniJPAS with redshift below 0.35 in detail. The method we developed make used of ANN trained with CALIFA and MaNGA in order to predict and detect the main emission lines in the J-spectrum: Hα, Hβ, [O III], and [N II].

We used a criterion based on the mass and color of the galaxy. We estimated that 83 % and 17 % in the sample are blue and red galaxies, respectively. With the ANN classifier, which is based on the EW of the emission lines, we found that 82 % of the sample are strong ELs amd 18 % are weak ELs.

We employed the BPT and WHAN diagrams to classify galaxies according to the main source of ionization and to select star-forming galaxies. We obtained that of the galaxies with reliable EW values (2000 galaxies in total), 72.8 ± 0.4 %, 17.7 ± 0.4 %, and 9.4 ± 0.2 % are SF, Seyfert, and passive or LINER galaxies, respectively, using the WHAN diagram and the Ka03 separation line. One hundred and fifty-four galaxies from the parent sample remain unclassified because of high uncertainties in the predictions of the emission lines. Ninety-four percent of the SF galaxies and 97 % of the LINER or passive galaxies are classified with the color criterion as blue and red, respectively.

The analyses of the properties of the stellar population performed in González Delgado et al. (2021) allowed us to compare and complement the information of the emission lines. For instance, we showed in color-mass diagrams that blue (red) galaxies are composed of a younger (older) stellar population, respectively, and present stronger (weaker) emission lines. This synergy between the properties of the gas and the stellar populations also appears in the BPT diagram, where galaxies become more massive as they evolve through the SF-wing.

We derived the SFR from the flux of Hα and relied on the Balmer decrement to correct for the extinction produced by interstellar dust. Subsequently, we fit the slope, zeropoint, and the intrinsic scatter of the SFMS, obtaining $0.90^{+0.02}_{-0.02}$, $-8.85^{+0.19}_{-0.20}$ and, $0.20^{+0.01}_{-0.01}$, respectively. We tested the turnover-mass hypothesis

by fitting a quadratic and a broken power law. However, we did not observe a flattening of the slope at high mass. We argue that this is likely produced by our selection criteria of SF galaxies together with the limitation of the method to detect very weak emission lines in comparison with spectroscopic surveys. The results we obtained are compatible with those of other studies.

Finally, we computed the cosmic evolution of the $\rho_{\rm SFR}$ within three redshift bins: $0 < z \leq 0.15$, $0.15 < z \leq 0.25$, and $0.25 < z \leq 0.35$. We found agreement with previous measurements based on the Hα emission line. Nevertheless, we found an offset compared to the studies that derived $\rho_{\rm SFR}$ from the SED fitting of the stellar continuum. We discussed the origin of this discrepancy in detail, which is most probably a combination of several factors, such as the correction for dust attenuation, the SFR tracer, or the escape of ionizing photons.

The work presented in this chapter builds the foundation upon which the analysis of ELGs in J-PAS will be conducted as soon as hundreds of squares degrees are mapped in the northern sky in the next years.

Chapter 5

Quasar selection in the AEGIS field with artificial neural networks and hybridisation

This chapter is based on the article:

The miniJPAS survey quasar selection II: Classification with artificial neural networks and hybridisation
by G. Martínez Solaeche, Carolina Queiroz, R. M. González Delgado, Natália V. N. Rodrigues, R. García-Benito et. al

Submitted to MNRAS on 25th July 2022

5.1 Introduction

The new era in modern astronomy goes hand in hand with the era of big data. Astronomical observations produce increasingly larger amounts of data. The new generation of surveys such as the Dark Energy Spectroscopic Instrument (DESI, Levi et al. 2013), the Large Synoptic Survey Telescope (LSST, Ivezić et al. 2019) or the Square Kilometer Array (SKA, Dewdney et al. 2009) will observe of the order of millions or even billions of objects. In particular, the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS, Benitez et al. 2014) will observe in 54 narrow-band filters thousands of deg^2 in the northen sky in the upcoming years, detecting more than 40 million objects. Consequently, it is necessary to automatise as much as possible all the tasks so as to process faster and more efficiently the astronomical information. Certainly, the identification and classification of astronomical objects is the first step prior to any further scientific analysis.

Traditionally, photometric surveys identified galaxies and stars based on their morphological structure and colour properties (see e.g. Baldry et al. 2010; Henrion et al. 2011; Saglia et al. 2012; López-Sanjuan et al. 2019). Typically, galaxies are extended objects while point-like sources are mainly either stars or quasi-stellar objects (quasars). Nevertheless, the lack of spatial resolution for the most distant and faint galaxies makes them look very similar to point-like sources. Furthermore, the colour space in multi-band photometric surveys becomes more and more complex as the number of filters increases, which makes it necessary to employ more sophisticated algorithms to fully exploit all the information encoded in it.

In the last years, machine learning (ML) algorithms have been used in plenty of applications within the astronomical field. From the estimation of photometric redshifts (Cavuoti et al. 2017; Pasquet et al. 2019; Ramachandra et al. 2022), the identification of low-metallicity stars (Whitten et al. 2019), the determination of the star formation rate (Delli Veneri et al. 2019; Bonjean et al. 2019), the classification of morphological types in galaxies (Domínguez Sánchez et al. 2018), the identification of causality in galaxy evolution (Bluck et al. 2022) to the measurement of the equivalent widths of emission lines in photometric data (Martínez-Solaeche et al. 2021). The problem of source identification in photometric surveys has also been addressed by ML either to distinguish between point-like and extended sources (Vasconcellos et al. 2011; Kim et al. 2015; Kim & Brunner 2017; Burke et al. 2019; Baqui et al. 2021) or even between galaxies, stars and quasars (Krakowski et al. 2016; Bai et al. 2019; Logan & Fotopoulou 2020; Xiao-Qing &

Jin-Meng 2021; He et al. 2021).

The goal of the this chapter is to classify the objects detected with the miniJPAS survey (Bonoli et al. 2021) into stars, galaxies, quasars at high redshift ($z \geq 2.1$), and quasars at low redshift ($z \leq 2.1$) by using artificial neural networks (ANN). The threshold at $z = 2.1$ corresponds to the limit at which quasars show the Lyman-α emission line within the J-spectra. The miniJPAS survey is part of the J-PAS project[1], which detected more than 60 000 objects within the All-wavelength Extended Groth Strip International survey (AEGIS; Davis et al. 2007) using 56 narrow-band J-PAS filters ($\sim$ 145 Å) and the four *ugri* broad-band filters. The separation of 100 Å among filters makes the J-PAS filter system equivalent to obtaining a low resolution spectrum with $R \sim 60$ (J-spectrum hereafter). Such unique characteristics make possible to observe and analyse galaxies and quasars in continuous redshift ranges, $0 \lesssim z \lesssim 1$ and $0.5 \lesssim z \lesssim 4$, respectively (Bonoli et al. 2021). In fact, different studies proved the capability of J-PAS to address several topics within the astrophysical field, e.g. the evolution of the stellar population properties of galaxies up to $z \sim 1$ (González Delgado et al. 2021), the properties of the nebular emission lines of galaxies down to $z \leq 0.35$ (Martínez-Solaeche et al. 2022), the measurement of black hole virial masses for the quasar population (Chaves-Montero et al. 2021) or the study of galaxy properties within galaxy clusters (Rodríguez Martín et al. 2022) and groups (González Delgado et al. 2022). Unfortunately, the data available for spectroscopically confirmed sources in the miniJPAS area are not sufficient to train and test ML algorithms for the present purpose. Therefore, we employ mock data developed by Queiroz et al. (2022), and use as a truth table the sources identified spectroscopically within the AEGIS field by the Sloan Digital Sky Survey (SDSS, York et al. 2000) in the DR12Q superset catalogue (Pâris et al. 2017).

Modern deep ANN are generally showing better performance than traditional methods. Still, most of the time they remain poorly calibrated (Goodfellow et al. 2014; Guo et al. 2017). The probabilities associated with the predicted label classes may suffer from overconfidence, as they do not correspond to true likelihoods. Consequently, objects are classified as part of one class or another with high probability regardless of the prediction accuracy. Realistic probability distributions are particularly important for spectroscopy follow-up programs which typically prioritise the observation of high probability objects of some particular class. Furthermore, some objects are indeed dual in its nature. Although we consider as quasars only those galaxies with an extremely luminous active galactic

[1] http://www.j-pas.org

nucleus (AGN), there are objects in which a significant fraction of the detected light comes from the stars in the host galaxy. In this scenario, ML algorithms should ideally provide a high probability in both classes.

In a recent paper, Zhang et al. (2017) proposed an original idea called `mix up` to enlarge the dataset and improve the generalisation of the trained model, thus increasing the robustness to adversarial examples. Latter on, Thulasidasan et al. (2019) showed that `mix up` or hybridisation, as we prefer to call it, also improves the calibration and predictive uncertainty of deep neural networks. In this work, we enlarge our training set by mixing features from stars, galaxies and quasars and we study the effect of hybridisation in both the performance and the calibration of the models. The ML classifiers used in this chapter will be combined with other ML algorithms. In Rodrigues et al. (in prep.) we trained convolutional neural networks (CNNs) proposing different approaches to incorporate photometry uncertainties as inputs in the training phase. Furthermore, we compare the performance of the CNNs with respect to decision tree algorithms. While in this work we focus our attention on the galaxy-quasar degeneracy and the ability of ANN to estimate realistic probability density distributions (PDF), in Rodrigues et al. we study other relevant aspects, e.g. the stellar types that are more frequently confused with the quasar population, the J-PAS feature importance or the stability of the CNNs predictions with respect to minor changes in the training set. Besides, in Pérez-Ràfols et al. (in prep. a) we adapt SQUEZE (Pérez-Ràfols et al. 2020) to work with J-PAS data, a code based on optical emission line identification to separate between quasars and non-quasars that also estimates the redshift of quasars. Ultimately, all these codes will be merged in a combined algorithm (Pérez-Ràfols et al. in prep. b) so as to classify more efficiently the miniJPAS sources and provide a high-redshift quasar target list for a spectroscopic follow-up with the WEAVE multi-object spectrograph survey (Dalton et al. 2014), which is planning to carry out a Lyman-α forest and metal line absorption survey (Pieri et al. 2016).

This chapter is organised as follows. In section 5.2, we present miniJPAS data, and we briefly summarise the processes employed in the construction of the mock catalogue. In section 5.3, we describe in detail the main characteristics of the classifiers, and how data augmentation is employed through hybridisation. We indicate the performance metrics used for testing purposes along the chapter in section 5.3.4, and we show the main results obtained in this chapter in section 5.4 and section 5.5. Finally, we summarise and conclude in section 5.6. Throughout the chapter, all magnitudes are presented in the AB system (Oke & Gunn 1983).

5.2 The miniJPAS survey and mocks

The miniJPAS survey includes data from four pointings scaning $\sim$ 1 deg^2 along the AEGIS field. The photometric system includes 56 bands, namely 54 narrow-band filters in the optical range plus two medium-bands – one in the near-UV (uJAVA band) and another in the NIR (J1007 band). With a separation of $\sim$ 100 Å, each narrow-band filter has a full width at half maximum (FWHM) of $\sim$ 145 Å, whereas the FWHM of the uJAVA band is 495 Å and J1007 is a high-pass filter. Additionally, four broad-bands u,g,r, and i were used to complement the observations. These were carried out with the 2.55 m telescope (T250) at the Observatorio Astrofísico de Javalambre, a facility developed and operated by CEFCA, in Teruel (Spain). The data were acquired using the pathfinder instrument, a single CCD direct imager ($9.2k \times 9.2k$, $10\mu m$ pixel) located at the centre of the T250 FoV with a pixel scale of 0.23 arcsec pix^{-1}, vignetted on its periphery, providing an effective FoV of 0.27 deg^2. The r band has been chosen as the detection band and the reference image in the 'dual-mode' catalogue. This image is used to define the position and sizes of the apertures from which the rest of the photometry is extracted. The miniJPAS survey is 99% complete up to $r \leq 23.6$ mag for point-like sources, and detected more than 60 000 objects[2] (Bonoli et al. 2021). In this chapter, we only analyse objects with `FLAGS`= 0 and `MASK_FLAGS`= 0 (46441 in total), i.e. they are free from detection issues such as contamination from bright stars, light reflections in the telescope or in its optical elements, etc. We refer the reader to Bonoli et al. (2021) for details on the flagging scheme. Removing flagged sources from the catalogue decreases our sample size but it does not introduce any bias, since the fraction of sources that are flagged is independent of their magnitude (Hernán-Caballero et al. 2021).

The algorithms presented in this chapter are trained and tested on mock data (Queiroz et al. 2022). The J-spectra of galaxies, stars and quasars are simulated by convolving SDSS spectra included in the SDSS DR12Q Superset catalogue (Pâris et al. 2017) with the transmission profiles of J-PAS photometric system (synthetic fluxes). The SDSS DR12Q Superset contains all objects targeted as quasars from the final data release of the Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013). Therefore, it contains also galaxies and stars whose broad-band colours are compatible with those from quasars. Since the SDSS sample is complete only up to $r \sim 20.5$, fainter objects are generated by adding random fluctuations (noise) to the synthetic fluxes. The mock catalogue

[2] http://archive.cefca.es/catalogues/minijpas-pdr201912

Sample	Galaxies	Stars	Quasars	All
Training	10^5	10^5	10^5	3×10^5
Validation	10^4	10^4	10^4	3×10^4
Test	10^4	10^4	10^4	3×10^4
1-deg^2	6410	2190	510	9110

Table 5.1: Number of objects in each data set contained in the mock catalogue.

includes several noise models that mimic the observed S/N in miniJPAS for the `APER_3` magnitude aperture[3]. We use model 1, which assume the noise distribution of miniJPAS point-sources in each filter is well described by a single Gaussian distribution (Queiroz et al. 2022). Nevertheless, the main results presented in this chapter are not affected by the noise model choice. In Pérez-Ràfols et al. in prep. b we will compare the performance of each algorithm in the mock test sample with different noise models and we will classify sources in miniJPAS for each one of them. Galaxies follow the magnitude-redshift distribution of SDSS and DEEP3 (Cooper et al. 2011, 2012) found in miniJPAS. Quasars follow the luminosity function of Palanque-Delabrouille et al. (2016), and stars are distributed according to the Besançon Model of stellar population synthesis of the Galaxy (Robin et al. 2003) and the SDSS miniJPAS spectroscopic sample.

The number of stars, galaxies and quasars are balanced to prevent biases in the classifiers towards over-represented classes. The same applies for the test set and the validation set, which is used to fine tune the hyper-parameter of the classifiers. Additionally, another test sample with the expected number of point-like sources within the miniJPAS area is generated to provide a more direct comparison. In Table 5.1 we summarise the number of objects contained in the mocks. Further details on how these synthetic data have been created can be found in Queiroz et al. (2022).

[3]This is the magnitude obtained within a three arcsec-aperture.

5.3 Star/galaxy/quasar classifier

In this section we describe in detail how the ANN classifiers were developed. These have been designed to distinguish between four different classes: stars, galaxies, quasars at high redshift ($z \geq 2.1$), and quasars at low redshift ($z < 2.1$) referred to as QSO-h and QSO-l, respectively.

5.3.1 Artificial neural networks

In this chapter, we used ANN coded with `Tensorflow` (Abadi et al. 2015) and `Keras` libraries (Chollet et al. 2015) in `Python`. The ANN has eight hidden layers with 200 neurons each. As a regularisation technique, we use weight constraints and impose a maximum value of two in each neuron (kernel constraint). We also drop out 15 % of the neurons in each layer. We use the Rectified Linear Unit (ReLU) as our activation function (Nair & Hinton 2010). Weights are initialised with the He initialisation strategy (Géron 2019). The loss function employed is the cross entropy.

We trained two models that use two different sets of inputs, dubbed as ANN_1 and ANN_2. The inputs of ANN_1 (59 in total) are relative fluxes, i.e. the flux in each filter is divided by the flux in the r band. Since the miniJPAS dual-mode catalogue used the r-band for detections, this normalisation is well defined for all the objects. The inputs of ANN_2 are the colours measured with respect to the m_{AB} in the r band plus the normalised magnitude in this band (60 inputs in total):

$$\begin{aligned} \mathrm{ANN}_1^j &= \frac{f_\lambda^j}{f_\lambda^r} \\ \mathrm{ANN}_2^j &= m_{AB}^j - m_{AB}^r, \qquad \mathrm{ANN}_2^r = \frac{m_{AB}^r - \max(m_{AB}^r)}{\min(m_{AB}^r) - \max(m_{AB}^r)} \end{aligned} \tag{5.1}$$

where m_{AB} and f_λ stand for the magnitude and the flux in the j-th filter, respectively, and $\min(m_{AB}^r)$ and $\max(m_{AB}^r)$ are the minimum and maximum values of the magnitude in the r-band within our training set. Both sets of inputs capture the shape of the spectrum but the ANN_2 has also information about the observed luminosity of each source, which anchors the SED to a particular magnitude.

Objects in the dual mode catalogue might be undetected in some bands (non-detection). This happens when the S/N values for these bands are very low and the measured fluxes are null or negative after the sky background subtraction. In the mock catalogues non-detection (ND) follows the pattern observed in miniJPAS.

For specific details to see how ND are modelled we refer the reader to Queiroz et al. (2022). We set to zero the inputs of the ANN_2 if the fluxes are negative because colours are otherwise undefined. However, we allow the inputs of the ANN_1 to be below zero. Such small differences might help the ANN_1 to better modelling the sky background.

5.3.2 Data augmentation via hybridisation

Data augmentation has been proven to be an excellent tool to increase the size of the training sample and consequently the performance of ML algorithms when only a limited training sample are available (Shorten & Khoshgoftaar 2019). Rotation, translation or scaling are among the most popular techniques for image classification (Yang et al. 2022). In the case of non-image features such as the J-spectra, the most common manner to perform data augmentation is via Gaussian noise. However, the benefit of this technique in our training sample would be limited because it was already used to generate objects at different magnitudes bins in the construction of the mock catalogue. Thus, we adapt the `mix up` technique proposed in Zhang et al. (2017) to our classifiers that aim to distinguish between four classes. This technique allows us to enlarge the training set by mixing features from different classes generating a new training set composed only of hybrid objects. The new set of hybrid objects (y_i^H) and their respective fluxes are generated as a linear combination of individual objects in the original training set:

$$\mathbf{y_i^H} = \alpha_i \mathbf{y_i} + \sum_{j=1}^{4} c_{ij}(1-\alpha_j)(1-\delta_{ij})\mathbf{y_j} \tag{5.2}$$

$$\mathbf{f_i^H}(\lambda) = \alpha_i \mathbf{f_i}(\lambda) + \sum_{j=1}^{4} c_{ij}(1-\alpha_j)(1-\delta_{ij})\mathbf{f_j}(\lambda) \tag{5.3}$$

where $\mathbf{y_1}$ ($\mathbf{f_1}(\lambda)$), $\mathbf{y_2}$ ($\mathbf{f_2}(\lambda)$), $\mathbf{y_3}$ ($\mathbf{f_3}(\lambda)$) and, $\mathbf{y_4}$ ($\mathbf{f_4}(\lambda)$) are the vectors (fluxes) of each one of the classes (stars, galaxies, QSO-l, QSO-h, respectively), α_j is the mixing coefficient which varies between zero and one according to an exponential distribution function which depends on *beta*, the scale parameters that control the level of mixing ($f(x;\beta) = 1 - (1/\beta)\exp(-x/\beta)$). Finally, δ_{ij} is the Kronecker delta, and $c_{ij} = N_j/(N_{\rm tot}-N_i)$ where N_i is the number of objects belonging to class i within each magnitude bin in the original training set, and $N_{\rm tot} = N_1+N_2+N_3+N_4$. For instance, if we generate an hybrid galaxy ($\mathbf{y_2^H}$) at magnitude $m^{\rm r}_{AB} = 22.5$,

Eq. 5.3 becomes:

$$\mathbf{f_2^H}(\lambda) = \alpha_2 \mathbf{f_2}(\lambda) + (1 - \alpha_2)(c_{21}\mathbf{f_1}(\lambda) + c_{23}\mathbf{f_3}(\lambda) + c_{24}\mathbf{f_4}(\lambda)). \quad (5.4)$$

For low values of β, α_2 is near one with a high probability. Therefore, the new hybrid galaxy is still a galaxy but it has some of level of contamination from the other classes. The probability of not being a galaxy $(1 - \alpha_2)$ is distributed among the other classes taking into account their relative amounts at $m^{\rm r}_{AB} = 22.5$. Since stars are less frequent at such magnitude c_{21} is near zero and the new hybrid galaxy is mixed mainly with QSO-l and QSO-h. In order to compute the c_{ij} coefficients, we split the training set in rSDSS magnitude bins that contain roughly 20 000 objects. In this way, only objects with similar brightness are mixed together. We enlarge the training set five times with $\beta = 0.1$ (we will discuss later this choice in Sec. 5.4.1). We warn the reader that hybridisation does not mix objects following a physical recipe but it is rather a mathematical transformation of the data. The resulting hybrid fluxes are normalised following Eq. 5.1.

5.3.3 Training strategy

Usually, the intrinsic randomness of the training procedure leads to solutions that are not optimal. Weights and biases are drawn from a distribution function that generates the initial state. Therefore, each time that the training is performed, the algorithm converges to a different local minimum of the loss-function. Furthermore, the training set itself is augmented in a random manner via hybridisation. The 'hybrid' space is filled in a slightly different way in each realisation. In the limit case where the hybrid set is much larger than the original one such effect would be negligible. However, a huge training set is less practical to handle and more difficult to train than a smaller one. For all these reasons, we followed the committee approach (Bishop 1995), i.e. we train several ANNs and compute the median to provide a final classification. Then, we renormalise the output probabilities to ensure that the sum is one. In order to find the optimal number of ANN or committee members needed, we started with two and we added an additional one at each step until the results did not improve for the validation sample in terms of the f_1 score (see section 5.3.4 for a definition of this metric). We dertermined that eight members are enough to reach convergence.

5.3.4 Performance metrics

We discuss here the metrics used to evaluate the performance and robustness of the classifiers.

Confusion matrix

The confusion matrix is especially useful in the context of a multi-label classification problem. The actual classification of each object is shown in the columns of the matrix while the predicted ones lie in the rows. Therefore, in the best (ideal) case scenario the matrix would be purely diagonal with every prediction coinciding with the actual classification. Non-diagonal terms indicate which classes are confused between each other, and provide a valuable information so as to improve the training set and fine tune the hyper-parameters of the model.

f1-score

Unlike the confusion matrix, the f1-score yields one single scalar for each one of the classes. It finds a compromise between purity (precision) and completeness (recall):

$$\text{Purity} = \frac{TP}{TP + FP}, \qquad \text{Completeness} = \frac{TP}{TP + FN} \tag{5.5}$$

$$f_1 \text{ score} = \frac{2 \cdot \text{Purity} \cdot \text{Completeness}}{\text{Purity} + \text{Completeness}} \tag{5.6}$$

where TP, FP and, FN are the true positive rate, the false positive rate and, the false negative rate, respectively. FP appears as non-diagonal terms in the columns of the confusion matrix, while TN lies on the rows. In the case of an unbalanced test set with one or more classes underrepresented, the average performance of the model can be estimated using the weighted f_1 score, i.e.

$$f_1^W \text{ score} = \sum_{i=0}^{N_{\text{class}}} \frac{n_i f_1^i \text{ score}}{n_{\text{obj}}} \tag{5.7}$$

where n_i is the number of objects belonging to the i-th class in the test set, and n_{obj} is the total number of objects.

Expected Calibration Error

A well calibrated probabilistic classifier is able to predict probabilities that coincide on average with the fraction of objects that truly belong to a certain class (accuracy). Let us suppose we take one hundred objects with probability 10% of being a star, if the classifier is well calibrated, about ten of them should actually be stars. If there are more, that means our classifier is under-confident, if there are less, then the classifier is over-confident. Probability calibration curves are normally employed to display this relationship, where we bin the probability estimates and plot the accuracy versus the mean probability in each bin. Let B_{mj} be the set of objects whose predicted probabilities of being class j fall into bin m. The accuracy and confidence of B_{mj} are defined as:

$$acc(B_{mj}) = \frac{N_{mj}}{|B_{mj}|} \tag{5.8}$$

$$conf(B_{mj}) = \frac{1}{|B_{mj}|} \sum_{i \in B_{mj}} P_{ij} \tag{5.9}$$

where P_{ij} is the probability of being class j for the i-th object, and N_{mj} is the number of objects of class j within bin m. The Expected Calibration Error (ECE) is then defined as:

$$\text{ECE} = \frac{1}{N_c} \sum_{j=0}^{N_c} \sum_{m=0}^{N_b} \frac{|B_{mj}|}{N_b} |acc(B_{mj}) - conf(B_{mj})| \tag{5.10}$$

where N_c is the number of classes and N_b the number of bins. The lower the value of the ECE, the better is the calibration of the model. However, the output of the ANN only represent true probabilities under the assumption that there is not essential difference between our mock sample and miniJPAS observations. In order to have a better estimate of the ECE of the model we would need to compute this metric in sufficiently large true table which is not possible yet. In the future, when more data will be gather, we will be able to employ this metric on observations and evaluate properly the ECE of the ANN. In the remaining of this chapter we will refer to the outputs of the ANN as probabilities keeping in mind they are simply a proxy of the true probabilities.

Entropy

The entropy is a measurement of disorder. In the context of ML, the entropy of a classifier's prediction can tell us how uncertain the classifier is. The entropy of

the i-th object can be written as follow:

$$S_i = -\sum_{j=0}^{N_c} P_{ij} \log_2(P_{ij} + \epsilon) \tag{5.11}$$

where P_{ij} is the probability that the i-th object is class j, N_c is the number of classes, and ϵ is an arbitrary small number (10^{-14}) to avoid the divergence of the logarithm in case the probability for a given class is zero. The entropy is maximum ($\log_2 N_c$) if each class has a probability of $1/N_c$ and zero if the probability of belonging to a particular class is one.

5.4 Results

In this section we test the performance of the algorithms on the mock test sample (section 5.4.1). We discuss in detail the effect of augmenting the data through hybridisation and we compare the differences between classifiers. Finally, we evaluate the classification obtained with SDSS objects observed by miniJPAS in the AEGIS field.

5.4.1 Test sets

The performance of a classifier changes as a function of the magnitude of the objects. Fainter objects are more difficult to classify because of their lower S/N. Hence, in order to quantify the potential bias of the classifiers at different magnitudes we split both the validation and the test samples in three different bins according to the r-band magnitude:

$$\begin{aligned} &\text{BIN 0} : 17 < r \leq 20 \\ &\text{BIN 1} : 20 < r \leq 22.5 \\ &\text{BIN 2} : 22.5 < r \leq 23.6 \end{aligned} \tag{5.12}$$

The number of objects in the test sample for BIN 0, 1, and 2 are 5002, 13376, and 9436, respectively. In Fig. 5.1 we show the f_1^W score for each of the magnitude bins defined above including the average performance for the full sample (ALL BIN). We compare the score of the ANN trained with the hybrid set (ANN_1 mix and ANN_2 mix) and with the original training set (ANN_1 and ANN_2). In Fig. 5.2 the f_1 score is also shown for each of the classes. Overall, both classifiers (ANN_1

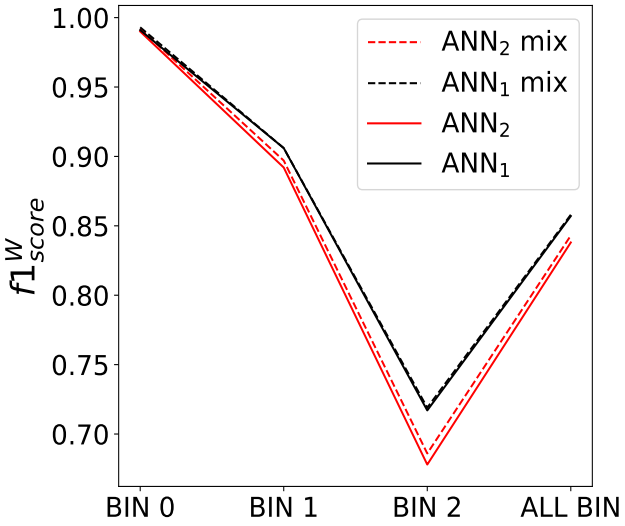


Figure 5.1: f_1^W score for different magnitude bins as defined in Eq. 5.12, and for the full sample (ALL BIN). Dashed (solid) lines represent the models trained with the hybrid (original) training set. ANN$_2$ and ANN$_2$ mix are trained with colours while ANN$_1$ and ANN$_1$ mix are trained with fluxes (see section 5.3.1).

and ANN$_2$) are very similar with small differences in each magnitude bin for each class. The fact that the ANN$_1$ classifier is slightly better might be related not only to the representation of the data (relative fluxes) but with the fact that it can capture better the sky background.

As expected, the accuracy of the classifiers decreases for fainter objects. The performance obtained with the hybrid set is very similar compared to the original training set, suggesting that the latter already contains all the variance needed, and more examples do not necessarily imply a better performance (but see the next section for a more detailed discussion).

In Fig. 5.3 we show the weighted f_1^W-score as a function of the median S/N ratio in the observed filters for the ANN$_1$ and the ANN$_2$. Each bin contains roughly 1000 objects. It is remarkable that even with a median S/N of 5 the f_1^W-score reaches 0.9.

The confusion matrices as a function of the magnitude bin for the ANN$_1$ model are shown in Fig. 5.4. Those from the remaining models are provided in Appendix C. QSO-l and galaxies are more difficult to distinguish between each other, especially at the faint end, where data are noisier. In fact, these objects do not belong to independent classes. Sometimes, the host galaxy of an AGN might have an

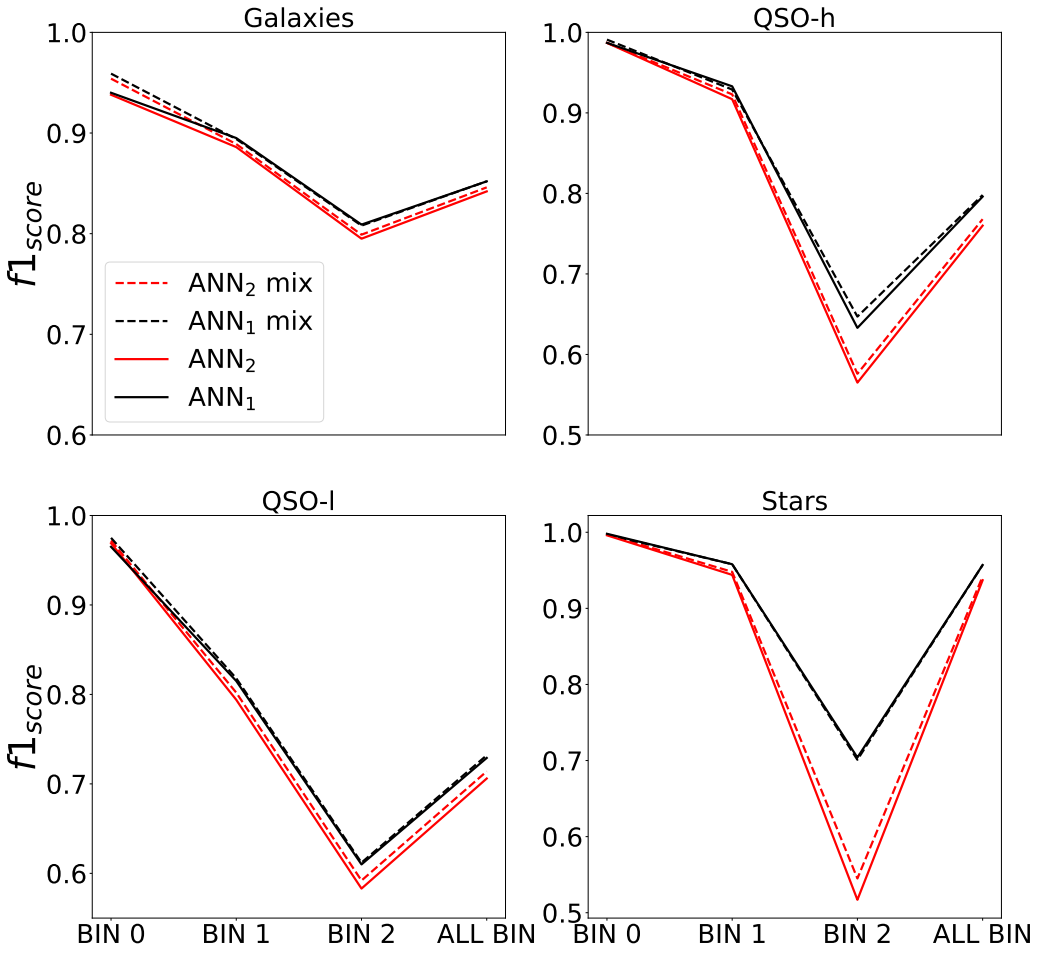


Figure 5.2: f_1 score for each of the classes: galaxies, QSO-h, QSO-l and, stars as a function of the magnitude bins defined in Eq. 5.12, and for the full sample (ALL BIN). Dashed (solid) lines represent the models trained with the hybrid (original) training set. ANN$_2$ and ANN$_2$ mix are trained with colours while ANN$_1$ and ANN$_1$ mix do with fluxes (see section 5.3.1)

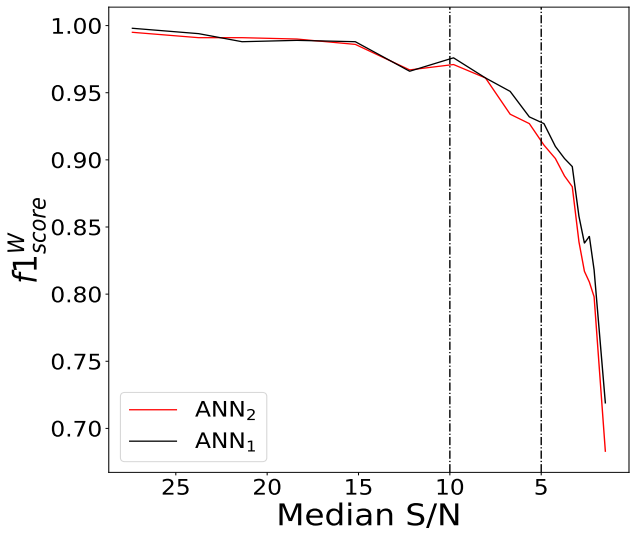


Figure 5.3: $^W_1$-score obtained with the ANN$_1$ and ANN$_2$ as a function of the median S/N. Dashed vertical lines indicate a S/N of 10 and 5.

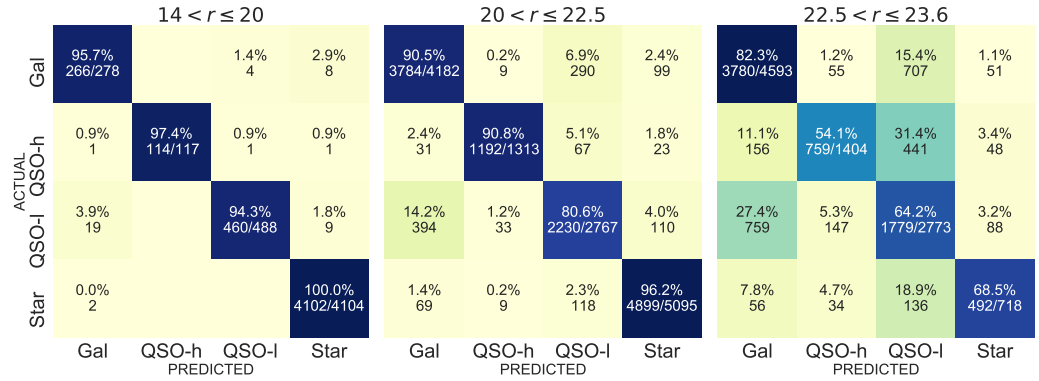


Figure 5.4: Confusion matrices obtained with the ANN$_1$ in the test sample.

important contribution to the SED. In Seyfert galaxies, the observed spectrum is usually a combination of thfe light coming from the AGN and the stellar populations within the galaxy. Therefore, we expect to have confusion between QSO-l and galaxies more often than between any of the other classes. Finally, in the faintest magnitude bin, 31.4 % of QSO-h are classified as QSO-l, and 18.9 % of the stars are confused with QSO-l.

In Fig. 5.5 we show examples of the most common missclassifications. The first row is composed of QSO-l that were classified as galaxies. In the second row we show galaxies that were identified as QSO-l, the third row corresponds to QSO-h confused with QSO-l, and the last row shows Stars classified as QSO-l. Despite of being unable to correctly predict the class of these objects, the second most likely class usually coincide with the actual class. Furthermore, it is important to emphasise that objects shown in Fig. 5.5 would be very difficult to identify via visual inspection even for a human expert. ML algorithms are indeed pushing the limits beyond the human capability.

It it expected that the ANN predictions for low S/N objects is more uncertain than the ones with high S/N. In Fig. 5.6 we show the median entropy as a function of the median S/N in bins of 1000 objects. While the predictions are very certain in the high S/N limit with the ANN_1 and ANN_2 classifiers, the entropy obtained by the ANNs trained in the hybrid sets remains almost constant from a S/N of 25 to 10 with a value of ~ 0.4 and then it increases slightly. In fact, the lowest entropy obtained with the ANN_1 mix and ANN_2 mix classifiers is governed by the mixing coefficient (α) used to generate the hybrid set in Eq. 5.3 and coincides with the median entropy of the hybrid classes in the training set.

In Fig. 5.7 we show the fraction of positive for each one of the classes as a function of the mean probability obtained with the ANNs in the mock test sample ($r \leq 23.6$). On the top (bottom) left panel we show the results of the ANN_1 (ANN_2) predictions trained with the original training set while on the right panel we show the predictions obtained with the hybrid set. The ECE for galaxies, QSO-h, QSO-l, and stars are shown on the top-left. Training with hybrid classes has a negative impact on the calibration. Once again, the ECE is a function of the mixing coefficient: as α increases the ECE increases. Overall, the ANN_1 is slightly better calibrated than the ANN_2, but ANN_2 mix is better than ANN_1 mix.

Finally, It is worth considering whether hybridisation improves the performance of the ANNs when the training set is smaller in size. The original training set in the mock catalogue is composed of 300 000 objects, and the hybrid set is five times larger. Now, let us assume that we have a training set ten times smaller than the original one (reduced set). After applying hybridisation we generate two new

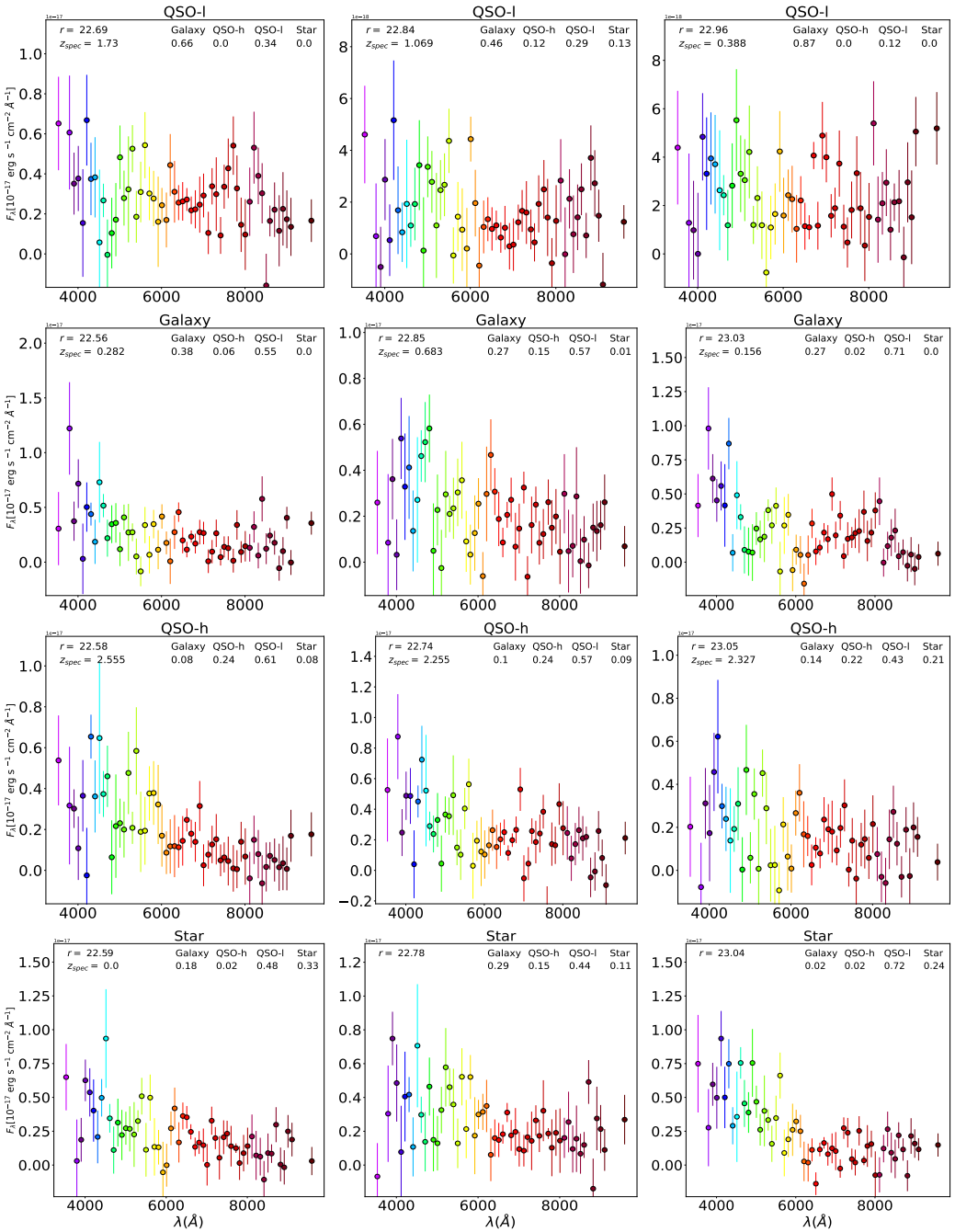


Figure 5.5: Examples of the most typical miss classification (mock test sample). First row shows QSO-l classified as galaxies, second row galaxies classified as QSO-l, third row QSO-h classified as QSO-l, and fourth row Star classified as QSO-l. From left to right objects are fainter. We indicate the AB magnitude in the *r*-band, the redshift (top-left) and the probabilities yielded by the ANN_1 classifier for each one of the classes (top-right).

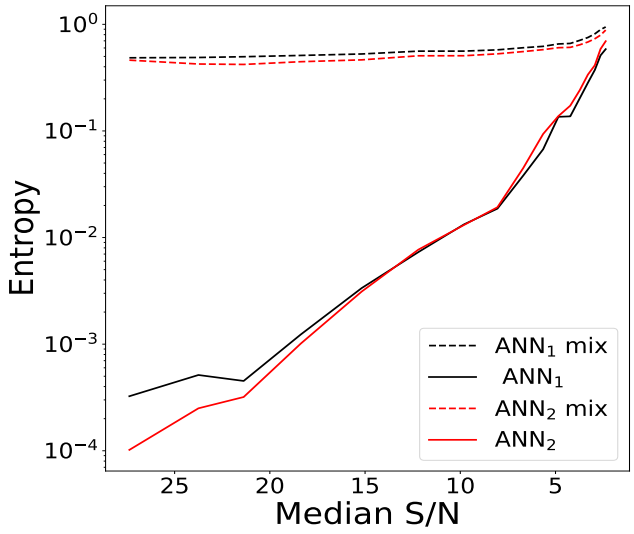


Figure 5.6: Median entropy as a function of the median S/N in bins of 1000 objects..

training sets that are five and ten times larger than the reduced set, respectively, known as the 'reduced hybrid set x5' and the reduced hybrid set x 10. Then, let us compare the performance of ANN_1 in the mock test sample. In Fig. 5.8 we show the difference between the f_1^W-score in each one of the mentioned training sets and the f_1^W-score obtained in the original training set as a function of the median S/N. We do not observe a significant improvement that might justify the use of hybridisation, at lest in the form we implemented in Eq. 5.3 for this particular data set.

5.4.2 SDSS versus miniJPAS

In this section we test the ANN classifiers on the SDSS test sample. Fig. 5.9 shows the f_1 score for each class and the f_1^W score. The performance of the algorithms trained with the hybrid set (ANN_1 mix and ANN_2 mix) are compared with the original training set (ANN_1 and ANN_2). Due to the limited number of objects we did not separate these samples in magnitude bins. Most of the objects (75%) belong to BIN 1 and only three are at the faint end (BIN 2). Therefore, the f_1 score is mostly representative of BIN 1. The results on the SSDS test sample are compatible with those obtained in the mock data (~ 0.9), suggesting that the simulations are reproducing fairly well the miniJPAS observations at least for

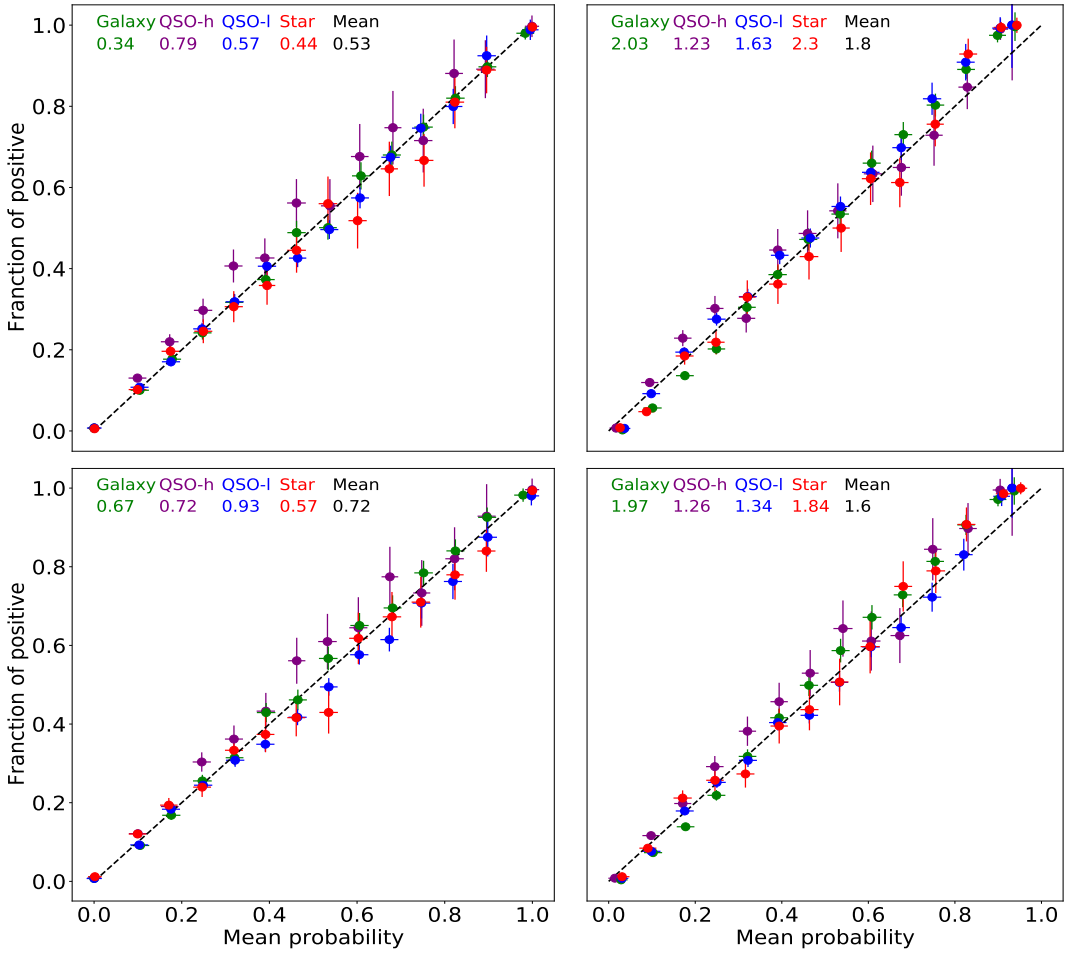


Figure 5.7: Fraction of positive for each one of the classes as a function of the mean predicted probability The ECE for galaxies, QSO-h, QSO-l, stars and the mean ECE are shown on the top left side of each panel. The errors bar represent the standard deviation in each one of the bins. Left panels are trained with the original training set while right panel used the hybrid set. Top (bottom) panels show the results of the ANN_1 (ANN_2).

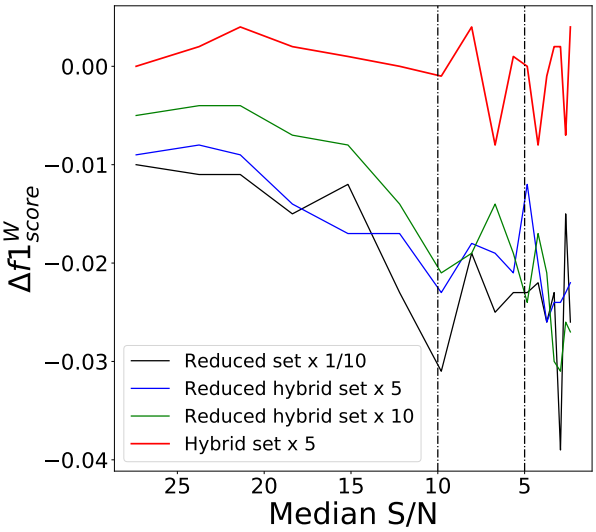


Figure 5.8: Difference between the f_1^W-score obtained with the ANN_1 trained in the original training set and the reduced set (ten times smaller), the hybrid set (five time larger), the reduced hybrid set x 5 (five times larger than the reduced set), and the reduced hybrid set x 10 (ten times larger than the reduced set) as a function of the median S/N. Dashed vertical lines indicate a S/N of 10 and 5.

magnitudes brighter than 22.5. Unfortunately, we do not have enough labeled objects fainter than 22.5 within the miniJPAS field, thus we need to rely on the mock results to give an expectation of the performance. As soon as WEAVE starts to observe the quasar target list provided by the J-PAS collaboration we will be able to fully asses the performance of the algorithms for the full range of magnitudes.

In Fig. 5.10 we show the confusion matrix obtained with ANN_1 for the SDSS test sample. The confusion matrices for the remaining models can be consulted in Appendix C. The sample of quasars predicted by the ANN and especially the subsample of QSO-h contain very few false positives (QSO columns in the confusion matrix), meaning the algorithms favour a pure rather than a complete sample. However, the sample of galaxies is more complete because $\sim$ 19 % of them are classified as QSO-l but only few SDSS galaxies are missing. Finally, stars are identified with very high accuracy with only five false positives and eight true negatives. We recall that these results are partially biased due to the small number of objects in it and the selection criteria that were used by the SDSS team to select the sample of quasar targets. Hence, it can only give us a a glimpse of the actual performance of the ANN in real data.

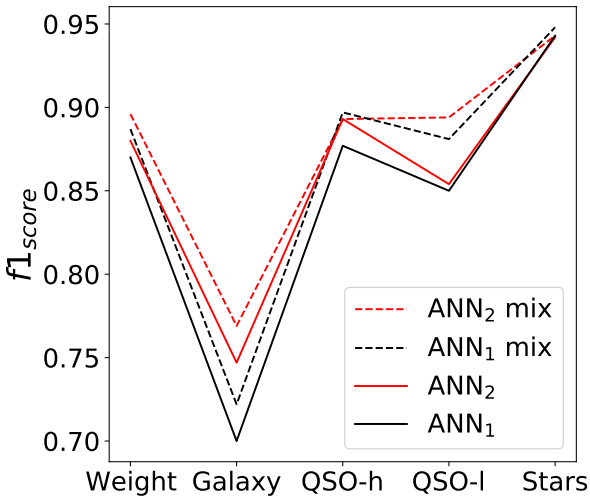


Figure 5.9: f_1^W score and f_1 score for each one of the classes obtained within the miniJPAS field observed and labeled by SDSS observations (see text in 5.4.2). Dashed (solid) lines represent the models trained with the hybrid (original) training set. ANN$_2$ and ANN$_2$ mix are trained with colours while ANN$_1$ and ANN$_1$ mix are trained with fluxes (see section 5.3.1). Note that the scales are different in y-axis for each class.

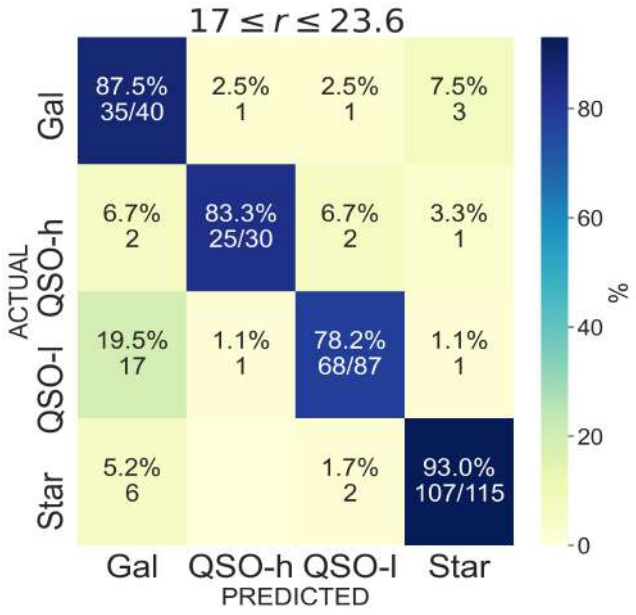


Figure 5.10: Confusion matrix obtained with ANN1 in the SDSS test sample.

In Fig. 5.11 we show some examples of objects observed simultaneously by SDSS and miniJPAS below redshift 1 that might present a spectrum composed of mixed features. In other words, the light coming from these objects has contributions from both the AGNs and the stellar populations within the galaxies. All the objects except 2470-3341 are classified in SDSS as QSO-l. However, `SExtractor` identified them as extended sources with a class-star value below 0.35 (CL in Fig. 5.11). Following the SDSS classification criterion, only 2470-3341 and 2241-1234 are correctly classified by the ANN_1. Nevertheless, 2406-15603 is rather a Seyfert 1 galaxy with broad emission lines such as Hα and Hβ. It also has a reddish spectrum and the extended structure of the galaxy can be seen very clearly in the image. Furthermore, while SDSS spectrum detects the broad emission line of Hα, the miniJPAS observation do not capture such emission, which is probably the most relevant feature to classify this object as a quasar. 2241-18615, 2406-2560, and 2406-7300 are classified as galaxies but the second preferred class is QSO-l. Indeed, those objects are not very different from 2470-3341, which is a Seyfert 2 galaxy according to the SDSS pipeline. Once again, the J-spectra miss two relevant features in 2241-18615 and 2406-2560, the Hα, and Hβ emission lines, respectively. Finally, 2241-1234 is correctly classified. Although it is an extended source according to `SExtractor`, the emission of the AGN dominates the spectrum. The high S/N obtained in this object makes the classification more certain.

5.5 miniJPAS quasar catalogue

We now focus our attention on the classifier predictions on miniJPAS data. In Fig. 5.12 we show the distribution of the confidence (probability) levels yielded by ANN_1 and ANN_2 classifiers for each class and each magnitude bin. We only predict the class for the objects that are considered point-like sources according to `SExtractor`. Both ANN classifiers predict roughly the same number of objects but they exhibit differences in the faintest magnitude bins, which are useful to build afterwards a combined algorithm that uses information from several classifiers (Pérez-Ràfols et al. in prep. b).

The total number of quasars predicted by ANNs for miniJPAS is compatible with previous estimates. Abramo et al. (2012) expected $\sim$ 240 quasars per square degree with the J-PAS photometric system for a limiting magnitude of g=23 assuming that quasars follow the luminosity function found by Croom et al. (2009) and the quasar selection is perfect. With the ANN_1 (ANN_2) we detect 163 (177)

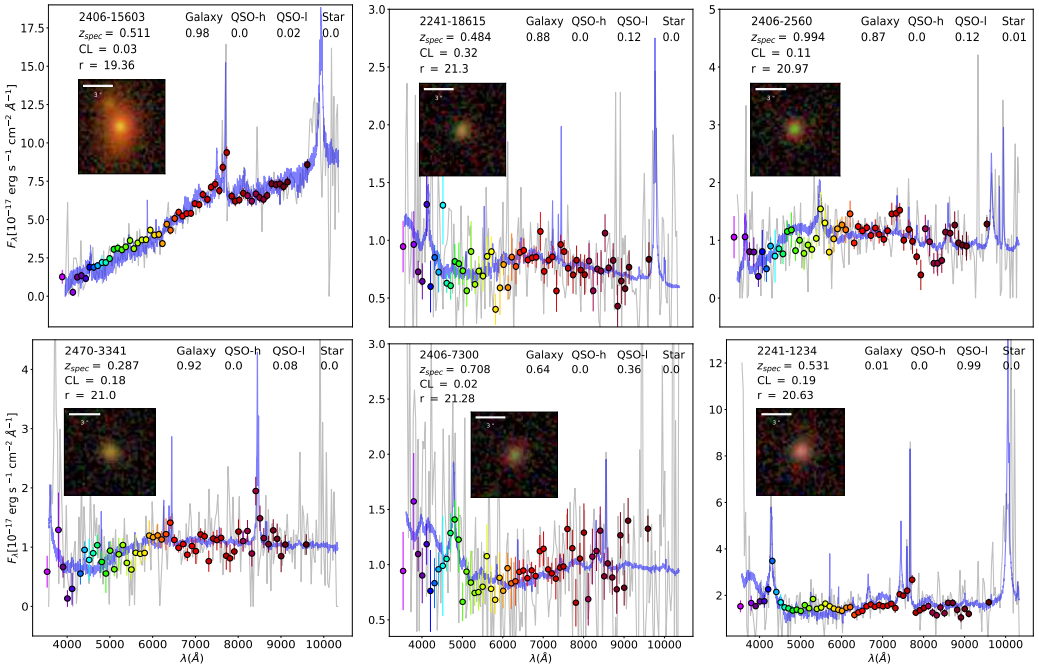


Figure 5.11: Seyfert galaxies observed both with miniJPAS and SDSS (see text in section 5.4.2). The SDSS spectra is scaled to match the miniJPAS *r*-band. Grey solid line represents the actual SDSS observation while blue line is a model developed by the SDSS team. We indicate in the legend the miniJPAS ID, the spectroscopic redshift of the object, the class-star yielded by `SExtractor` (CL), and the AB magnitude in the *r*-band. We also show the probabilities obtained by the ANN_1 for each one of the classes, and we attach a multi-colour RGB image centred on the object covering 6.5 arcsec across. All objects except 2470-3341 are classified by SDSS pipeline as quasars.

quasars with a probability greater than 0.5 and $g < 23$ in the point-like source catalogue (CL > 0.5). Even though 0.25 is the threshold to have an object classified as one particular class, we impose that the probabilities of being a quasar have to be greater than the contrary (P(QSO-h) + P(QSO-l) > P(Galaxy) + P(Star)), which implies P(QSO-h) + P(QSO-l) > 0.5. However, this is a conservative approach since we are only considering high probability objects. If instead we sum all over the probabilities of being quasar, we obtain 182.1 and 195.1 quasars with the ANN_1, and ANN_2, respectively.

In Fig. 5.13 we show the observed $(g - r)$ vs. $(u - g)$ colour-colour diagram for all the objects presented in the 1-deg^2 mock sample (first row) and in the miniJPAS observations (second row). The positions of quasars, stars and galaxies are consistent in each magnitude bin. We include all objects in miniJPAS with `FLAGS`= 0 and `MASK_FLAGS`= 0. There is a population of stars in the 1-deg^2 mock sample that are not present in miniJPAS observations (bottom left side in BIN 0 and 1). Those stars correspond to the most massive and bluest ones (O-type) which are usually found in regions of high activity of star-formation. The luminosity functions for O and B stars are estimated by extrapolating the prediction of the Besançon Model together with the stars within the miniJPAS SDSS Superset sample. Therefore, those populations might be overestimated in the 1-deg^2 mock sample. However, even if that were the case, the fraction of these stars is still low compared with those in the main sequence. Therefore, the impact that this effect has on a classifier whose main goal is to identify quasar candidates is very limited.

In the last row of Fig. 5.13 we colour-coded the miniJPAS observations with the CL probability. Quasars and stars predicted by the ANN_1 are classified by `SExtractor` as point-like sources (CL> 0.5) while galaxies are predicted as extended-sources (CL< 0.5). In the faintest magnitude bin extended and point-like sources are more difficult to distinguish in the colour space since they both overlap. The ratio between the number of point-like sources according to the CL and the ANN_1 ($R_{\mathrm{point}} = N_{\mathrm{point}}$ (CL) / N_{point} (ANN_1) for BIN 0, 1, and 2 are $R_{\mathrm{point}} = 0.93, 0.72, 0.18$, respectively, and the ratio between the number of extended sources are $R_{\mathrm{ext}} = 1.12, 1.06, 1.49$, respectively. We assume point-like sources to be quasars and stars while galaxies are considered extended sources. In summary, our predictions are in agreement with `SExtractor` considering that the performance of both classifiers decrease as a function of the observed magnitude. What is more, we are predicting on a sample (miniJPAs observations) that include extended sources while our training sample exclude these type of objects.

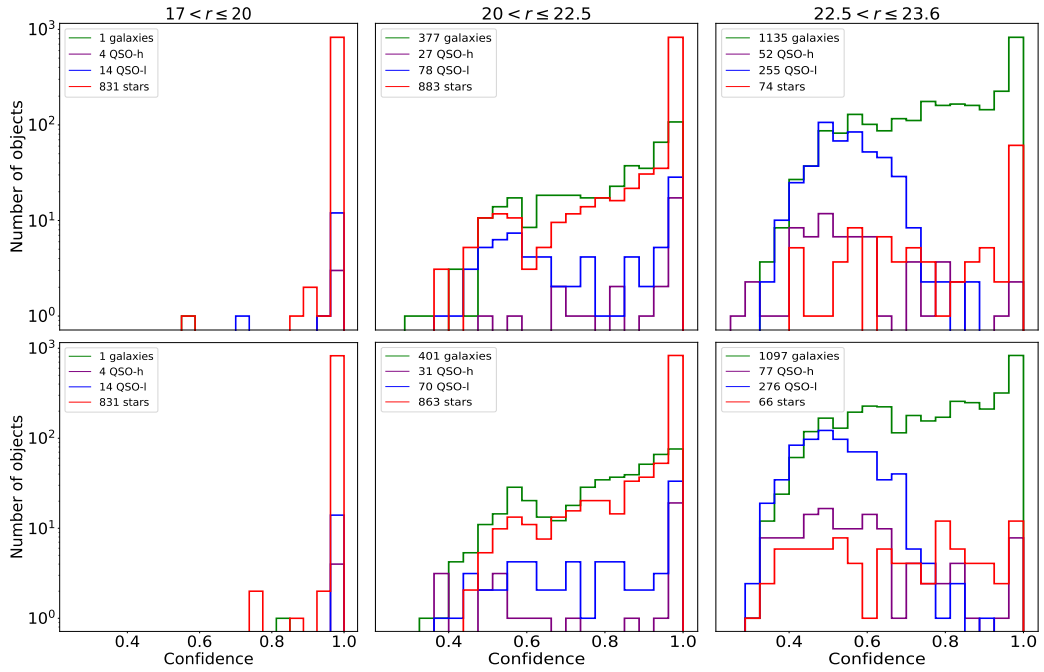


Figure 5.12: Confidence (probability) yielded by the ANN_1 (top) and ANN_2 (bottom) classifiers for each class and magnitude BIN in miniJPAS observations for point-like sources (CL > 0.5). The numbers of classified objects are shown in the legend.

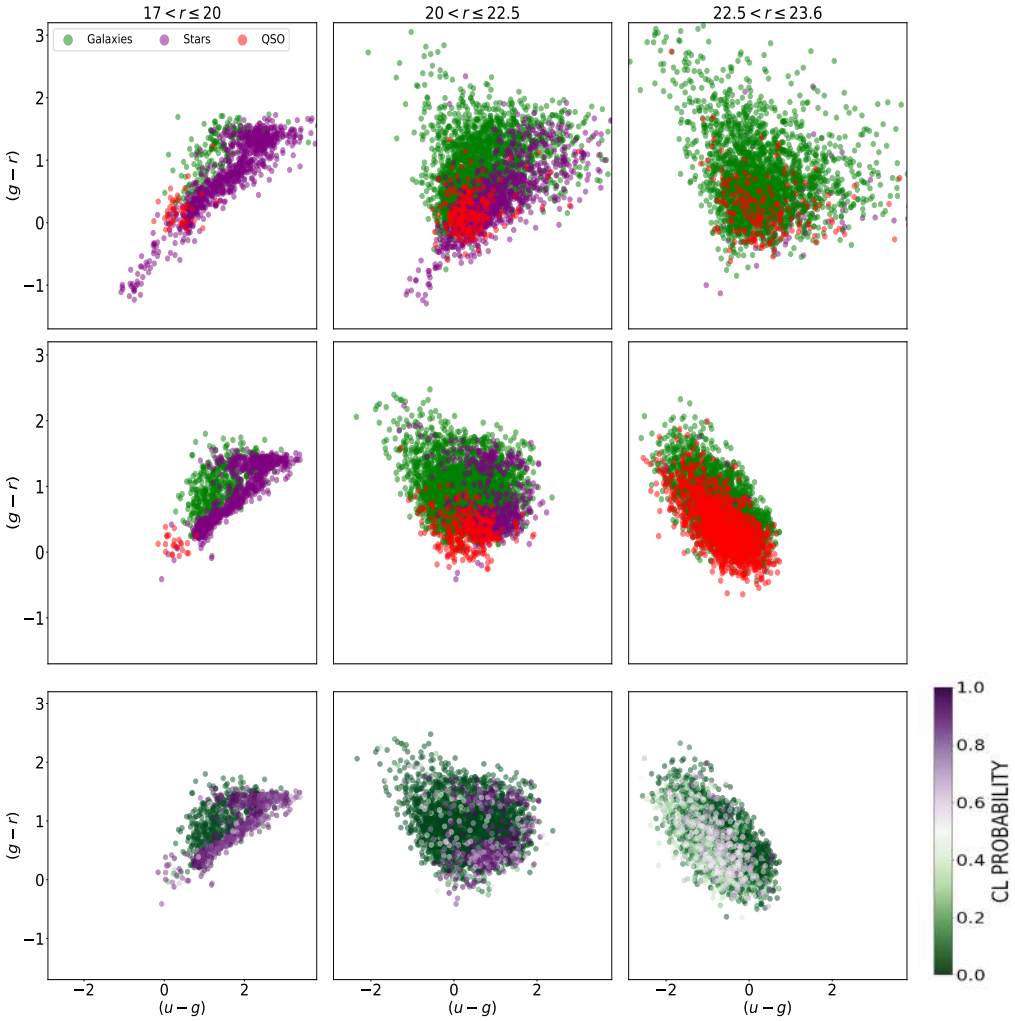


Figure 5.13: Observed $(g-r)$ vs. $(u-g)$ colour-colour for the 1-deg^2 mock sample (first row) and miniJPAS observations (second and third rows). Stars, galaxies and quasars are predicted classes with the ANN$_1$ in miniJPAS observations while they are true classes in the 1deg^2 mock sample. The dots in the third row are colour-coded according to the `SExtractor` probability developed to separate between point-like sources (CL > 0.5) and extended ones (CL < 0.5). Each column includes objects at different magnitude bins.

5.6 Summary and conclusions

In this chapter we present a method based on ANN to classify J-spectra in four categories: stars, galaxies, quasars at high redshift ($z \geq 2.1$), and quasars at low redshift ($z < 2.1$). The algorithms are trained and tested in mock data developed by Queiroz et al. (2022). We employ two different representations of miniJPAS photometry in order to train the algorithms. The ANN_1 uses as input photometric fluxes normalised to the detection band (r) while ANN_2 employs colours plus the magnitude in the *r*-band. Therefore, ANN_1 only has information of the shape of the spectrum while ANN_2 also has access to the observed luminosity.

We enlarge the training set by mixing features from four different classes adapting the `mix up` technique. We do not observe significant differences in the performance of the algorithms when an hybrid set is used for the training. A fundamental difference between other works where `mix up` has been employed with success and this work is probably the complexity of astronomical data. Observations have errors associated to them that depend at first order on the luminosity of the observed objects. Therefore, features do not encode the same information if objects are brighter or fainter. In other words, mixing between classes appear as a natural outcome in the feature space as the errors increase, which make faint objects indistinguishable from hybrid bright objects. Thus, hybridisation has an impact on the probabilities yielded by the ANN as they becomes less realistic if the level of mixing is increased in the training set. Having well calibrated algorithms is as important as obtaining a high performance, otherwise the outputs cannot be interpreted as a probability estimation.

We test the algorithm in the SDSS test sample, and we obtain a performance compatible with the prediction in the mock test sample. The main source of confusion appears between galaxies and low redshift quasars. We argue that there is an inherent physical mixing between these two classes, and we provide some examples with SDSS spectra where the host galaxy of the quasar has a non-negligible contribution to the total observed light, showing its dual nature. In such cases, most of the time the classifiers yield quasar and galaxy as the two preferred classes. Nevertheless, the SDSS test samples is relatively small set and it is only representative of objects brighter than 22.5, therefore the results should be treated with caution. The actual performance for fainter objects is still unknown and the estimation we provided are based on our current physical knowledge about quasars, galaxies and stars together with our capability to generate simulated J-spectra that mimic miniJPAS observation in the best possible way.

In the last section of this chapter, we estimated the number of quasars, galax-

ies and stars in the miniJPAS observations, and we showed that our predictions are compatible with previous estimates as well as with `SExtractor`, which separates between point-like and extended sources. The algorithms presented in this chapter are part of a combined algorithm that unifies the outcomes of several classifiers (Pérez-Ràfols et al. in prep. b). In the future, we will provide a quasars target list for a spectroscopic follow-up with the WEAVE survey. This will give us a valuable information of the strengths and weakness of our classifiers. Indeed, WEAVE and J-PAS collaboration will enter a feedback phase, where the knowledge acquired by one survey would be transfer to the other in an interactive process. A natural extension of this work when enough spectroscopic confirmed sources are available is to use transfer learning and retrained the algorithms with observations in order to capture better the structure of J-PAS data.

Chapter 6

Conclusions and future works

During the course of the present thesis we have developed new techniques based on machine learning (ML) in order to identify and characterise emission lines objects in J-PAS. By making use of legacy data from spectroscopic surveys we generated mock J-spectra for training and testing purposes. When it comes to galaxies, we trained artificial neural networks (ANN) with synthetic J-PAS fluxes from CALIFA and MaNGA spectra to first identify emission line galaxies (ELG) and second provide predictions of the EWs of their main optical emission lines such as Hα, Hβ, [O III], and [N II]. Traditional methods to measure emission lines with photometric surveys are limited in many aspects. Firstly, they cannot disentangle the contribution of several emission lines that are very close to each other to the total flux observed in one single filter. This is specially important for the case of Hα and [N II] lines because the ratio [N II]/Hα can be used to distinguish the main ionisation mechanism of galaxies. Secondly, the minimum EW measurable is limited by the photometric contrast of the filter measuring the line. Instead, ML algorithms are able to find complex relations between features. Thus, the EW of one particular line is a function of the flux in the filter tracing the line (the photometric contrast) but also depends on any other information provided as inputs to the algorithms such as the color of the galaxy or the fluxes of other emission lines. Therefore, the accuracy of the ANN developed in chapter 3 outperform previous estimates.

With the data observed by the J-PAS-*pathfinder* camera, the miniJPAS survey, we have conducted a preliminary study (chapter 4) to test our tools and prove that these techniques can be used to understand better the properties of ELGs. We were able to make a clean selection ELG with little contamination of AGN, estimate the SFR in galaxies via the flux of Hα, recover the star formation main sequence or

constrain the evolution of the cosmic star formation density down to redshift 0.35.

Finally, in chapter 5, we focused on the source classification problem with special attention to distinguish quasars from galaxies and stars. Simulated J-pectra of quasars, stars and galaxies obtained from the SDSS survey were used to train and test the ANN. We investigated the effect of data augmentation via hybridisation. This technique consists in mixing features from different astronomical sources in order to generate hybrid objects with mixing probabilities. Nevertheless, we did not observe a global improvement in the performance of the algorithms. Unlike other works outside the astronomical field where hybridsation have been proven to have a positive impact for calibrating the probability estimates, we observed that the ANN becomes under-confidence in their prediction. We believe this is likely due to the nature of astronomical data where errors are intrinsic to observation, therefore 'hybridsation' appears as natural outcome as the S/N of the objects decreases.

ML algorithms have been proven to be very useful to address many different problems where traditional methods are either inefficient because of their computational cost or unable to provide satisfactory solutions. Nevertheless, ML algorithms require large data sets to be trained and these data should be representative of the target population. In this thesis, we made used of simulated J-PAS data to train the algorithms either with galaxies from the nearby universe to estimate the EWs of the emission lines or with more distance objects to classify sources. Thus, our capability to success predicting in unseen data lies in two non-negligible assumptions. Firstly, we assume that the generated synthetic J-spectra are a fairy representation of future J-PAS observations. To a lesser extend this is not completely true. For instance, the JPCam will take pictures of galaxies with trays of 14 CCDs that will remain unchanged for a given period of time during the observing campaign. Therefore, the final reduced SED of any galaxies will be the results of correcting the effect of different observational conditions. The residual of such corrections might lead to small variation of the SED that are difficult to account in the error budget. Furthermore, the error estimates of photometric fluxes might deviate from a Gaussian behaviour under some particular conditions, for example in the low S/N regime. Secondly, although we have trained the ANNs with millions of spaxels from CALIFA and MANGA galaxies, which include plenty of diverse physical states, i.e. regions with different gas-phase metallicity, high and low star-formation activity or different dust distributions, by construction peculiar objects are always underrepresented. Consequently, it is very unlikely that we will be able to predict well the EW or any other physical quantity for these objects. For example, the EWs of extreme emission lines galaxies or very metal-poor galaxies

will be underestimated as very few are presented in the training sample.

The main conclusion of this thesis are listed below.

- Mock J-PAS data can be generated from surveys such as SDSS, CALIFA or MaNGA and ML codes can be trained for different purposes: identifying ELG galaxies, estimating the EWs of the main emission lines in the optical spectrum or distinguishing between stars, galaxies, and quasars.

- With a simulated J-PAS training set based on MaNGA and CALIFA spectra (the CALMa training set) we are able to predict the EW of Hα, Hβ, [N II], and [O III] in a testing set based on SDSS spectra with a relative standard deviation of 8.4 %, 13.7 %, 14.8 %, and 15.7 %,, respectively. The Hα, Hβ, [N II], and [O III] lines present a relative bias of 0.03 %, 5.0 %, 4.8 %, and -6.4 % respectively.

- The [N II]/Hα can be constrained within 0.092 dex and a bias of -0.02 dex and the [O III] Hβ ratio with no bias and a dispersion of 0.078 dex predicting with the CALMA training set in the SDSS testing set. The O 3N 2 is recovered within 0.108 dex and a bias of 0.04 dex.

- According to our simulation, the minimum S/N that we need in the photometry to measure an emission line with an EW of 10 Å in Hα, Hβ, [N II], and [O III] is 5, 1.5, 3.5, and, 10 respectively. However, methods based on the photometry contrast need for the same EW a S/N in the photometry of at least 15.5.

- A comparison of our predicted EW of Hα, Hβ and, [N II] with miniJPAS data and direct measurements of the same lines with SDSS spectra in $\sim$ 50 galaxies show an overall agreement. Although, the correlation in the EW of [O III] is less strong, more data need to be gathered to unveil the origin of such discrepancy.

- A sample of 2154 galaxies observed by miniJPAS in the range $0 < z < 0.35$ has been studied both from the point of view of the stellar populations and the properties of the ionized gas. Our results show that blue (red) galaxies are composed of a younger (older) stellar population and present stronger (weaker) emission lines as it has been found in previous studies. With a criterion based on the mass and color of the galaxy we estimated that 83 % and 17 % in the sample are blue and red galaxies, respectively. With the

ANN classifier, which is based on the EW of the emission lines, we found that 82 % of the sample are strong ELGs and 18 % are weak ELGs.

- By means of the BPT and WHAN diagrams we are able to classify galaxies according to the main source of ionization and make a selection of SF galaxies. We obtained that galaxies with reliable EW values (2000 galaxies in total), 72.8 ± 0.4 %, 17.7 ± 0.4 %, and 9.4 ± 0.2 % are SF, Seyfert, and passive/LINER galaxies, respectively. Among the SF galaxies, 94 % of them are blue while and 97 % of the LINER/passive galaxies are red.

- We are able to retrieve the SFMS by predicting the SFR with the Hα luminosity corrected from extinction with the Balmer decrement. We fit the SFMS with a power law and we obtained a slope of $0.90^{+0.02}_{-0.02}$ [yr^{-1}], a zero-point of $-8.85^{+0.19}_{-0.20}$ [M$_\odot$ yr^{-1}], and intrinsic scatter of $0.20^{+0.01}_{-0.01}$. Our results do not show a flattening of the SFR at high mass.

- We estimated the cosmic evolution of the $\rho_{\rm SFR}$ within three redshift bins: $0 < z \leq 0.15$, $0.15 < z \leq 0.25$, and $0.25 < z \leq 0.35$ founding agreement with previous measurements based on the Hα emission line. However, we found an offset compared to the works that derived $\rho_{\rm SFR}$ from the SED fitting of the stellar continuum. The origin of this discrepancy is still unknown but it is most probably due to a combination of several factors, the assumptions regarding the SFH, the correction for dust attenuation or the escape of ionizing photons among others.

- We developed a method based on ANN to classify J-spectra in four categories: stars, galaxies, quasars at high redshift ($z \geq 2.1$), and quasars at low redshift ($z < 2.1$). The algorithms are trained and tested in mock data developed by Queiroz et al. (2022). We enlarge the training set by mixing features from four different classes. Our results suggest that hybridisation does not improve or worsen the performance. However, training with hybrid objects has a negative impact on the probabilities estimated by the algorithms.

- We test the ANN classifiers in a small subset of miniJPAS data of which SDSS spectra of galaxies, quasars, and stars are available (a true table). The performance that we obtained is compatible with the predictions performed in the mock test sample. The main source of confusion appears between galaxies and low redshift quasars.

- We estimated that J-PAS will be able to detect ~ 450 quasars per deg^2 with $r < 23.6$ mag and redshift $0.4 \leq z \leq 4$.

As as soon as observations of the JPCam are completed for tens or hundreds of deg^2, we might be in a position to conduct a number of different studies. As we pointed out, there is a unavoidable distance between the simulated data used to train ML codes and observations, which might bias our predictions to a certain extend. One possible way to reduce this gap is to retrain the algorithms with J-PAS data of galaxies already observed by other spectroscopic surveys that contain information that we are interested in, e.g. the type of source or the EW of the emission lines. As it was proven in Domínguez Sánchez et al. (2019), using *transfer learning* one can reduce the size of the training size by one order of magnitude so we do not need to train the models from scratch. Thus, for the problem of source classification we might need the order of $\sim$ 30 000 sources with spectroscopic information to make this possible. This might seem like a lot. However, the miniJPAS survey observed $\sim$ 5 000 galaxies that had a spectroscopic counterpart in only 1 deg^2. Therefore, is not unthinkable that we can reach such a number soon. When it comes to the predictions of EWs of ELG, using *transfer learning* might be more challenging as we need different training for each redshift range. Nevertheless, well resolved galaxies observed by IFU-like surveys such as CALIFA or MaNGA contains hundred of spaxels each which reduce significantly the number of objects needed.

For now we can only characterize ELGs that are below $z = 0.35$. However, J-PAS will be able to detect galaxies up to $z \sim 1$. Other emission lines, such as the [O II] $\lambda\lambda$ 3726,3729 doublet, are visible in the optical range up to redshift $z < 1.6$ and can be used to trace the star formation (Kewley et al. 2004; Sobral et al. 2012). In the future, we would like to include this line in our model. Although this might be challenging for galaxies in the nearby universe because the [O II] line is at the edge of the wavelength coverage of MaNGA and CALIFA, it is feasible elsewhere. In particular, we might train an ANN with galaxies above $z = 0.35$ in order to predict the EW of [O II] but also the EW of Hβ, and [O III] emission lines.

It is important to remember that we do not use all J-PAS filters to predict the EW of the emission lines, thus the information of the SED is limited to a shorter wavelength range. Since synthetic J-spectra were generated with a collection of MaNGA, CALIFA, and SDSS spectra which have different wavelength coverage, this choice was the most straightforward way to proceed. Alternatively, we might rely on semi-empirical models of galaxy spectra anchoring the SED beyond the observational wavelength coverage of the aforementioned surveys. Nevertheless,

this is not a simple task as we would need to model the UV and far UV emission, which is not yet fully constrained from optical observations (López Fernández et al. 2016).

In chapter 4 we found discrepancies in the number of ionising photons that are derived from the Hα luminosity and the ones from the analysis of the stellar populations. We argue that this might be caused by a combination of several factors. Briefly, we might be underestimating the nebular extinction or perhaps a significant fraction of ionizing photons are escaping from the H II regions, thus being unable to ionize the interstellar gas. It is also possible that the delayed-τ model of the SFHs that we assumed may need to over-shoot to account for a SFH that is instead exponentially rising, at least for an important fraction of the galaxies in our sample. Are the results more consistent if other SFHs are used to fit the SED of galaxies? What is more, is a model that fit the stellar populations and the emission lines simultaneously more in agreement with the results we obtained with the ANN? If instead, there is ionizing radiation that leaks from the H II regions or in fact there are more interstellar dust around the gas, is the radiation at IR wavelength enhanced? Certainly, all these possibilities need to be investigated further. Multi-wavelength data of galaxies will be available for many in galaxies in J-PAS, therefore we will be able to shed light on this issue.

The ML techniques used in this thesis belong to a branch of artificial intelligence called *supervised learning*. Nonetheless, many works in astronomy are using now unsupervised learning to address many different problems. One concept that I find particularly interesting to explore in the future is the notion of *similarity distance*. In principle, objects that are similar according to observations should share the same physical properties. Of course, the observations might be incomplete and therefore the physical differences between these two objects might be hidden in some unobserved variable. This is what we call a degeneracy. Physical models deal differently with a degeneracy. For example, a SED fitting code that fit the stellar populations of galaxies might estimate more dust to explain the reddening of the observed spectrum or it might assume that the stars are indeed more metal-rich. Be that as it may, we make use of our preferred physical model in the market from which we draw the conclusions about what is the most likely physical phenomena that govern the observations we are analyzing. So the question lies in whether we are able to come to the same conclusions with only a data-driven approach. With current and upcoming surveys such as J-PAS we might answer this question as we could make use of large astronomical data to find the set of features that better define a *similarity distance*. Certainly, two population of galaxies that are the result of different formation scenarios should have different physical prop-

erties but they should also be distanced in the 'similarity space'. Nevertheless, we should pay special attention to the dissimilarity produced by non-physical differences (Sarmiento et al. 2021). For instance, galaxies that share the same physical properties might look different not because they are intrinsically different but because of instrumental effects. Furthermore, this concept might be useful to define regions (spaxels or pseudo-spaxels) within galaxies that can be grouped together. This is specially useful if we need to increase the S/N or if we want to analyze an H II region.

The future is exciting, certainly J-PAS will change our view of the cosmos. However 'as our circle of knowledge expands, so does the circumference of darkness surrounding it'.

Appendix A

SDSS training set

In this section, we show how the SDSS training set scores in the SDSS testing sample. This represents the ideal situation where the testing set is included within the parameter space of the training set. In other words, the testing sample is a subset of the training set and consequently the only uncertainties found in the targets variables (EWs) area associated to the capability of the ANN_R algorithm to decode the information provided by the inputs (J-spectrum). Nonetheless, we cannot infer from that the actual potential of the ANN_R to predict in J-PAS data. As we discussed in Sect. 3.3.2, herein lies the reason why the ANN_R must be tested with data with different observational setup and calibrations.

In Fig. A.1, we show the EWs predicted by the ANN_R versus the EWs provided by the SDSS testing sample from the MPA-JHU DR8 catalog. This plot follows the same scheme of Fig. 3.4. We are able to constrain better the EW of Hα followed by Hβ, [O III] and [N II]. However, the [N II] line is recovered with no bias and it does not saturate at high values. This is an important difference respect to what we found training with the CALMa training set.

In Fig. A.2, we show the comparison between the logarithmic ratios of [N II]/Hα, [O III]/Hβ and O 3N 2 in a similar way as we did in Fig. 3.5. The [N II]/Hα ratio is predicted within 0.089 dex and a bias of 0.019 dex and the [O III]/Hβ ratio within 0.08 dex and a bias of 0.027 dex. As a result, the O 3N 2 is recovered within 0.12 dex and a bias of 0.014.

Finally, we show in Fig. A.3 a comparison of the BPT diagram recovered by the ANN_R (left plot) and the one obtained from the SDSS testing sample (right plot) following, once again, the same scheme of Fig. 3.6. The similarity between those diagrams is remarkable. We are not only able to recover properly the SF-wing but

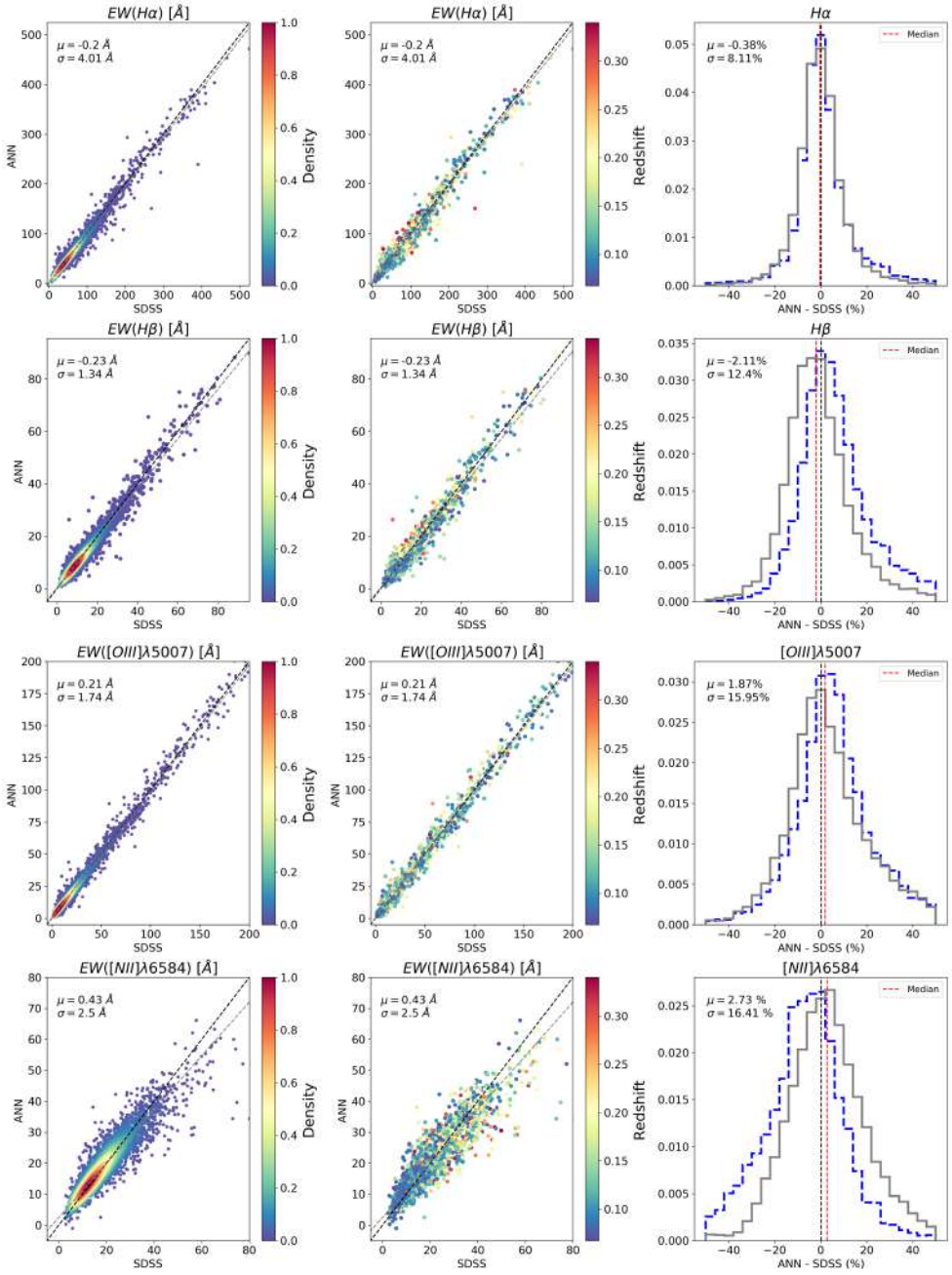


Figure A.1: EWs of Hα, Hβ, [N II] and [O III] predicted by the ANN$_R$ compared to SDSS testing sample.The ANN$_R$ is trained with the SDSS training set. The color-code represents the density in arbitrary units (right panel) and the redshift (left panel). The grey histograms show the relative difference between both values. The blue histograms are the ones in Fig. 3.4 and are shown for a visual comparison. Black and blue numbers are the median and the median absolute deviation of the difference. Black and blue numbers are the median and the MAD of the difference. Black line is the 1:1 relation and grey dashed lines represents the best linear fit. The red dashed line represents the median.

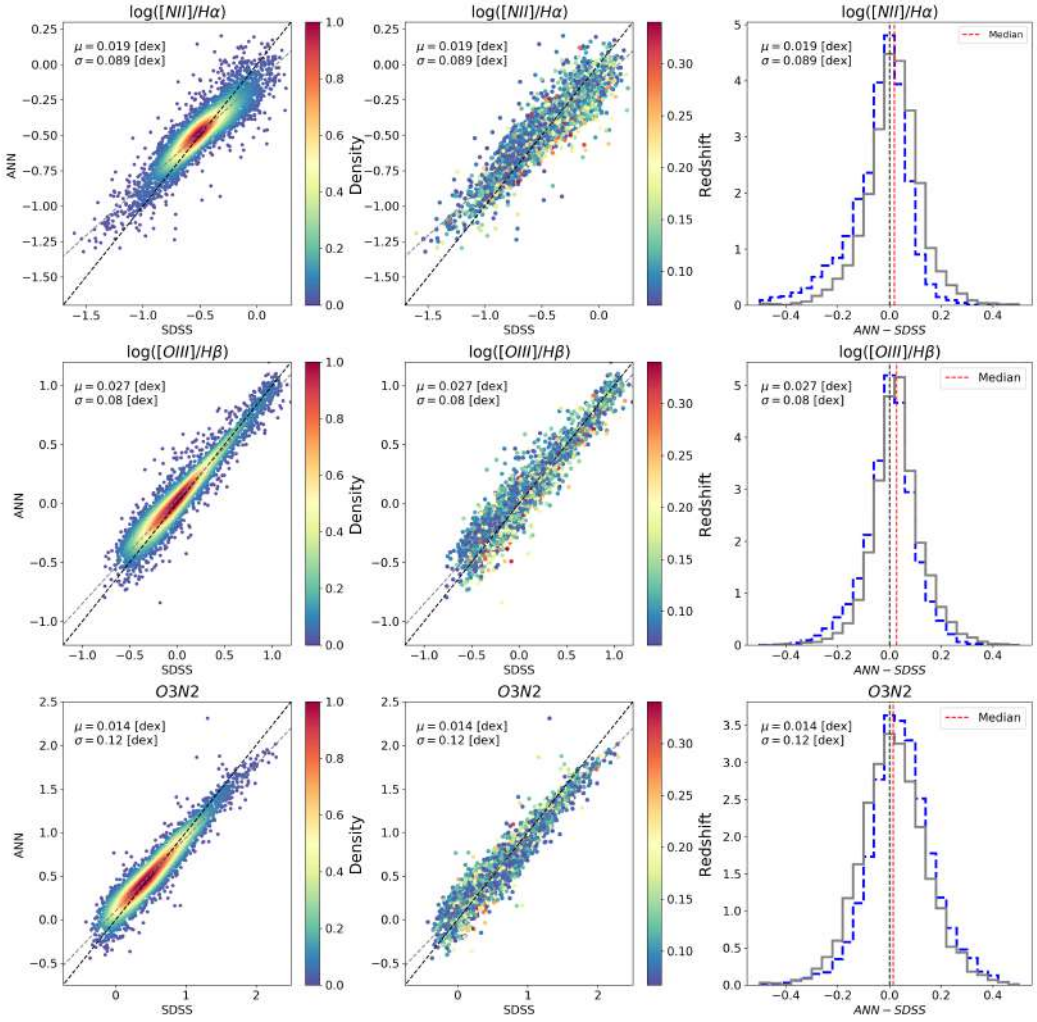


Figure A.2: Comparison between [N II]/Hα, [O III]/Hβ and O 3N 2 ratios estimated by the ANN$_R$ and SDSS testing sample. The ANN$_R$ is trained with the SDSS training set Same scheme of Fig. A.1.

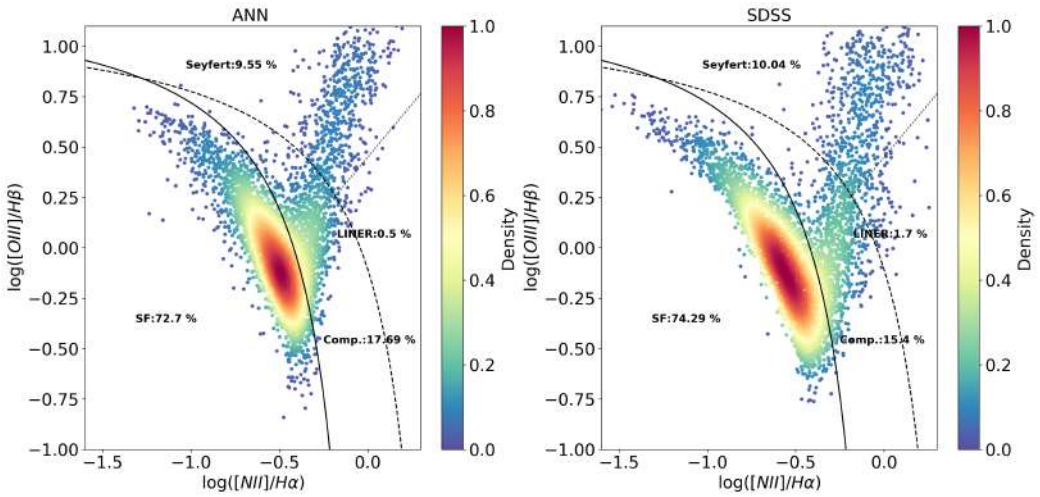


Figure A.3: BPT diagram obtained with the ANN$_R$ and SDSS MPA-JHU DR8 catalog where the color-code indicates the density of points. The ANN$_R$ is trained with the SDSS training set. The solid (ka03), dashed (Ke01) and dotted lines (S07) define the regions for the four main ionization mechanism of galaxies. The percentage for each group is shown in black.

also the AGN branch, obtaining similar percentages of galaxies in all the regions.

Appendix B

AGN selection criteria

In Table B.1 we show the best-fitting parameter as a function of the separation curves, the redshift bin, and the fitting equation we used to fit the SFMS. The results are discussed in the main text (Sect. 4.5.6)

Sample	Size	[N II]/ Hα	Eq.	α	β	σ_{int}	γ	M_0
$0 < z \leq 0.35$	1361	≤ 0.79	PW	$0.88^{+0.02}_{-0.02}$	$-8.69^{+0.18}_{-0.18}$	$0.20^{+0.01}_{-0.01}$	-	-
			BPW	$0.84^{+0.03}_{-0.03}$	$-0.89^{+0.09}_{-0.07}$	$0.20^{+0.01}_{-0.01}$	-	$10.75^{+0.18}_{-0.14}$
			QPW	$2.49^{+0.34}_{-0.35}$	$-15.50^{+1.71}_{-1.76}$	$0.20^{+0.01}_{-0.01}$	$0.09^{+0.02}_{-0.02}$	-
	1178	≤ 0.48	PW	$0.90^{+0.02}_{-0.02}$	$-8.85^{+0.19}_{-0.20}$	$0.20^{+0.01}_{-0.01}$	-	-
			BPW	$0.82^{+0.03}_{-0.03}$	$-0.99^{+0.12}_{-0.09}$	$0.20^{+0.01}_{-0.01}$	-	$10.93^{+0.22}_{-0.17}$
			QPW	$2.21^{+0.33}_{-0.33}$	$-14.18^{+1.61}_{-1.61}$	$0.20^{+0.01}_{-0.01}$	$0.08^{+0.02}_{-0.02}$	
	1026	≤ 0.40	PW	$0.92^{+0.02}_{-0.02}$	$-8.99^{+0.20}_{-0.20}$	$0.20^{+0.01}_{-0.01}$	-	-
			BPW	$0.82^{+0.03}_{-0.04}$	$-1.08^{+0.18}_{-0.13}$	$0.20^{+0.01}_{-0.01}$	-	$11.01^{+0.32}_{-0.21}$
			QPW	$2.11^{+0.37}_{-0.36}$	$-13.80^{+1.75}_{-1.75}$	$0.19^{+0.01}_{-0.01}$	$0.07^{+0.02}_{-0.02}$	-
$0 < z \leq 0.15$	220	≤ 0.79	PW	$0.84^{+0.04}_{-0.03}$	$-8.40^{+0.33}_{-0.34}$	$0.20^{+0.02}_{-0.02}$	-	-
	197	≤ 0.48		$0.85^{+0.04}_{-0.04}$	$-8.54^{+0.34}_{-0.38}$	$0.21^{+0.02}_{-0.02}$	-	-
	171	≤ 0.40		$0.90^{+0.04}_{-0.04}$	$-8.97^{+0.41}_{-0.42}$	$0.21^{+0.02}_{-0.02}$	-	-
$0.15 < z \leq 0.25$	461	≤ 0.79	PW	$0.77^{+0.04}_{-0.04}$	$-7.52^{+0.36}_{-0.37}$	$0.18^{+0.02}_{-0.02}$	-	-
	384	≤ 0.48		$0.77^{+0.04}_{-0.03}$	$-7.54^{+0.36}_{-0.37}$	$0.17^{+0.02}_{-0.02}$	-	-
	336	≤ 0.40		$0.81^{+0.04}_{-0.04}$	$-7.88^{+0.39}_{-0.42}$	$0.17^{+0.02}_{-0.02}$	-	-
$0.25 < z \leq 0.35$	641	≤ 0.79	PW	$0.81^{+0.04}_{-0.04}$	$-7.94^{+0.35}_{-0.38}$	$0.06^{+0.04}_{-0.06}$	-	-
	561	≤ 0.48		$0.85^{+0.03}_{-0.03}$	$-8.26^{+0.35}_{-0.36}$	$0.00^{+0.06}_{-0.00}$	-	-
	488	≤ 0.40		$0.82^{+0.04}_{-0.04}$	$-7.98^{+0.41}_{-0.42}$	$0.00^{+0.01}_{-0.00}$	-	-

Table B.1: Parameters of the SFMS derived in different redshift bins with the models described in Sects. 4.5.3 and 4.5.5 using different selection criteria (see Sect. 4.5.6). PW, BPW, and QPW stand for power law, broken power law, and quadratic power law, respectively.

Appendix C

Confusion matrices

In this section, we show the confusion matrices obtained in the test sample with the ANN_1 mix (C.4), ANN_2 (C.2), and ANN_2 mix (C.3), and the confusion matrices obtained in the SDSS test sample with the ANN_1 mix (C.1), ANN_2 (C.2), and ANN_2 mix (C.3).

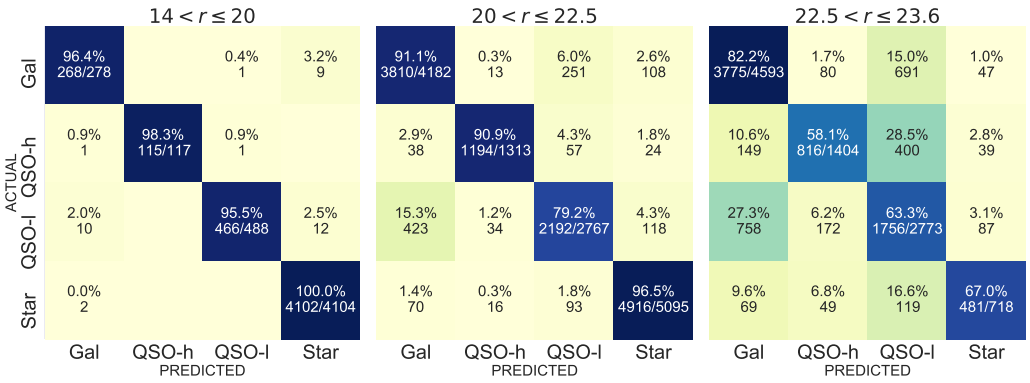


Figure C.1: Confusion matrices obtained with the ANN_1 mix in the test sample.

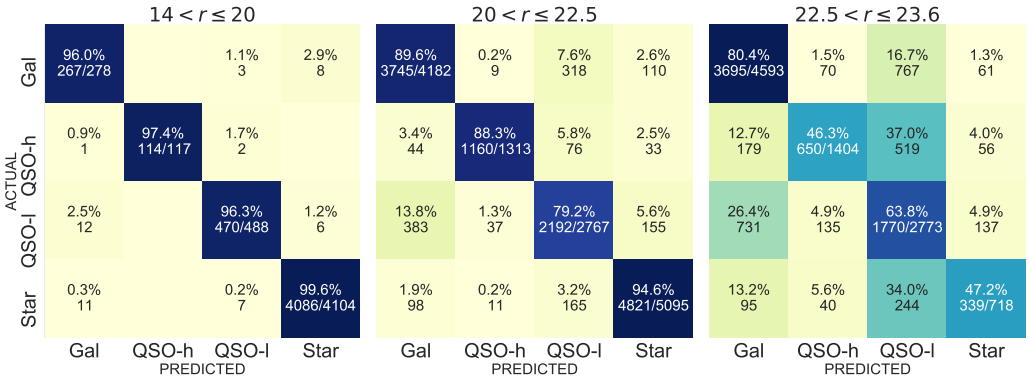


Figure C.2: Confusion matrices obtained with the ANN_2 in the test sample.

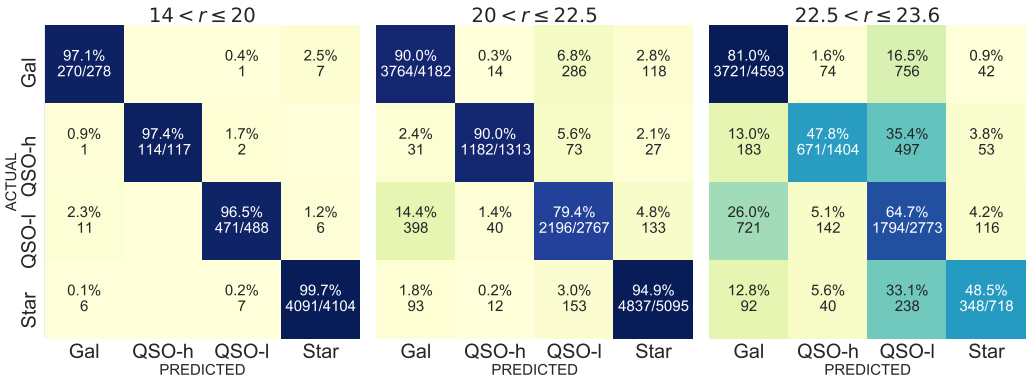


Figure C.3: Confusion matrices obtained with the ANN_2 mix in the test sample.

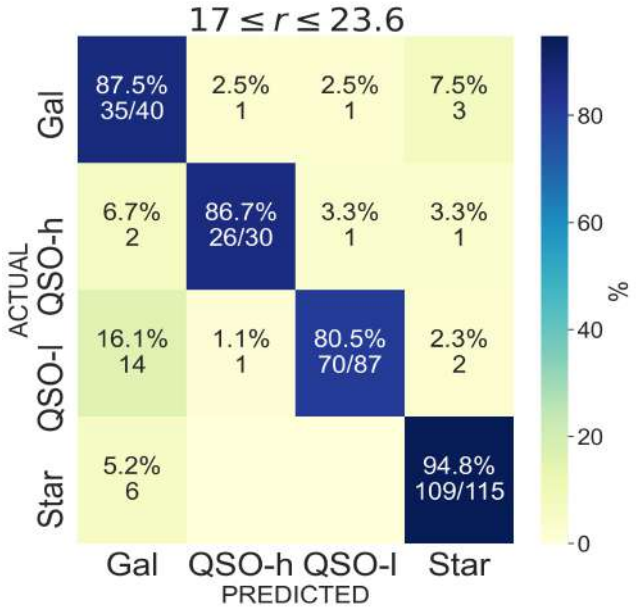


Figure C.4: Confusion matrix obtained with ANN1 mix in the SDSS test sample.

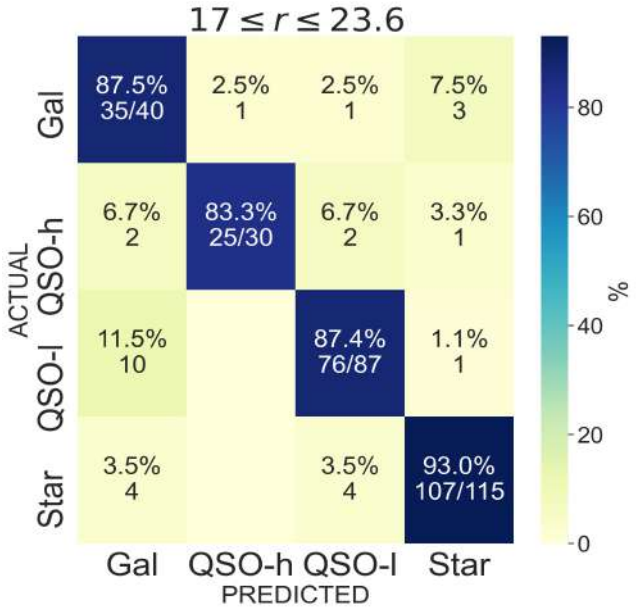


Figure C.5: Confusion matrix obtained with ANN2 mix in the SDSS test sample.

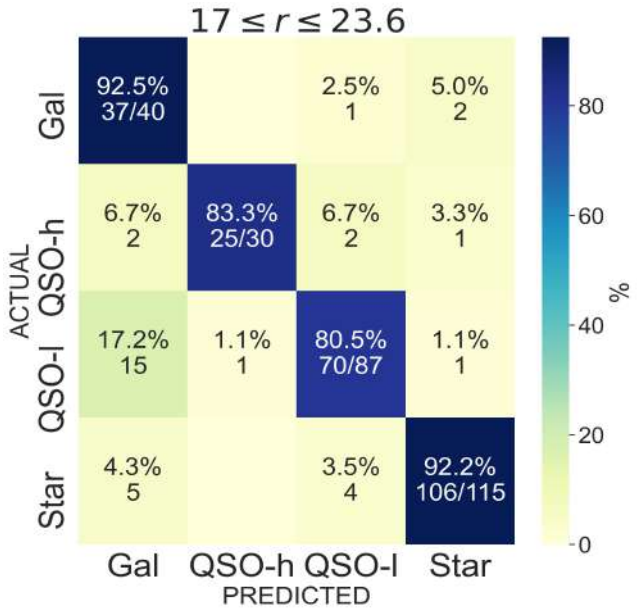


Figure C.6: Confusion matrix obtained with ANN2 in the SDSS test sample.

Bibliography

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org

Abramo, L. R., Strauss, M. A., Lima, M., et al. 2012, MNRAS, 423, 3251, [1108.2657].

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4, [1704.05858].

Ali, A., Shamsuddin, S. M., & Ralescu, A. 2015, in SOCO 2015

Alsing, J., Peiris, H., Leja, J., et al. 2020, ApJS, 249, 5, [1911.11778].

Anderson, L. D., Deharveng, L., Zavagno, A., et al. 2015, ApJ, 800, 101, [1412.6470].

Arnouts, S. & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate

Asari, N. V., Cid Fernandes, R., Stasińska, G., et al. 2007, MNRAS, 381, 263, [0707.3578].

Bacon, R., Accardo, M., Adjali, L., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III, ed. I. S. McLean, S. K. Ramsay, & H. Takami, 773508

Bacon, R., Copin, Y., Monnet, G., et al. 2001, MNRAS, 326, 23, [astro-ph/0103451].

Bai, Y., Liu, J., Wang, S., & Yang, F. 2019, AJ, 157, 9, [1811.03740].

Baldry, I. K., Glazebrook, K., Brinkmann, J., et al. 2004, ApJ, 600, 681, [astro-ph/0309710].

Baldry, I. K., Robotham, A. S. G., Hill, D. T., et al. 2010, MNRAS, 404, 86, [0910.5120].

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5, [astro-ph].

Bamford, S. P., Rojas, A. L., Nichol, R. C., et al. 2008, Monthly Notices of the Royal Astronomical Society, 391, 607, [http://oup.prod.sis.lan/mnras/article-pdf/391/2/607/5767752/mnras0391-0607.pdf].

Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87, [2007.07622].

Baron, D. 2019, arXiv e-prints, arXiv:1904.07248, [1904.07248].

Baron, D. & Poznanski, D. 2017, MNRAS, 465, 4530, [1611.07526].

Bastian, N., Covey, K. R., & Meyer, M. R. 2010, ARA&A, 48, 339, [1001.2965].

Belfiore, F., Maiolino, R., Bundy, K., et al. 2018, MNRAS, 477, 3014, [1710.05034].

Belfiore, F., Maiolino, R., Maraston, C., et al. 2016, MNRAS, 461, 3111, [1605.07189].

Bellstedt, S., Robotham, A. S. G., Driver, S. P., et al. 2020, MNRAS, 498, 5581, [2005.11917].

Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv e-prints, arXiv:1403.5237, [1403.5237].

Bertin, E. 2010, SCAMP: Automatic Astrometric and Photometric Calibration, Astrophysics Source Code Library, record ascl:1010.063

Best, P., Smail, I., Sobral, D., et al. 2013, in Thirty Years of Astronomical Discovery with UKIRT, Vol. 37, 235

Bishop, C. M. 1995, Neural Networks for Pattern Recognition (USA: Oxford University Press, Inc.)

Bluck, A. F. L., Maiolino, R., Brownson, S., et al. 2022, A&A, 659, A160, [2201.07814].

Bonjean, V., Aghanim, N., Salomé, P., et al. 2019, A&A, 622, A137, [1901.01932].

Bonoli, S., Marín-Franch, A., Varela, J., et al. 2021, A&A, 653, A31, [2007.01910].

Boogaard, L. A., Brinchmann, J., Bouché, N., et al. 2018, A&A, 619, A27, [1808.04900].

Boquien, M., Kennicutt, R., Calzetti, D., et al. 2016, A&A, 591, A6, [1603.09340].

Boselli, A., Fossati, M., & Sun, M. 2022, A&A Rev., 30, 3, [2109.13614].

Breda, I., Papaderos, P., & Gomes, J.-M. 2020, A&A, 640, A20, [2006.02307].

Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151, [astro-ph/0311060].

Bruzual, G. & Charlot, S. 2003, MNRAS, 344, 1000, [astro-ph/0309134].

Bundy, K. 2015, in IAU Symposium, Vol. 311, Galaxy Masses as Constraints of Formation Models, ed. M. Cappellari & S. Courteau, 100–103

Bundy, K., Bershady, M. A., Law, D. R., et al. 2015, ApJ, 798, 7, [1412.1482].

Burgarella, D., Buat, V., Gruppioni, C., et al. 2013, A&A, 554, A70, [1304.7000].

Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, MNRAS, 490, 3952, [1908.02748].

Byun, W., Sheen, Y.-K., Seon, K.-I., et al. 2021, arXiv e-prints, arXiv:2106.14363, [2106.14363].

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682, [astro-ph/9911459].

Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582, [astro-ph].

Cano-Díaz, M., Ávila-Reese, V., Sánchez, S. F., et al. 2019, MNRAS, 488, 3929, [1907.04386].

Cano-Díaz, M., Sánchez, S. F., Zibetti, S., et al. 2016, ApJ, 821, L26, [1602.02770].

Cantalupo, S. 2010, MNRAS, 403, L16, [0912.4149].

Cappellari, M., Emsellem, E., Krajnović, D., et al. 2011, MNRAS, 413, 813, [1012.1551].

Catalán-Torrecilla, C., Gil de Paz, A., Castillo-Morales, A., et al. 2015, A&A, 584, A87, [1507.03801].

Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959, [1611.02162].

Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, A&A, 622, A176, [1804.02667].

Chabrier, G. 2003, PASP, 115, 763, [astro-ph/0304382].

Chaves-Montero, J., Bonoli, S., Trakhtenbrot, B., et al. 2021, arXiv e-prints, arXiv:2111.01180, [2111.01180].

Chen, Y., Bressan, A., Girardi, L., et al. 2015, MNRAS, 452, 1068, [1506.01681].

Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2021a, MNRAS, 507, 4425, [2107.10210].

Cheng, T.-Y., Huertas-Company, M., Conselice, C. J., et al. 2021b, MNRAS, 503, 4446, [2009.11932].

Chevance, M., Kruijssen, J. M. D., Krumholz, M. R., et al. 2022, MNRAS, 509, 272, [2010.13788].

Chollet, F. et al. 2015, Keras, https://keras.io

Chung, A., van Gorkom, J. H., Kenney, J. D. P., & Vollmer, B. 2007, ApJ, 659, L115, [astro-ph/0703338].

Cicone, C., Maiolino, R., Sturm, E., et al. 2014, A&A, 562, A21, [1311.2595].

Cid Fernandes, R., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, MNRAS, 358, 363, [astro-ph/0412481].

Cid Fernandes, R., Stasińska, G., Mateus, A., & Vale Asari, N. 2011, MNRAS, 413, 1687, [1012.4426].

Cid Fernandes, R., Stasińska, G., Schlickmann, M. S., et al. 2010, MNRAS, 403, 1036, [0912.1643].

Ciesla, L., Elbaz, D., & Fensch, J. 2017, A&A, 608, A41, [1706.08531].

Collister, A. A. & Lahav, O. 2004, PASP, 116, 345, [astro-ph/0311058].

Conroy, C. 2013, ARA&A, 51, 393, [1301.7095].

Cooper, M. C., Aird, J. A., Coil, A. L., et al. 2011, ApJS, 193, 14, [1101.4018].

Cooper, M. C., Griffith, R. L., Newman, J. A., et al. 2012, MNRAS, 419, 3018, [1109.5698].

Cortijo-Ferrero, C., González Delgado, R. M., Pérez, E., et al. 2017, A&A, 607, A70, [1707.05324].

Costantin, L., Iovino, A., Zibetti, S., et al. 2019, A&A, 632, A9, [1910.01647].

Coughlin, A., Rhoads, J. E., Malhotra, S., et al. 2018, ApJ, 858, 96, [astro-ph].

Croom, S. M., Owers, M. S., Scott, N., et al. 2021, MNRAS, 505, 991, [2101.12224].

Croom, S. M., Richards, G. T., Shanks, T., et al. 2009, MNRAS, 399, 1755, [0907.2727].

Cucciati, O., Tresse, L., Ilbert, O., et al. 2012, A&A, 539, A31, [1109.1005].

Daddi, E., Cimatti, A., Renzini, A., et al. 2004, ApJ, 617, 746, [astro-ph/0409041].

Dale, D. A., Barlow, R. J., Cohen, S. A., et al. 2010, ApJ, 712, L189, [1003.0463].

Dalton, G., Trager, S., Abrams, D. C., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, ed. S. K. Ramsay, I. S. McLean, & H. Takami, 91470L

Darvish, B., Mobasher, B., Sobral, D., et al. 2016, ApJ, 825, 113, [1605.03182].

Davies, L. J. M., Driver, S. P., Robotham, A. S. G., et al. 2016, MNRAS, 461, 458, [1606.06299].

Davies, L. J. M., Lagos, C. d. P., Katsianis, A., et al. 2019, MNRAS, 483, 1881, [1811.03712].

Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJ, 660, L1, [astro-ph/0607355].

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10, [1208.0022].

de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247, [1230.002].

Delli Veneri, M., Cavuoti, S., Brescia, M., Longo, G., & Riccio, G. 2019, MNRAS, 486, 1377, [1902.02522].

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv e-prints, arXiv:1611.00036, [1611.00036].

Dewdney, P. E., Hall, P. J., Schilizzi, R. T., & Lazio, T. J. L. W. 2009, IEEE Proceedings, 97, 1482, [1630.002].

Díaz-García, L. A., Cenarro, A. J., López-Sanjuan, C., et al. 2019a, A&A, 631, A156, [1711.10590].

Díaz-García, L. A., Cenarro, A. J., López-Sanjuan, C., et al. 2015, A&A, 582, A14, [1505.07555].

Díaz-García, L. A., Cenarro, A. J., López-Sanjuan, C., et al. 2019b, A&A, 631, A158, [1901.05983].

Dobbels, W. & Baes, M. 2021, A&A, 655, A34, [2110.01704].

Domínguez, A., Siana, B., Brooks, A. M., et al. 2015, MNRAS, 451, 839, [1408.5788].

Domínguez, A., Siana, B., Henry, A. L., et al. 2013, ApJ, 763, 145, [1206.1867].

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93, [1807.00807].

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661, [1711.05744].

Donnari, M., Pillepich, A., Joshi, G. D., et al. 2021, MNRAS, 500, 4004, [2008.00005].

Drake, A. B., Simpson, C., Collins, C. A., et al. 2013, MNRAS, 433, 796, [1305.1305].

Driver, S. P., Andrews, S. K., da Cunha, E., et al. 2018, MNRAS, 475, 2891, [1710.06628].

Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971, [1009.0614].

Driver, S. P. & Robotham, A. S. G. 2010, MNRAS, 407, 2131, [1005.2538].

Duarte Puertas, S., Vilchez, J. M., Iglesias-Páramo, J., et al. 2017, A&A, 599, A71, [1611.07935].

Dutton, A. A., van den Bosch, F. C., & Dekel, A. 2010, MNRAS, 405, 1690, [0912.2169].

Ebeling, H., Stephenson, L. N., & Edge, A. C. 2014, ApJ, 781, L40, [1312.6135].

Emami, N., Siana, B., Weisz, D. R., et al. 2019, ApJ, 881, 71, [1809.06380].

Eriksen, M., Alarcon, A., Cabayol, L., et al. 2020, MNRAS, 497, 4565, [2004.07979].

Faber, S. M., Willmer, C. N. A., Wolf, C., et al. 2007, ApJ, 665, 265, [astro-ph/0506044].

Falcón-Barroso, J., Sánchez-Blázquez, P., Vazdekis, A., et al. 2011, A&A, 532, A95, [1107.2303].

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306, [1202.3665].

Förster Schreiber, N. M. & Wuyts, S. 2020, ARA&A, 58, 661, [2010.10171].

Furlong, M., Bower, R. G., Theuns, T., et al. 2015, MNRAS, 450, 4486, [1410.3485].

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1, [1804.09365].

Gallego, J., Zamorano, J., Aragon-Salamanca, A., & Rego, M. 1995, ApJ, 455, L1, [astro-ph].

García-Benito, R., Zibetti, S., Sánchez, S. F., et al. 2015, A&A, 576, A135, [1409.8302].

Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, Space Sci. Rev., 123, 485, [astro-ph/0606175].

Garn, T. & Best, P. N. 2010, MNRAS, 409, 421, [1007.1145].

Garn, T., Sobral, D., Best, P. N., et al. 2010, MNRAS, 402, 2017, [0911.2511].

Géron, A. 2019, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (O'Reilly Media)

Giammanco, C., Beckman, J. E., & Cedrés, B. 2005, A&A, 438, 599, [astro-ph/0504234].

Gil de Paz, A., Carrasco, E., Gallego, J., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, ed. C. J. Evans, L. Simard, & H. Takami, 99081K

Glorot, X., Bordes, A., & Bengio, Y. 2011, in Proceedings of Machine Learning Research, Vol. 15, Deep Sparse Rectifier Neural Networks, ed. G. Gordon, D. Dunson, & M. Dudík (Fort Lauderdale, FL, USA: JMLR Workshop and Conference Proceedings), 315–323

Gomes, J. M., Papaderos, P., Vílchez, J. M., et al. 2016, A&A, 586, A22, [1511.01300].

Gonçalves, T. S., Martin, D. C., Menéndez-Delmestre, K., Wyder, T. K., & Koekemoer, A. 2012, ApJ, 759, 67, [1209.4084].

González, R. E., Muñoz, R. P., & Hernández, C. A. 2018, Astronomy and Computing, 25, 103, [1809.01691].

González Delgado, R. M., Díaz-García, L. A., de Amorim, A., et al. 2021, arXiv e-prints, arXiv:2102.13121, [2102.13121].

González Delgado, R. M., García-Benito, R., Pérez, E., et al. 2015, A&A, 581, A103, [1506.04157].

González Delgado, R. M., Rodríguez-Martín, J. E., Díaz-García, L. A., et al. 2022, arXiv e-prints, arXiv:2207.05770, [2207.05770].

Goodfellow, I. J., Shlens, J., & Szegedy, C. 2014, arXiv e-prints, arXiv:1412.6572, [1412.6572].

Graham, M. L., Connolly, A. J., Ivezić, Ž., et al. 2018, AJ, 155, 1, [1706.09507].

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35, [1105.3753].

Gu, Y., Fang, G., Yuan, Q., Lu, S., & Liu, S. 2021, ApJ, 921, 60, [2109.11261].

Gunawardhana, M. L. P., Hopkins, A. M., Bland-Hawthorn, J., et al. 2013, MNRAS, 433, 2764, [1305.5308].

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. 2017, arXiv e-prints, arXiv:1706.04599, [1706.04599].

Hahn, C. & Melchior, P. 2022, arXiv e-prints, arXiv:2203.07391, [2203.07391].

Hao, C.-N., Kennicutt, R. C., Johnson, B. D., et al. 2011, ApJ, 741, 124, [1108.2837].

Hayashi, M., Shimakawa, R., Tanaka, M., et al. 2020, PASJ, 72, 86, [2007.07413].

Hayashi, M., Tanaka, M., Shimakawa, R., et al. 2018, PASJ, 70, S17, [1704.05978].

He, Z., Qiu, B., Luo, A. L., et al. 2021, MNRAS, 508, 2039, [12990.002].

Heckman, T. M. 1980, A&A, 500, 187, [astro-ph].

Heckman, T. M. 1987, in Observational Evidence of Activity in Galaxies, ed. E. E. Khachikian, K. J. Fricke, & J. Melnick, Vol. 121, 421

Henrion, M., Mortlock, D. J., Hand, D. J., & Gandy, A. 2011, MNRAS, 412, 2286, [1011.5770].

Hernán-Caballero, A., Varela, J., López-Sanjuan, C., et al. 2021, A&A, 654, A101, [2108.03271].

Hirano, S., Yoshida, N., Sakurai, Y., & Fujii, M. S. 2018, ApJ, 855, 17, [1711.07315].

Hopkins, P. F., Kereš, D., Oñorbe, J., et al. 2014, MNRAS, 445, 581, [1311.2073].

Hornik, K., Stinchcombe, M., & White, H. 1989, Neural Networks, 2, 359, [1230.082].

Hosokawa, T., Hirano, S., Kuiper, R., et al. 2016, ApJ, 824, 119, [1510.01407].

Hsieh, B. C., Lin, L., Lin, J. H., et al. 2017, ApJ, 851, L24, [1711.09162].

Hubble, E. P. 1926, ApJ, 64, 321, [1230.662].

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

Huchra, J. & Sargent, W. L. W. 1973, ApJ, 186, 433, [astro-ph].

Ichinohe, Y. & Yamada, S. 2019, MNRAS, 487, 2874, [1905.13434].

Iglesias-Páramo, J., Vílchez, J. M., Rosales-Ortega, F. F., et al. 2016, ApJ, 826, 71, [1605.03490].

Ilbert, O., Arnouts, S., Le Floc'h, E., et al. 2015, A&A, 579, A2, [1410.4875].

Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55, [1301.3157].

Ilbert, O., Salvato, M., Le Floc'h, E., et al. 2010, ApJ, 709, 644, [0903.0102].

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, [0805.2366].

Jarvis, M. J., Bonfield, D. G., Bruce, V. A., et al. 2013, MNRAS, 428, 1281, [1206.4263].

Jiménez, M., Torres Torres, M., John, R., & Triguero, I. 2020, IEEE Access, 8, 47232, [1230.112].

Kalinova, V., Colombo, D., Sánchez, S. F., et al. 2021, A&A, 648, A64, [2101.10019].

Karim, A., Schinnerer, E., Martínez-Sansigre, A., et al. 2011, ApJ, 730, 61, [1011.6370].

Kashino, D., Silverman, J. D., Rodighiero, G., et al. 2013, ApJ, 777, L8, [1309.4774].

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003a, MNRAS, 346, 1055, [astro-ph/0304239].

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003b, MNRAS, 341, 33, [astro-ph/0204055].

Kelz, A., Verheijen, M. A. W., Roth, M. M., et al. 2006, PASP, 118, 129, [astro-ph/0512557].

Kennicutt, Robert C., J. 1998, ARA&A, 36, 189, [astro-ph/9807187].

Kennicutt, Robert C., J., Tamblyn, P., & Congdon, C. E. 1994, ApJ, 435, 22, [astro-ph].

Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., & Greig, B. 2017, ApJ, 848, 23, [1705.04688].

Kerzendorf, W. E., Vogl, C., Buchner, J., et al. 2021, ApJ, 910, L23, [2007.01868].

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, ApJ, 556, 121, [astro-ph/0106324].

Kewley, L. J., Geller, M. J., & Jansen, R. A. 2004, AJ, 127, 2002, [astro-ph/0401172].

Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961, [astro-ph/0605681].

Kewley, L. J., Maier, C., Yabe, K., et al. 2013, ApJ, 774, L10, [1307.0514].

Kewley, L. J., Nicholls, D. C., & Sutherland, R. S. 2019, ARA&A, 57, 511, [1910.09730].

Khostovan, A. A., Malhotra, S., Rhoads, J. E., et al. 2021, MNRAS, 503, 5115, [2103.10959].

Khostovan, A. A., Malhotra, S., Rhoads, J. E., et al. 2020, MNRAS, 493, 3966, [2001.04989].

Kim, E. J. & Brunner, R. J. 2017, MNRAS, 464, 4463, [1608.04369].

Kim, E. J., Brunner, R. J., & Carrasco Kind, M. 2015, MNRAS, 453, 507, [1505.02200].

Kouroumpatzakis, K., Zezas, A., Maragkoudakis, A., et al. 2021, arXiv e-prints, arXiv:2107.00974, [2107.00974].

Kovač, K., Lilly, S. J., Knobel, C., et al. 2010, ApJ, 718, 86, [0909.2032].

Koyama, Y., Shimakawa, R., Yamamura, I., Kodama, T., & Hayashi, M. 2019, PASJ, 71, 8, [1809.03715].

Krakowski, T., Małek, K., Bilicki, M., et al. 2016, A&A, 596, A39, [1607.01188].

Kroupa, P. 2001, MNRAS, 322, 231, [astro-ph/0009005].

Lacerda, E. A. D., Cid Fernandes, R., Couto, G. S., et al. 2018, MNRAS, 474, 3727, [1711.07844].

Lacerda, E. A. D., Sánchez, S. F., Cid Fernandes, R., et al. 2020, MNRAS, 492, 3073, [2001.00099].

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193, [1110.3193].

Law, D. R., Yan, R., Bershady, M. A., et al. 2015, AJ, 150, 19, [1505.04285].

Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845, [astro-ph/0409133].

Lee, J. C., Gil de Paz, A., Tremonti, C., et al. 2009, ApJ, 706, 599, [0909.5205].

Lee, J. C., Veilleux, S., McDonald, M., & Hilbert, B. 2016, ApJ, 817, 177, [1601.00201].

Lee, N., Sanders, D. B., Casey, C. M., et al. 2015, ApJ, 801, 80, [1501.01080].

Leja, J., Carnall, A. C., Johnson, B. D., Conroy, C., & Speagle, J. S. 2019, ApJ, 876, 3, [1811.03637].

Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv e-prints, arXiv:1308.0847, [1308.0847].

Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007, ApJS, 172, 70, [astro-ph/0612291].

Logan, C. H. A. & Fotopoulou, S. 2020, A&A, 633, A154, [1911.05107].

Logroño-García, R., Vilella-Rojo, G., López-Sanjuan, C., et al. 2019, A&A, 622, A180, [1804.04039].

López Fernández, R., Cid Fernandes, R., González Delgado, R. M., et al. 2016, MNRAS, 458, 184, [1602.01123].

López Fernández, R., González Delgado, R. M., Pérez, E., et al. 2018, A&A, 615, A27, [1802.10118].

López-Sanjuan, C., Vázquez Ramió, H., Varela, J., et al. 2019, A&A, 622, A177, [1804.02673].

Lumbreras-Calle, A., Muñoz-Tuñón, C., Méndez-Abreu, J., et al. 2019, A&A, 621, A52, [1803.08045].

Ly, C., Malkan, M. A., Kashikawa, N., et al. 2007, ApJ, 657, 738, [astro-ph/0610846].

Madau, P. & Dickinson, M. 2014, ARA&A, 52, 415, [1403.0007].

Magris C., G., Mateu P., J., Mateu, C., et al. 2015, PASP, 127, 16, [1411.7029].

Maiolino, R. & Mannucci, F. 2019, A&A Rev., 27, 3, [1811.09642].

Marigo, P., Bressan, A., Nanni, A., Girardi, L., & Pumo, M. L. 2013, MNRAS, 434, 488, [1305.4485].

Marin-Franch, A., Taylor, K., Cenarro, J., Cristobal-Hornillos, D., & Moles, M. 2015, in IAU General Assembly, Vol. 29, 2257381

Mármol-Queraltó, E., McLure, R. J., Cullen, F., et al. 2016, MNRAS, 460, 3587, [1511.01911].

Martin, D. C., Fanson, J., Schiminovich, D., et al. 2005, ApJ, 619, L1, [astro-ph/0411302].

Martínez-Solaeche, G., González Delgado, R. M., García-Benito, R., et al. 2021, A&A, 647, A158, [2008.04287].

Martínez-Solaeche, G., González Delgado, R. M., García-Benito, R., et al. 2022, arXiv e-prints, arXiv:2204.01698, [2204.01698].

Matthee, J. & Schaye, J. 2019, MNRAS, 484, 915, [1805.05956].

Matthee, J., Sobral, D., Best, P., et al. 2017, MNRAS, 471, 629, [1702.04721].

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., et al. 2019, MNRAS, 489, 241, [1907.01567].

Mendez, A. J., Coil, A. L., Lotz, J., et al. 2011, ApJ, 736, 110, [1101.3353].

Mitchell, P. D., Lacey, C. G., Cole, S., & Baugh, C. M. 2014, MNRAS, 444, 2637, [1403.1585].

Moles, M., Benítez, N., Aguerri, J. A. L., et al. 2008, AJ, 136, 1325, [0806.3021].

Moles, M., Sánchez, S. F., Lamadrid, J. L., et al. 2010, PASP, 122, 363, [0912.3762].

Molina, M., Eracleous, M., Barth, A. J., et al. 2018, ApJ, 864, 90, [1804.06888].

Molino, A., Benítez, N., Ascaso, B., et al. 2017, MNRAS, 470, 95, [1705.02265].

Molino, A., Benítez, N., Moles, M., et al. 2014, MNRAS, 441, 2891, [1306.4968].

Molino, A., Costa-Duarte, M. V., Mendes de Oliveira, C., et al. 2019, A&A, 622, A178, [1804.03640].

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613, [astro-ph/9510034].

Moster, B. P., Somerville, R. S., Newman, J. A., & Rix, H.-W. 2011, ApJ, 731, 113, [1001.1737].

Muzzin, A., Marchesini, D., Stefanon, M., et al. 2013, ApJ, 777, 18, [1303.4409].

Nair, V. & Hinton, G. E. 2010, in Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10 (Madison, WI, USA: Omnipress), 807–814

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5, [1203.3192].

Noirot, G., Sawicki, M., Abraham, R., et al. 2022, MNRAS, 512, 3566, [2203.06185].

Oke, J. B. & Gunn, J. E. 1983, ApJ, 266, 713, [11110.002].

Oliver, S., Frost, M., Farrah, D., et al. 2010, MNRAS, 405, 2279, [1003.2446].

Oteo, I., Sobral, D., Ivison, R. J., et al. 2015, MNRAS, 452, 2018, [1506.02670].

Otí-Floranes, H. & Mas-Hesse, J. M. 2010, A&A, 511, A61, [0912.0833].

Paccagnella, A., Vulcani, B., Poggianti, B. M., et al. 2016, ApJ, 816, L25, [1512.04549].

Palanque-Delabrouille, N., Magneville, C., Yèche, C., et al. 2016, A&A, 587, A41, [1509.05607].

Pâris, I., Petitjean, P., Ross, N. P., et al. 2017, A&A, 597, A79, [1608.06483].

Pascual, S., Gallego, J., & Zamorano, J. 2007, PASP, 119, 30, [astro-ph/0611121].

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26, [1806.06607].

Pellegrini, E. W., Oey, M. S., Winkler, P. F., et al. 2012, ApJ, 755, 40, [1202.3334].

Peng, Y., Maiolino, R., & Cochrane, R. 2015, Nature, 521, 192, [1505.03143].

Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, ApJ, 721, 193, [1003.4747].

Penny, S. J., Masters, K. L., Smethurst, R., et al. 2018, MNRAS, 476, 979, [1710.07568].

Pérez-González, P. G., Cava, A., Barro, G., et al. 2013, ApJ, 762, 46, [1207.6639].

Pérez-Ràfols, I., Pieri, M. M., Blomqvist, M., Morrison, S., & Som, D. 2020, MNRAS, 496, 4931, [1903.00023].

Pettini, M. & Pagel, B. E. J. 2004, MNRAS, 348, L59, [astro-ph/0401128].

Pieri, M. M., Bonoli, S., Chaves-Montero, J., et al. 2016, in SF2A-2016: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. C. Reylé, J. Richard, L. Cambrésy, M. Deleuil, E. Pécontal, L. Tresse, & I. Vauglin, 259–266

Poggianti, B. M., Fasano, G., Omizzolo, A., et al. 2016, AJ, 151, 78, [1504.07105].

Poggianti, B. M., Moretti, A., Gullieuszik, M., et al. 2017, ApJ, 844, 48, [1704.05086].

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, AJ, 160, 45, [2002.10464].

Pozzetti, L., Bolzonella, M., Zucca, E., et al. 2010, A&A, 523, A13, [0907.5416].

Prugniel, P., Vauglin, I., & Koleva, M. 2011, A&A, 531, A165, [1104.4952].

Puglisi, A., Rodighiero, G., Franceschini, A., et al. 2016, A&A, 586, A83, [1507.00005].

Qin, J., Zheng, X. Z., Wuyts, S., Pan, Z., & Ren, J. 2019, ApJ, 886, 28, [1909.13505].

Queiroz, C., Abramo, L. R., Rodrigues, N. V. N., et al. 2022, arXiv e-prints, arXiv:2202.00103, [2202.00103].

Ramachandra, N., Chaves-Montero, J., Alarcon, A., et al. 2022, MNRAS, 515, 1927, [127260.002].

Renzini, A. & Peng, Y.-j. 2015, ApJ, 801, L29, [1502.01027].

Robin, A. C., Reylé, C., Derrière, S., & Picaud, S. 2003, A&A, 409, 523, [1292920.102].

Robotham, A. S. G. & Obreschkow, D. 2015, PASA, 32, e033, [1508.02145].

Rodighiero, G., Daddi, E., Baronchelli, I., et al. 2011, ApJ, 739, L40, [1108.0933].

Rodríguez Martín, J. E., González Delgado, R. M., Martínez-Solaeche, G., et al. 2022, arXiv e-prints, arXiv:2207.10101, [2207.10101].

Roth, M. M., Kelz, A., Fechner, T., et al. 2005, PASP, 117, 620, [astro-ph/0502581].

Saglia, R. P., Tonry, J. L., Bender, R., et al. 2012, ApJ, 746, 128, [1109.5080].

Salim, S., Rich, R. M., Charlot, S., et al. 2007, ApJS, 173, 267, [0704.3611].

Salmi, F., Daddi, E., Elbaz, D., et al. 2012, ApJ, 754, L14, [1206.1704].

Sánchez, S. F., Avila-Reese, V., Hernandez-Toledo, H., et al. 2018, Rev. Mexicana Astron. Astrofis., 54, 217, [1709.05438].

Sánchez, S. F., Avila-Reese, V., Rodríguez-Puebla, A., et al. 2019, MNRAS, 482, 1557, [1807.11528].

Sánchez, S. F., García-Benito, R., Zibetti, S., et al. 2016a, A&A, 594, A36, [1604.02289].

Sánchez, S. F., Kennicutt, R. C., Gil de Paz, A., et al. 2012, A&A, 538, A8, [1111.0962].

Sánchez, S. F., Pérez, E., Sánchez-Blázquez, P., et al. 2016b, Rev. Mexicana Astron. Astrofis., 52, 171, [1602.01830].

Sánchez, S. F., Pérez, E., Sánchez-Blázquez, P., et al. 2016c, Rev. Mexicana Astron. Astrofis., 52, 21, [1509.08552].

Sánchez-Blázquez, P., Peletier, R. F., Jiménez-Vicente, J., et al. 2006, MNRAS, 371, 703, [astro-ph/0607009].

Santos, S., Sobral, D., Matthee, J., et al. 2020, MNRAS, 493, 141, [1910.02959].

Sarmiento, R., Huertas-Company, M., Knapen, J. H., et al. 2021, ApJ, 921, 177, [2104.08292].

Schawinski, K., Thomas, D., Sarzi, M., et al. 2007, MNRAS, 382, 1415, [0709.3015].

Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, MNRAS, 440, 889, [1402.4814].

Schmidt, M. 1968, ApJ, 151, 393, [astro-ph].

Schreiber, C., Pannella, M., Elbaz, D., et al. 2015, A&A, 575, A74, [1409.5433].

Searle, L., Sargent, W. L. W., & Bagnuolo, W. G. 1973, ApJ, 179, 427, [1245560.002].

Serrano, S., Gaztañaga, E., Castander, F. J., et al. 2022, arXiv e-prints, arXiv:2206.14022, [2206.14022].

Sérsic, J. L. 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41, [123.8726].

Sharma, K., Kembhavi, A., Kembhavi, A., et al. 2020, MNRAS, 491, 2280, [1909.05459].

Sharma, K., Prugniel, P., & Singh, H. P. 2016, A&A, 585, A64, [1512.04882].

Shin, K., Ly, C., Malkan, M. A., et al. 2021, MNRAS, 501, 2231, [1910.10735].

Shioya, Y., Taniguchi, Y., Sasaki, S. S., et al. 2008, ApJS, 175, 128, [0709.1009].

Shorten, C. & Khoshgoftaar, T. 2019, Journal of Big Data, 6

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163, [198.30001].

Sobral, D., Best, P. N., Matsuda, Y., et al. 2012, MNRAS, 420, 1926, [1109.1830].

Sobral, D., Matthee, J., Best, P. N., et al. 2015, MNRAS, 451, 2303, [1502.06602].

Sobral, D., Smail, I., Best, P. N., et al. 2013, MNRAS, 428, 1128, [1202.3436].

Sobral, D., Stroe, A., Koyama, Y., et al. 2016, MNRAS, 458, 3443, [1603.00462].

Sparre, M., Hayward, C. C., Springel, V., et al. 2015, MNRAS, 447, 3548, [1409.0009].

Speagle, J. S., Steinhardt, C. L., Capak, P. L., & Silverman, J. D. 2014, ApJS, 214, 15, [1405.2041].

Stasińska, G., Vale Asari, N., Cid Fernandes, R., et al. 2008, MNRAS, 391, L29, [0809.1341].

Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861, [astro-ph/0107201].

Stroe, A. & Sobral, D. 2015, MNRAS, 453, 242, [1507.02687].

Tammour, A., Gallagher, S. C., Daley, M., & Richards, G. T. 2016, MNRAS, 459, 1659, [1603.03318].

Tasca, L. A. M., Le Fèvre, O., Hathi, N. P., et al. 2015, A&A, 581, A54, [1411.5687].

Taylor, K., Marín-Franch, A., Laporte, R., et al. 2014, Journal of Astronomical Instrumentation, 3, 1350010, [1301.4175].

Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., & Bluck, A. F. L. 2022, AJ, 163, 71, [2112.03425].

Thorne, J. E., Robotham, A. S. G., Davies, L. J. M., et al. 2020, arXiv e-prints, arXiv:2011.13605, [2011.13605].

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., & Michalak, S. 2019, arXiv e-prints, arXiv:1905.11001, [1905.11001].

Tinsley, B. M. 1968, ApJ, 151, 547, [1223.00].

Tinsley, B. M. & Gunn, J. E. 1976, ApJ, 203, 52, [17770.2302].

Tomczak, A. R., Quadri, R. F., Tran, K.-V. H., et al. 2016, ApJ, 817, 118, [1510.06072].

Tomczak, A. R., Quadri, R. F., Tran, K.-V. H., et al. 2014, ApJ, 783, 85, [1309.5972].

Turner, S., Kelvin, L. S., Baldry, I. K., et al. 2019, MNRAS, 482, 126, [1810.00887].

Urrutia, T., Wisotzki, L., Kerutt, J., et al. 2019, A&A, 624, A141, [1811.06549].

Valdes, F., Gupta, R., Rose, J. A., Singh, H. P., & Bell, D. J. 2004, ApJS, 152, 251, [astro-ph/0402435].

Van Sistine, A., Salzer, J. J., Sugden, A., et al. 2016, ApJ, 824, 25, [astro-ph].

Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., et al. 2011, AJ, 141, 189, [1011.1951].

Vilella-Rojo, G., Logroño-García, R., López-Sanjuan, C., et al. 2021, arXiv e-prints, arXiv:2101.04062, [2101.04062].

Vilella-Rojo, G., Viironen, K., López-Sanjuan, C., et al. 2015, A&A, 580, A47, [1505.07115].

Weinberger, R., Springel, V., Pakmor, R., et al. 2018, MNRAS, 479, 4056, [1710.04659].

Werle, A., Cid Fernandes, R., Vale Asari, N., et al. 2019, MNRAS, 483, 2382, [1811.11255].

Westra, E., Geller, M. J., Kurtz, M. J., Fabricant, D. G., & Dell'Antonio, I. 2010, ApJ, 708, 534, [0911.0417].

Whitaker, K. E., Franx, M., Leja, J., et al. 2014, ApJ, 795, 104, [1407.1843].

Whitaker, K. E., van Dokkum, P. G., Brammer, G., & Franx, M. 2012, ApJ, 754, L29, [1205.0547].

Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2019, A&A, 622, A182, [1811.02279].

Willett, K. W., Schawinski, K., Simmons, B. D., et al. 2015, MNRAS, 449, 820, [1502.03444].

Williams, J. P., Blitz, L., & McKee, C. F. 2000, in Protostars and Planets IV, ed. V. Mannings, A. P. Boss, & S. S. Russell, 97

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868, [1008.0031].

Xiao-Qing, W. & Jin-Meng, Y. 2021, Chinese Journal of Physics, 69, 303, [3330.789].

Yang, S., Xiao, W., Zhang, M., et al. 2022, arXiv e-prints, arXiv:2204.08610, [2204.08610].

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579, [astro-ph/0006396].

Zahid, H. J., Dima, G. I., Kewley, L. J., Erb, D. K., & Davé, R. 2012, ApJ, 757, 54, [1207.5509].

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. 2017, arXiv e-prints, arXiv:1710.09412, [1710.09412].

Zhou, X., Gong, Y., Meng, X.-M., et al. 2022, MNRAS, 512, 4593, [2112.08690].

Acronyms

AGN Active galaxy nuclei
ANN Artificial neural networks
BAO Baryon acoustic oscillations
CALIFA Calar Alto Legacy Integral Field Area
CCD Charge-coupled device
CEL Collisionally excited emission lines
CEFCA Centro de Estudios de Física del Cosmos
CNN Convolutional neural networks
DT Decision tree
ELG Emission Line galaxy
FWHM full-width-half-maximum
GALEX Galaxy Evolution Explorer
HLR Half-light radius
HST Hubble Space Telescope
IMF Initial mass function
IR Infrared
ISM Interstellar medium
J-PAS Javalambre Physics of the Accelerating Universe Astrophysical Survey
JPCam Javalambre Panoramic Camera
J-PLUS Javalambre Photometric Local Universe Survey
LRG Luminous Red galaxy
MAD Median absolute deviation
MaNGA Mapping Nearby Galaxies at Apache Point Observatory
ML Machine learning
OAJ Observatorio Astrofísico de Javalambre
PSF Point spread function
QSO Quasi-stellar object
RF Random forest

SDSS Sloan Digital Sky Survey
SED Spectra energy distribution
SF Star formation
SFH Star formation history
SFMS Star formation main sequence
SFR Star formation rate
SMBH Supermassive black holes
sSFR Specific star formation rate
UV Ultraviolet
WISE Wide-field Infrared Survey Explorer

List of Tables

List of Figures