What drives the scatter of local star-forming galaxies in the BPT diagrams? A Machine Learning based analysis

Mirko Curti,[1,2] ⋆ Connor Hayden-Pawson,[1,2] Roberto Maiolino,[1,2,3] Francesco Belfiore,[4]
Filippo Mannucci,[4] Alice Concas,[5,4] Giovanni Cresci,[4] Alessandro Marconi,[5,4]
and Michele Cirasuolo[6]

[1] *Cavendish Laboratory, University of Cambridge, 19 J. J. Thomson Ave., Cambridge CB3 0HE, UK*
[2] *Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[3] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[4] *INAF - Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125, Firenze, Italy*
[5] *Dipartimento di Fisica e Astronomia, Universitá di Firenze, Via G. Sansone 1, 50019, Sesto Fiorentino (Firenze), Italy*
[6] *European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching bei Muenchen, Germany*

Accepted XXX. Received XXX

ABSTRACT

We investigate which physical properties are most predictive of the position of local star forming galaxies on the BPT diagrams, by means of different Machine Learning (ML) algorithms. Exploiting the large statistics from the Sloan Digital Sky Survey (SDSS), we define a framework in which the deviation of star-forming galaxies from their median sequence can be described in terms of the relative variations in a variety of observational parameters. We train artificial neural networks (ANN) and random forest (RF) trees to predict whether galaxies are offset above or below the sequence (via classification), and to estimate the exact magnitude of the offset itself (via regression). We find, with high significance, that parameters associated to variations in the nitrogen-over-oxygen abundance ratio (N/O) are the most predictive for the [N ii]-BPT diagram, whereas properties related to star formation (like variations in SFR or EW(Hα)) perform better in the [S ii]-BPT diagram. We interpret the former as a reflection of the N/O-O/H relationship for local galaxies, while the latter as primarily tracing the variation in the effective size of the S^+ emitting region, which directly impacts the [S ii] emission lines. This analysis paves the way to assess to what extent the physics shaping local BPT diagrams is also responsible for the offsets seen in high redshift galaxies or, instead, whether a different framework or even different mechanisms need to be invoked.

Key words: galaxies: ISM – galaxies: abundances – galaxies: evolution

1 INTRODUCTION

Rest-frame optical emission lines contain a wealth of information about the physics of gas and stars in star-forming galaxies. The relative intensity of both collisionally excited and recombination lines indeed reflects the properties of the ionising radiation source, dust content, density, temperature, chemical abundances, and kinematics of the gas within the emitting HII regions (Kewley et al. 2019). Classical diagnostic diagrams based on optical emission lines, such as the [O iii]λ5007/Hβ versus [N ii]λ6584/Hα (Baldwin et al. 1981) and [O iii]λ5007/Hβ versus [S ii]$\lambda\lambda$6717, 31/Hα (Veilleux & Osterbrock 1987), also known as the 'BPT' diagrams, have been widely used in the literature to discriminate between different ionising sources and excitation mechanisms in galaxies, in order to separate, for instance, galaxies ionised by star formation processes from those whose spectra are dominated by the presence active galactic nuclei (AGNs) in their centre. Both Kewley et al. (2001) and Kauffmann et al. (2003c) provided classification schemes to separate star-forming galaxies from AGNs, the former based on predictions from photoionization models, while the latter more empirical. Interestingly, star-forming galaxies in the local universe are observed to follow a remarkably tight sequence in these diagrams, which is generally interpreted as a result of the correlation between metallicity and ionization parameter (U) in star-forming galaxies (McCall et al. 1985; Dopita & Evans 1986; Mingozzi et al. 2020). Indeed, strong-line metallicity diagnostics widely adopted in large statistical studies are often based on calibrating the position of galaxies

⋆ E-mail: mc2041@cam.ac.uk

on such diagrams against their oxygen abundance (see Maiolino & Mannucci 2019, for a review).

The advent of integral field spectroscopic surveys of local galaxies like CALIFA(refs), MaNGA (Bundy et al. 2015) and SAMI(refs), provided the chance to review the standard classification schemes by leveraging on the information about the spatial variation of emission line ratios across galaxies. For instance, many studies have shown that spectra from low-ionisation emission line regions (LINER) are not necessarily associated to a nuclear origin (e.g., Yan & Blanton 2012; Belfiore et al. 2016; Hsieh et al. 2017). Various authors have also explored and modelled different multi-dimensional reprojections of the standard line ratios diagnostics to attempt breaking some of the degeneracies in the determination of seyfert-like, shock-like and star-forming spaxels in integral field data (e.g., D'Agostino et al. 2019; Ji et al. 2020; Law et al. 2021b). Aside from discriminating between different ionising sources, it is interesting to note that the distribution of star-forming galaxies itself in the BPT diagrams presents a non-negligible amount of scatter, which is shown to correlate with different physical properties (e.g., Masters et al. 2016; Faisst et al. 2018). Therefore, these diagrams are a valuable source of information, as the relative position of sources within the plane can be used to get more insights on the difference in their underlying physical conditions.

Moreover, in the last decade it has been widely demonstrated that star-forming galaxies at high redshift (i.e., 1 <z< 3) occupy a slightly different position on the classical BPT diagrams, showing, on average, a clear offset towards higher [O III]λ5007/Hβ and/or [N II]λ6584/Hα compared to their local counterparts. In general, such deviation is attributed to a combined effect of the evolution in the underlying stellar populations associated to, e.g., a hardening of the far ultraviolet (FUV) ionising spectrum, alpha-enhancement, contribution from binarity and rotation (e.g., Steidel et al. 2014; Strom et al. 2017; Topping et al. 2020b; Topping et al. 2020a) and/or in the physical properties of the ISM like density, ionisation parameter and gas chemical abundances (e.g., Shapley et al. 2015; Yabe et al. 2015; Kashino et al. 2017; Masters et al. 2016). Several attempts have been made to theoretically model the emission line ratios in the BPT diagrams and reproduce their variation with cosmic time, by coupling the evolution of galaxy properties from cosmological simulations with state-of-the-art stellar and nebular emission models (e.g., Kewley et al. 2013; Byler et al. 2017; Hirschmann et al. 2017; Xiao et al. 2018). However, although variations in emission line ratios can be theoretically reproduced by means of the interplay of many different parameters, it is often difficult to break the degeneracy and disentangle their true relative contribution, which requires both a large variety of independent observational constraints as well as a careful assessment of the underlying model assumptions. For instance, some models often assume fixed, underlying correlations between different abundance patterns with zero-scatter (e.g., N/O and C/O with O/H), and hence cannot grasp the direct impact of variations of such abundances at fixed metallicity on the modelling of the emission lines.

With this scenario in mind, in this work we present a complementary, self-consistent and fully data-based framework which exploits machine learning algorithms to quantitatively describe how the distribution of local star-forming galaxies across the BPT diagrams is connected to different observational properties and what we can infer about the relationships between the observed variations in line ratios and the underlying physics of star-forming galaxies. In particular, we leverage on the large statistics provided by the Sloan Digital Sky Survey (SDSS, York et al. 2000) to perform a detailed analysis of the dependencies between the deviation of galaxies from the mean star formation (SF) locus and a variety of key observational parameters, directly or indirectly tracing different physical properties of galaxies.

In recent years in fact, machine learning techniques have seen an increasingly significant impact on astronomical studies, in response to the undergoing rapid growth in size and complexity of datasets as provided by current large surveys like SDSS, MANGA or GAIA (Gaia Collaboration et al. 2016), and in preparation for future ones like DESI (Levi et al. 2013), SKA (Dewdney et al. 2009) and LSST (Ivezic et al. 2008). Such algorithms are successfully implemented to solve a variety of different problems, including the classification of galaxy morphological types (e.g., de la Calleja & Fuentes 2004; Barchi et al. 2020; Vavilova et al. 2021; Reza 2021), the identification of transients (Sooknunan et al. 2021), or the multi-parametric analysis of very large databases of galaxy properties (e.g., Teimoorinia et al. 2016, 2021; Ho 2019; Bluck et al. 2019, 2020). Inspired especially by the latter works, in this paper we train and test artificial neural networks (ANN) and random forest decision trees (RF) to assess the performance of a set of carefully selected parameters (both individually and as a whole) and identify which properties are the most relevant in predicting the observed deviation of star-forming galaxies from their average sequence in both the [N II]- and [S II]-BPT diagrams. In a forthcoming paper of this series, we will expand on the present work by exploiting the information provided by MaNGA in order to compare trends on global/integrated and local/spatially resolved scales. This approach, if successful in describing what observed in the local Universe, could then be tested on high redshift galaxy samples to assess to what extent the physics that govern the scatter in local BPT diagrams is the same causing the observed evolution in the emission line properties at high-z or, instead, whether a different framework or even different physical mechanisms need to be involved.

The current paper is structured as follows. In Section 2 we describe the observational dataset, the sample selection and we introduce the set of parameters adopted in this work. In Section 3 we describe framework and metrics adopted for describing the scatter within the [N II]-BPT diagram, whereas in Section 4 the proper machine learning analysis is performed. In Section 5, we repeat the same analysis for the [S II]-BPT diagram. We summarise the results and present our conclusions in Section 6. Throughout this paper, we assume a standard ΛCDM cosmology with with H$_0$ = 70 km s^{-1}, Ω_m=0.3, and Ω_Λ = 0.7.

2 DATA

2.1 SDSS data

2.1.1 *Sample Selection*

The sample is drawn from the seventh data release (DR7) of the Sloan Digital Sky Survey (SDSS) (Abazajian et al. 2009), whose galaxy properties and emission line fluxes are provided by the MPA/JHU catalog[1]. We selected galaxies classified as star forming according to their position on the [N II]-BPT diagram, following the more robust classification scheme by Kauffmann et al. 2003b and further requiring the equivalent width of the Hα to be higher than 6, in order to set a more stringent limit to contributions from low ionisation gas powered by different types of stellar populations (e.g., Cid Fernandes et al. 2011; Zhang et al. 2017; Lacerda et al.

[1] available at http://www.mpa-garching.mpg.de/SDSS/DR7/

2018). We applied a redshift cut on z > 0.035 in order to ensure the presence of the [O II]λ3727 emission line within the wavelength coverage of the SDSS spectrograph and sample at least the inner 2 kpc of galaxies. In addition, we discarded all galaxies whose catalogue flags indicates unreliable stellar mass and star-formation rate (SFR) estimates, which includes also all those galaxies showing non-physical aperture correction factors lower than 1. Moreover, we applied a signal-to-noise (S/N) threshold[2] of 5 on Hα and 3 on all the emission lines involved in the analysis, namely Hβ, [O II]λ3726, 3729, [O III]λ5007, [N II]λ6584 and [S II]λ6718, 6732. All emission lines were corrected for reddening, where required, from the measured Balmer Decrement (assuming an intrisinc value of Hα/Hβ=2.87, as given by the case B recombination) and adopting the Cardelli et al. (1989) extinction law. Finally, we removed galaxies affected by poor photometric deblending by selecting on the DEBLEND_NOPEAK and DEBLEND_AT_EDGE flags, as well as galaxies whose aperture correction factors are lower than 1 (e.g., where the stellar mass derived from the total photometry is lower that the stellar mass derived within the fibre). After applying all these criteria, the total analysed sample is reduced to 128,120 galaxies.

2.1.2 *Observational parameters and physical properties*

In this work we aim at quantitatively assessing which physical properties are most connected with the position of galaxies in the BPT diagrams. Therefore, we consider direct measurements of physical quantities, as well as a variety of different observational proxies, as the main parameters in our analysis. The full list of involved parameters is described in the following and also reported in Table 1 and 2.

The total stellar mass for our galaxies is provided by the MPA/JHU catalog and have been estimated from fits to the photometry, following the prescription of Kauffmann et al. (2003a) and Salim et al. (2007). Star formation rates used in this work are derived from the extinction corrected Hα luminosity inside the fibre, adopting the calibration proposed by Kennicutt & Evans (2012). We then apply the aperture corrections provided by the MPA/JHU catalog, which build on the work of Salim et al. (2007) to improve those originally provided by Brinchmann et al. (2004), to compute the total SFR for our galaxies. Both stellar masses and SFRs estimates are re-scaled to a common Chabrier (2003) IMF.

Spectral indices like $D_N(4000)$ and EW($H\alpha$) are provided for SDSS galaxies by the MPA/JHU catalog too. In particular, EW($H\alpha$) is a model-independent tracer of the specific star formation rate (sSFR=$M_\star$/SFR), quantifying the relative contribution of recent star formation on the integrated star formation history (SFH) of galaxies, whereas $D_N(4000)$ is a sensitive probe of the overall ageing of the stellar population. We also consider the central velocity dispersion of the Balmer lines (e.g., $\sigma_{H\alpha}$) as a tracer of the gas kinematics and potentially revelatory of non-virial motions (as shocks are known to produce kinematic components with velocity dispersion significantly larger than those of HII regions, see e.g., Rich et al. 2010; D'Agostino et al. 2019; Law et al. 2021b).

In terms of properties derived from emission line ratios, we measure the gas-phase metallicity exploiting the calibrations presented in Curti et al. (2017, 2020) (which are defined on the T_e-based abundance scale). We refer to Curti et al. (2020) for a detailed description of the procedure, where metallicity is constrained by simultaneously adopting several emission line ratios in order to minimise the degeneracies and biases intrinsic to each individual calibration. The [N II]/[O II] and [N II]/[S II] ratios are taken as an observational proxy of the nitrogen-over-oxygen (N/O) abundance. More specifically, [N II]/[O II] traces the N^+/O^+ ionic abundance ratio, which closely matches the N/O abundance ratio because of the similar ionization structures of the two elements. The [N II]/[S II] ratio works well in tracing N/O too, given the close ionisation potential of S^+ and O^+ ions, although presenting a small residual dispersion compared to [N II]/[O II], which is shown to correlate for instance with the star formation rate (Hayden-Pawson et al., in prep.). Both line ratios can be easily converted to N/O following a variety of different calibrations (e.g., Hayden-Pawson et al., in prep., based on the SDSS stacked spectra described in Curti et al. 2017, or Pérez-Montero & Contini 2009).

The ionisation parameter instead is mapped on the [O III]λ5007/[O II]λ3727, 29 and [Ne III]λ3869/[O II]λ3727, 29 line ratios, which can be converted to U following the calibrations presented in Kewley et al. (2019). The former (Aller 1942) is easily observed across the full sample of selected star-forming galaxies, but presents a secondary dependence on metallicity (Kewley & Dopita 2002), while the latter (Levesque & Richardson 2014) is mildly affected by dust extinction, but requires the detection of the faint [Ne III]λ3869 in individual sources. Finally, the electron density of the gas (N_e) is traced by the observed ratio between the lines of the sulfur doublet, i.e. [S II]λ6718/[S II]λ6732, which is a widely adopted diagnostic in star-forming HII regions as it is highly sensitive to N_e in the regime between the critical densities of the two lines (Osterbrock & Ferland 2006).

3 FRAMEWORK

3.1 Variation of physical properties across the [N II]-BPT star-forming sequence

We will initially focus our investigation on the [N II]-BPT diagram, where star forming galaxies in the local Universe are observed to form quite a tight sequence. In this work, we adopt a polynomial fit as representative of the locus of highest density of star-forming galaxies along the sequence (hereinafter, SF sequence, or SF locus), as originally provided by Kewley et al. (2013) and given by

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) = 0.61/(\log([\mathrm{N\,II}]/\mathrm{H}\alpha) + 0.08) + 1.1\,. \quad (1)$$

In Figure 1, the selected star-forming SDSS galaxies are plotted in the [N II]-BPT diagram and colour-coded in each panel by the different galaxy properties and parameters described in Section 2.1.2. To aid the visualisation of the underlying trends, galaxies are binned in small hexagons, and the average value of each parameter in such bins is considered. In addition, lines of constant value (i.e., iso-contours) in each parameter are marked in white, while the polynomial fit to the SF sequence of equation 1 is shown by the red curve. Finally, the histogram of values in a given parameters is shown, together with the average and standard deviation of the distribution, within its reference panel.

By visually inspecting Fig. 1, it is immediately evident that the relative position of galaxies on the [N II]-BPT diagram is strongly correlated with different physical properties. In particular, moving along the sequence of star-forming galaxies (e.g., from the bottom-right to the upper-left) we

[2] applying the re-scaled uncertainties provided by the MPA/JHU group, which include both the uncertainties on the spectrophotometry and continuum subtraction

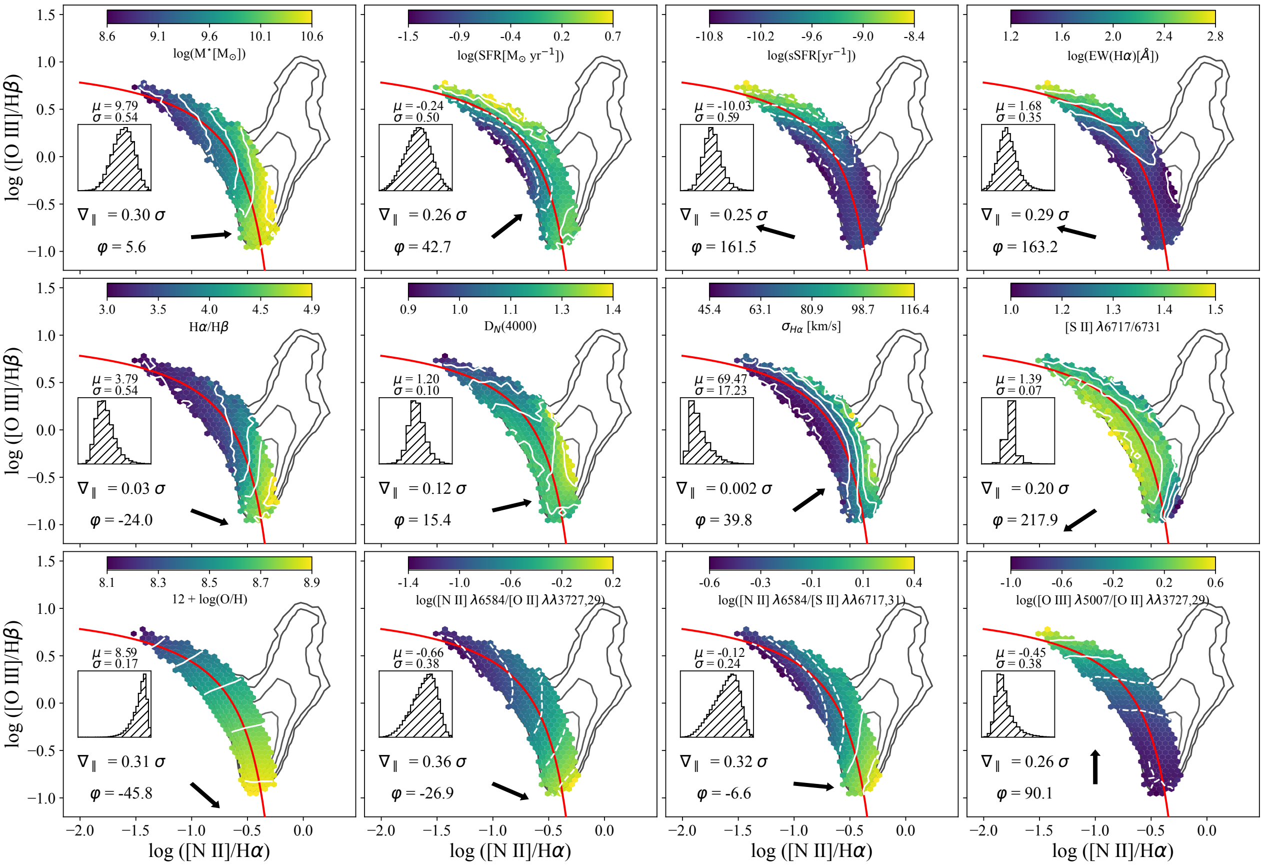

Figure 1. Distribution of the selected star-forming galaxies from the SDSS in the [N ii]-BPT diagram. In each panel, the colour-coding reflects the average value computed in small hexagonal bins for the following parameters (moving along rows from the upper left to the bottom right): log($M_\star$), log(SFR), log(sSFR), log(EW[Hα]), Hα/Hβ, Dn4000, $\sigma_{H\alpha}$, [S ii]λ6717/λ6731,12+log(O/H), log([N ii]λ6584/[O ii]$\lambda\lambda$3727, 29), log([N ii]λ6584/[S ii]$\lambda\lambda$6717, 31), log([O iii]λ5007/[O ii]$\lambda\lambda$3727, 29). White contours indicate lines of constant values in each parameter. The best-fit to the median SF locus from Eq. 1 is indicated by the red curve. In each panel, we also plot an histogram of the distribution of values for the reference parameter, reporting its average and standard deviation, and the $\nabla_{||}$ statistics, which quantifies the magnitude of the gradient in that parameter along the best-fit curve, in units of σ. Finally, in each panel we show the vector representative of the direction of maximum variation across the diagram in the reference parameter, whose angle is computed (positive counter-clock wise from the x-axis) from the ratio of the partial correlation coefficients between that given parameter and each individual BPT-axis, i.e. with [O iii]λ5007/Hβ at fixed [N ii]λ6584/Hα, and viceversa.

can recognise clear trends in stellar mass, specific star formation rate (or, equivalently, EW(Hα)), gas-phase metallicity, [N ii]λ6584/[S ii]λ6717, 31 and [N ii]λ6584/[O ii]λ3727, 29 (both tracing N/O), [O iii]λ5007/[O ii]λ3727, 29 (tracing primarily U) and Hα/Hβ (tracing dust extinction).

However, a more careful assessment provides deeper insights about the nature of such dependencies. In particular, the distribution of star-forming galaxies in the diagram form a very smooth sequence in oxygen abundance, with variations in log(O/H) closely following the shape of the SF locus, and lines of constant metallicity which are, instead, almost orthogonal to the best-fit line at any given point. Not surprisingly, the [N ii]-BPT has been modelled and calibrated for a long time against oxygen abundance to serve as a metallicity diagnostic, and likewise, have been the individual line ratios upon which the diagram is defined (e.g., Pettini & Pagel 2004; Maiolino et al. 2008). Different properties (and their tracers), which are physically connected to metallicity like $M_\star$, N/O and U, although characterised by the presence of a strong gradient along the sequence, do show iso-contours at various levels of inclination with respect to the best-fit line. For instance, lines of constant [O iii]λ5007/[O ii]λ3727, 29 (i.e., constant U) are almost everywhere parallel to the x-axis (hence spanning a variety of different inclinations from the best-fit line of the SF of equation 1), demonstrating the strong correlation between ionisation parameter and [O iii]λ5007/Hβ. For parameters like SFR and $\sigma_{H\alpha}$ instead, it is more difficult to identify a clear trend along the SF sequence, whereas clear segregation in such parameters can be seen between galaxies lying leftmost and rightmost the best-fit curve.

We can try to quantitatively estimate the amplitude of the variation in each parameter along the SF sequence as described below. First, we perform a bi-variate spline interpolation over the underlying (binned) distribution of star-forming galaxies, so to infer the values assumed by each parameter at any discrete (sampling) point along the best-fit curve of the SF locus.From the array of values assumed by each parameter on the SF sequence best-fit curve, we can then compute a 'gradient array' (from second order accurate central differences in the interior points of the original array), whose amplitude (i.e., the square root of the sum of its elements, taken in quadrature) is reported in each panel as $\nabla_{||}$: such statistics are useful to quantify how strongly each parameter varies as we move along the the best-fit line of the SF locus or, in other words, to what extent the sequence of star-forming galaxies in the [N ii]-BPTdiagram can be interpreted as a sequence in a given physical property. This quantity is further normalised by the 1σ dispersion of values in each parameter, in order to account for the different dynamic ranges and allow a meaningful comparison between different quantities.

The highest $\nabla_{\|}$ values ($> 0.30\sigma$) are found for $M_\star$, O/H, [N II]/[O II] and [N II]/[S II], confirming that the sequence of star-forming galaxies in the [N II]-BPT is primarily a sequence in stellar mass, metallicity and nitrogen-over-oxygen abundance (i.e., the variation in such parameters along the SF sequence is relatively large compared to the overall distribution of values within the entire diagram). Relatively high scores in $\nabla_{\|}$ are marked also by [O III]/[O II], (s)SFR and EW(Hα). However, we note that although for some properties like O/H a large $\nabla_{\|}$ is actually associated with a smooth and monotonic variation along the sequence, for others (like e.g., SFR) it is the result of having the best-fit line crossing more irregular patterns within the diagram, hence varying even rapidly but not necessarily monotonically along the curve.

Another potentially interesting aspect to consider is how much each parameter is correlated individually with the line ratios of the [N II]-BPT, i.e., with the two axis of the diagram, if taken separately. To estimate this we compute, for each given parameter, the Spearman partial correlation coefficients with both [N II]λ6584/Hα and [O III]λ5007/Hβ; a partial correlation coefficient quantifies the strength of correlation between two variables while keeping fixed the third (and/or further variables in case of higher dimensionality problems), and is defined as

$$\rho_{AB,C} = \frac{\rho_{AB} - \rho_{AC} \cdot \rho_{BC}}{\sqrt{1-\rho_{AC}^2}\sqrt{1-\rho_{BC}^2}} \tag{2}$$

where ρ_{AB} indicates, in general, the Spearman correlation coefficient between the two variables A and B. We then follow, e.g., Bluck et al. (2019), and define the vector representing the preferential direction of variation in a given parameter across the diagram, i.e., which is the average direction one should follow across the diagram in order to maximise the variation in that given parameter. The inclination of such vector with respect to the horizontal axis can be derived from the arctangent of the ratio of its two components, i.e., from the ratio of the partial correlation coefficients of its specific reference parameter with the individual BPT axis:

$$\varphi = \tan^{-1}\left(\frac{\rho_{Y\,p,X}}{\rho_{X\,p,Y}}\right) \tag{3}$$

where p is any of the parameters in our set and Y, X are the log([O III]λ5007/Hβ), log([N II]λ6584/Hα) line ratios, respectively. Such ‘correlation vector’ and its associated φ angle are shown for all parameters in the corresponding panel of Fig. 1.The analysis is performed on binned data in order to avoid biases introduced by the non homogeneous density distribution of individual galaxies across the diagram, as well as to remove strong outliers.

The introduction of the ‘correlation vector’ further confirms that the direction of preferred variation in metallicity is closely aligned with the shape of the SF sequence ($\varphi = -46°$, pointing from the upper-left to the bottom-right in the diagram), being positively correlated with the x-axis while negatively with the y-axis. We further note that for N/O tracers, the gradient vector is more inclined towards the x-axis for [N II]/[O II] than it is for metallicity, whereas it has almost a flat inclination ($\varphi = -6.6°$) for [N II]/[S II], as possibly driven by the secondary, additional SFR-dependence of such line ratio. For ionisation parameter tracers like log([O III]λ5007/[O II]λ3727, 29) instead, the vector is almost perfectly vertical ($\varphi = 90.1°$), showing that such parameter in star-forming galaxies is almost entirely (positively) correlated with the y-axis and basically uncorrelated with the x-axis (when the other axis is fixed) in the [N II]-BPT. For the other parameters, the direction of the gradient vector presents different levels of inclination with respect to the SF locus: $D_N(4000)$ is well correlated with [N II]λ6584/Hα (low φ values), whereas $\sigma_{H\alpha}$ and SFR present gradient vectors whose inclination (close to $\varphi \sim 45°$) suggests a comparable level of correlation with both [N II]-BPT axis, taken individually.

3.2 Metrics: Δlog(p), distance D and angle θ

Following the observations and the analysis presented in the previous Section, we now take a step further and try to build a relatively straightforward modelling of the observed scatter in the BPT diagrams, which is based on the two main assumptions described below:

i) each galaxy which is shifted from the best-fit curve (i.e., which does not follow the bulk of galaxy distribution along the SF sequence) experiences an offset which we describe to occur *orthogonal to the curve* at any point (hence, it experiences the minimum possible offset);

ii) such an offset correlates with *relative variations* in, either one or more, physical parameters, when compared to the average values pertaining to galaxies which closely follow the SF sequence.

The aim of the subsequent analysis is therefore to connect the offset from the best-fit line of the SF sequence with the observed variation in different physical parameters, quantify the amount of underlying correlation and assess which parameters are the most useful in predicting the observed deviation of galaxies from the median loci across the diagram. For each galaxy in the sample, we thus introduce the $\Delta\log(p)$ metric, defined as the difference between the (logarithm of the) value assumed by a given galaxy in the parameter p and the average value assumed by galaxies which lie on the closest point along the best-fit curve of the SF sequence (i.e., assuming a purely orthogonal offset):

$$\Delta\log(p) = \log(p) - <\log(p)>_{\text{SF locus}} . \tag{4}$$

We note here that considering the logarithm of each quantity makes the comparison between different parameters more meaningful and straightforward.

In Fig. 2 we replicate the scheme of Fig. 1, but in this case the small hexagons in each panel are colour-coded by the average variation in the logarithm of the relative parameter (i.e., the average $\Delta\log(p)$ in the bin), as defined in equation 4; the best-fit line of the SF sequence is instead coloured according to the typical value assumed by each parameter at any given point along the curve, as inferred from interpolating over the underlying galaxy distribution (we refer to the previous subsection for more details). The colour scheme (centred on zero) helps to identify trends between the relative location of galaxies and the magnitude of variations in the various parameters: for instance, whether a galaxy occupies the region above or below the best-fit curve is visually seen to correlate overall very well with different properties, e.g. both N/O tracers, Δlog([O III]/[O II]), $\Delta\sigma_{H\alpha}$, whereas the strength of the correlation with variations in other parameters like SFR or EW(Hα) appears more limited to specific regions of the diagram.

We note here that an important corollary following directly from our framework and main assumptions is that the location of any given galaxy lying on the best-fit curve can be, in principle, predicted by the only knowledge of its gas-phase oxygen abundance, whereas any offset can be considered to occur at fixed O/H, as iso-contours in this quantity appear orthogonal everywhere to the SF sequence (see Fig. 1). Indeed, the amplitdue of Δlog(O/H) is basically zero (or very mild) almost everywhere across the diagram, meaning that for any given point on the SF sequence, all galaxies located along an orthogonal line that originates from that point can be assumed to have the same metallicity.

We first attempt to quantify the amount of variation in each parameter across the SF sequence by means of the $\Delta_{\perp}$ statistics, defined as the difference between the average Δlog(p) computed in galaxies lying in the regions upward and downward the best-fit line, normalised by the standard deviation of (the logarithm of) values spanned by each parameter; this quantity is reported for each feature within the corresponding panel of Fig. 2. In terms of absolute dynamic range, the amplitude of Δlog(p) is maximum for $\sigma_{\mathrm{H}\alpha}$(0.74 σ), and relatively high also for $\mathrm{M}^{\star}$, SFR and both N/O tracers, whereas it is minimum (as expected) for metallicity. We note here that Δlog(p) statistics grasp well what could be already inferred by visually inspecting Fig. 1 and 2, quantifying how the average orthogonal variation in a given parameter with respect to the SF sequence best-fit compares with the 'width' of the overall distribution of values in that parameter (i.e., with σ). However, we also stress that a high value in Δlog(p) does not necessarily imply a stronger causal connection with the offset from the SF sequence, as some parameters might be intrinsically more connected with it even if characterised by a lower dynamical range in their logarithmic variation. In order to properly ascertain which parameters in our set are of most impact on the level of scatter in the diagram, we will therefore exploit a number of machine learning (ML) techniques, as outlined in the following Sections. The full set of parameters and the associated properties and statistics discussed in this Section are summarised in Table 1.

The ML analysis will be targeted at reproducing (with the highest possible accuracy) the offset of galaxies from the SF sequence in the BPT diagrams. We can thus introduce a few more parameters, whose definitions are based on the framework described above, which will help us in identifying the target labels for the ML algorithms; such quantities are here described for the [N II]-BPT, but are defined in an equivalent way for the [S II]-BPT, as discussed later in Section 5. Firstly, for each galaxy in the diagram we can define the *offset vector* as the vector pointing to the same galaxy and originating from the closest point on the best-fit line of the SF sequence (i.e., the vector is orthogonal to the curve at any point and its amplitude is the minimum possible). Its length **D** is then simply given by the Euclidean distance of the galaxy from the best-fit curve defined by equation 1 in the [N II]-BPT diagram parameter space, and can be written, in terms of its components, as :

$$\mathbf{D} = \sqrt{\sum_i (\Delta q_i)^2}\,, \tag{5}$$

where Δq is the difference between the q-coordinate of a given galaxy in the diagram and the q-coordinate of the nearest point on the best-fit curve, with $q \in$ [log([N II]λ6584/Hα), log([O III]λ5007/Hβ)] for the [N II]-BPT. In the left-hand panel of Fig. 3, the distribution of star-forming galaxies in the diagram is now colour-coded by **D** (again, averaged in small hexagonal bins to aid visualisation): galaxies lying below the SF locus best-fit are assigned a negative value of **D**, in order to distinguish them from galaxies located above. This quantity represents one of the target labels for the machine learning analysis presented in Section 4, but can be also simply used to identify whether a galaxy is located above or below the SF sequence.

In the right-hand panel of Fig. 3 instead, each hexagonal bin is colour-coded according to the (average) angle formed by the offset vector of a galaxy with the horizontal axis (increasing positive counterclockwise), as given by :

$$\theta = \tan^{-1}\left(\frac{\Delta\log([\mathrm{O\,III}]/\mathrm{H}\beta)}{\Delta\log([\mathrm{N\,II}]/\mathrm{H}\alpha)}\right). \tag{6}$$

This parameter is useful to quantify which is the predominant component of the offset vector, i.e. whether the offset occurs preferentially along the [N II]/Hα- or the [O III]/Hβ-axis. Given the shape of the distribution of star forming galaxies within the [N II]-BPT, offsets in the bottom-right part of the sequence occur preferentially along [N II]/Hα (i.e., low values of θ), whereas θ increases (hence deviations in [O III]/Hβ becomes increasingly more relevant) as we move along the SF sequence towards the upper-left region. Whether this has an impact on the results of the ML analysis will be specifically addressed in Section 4.4.

4 MACHINE LEARNING ANALYSIS

4.1 Algorithms, problems and parameters

In the previous section, we have attempted to assess how the position of star-forming galaxies within the [N II]-BPT diagram correlates with a variety of physical parameters, by visually inspecting the distribution of such parameters within the diagram and introducing statistics aimed at quantifying the amplitude of variations along and across the SF sequence. Here, we move a step forward and implement different machine learning algorithms in order to provide a more robust and quantitative assessment of the drivers of the scatter across the star-forming galaxy sequence in the diagram. The ultimate goal is to provide a method to robustly identify which physical parameters are statistically more connected with the deviation from the SF locus, adopting a framework which is completely based on observational data and independent on any of the standard prescriptions included in the majority of photoionisation models. In practice, Artificial Neural Networks (ANN) and Random Forest (RF) of decision trees are trained and tested on our large sample of selected SDSS star-forming galaxies in order to solve both a *classification* and a *regression* problem. The former, aimed at describing which parameters perform better in predicting whether a galaxy is simply located above or below the best-fit curve representative of the SF sequence. The latter, instead, to assess the ability of each variable (and of the whole set) in predicting the exact distance (i.e., the amplitude **D** of the offset vector described in equation 5) from the SF sequence itself.

The implementation of both ANN and RF algorithms allows us to tackle these two problems from rather different angles. With the ANN, we aim at exploring the performance of each parameter individually, as well as the maximum potential of the full set as a whole, by means of a model-independent approach free of any underlying assumptions about correlations, linearity and monotonicity within the data. Unfortunately, when fed with a set of multiple parameters, ANN do not provide information about the relative impact that each individual parameter has in contributing to its overall predictive power, i.e., one might ask whether the full set of parameters is really required to achieve the highest level of accuracy or even a subset could provide comparable results and, ultimately, which parameters specifically contain the informations that maximise the predictivity of the model. For this reason, we perform the same analysis implementing also RF decision trees, which intrinsically allows us to disentangle the relative importance of even highly correlated features involved in the prediction algorithm. In other words, by means of the RF analysis we aim at assessing which

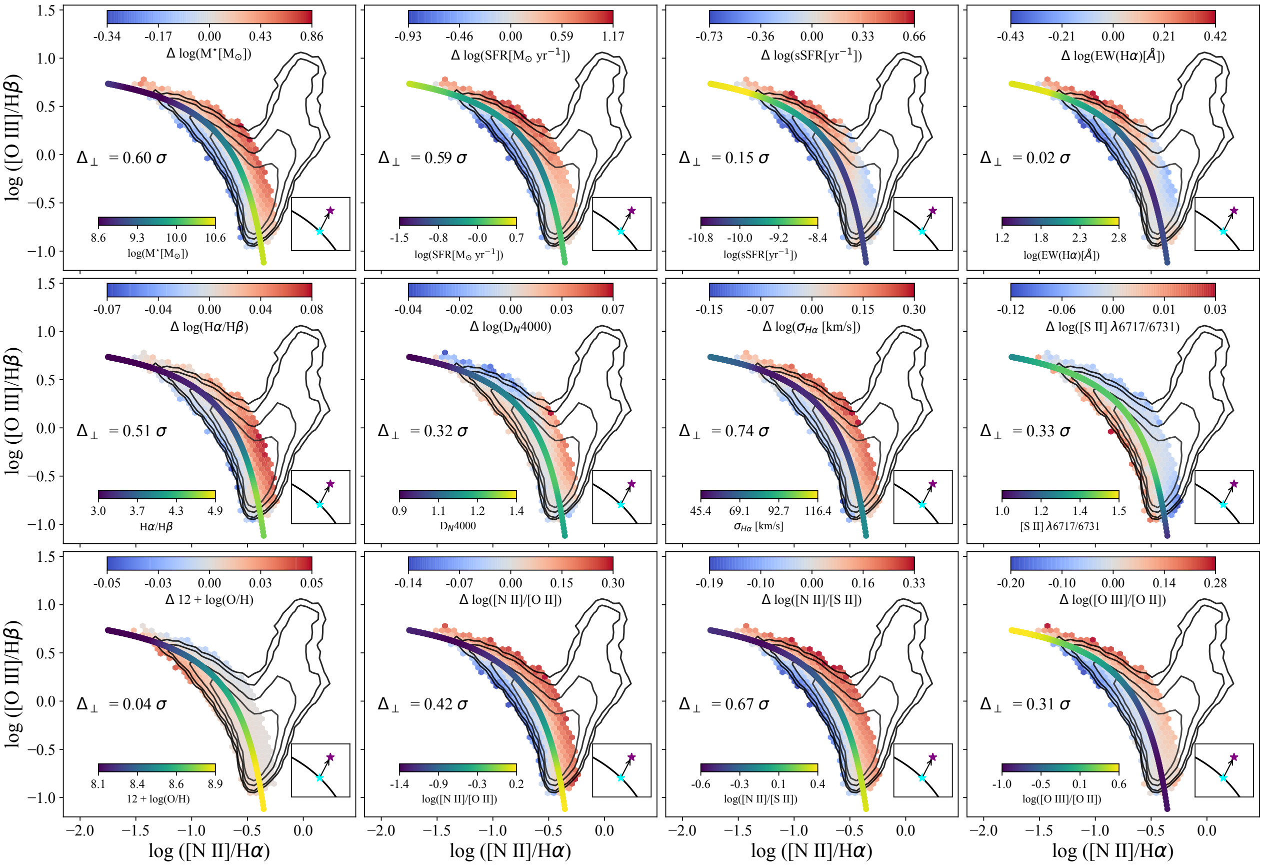

Figure 2. In this figure, the panels are organised as in Fig. 1, but now the hexagonal bins are colour-coded by the average Δlog(p), as defined for equation 4 for each parameter and assuming a purely orthogonal 'offset vector' from the best-fit curve of the SF locus. Within each panel, such curve is colour-coded by the average value assumed in that parameter by galaxies lying exactly on the SF sequence, at any given point. We also report the $\Delta_\perp$ statistics, which quantifies the difference in Δlog(p) between galaxies lying above and below the SF locus curve, reported in units of standard deviation.

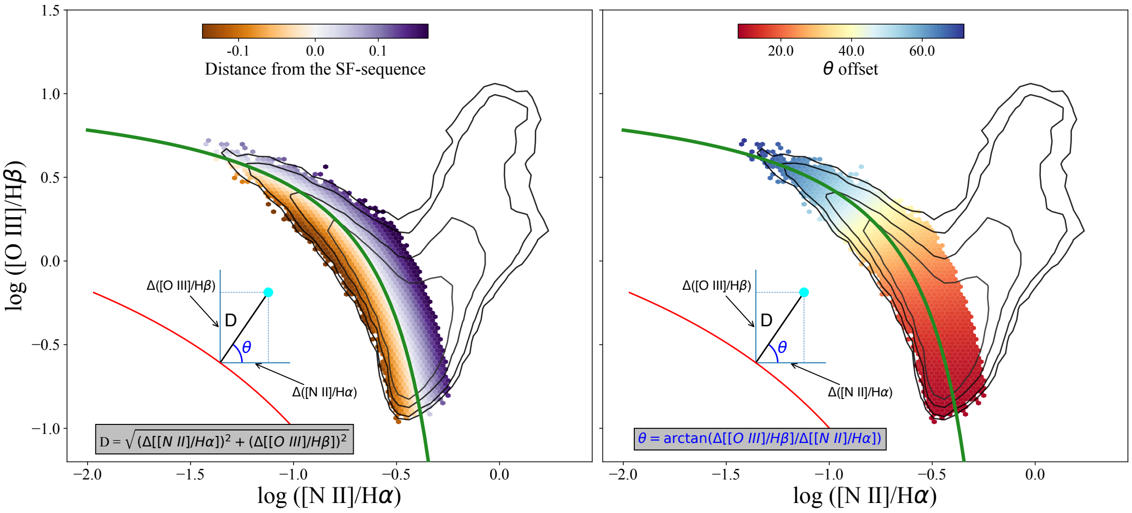

Figure 3. The distribution of SDSS star forming galaxies in the [N II]-BPT diagram is colour coded by the distance metric defined by equation 5 (*left panel*) and by the corresponding angle θ of the 'offset vector' with respect to the horizontal axis (*right panel*), according to equation 6.

Parameter	Physical property	$\nabla_{\parallel}$	φ	$\Delta_{\perp}$	multi-parameter set
$\log(M^{\star}[M_{\odot}])$	Stellar mass	0.3σ	5.6°	0.6σ	✓
$\log(\mathrm{SFR}[M_{\odot}\ \mathrm{yr}^{-1}])$	Star formation rate	0.26σ	42.7°	0.59σ	✓
$\log(\mathrm{sSFR}[\mathrm{yr}^{-1}])$	Specific SFR	0.25σ	161.5°	0.15σ	✗
EW(Hα)[]	Specific SFR	0.29σ	163.2°	0.02σ	✓
Hα/Hβ	Dust extinction	0.03σ	−24.0°	0.51σ	✗
$D_N(4000)$	Age of stellar populations	0.12σ	15.4°	0.32σ	✓
$\sigma_{H\alpha}$ [km/s]	Gas velocity dispersion	0.002σ	39.8°	0.74σ	✓
[S ii]λ6717/6731	Gas density	0.2σ	217.9°	0.33σ	✓
12 + log(O/H)	Oxygen abundance	0.31σ	−45.8°	0.04σ	✗
log([N ii] λ6584/[O ii] $\lambda\lambda$3727, 29)	N/O abundance	0.36σ	−26.9°	0.42σ	✗
log([N ii] λ6584/[S ii] $\lambda\lambda$6717, 31)	N/O abundance	0.32σ	−6.6°	0.67σ	✓
log([O iii] λ5007/[O ii] $\lambda\lambda$3727, 29)	Ionisation parameter (U)	0.26σ	90.1°	0.31σ	✓

Table 1. Full list of the parameters of interest for the analysis of the [N ii]-BPT diagram. For each quantity, we report the values of the statistics introduced in Secion 3 and reported in each panel of Fig. 1 and 2. In the last column, we mark the parameters which are included in the set for the multi-parametric ANN and RF analysis; for more details about the justification of such selection, we refer to Section 4.

of the involved (and intercorrelated) parameters are intrinsically the most informative in predicting our target variables when the full set is used in concert.

The set of parameters to be considered in the analysis is taken from the list of observables and properties discussed in Section 2.1; in particular, in the last column of Table 1 we mark which parameters are included in the multi parametric ML analysis. In fact, although each parameter is assessed through the ANN individually (i.e., by feeding the network with the data relative to only one parameter at a time), when evaluating the performances of the algorithms considering multiple parameters altogether it is warranted to perform a careful selection of the quantities to be included in the final set, in order to avoid nuisance parameters which either duplicate the physical information and/or are trivially correlated with others, or with the target labels. From now on, we refer to the analysis performed on such list of parameters as the 'multi-parameter' run(s), and to the list itself as the 'multi-parameter' set ; accordingly, the RF analysis will also be based upon the subset of parameters included in this list.

First, we decide not to include metallicity at all in the ML analysis. In fact, being oxygen abundance mostly derived from the combination of several strong line ratios (including the line ratios which constitute the BPT-axis, see Section 2 and Curti et al. 2020), such quantity can be trivially recovered from a combination other emission line-based parameters and the BPT line ratios themselves (which are at the basis of the definition of the distance target label **D**); hence, in our framework log(O/H) can be treated as a nuisance parameter, not independent from the others, which could bias the performances of both the 'multi-parameter' ANN run and the relative feature importance assessment performed by the RF. However, based on what is shown in Fig. 1 and on the assumptions **i)** and **ii)** discussed in Section 3, the contribution from metallicity to setting the offset from the best-fit line is likely to be negligible, being iso-O/H lines orthogonal to the SF sequence at any point (as also quantified by a $\Delta_{\perp}$ statistics ~ 0 for log(O/H)). Hence, we can add a third assumption to our framework, that is **iii)** any contribution to the observed offset from the SF sequence from any of the involved parameters is assumed here to occur at fixed metallicity. The validity of such assumption is further discussed later in the text.

Then, we chose to adopt [N ii]/[S ii] instead of [N ii]/[O ii] as a tracer of the N/O abundance in the 'multi-parameter' analysis of the [N ii]-BPT diagram. As stated before, this choice is primarily motivated by the fact that we aim to provide the network with a set of parameters which are as much as possible independent from one another and whose linear combinations are not trivially connected, from a mathematical point of view, to the position on the [N ii]-BPTdiagrams itself and to our target labels. We have tested in fact, that including both [N ii]/[O ii] and [O iii]/[O ii] together (even in their $\Delta\log(p)$ form) the ANN can reconstruct something very similar to the [N ii]/[O iii] ratio (which is closely related mathematically to **D**), 'artificially' boosting its performances. For the same reason, the RF would be strongly biased towards the choice of these two parameters in its relative feature importance computation, well beyond the underlying physical connection of such parameters to the target variable, and hiding potential contributions from other quantities. Nonetheless, we acknowledge that [N ii]/[S ii] intrinsically accounts also for small residual dependencies on top of N/O (see e.g., Hayden-Pawson et al. 2021), as sulphur abundance could not exactly trace the oxygen one in case of strong variations the ionising conditions; hence, such a parameter is less reflective of the 'true' nitrogen-over-oxygen abundance than [N ii]/[O ii] is. Therefore, although our 'fiducial' analysis of the 'multi-parameter' runs is based on [N ii]/[S ii] as a tracer of N/O, within the text, and more specifically in Appendix A, we discuss also different combination of parameters, including [N ii]/[O ii] and modifying the list of the other emission lines-based parameters in the set accordingly (for instance assuming [Ne iii]/[O ii] instead of [O iii]/[O ii] to trace the ionisation parameter). However, we anticipate and reassure that none of the main results presented in this paper is affected by the choice of different N/O tracers.

Furthermore, we choose EW(Hα) as an independent probe of the sSFR (in order to avoid trivial correlations between sSFR, SFR and $M_{\star}$), and log([O iii]λ5007/[O ii]λ3727, 29) as a tracer of the ionisation parameter because requiring even low-significance (e.g., 2.5σ) detections of the [Ne iii]λ3869 emission line would introduce significant sample selection biases (i.e., preferentially removing galaxies in the high-mass, high-metallicity, bottom-right region of the diagram). Finally, we also remove Hα/Hβ from the 'multi-parameter' runs, as the BPT diagrams are, by definition, insensitive to dust extinction (thanks to the small wavelength separation of their lines), hence any correlation between such parameter and the location of galaxies in the diagram would necessarily follow from the correlation between the dust content and other physical parameters; moreover, this would further limit the algorithms to perform any trivial mathematical operation between emission line-based parameters. Before proceeding, each feature in the dataset is properly rescaled by subtracting its average value and normalised by the interquartile (i.e, $25th - 75th$ percentile) range.

4.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are a set of algorithms with structures that are inspired by the neural networks that constitute the human brain, and whose flexible structure and non-linearity allows to perform a wide variety of tasks (Baron 2019). For the purposes of the present work, a multilayered neural network is designed exploiting the TENSORFLOW[3] package within a PYTHON environment. The baseline structure of the network is very similar for both the classification and the regression task, however the details (and the relative differences) are described for each of the two cases within the dedicated subsections. In brief, a typical network consists of an input layer, output layer, and several hidden layers, where each of these contain neurons that transmit information to the neurons in the succeeding layer. The input data is transmitted from the input layer through the hidden layers, and reaches the output layer, where the target variable is predicted. The value of every neuron in the network (except those in the the input layer) is a linear combination of the neurons in the previous layer, followed by the application of a (typically non linear) activation function. The weights of the network are model parameters which are optimized during the training stage via back-propagation.

For the purposes of training the network, we randomly select the two-thirds ($\sim$ 67 per cent) of the dataset to define a *training sample*, with the remaining one-third ($\sim$ 33 per cent) that constitutes the *test sample* (and which the network does not interact with at all during the training stages) over which the performances of the ANN are evaluated. Given the large available statistics, this choice provides a sufficiently large set to perform an extensive training of the network without sacrificing its ability to generalise; moreover, both sub-samples are large enough to be fully representative of the distribution of galaxies in the BPT of the whole parent population. Nonetheless, we stress here that none of the results presented in this paper are affected by a different choice in sample splitting (e.g, a 50-50 per cent or a 75-25 per cent splitting are two widely adopted approaches).

We perform the analysis by either feeding the network with one parameter at the time, to evaluate their individual connection with the galaxy offset from the SF locus, and with a set of multiple parameters simultaneously (see previous section), in order to explore the maximum predictivity potential of the data. Because of the increased impact of overfitting in the 'multi-parameter' run compared to the individual runs, the structure of the network is slightly different in the former case than in the latter and its overall complexity is reduced, for both classification and regression analysis, as described more in detail in the following sections.

4.2.1 Classification

We first start with a rather sample classification analysis. The goal is then to determine which parameters are best in predicting whether a galaxy is offset above or below the SF locus in the BPT diagrams. In principle, in this case we do not need to assume any particular direction of the 'offset vector', but galaxies are just assigned either 1 or 0 label according to their position above or below the SF locus (i.e., according to positive/negative values of **D**), defining a simple binary classification problem. However, we recall that the Δlog(p) values of equation 4 which go into the network are actually computed by assuming a purely orthogonal deviation.

We design a multilayered network composed of two hidden layers (with 12 and 6 neurons, respectively), with a *rectified linear unit* (*ReLu*) activation function for the hidden layers and a *sigmoid* function for the one-dimensional (i.e., a binary 0/1 value) output layer. The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time (i.e., the neurons will only be activated if the output of the linear transformation is larger than zero), whereas the *sigmoid* function is largely used in models where the output layer should return a probability (in this case, the probability of belonging to a given class), since it maps any input values onto the [0, 1] range. The model is compiled implementing the ADAM solver (with a learning rate = 0.001) and optimising the standard *binary crossentropy* loss function. The 'Accuracy' (i.e., the fraction of galaxies correctly classified) is the metric assumed by the model to assess its performance during the training procedure and when applying its predictions to the test sample.

Such network structure is the result of an extensive direct experimentation with the dataset aimed at maximising the accuracy while keeping overfitting at a minimum. As an uncontrolled increase in the complexity of the network can lead to significant overfitting (i.e., the network performing significantly better on the training sample than on the test sample), we require the difference in accuracy between the performance of the model on the training and test samples to be within a few per cent, and we tune the network hyperparameters accordingly. As briefly metioned above, in the 'multi-parameter' run we decide to reduce the complexity of the network by implementing a single-hidden layer with 10 neurons only, in order to minimise the impact of overfitting. For this binary classification problem, the two classes (i.e., *above* and *below* the best-fit line) are also randomly re-sampled in order to be equally represented in both the training and test set (i.e., to have 50 per cent of galaxies lying *above* and 50 per cent lying *below* the SF locus in both training and test samples). In any case, we also consider here the area under the true positive rate (TPR)–false-positive rate (FPR) curve (known simply as the area-under-the-curve, 'AUC') as an additional metric to evaluate the network performance; one of the advantages of the AUC statistic in fact is that it is insensitive to the fraction of each class provided to the network. Furthermore, for the classification problem we focus only on galaxies with values of $|D| > 0.025$, i.e., we remove galaxies which lie so close to the best-fit line to be potentially misclassified given the typical uncertainties on their measured emission line ratios.

In the left panel of Fig. 4 we present the results of the binary classification analysis for the set of parameters described in Section 3. The fraction of correctly classified galaxies is shown on the y-axis, and the parameters used to train the network are shown on the x-axis and ordered from the most to the least predictive. The first bar in the plot refers to the run performed with the 'multi-parameter' set, which contains only a sub-set of the full list of parameters, as listed in the last column of Table 1 and according to what is discussed in Section **??**.

When all the parameters from the 'multi-parameter' set are fed together into the network, the model achieves an impressive classification accuracy of 90.57 ± 0.11 per cent (AUC=96.72 ± 0.07 per cent) on the test sample. Therefore, the position of a galaxy with respect to the SF sequence in the [N II]-BPT (i.e., whether a galaxy is offset above or below it) can, in principle, be predicted with excellent accuracy by knowing no more than the set of parameters

[3] https://www.tensorflow.org/overview

adopted here [4]. No significant variation on the performances is obtained from either increasing the network complexity or slightly varying the values of the hyperparameters, further confirming the stability of the result.

In terms of performances of individual parameters (i.e., when the network is fed with only one parameter at the time), Δlog([N ii]/[S ii]) achieves the best performance compared to the rest of the set, with an accuracy of 87.07 ± 0.14 per cent; adopting Δlog([N ii]/[O ii])provides a comparable (though slightly lower) accuracy of 80.85 ± 0.18 per cent. This means that deviations in the N/O abundance from the average value pertaining to galaxies along the SF locus (traced either by [N ii]/[O ii] or [N ii]/[S ii]) are extremely informative in predicting whether galaxies are offset above or below the SF sequence itself, and perform better than any other individual parameter in our set.

Among the other parameters, deviations in $M_\star$ and $\sigma_{H\alpha}$rank immediately after N/O tracers, although scoring significantly lower accuracies, followed by parameters tracing deviations in the ionisation parameter, SFR and dust extinction. We note finally that deviations in sSFR (probed either by Δ EW(Hα) or directly by the ratio between SFR and $M^\star$) and electron density (probed by the [S ii]λ6717/λ6731 ratio) achieve instead a result only ~ 10 percent better than a purely random variable (reported by the last bar and equivalent, in a balanced-sample binary classification problem, to tossing a coin with 50% probability of success), hence proving to be not very informative overall at describing the relative position of galaxies with respect to the SF locus within this diagram.

4.2.2 Regression

We now move to a different part of the analysis, which shares the same goal as the previous one (i.e., describing the connection between relative variations in different physical parameters and the scatter in the BPT diagrams) but set a different target label for the ANN. In particular, we now want to test the ability of our group of parameters to predict the *magnitude of the offset* (i.e., the length **D** of the offset vector, taken positive if pointing above the best-fit line) from the sequence of local star-forming galaxies, in a standard regression analysis. Here, following what is discussed in Section 3, and differently from the classification analysis, the offset vector is assumed to be exactly orthogonal to the best-fit curve of the SF sequence, at any given point. In principle then, there is no reason to assume a priori that the classification and the regression analysis should provide the same results, although the two problems are clearly closely related to each other.

Similar to the previous case, we create a network with two hidden layers (12 and 6 neurons, respectively) and a *ReLu* activation function. The model is compiled with the ADAM optimiser (with a learning rate = 0.001) and minimises the *mean squared error* (mse) as the loss function. Again, for the ‘multi-parameter’ run the complexity of the network is reduced to a single-hidden layer with only 10 neurons, in order to control the impact of overfitting. Extensive testing of the network outputs and performances suggests adoption of a *mini-batch gradient descent*[5] algorithm with a batch size of 128 and 32 for the ‘individual’ and ‘multi-parameter’ runs respectively, and to train the network over a total of 100 epochs.

In the bottom panel of Fig. 4, we report the results of the ANN regression analysis, where the performance of each individual parameter in our set (and of the ‘multi-parameter‘ run) is assessed on the basis of that of a purely random variable. Following, e.g., Bluck et al. (2019), we define in fact the ‘improvement over random’ metric as

$$\mathrm{IoR}_i = \frac{\mathrm{RMSE}_i - \mathrm{RMSE}_{\mathrm{rand}}}{0 - \mathrm{RMSE}_{\mathrm{rand}}}\ , \qquad (7)$$

where $\mathrm{RMSE}_{i,\mathrm{rand}}$ is the root-mean-squared-error of the i-th variable and of a purely random variable respectively, whereas zero represents, by definition, the best possible performance in terms of RMSE in a regression problem (i.e., the target variable is predicted with 100 per-cent accuracy).

When trained with the ‘multi-parameter’ set, the network achieves an IoR = 44.81 ± 0.19% in predicting the exact distance **D** from the SF sequence in the test sample. The values of **D** predicted for the test sample by the network in the ‘multi-parameter’ run are compared to the *true* target **D** values as shown in the inset, upper-right panel of Fig. 4; we report a median of the errors on the predictions of $\mu = 0.002$ and a standard deviation of $\sigma = 0.038$. Similar results are found on the training sample, with no significant overfitting reported.

Individual parameters rank in a very similar order as in the classification problem, with Δlog([N ii]/[S ii]) and Δlog([N ii]/[O ii]) (associated with relative deviations in the N/O abundance) being the most predictive quantities (IoR = 33.45 ± 0.31% and 22.8 ± 0.26%, respectively) of the distance from the best-fit line of the SF sequence in the [N ii]-BPT diagram. It is interesting to note that, such parameters aside, none of the included quantities scores above 20 per cent in IoR, with six out of nine parameters marking an improvement below 10 per cent. This is somehow expected, and confirms that predicting the exact distance from the SF sequence (in regression) is much more difficult than just classifying a galaxy as belonging to the region above or below it; indeed, no individual parameter, except for those associated to variations in N/O, is really capable of providing enough information to predict our target variable **D** with a high level of accuracy. However, when the information from multiple parameters is provided, the predictive power is increased and the network can reproduce the offset of star-forming galaxies in the [N ii]-BPT with significantly higher accuracy.

4.3 Random Forest

In this section, we exploit a random forest (RF) of decision trees in order to determine how effective a given parameter is in solving the classification and regression problems addressed before, when considered in direct comparison with the other parameters. In fact, the RF treats multiple parameters as if they were in a competition, selecting the most useful for each decision node.

In general, a decision tree is a set of consecutive nodes, where each node represents a condition on one feature in the dataset. The conditions are of the form $X_j > X_{j,th}$, where X_j is the value of

[4] this, however, does not automatically imply that different parameters would not perform equally well, or perhaps even better

[5] A gradient descent is an optimization technique used to find the weights of machine learning algorithms. It works by exploiting the error associated to model predictions on the training data to update the parameters in order to reduce the discrepancies at the following steps. The ‘mini-batch’ gradient descent is a variation of this approach, which splits the training dataset into small batches that are used to calculate the errors and update the model coefficients. Its main advantages over the standard gradient descent are that the model is updated more frequently (which allows for a more robust convergence), an increased computational efficiency, and that it does not require to maintain all the training data in memory at once.

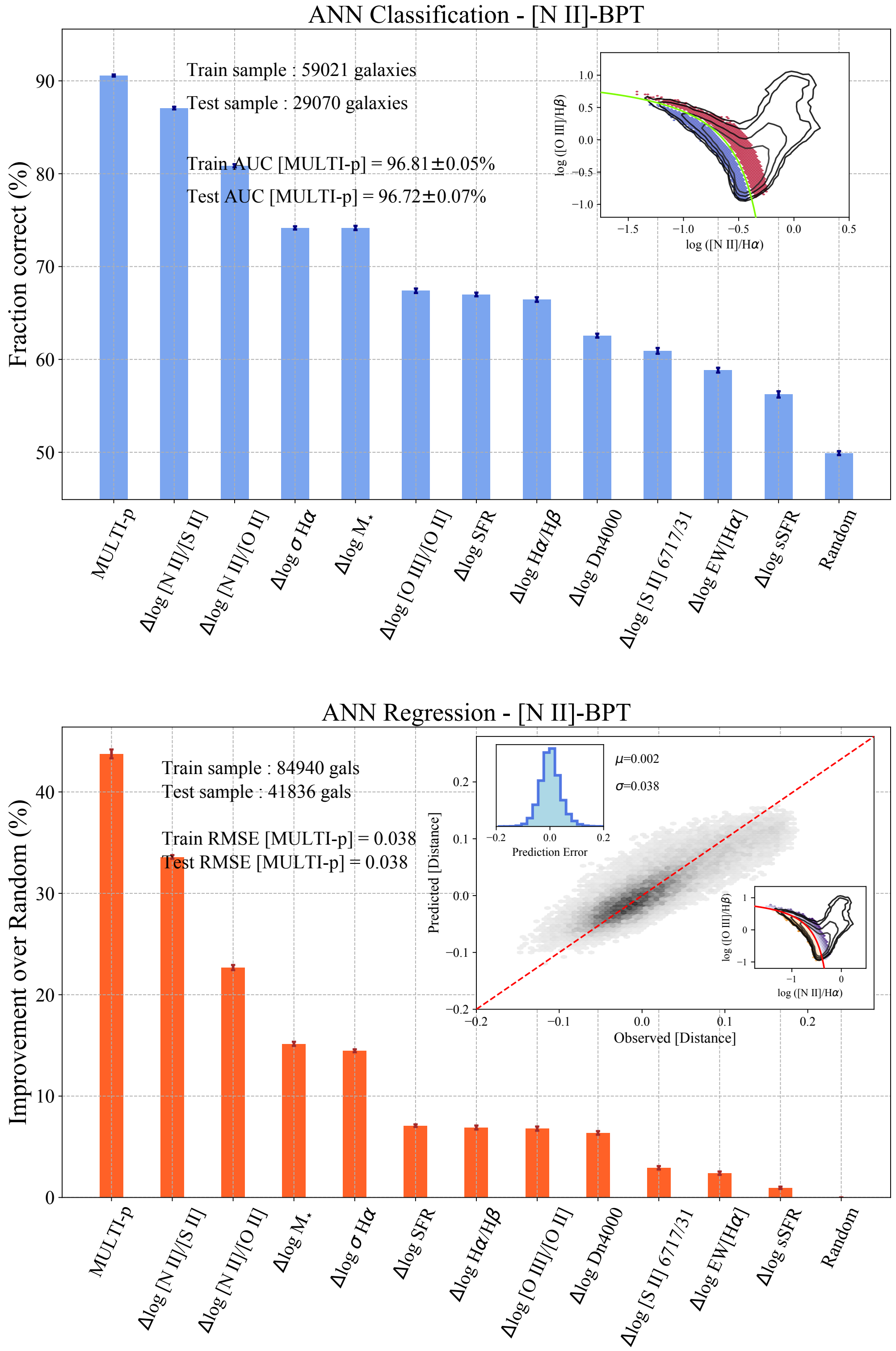

Figure 4. *Upper Panel:* Results from the ANN classification analysis aimed at predicting whether galaxies are located above or below the best-fit line of the star-forming galaxies sequence in the [N ii]-BPT diagram (as schematised in the upper-right inset panel). Histogram bars report the absolute fraction of correctly classified galaxies for both the 'multi-parameter' set (see last column of Table 1), and each parameter taken individually (ordered from most to least predictive); the last bar report, for comparison, the performance of a random variable (which is equivalent to 50 per cent accuracy in a binary classification problem). *Bottom Panel:* Results from the ANN regression analysis aimed at predicting the exact magnitude **D** of the offset from the best-fit line of the star-forming galaxies sequence in the [N ii]-BPT diagram, as defined in Eq. 5. Histogram bars in this case report the 'improvement-over-random' (IoR) statistics (ordered from most to least predictive parameter, with the a random variable scoring, by definition, 0% IoR). Within the top-right, inset panel, we compare the **D** values predicted by the network on the test sample in the 'multi-parameter' run to the 'true' **D** target values for the same galaxies. In both panels, we also report the number of galaxies in the training and test sub-samples and the relative AUC and RMSE scores of the 'multi-parameter' runs.

the feature at index j and $X_{j,th}$ is a threshold, which is determined during the training stage. The lowest nodes in the tree are usually called ‘leaves’, and carry the final assigned label of a particular path within the tree (e.g., in our classification case, whether a galaxy is labeled as *above* or *below*). A RF, then, is simply a collection of decision trees, where different decision trees are trained on different bootstrapped, randomly-selected subsets of the original training set (and where, if desired, random subsets of the input features can be selected during the training of each individual tree to construct the conditions in individual nodes). The final RF prediction is just an aggregate of individual predictions of the trees in the forest, in the form of a majority vote; the main advantage is that, while a single decision tree tends to overfit the training data, the combination of many decision trees in a RF generalizes well to previously unseen datasets. Furthermore, by quantifying the decrease in impurity provided by each parameter in each fork and within each tree of the RF, the relative importance of the various parameters is established. This competitive approach is especially useful when the parameters considered are highly inter-correlated in a complex and highly non-linear manner.

We recall here that the following RF analysis is based on the ‘multi-parameter’ subset defined in the last column of Table 1, whose selection is justified in detail in Section 4. To implement the RF into our analysis, we adopt the RANDOMFORESTCLASSIFIER and RANDOMFORESTREGRESSOR classes from the SCIKIT-LEARN package in PYTHON.

4.3.1 Classification

In the binary classification scheme, we set up a forest of 100 independent estimators, allowing each tree to grow indefinitely but setting a minimum threshold to the number of samples allowed at each leaf-node equal to the number of galaxies in the training sample divided by 250 (i.e., ~ 350 samples in our case). This choice allows us to control overfitting, which is assessed by requiring the difference in performances between the training and the test sample to be limited to a few percent. The RF Classifier is set to minimise the *Gini* impurity as the loss-function at each decision node. The accuracy of the RF classification task is assessed by evaluating the AUC on both the training and the test sample. We perform 30 independent runs (randomised at the training-test sample split level), hence evaluating the average and standard deviation of the results over $30 \times 100 = 3000$ independent estimators. Consistently with the ANN analysis, only galaxies with $|\mathbf{D}| > 0.025$ are included in the RF classification analysis.

In the following, we also explore and discuss two different ways for computing the relative feature importance. Firstly, we leave the RF free to consider the entire set of parameters at each decision split (i.e., what we call the ’All features’ case, setting the *max features* hyperparameter of the RF accordingly). In this way, the algorithm is capable to fully handle the inter-correlations between the different features and find the one (or the group of parameters) which is most intrinsically connected with the target variable. In Fig. 5, the results of the RF classification analysis in this first case are shown by the filled bar chart. The parameters are ranked in terms of their relative importance (from the most to the least relevant), which is reported on the y-axis. The overall performance of the RF model on both the training and test sample is reported in terms of AUC: the RF achieves an AUC = 96.35 ± 0.09 per cent in the binary classification task, a performance comparable to that scored by the ANN when trained with the ‘multi-parameter’ set. However, although at first sight the ranking in the relative importance of the various parameters resemble that obtained in Fig. 4 for the ANN analysis (in terms of accuracy of individual features), there are a number of remarkable differences. In particular, the relative importance of Δlog([N II][S II]) (hence deviations in N/O abundance) is strongly dominant over the other parameters, accounting for more than 80% of the total predictive power, whereas Δlog(EW[Hα]) and Δlog(D_N(4000)) (tracing variations in the specific star formation rate and age of stellar populations) are ranked second and third, respectively, retaining together about 10% of the residual relative importance. Interestingly, although these parameters were among the least performing, individually, in the ANN analysis, the RF highlights how their information is more complementary to Δlog([N II]/[S II]) than any other parameter in the set. On the contrary, the importance of all the remaining parameters is strongly suppressed, revealing how their individual predictive power (as measured by the ANN) was likely due to underlying correlations with one of the best-ranked features.

In addition, we have also explored the case in which only a fixed number of (randomly selected) features are considered at each node of the RF trees, by setting the *max features* hyperparameter equal to the square root of the total number of parameters in the set (what we label the ‘$\sqrt{N_{\rm features}}$’ case). Although, in this second approach, the correlations between parameters are not fully accounted for in computing the feature importance ranking, this analysis provides an insightful estimate of which parameters perform better in case the most important one(s) is(are) not available. The results of this further classification analysis are shown in Fig. 5 by the empty, hatched bar chart. The algorithm is now forced to take into consideration also different features than the most important ones, spreading the final relative importance among a larger number of parameters; in fact, log(M*) and $\sigma_{\rm H\alpha}$ are now ranked higher than Δlog(EW[Hα]) and Δlog(D_N(4000)). Nevertheless, the RF still robustly identifies Δlog([N II]/[S II]) (hence, variations in N/O) as the most important parameter, which corroborates the interpretation of its role of primary physical driver of the scatter in the [N II]-BPT. A result fully consistent with such interpretation is also recovered when considering Δlog([N II]/[O II])in place of Δlog([N II]/[S II])to trace variations in N/O, and is presented in Appendix A.

4.3.2 Regression

For the regression problem, we design a very similar Random Forest structure as for the classification task, and only change the loss-function to the *mean squared error*. The results of the RF regression analysis are shown, for both the ‘All features’ and ‘$\sqrt{N_{\rm features}}$’ cases described in the previous section, in the bottom panel of Fig. 5. Overall, the performance of the RF in predicting the exact distance from the SF sequence is comparable to that achieved by the ANN, with a median and standard deviation of the residuals of 0.0002 and 0.039, respectively. The parameters’ ranking closely traces what seen already for the classification problem, with Δlog([N II]/[S II]) being, by far, the most important parameter and retaining more than 80% of the total predictive power, which increases to > 90% when Δlog(D_N(4000)) and Δlog(EW[Hα]) are used in conjunction. This means that in principle, modulo the assumptions discussed in Section 3 and within the residual uncertainties, almost no further information is needed to quantify the magnitude of the offset from the SF locus which a galaxy resides at in the [N II]-BPT diagram (we recall here that we are implicitly assuming these variations to occur at fixed metallicity). Finally, similar to that discussed before, we note that when considering only the square root of the number of features at each node, the relative importance of parameters that are

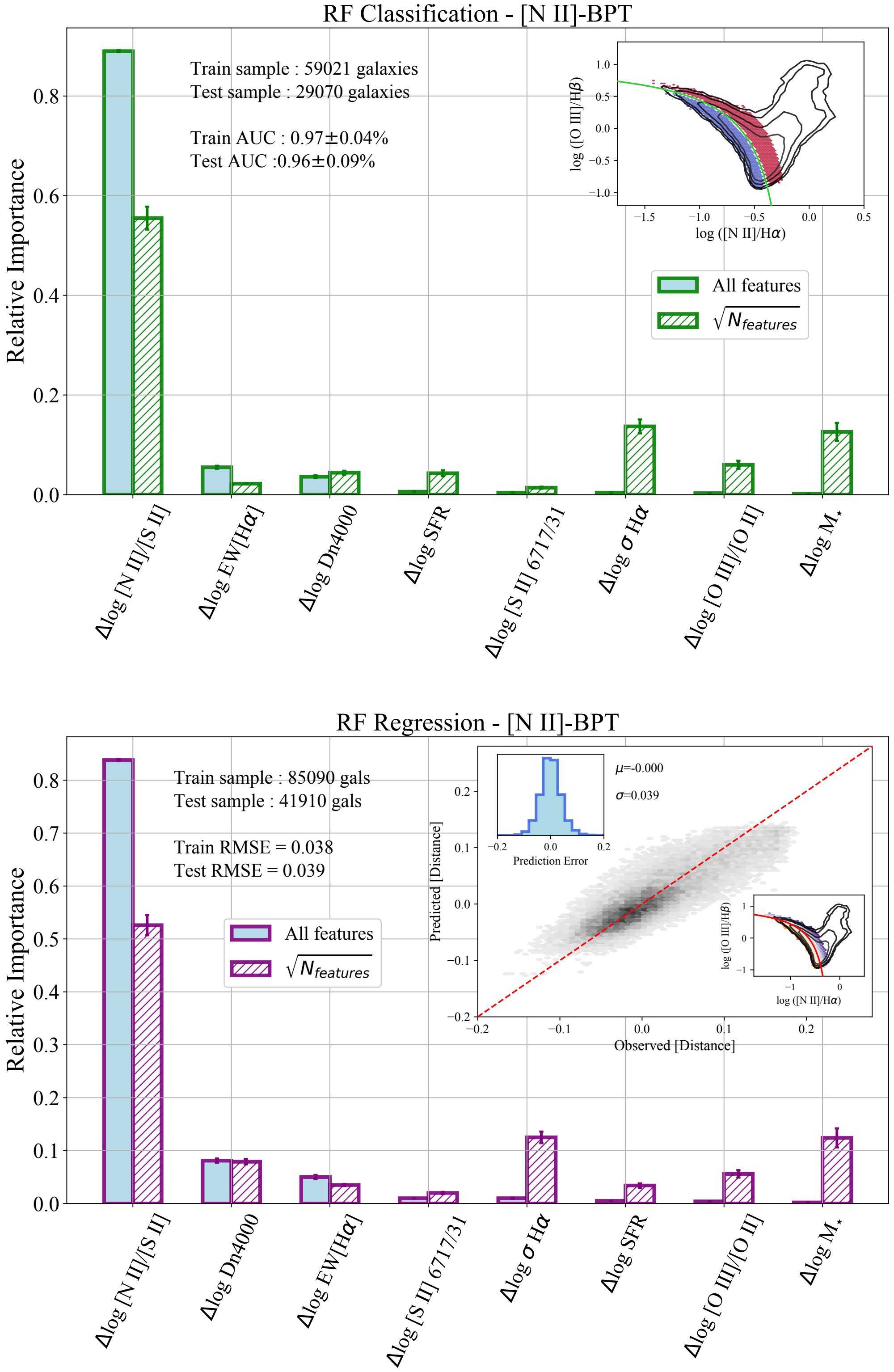

Figure 5. Results of the RF classification (*upper panel*) and regression (*bottom panel*) analysis, adopting the 'multi-parameter' set under study as listed in the last column of Table 1. The y-axis reports the relative importance of each of the parameters labelled on the x-axis (with error bars derived from the standard deviation of 30 independent runs). Colour-filled bars refer to the RF analysis conducted allowing 'all features' to be considered at each node of the trees, whereas empty, hatched bars refer to the case where only the square root of the number of features are picked up at each splitting node. The former fully disentangles the inter-dependencies between the various parameters, providing the best possible combination of features to use in concert to maximise the performances of the RF, whereas in the latter a relatively higher importance is retained also by parameters which are (to some extent) correlated with the best ranked ones, being valuable alternatives in case these are not available. In both cases, Δlog([N ɪɪ]/[S ɪɪ]), tracing relative variations in the N/O abundance compared to the median SF locus, is robustly identified as the most predictive parameter in either classification and regression tasks. A (small) residual, complementary importance is retained by $D_N(4000)$ and EW(Hα).

closely connected to Δlog([N ii]/[S ii]), which can thus perform as good substitutes of such parameter, increases to a level that matches that of the second and third parameters in the ranking.

Summarising, from the joint ANN and RF analysis presented in the previous sections, we can robustly claim that deviations in the N/O abundance (traced in this case by Δlog([N ii]/[S ii])) with respect to the average of galaxies along the SF locus are the main drivers of the deviation of star forming galaxies from their median loci in the [N ii]-BPT diagram, once the offset is considered orthogonal at any point to the best-fit line. This result is further confirmed in case Δlog([N ii]/[O ii]) is included (in place of Δlog([N ii]/[S ii])) in the RF analysis, although in that case the relative importance of the other parameters is impacted. In fact, we stress again here that because the RF disentangles the relative importance of a set of features used *in conjunction* with each other, changing even one parameter only within the set can have an impact on the relative importance retained by all of the remaining variables too. For more details, we refer to Appendix A and to the discussion of section 4.5.

4.4 Does the relative parameter importance change across the diagram ?

In the previous sections, we have analysed the connection between the scatter of galaxies in the [N ii]-BPT and different physical parameters, and found Δlog([N ii]/[S ii]) as the most predictive parameter of both the direction (classification task) and amplitude (regression task) of the offset vector from the best-fit curve of the SF sequence. However, the distribution of star-forming galaxies in the diagram is not homogeneous, with the highest density of galaxies concentrated in the high-metallicity, bottom-right region, where the offset vector is primarily directed along [N ii]λ6584/Hα. As shown already in Fig. 3 in fact, the relative strength of the two components of an orthogonal 'offset vector' changes as we move along the sequence of star-forming galaxies in the diagram. This effect can be parametrised in terms of the arctangent of the angle (positive counterclockwise) formed by the 'offset vector' with the x-axis, and indicated with θ in equation 6: moving from the bottom-right to the upper-left region of the sequence θ increases, and so it does the relative strength of the offset along [O iii]λ5007/Hβ compared to that along [N ii]λ6584/Hα. Therefore, it is worth asking if either the individual absolute performance and/or the relative importance of the various parameters involved in the ML analysis changes, as a function of the location considered along the star-forming sequence.

For this reason, in this Section we repeat the analysis presented in Section 4.2 and 4.3 by splitting the [N ii]-BPT diagram in four *sectors*, defined on the basis of the different inclinations of the offset vector with respect to the horizontal axis, i.e., of different intervals spanned by the θ angle. The choice of the number and 'width' of these sectors is empirical, and driven by the aim, on the one hand, to obtain a segregation of the diagram as homogeneous as possible (i.e., to avoid having sectors spanning too different ranges in θ), while on the other, to have a minimum reasonable number of galaxies within each sector in order to perform a meaningful statistical analysis.

The partition of the [N ii]-BPTdiagram in four sectors is graphically represented in Fig. 6. Because of the (even very) different numbers of galaxies within each sector (with the bottom-right ones, i.e. at low $< \theta >$ values, being much more populated than the others), the input values for the hyper-parameters of both ANN and RF models are tuned to adapt to the varying sampling, especially in the upper-left sector $<\theta \sim 45°>$ where the total number of sources falls below 10, 000. For instance, for the ANN analysis of individual parameters in such sector, the sample is more unevenly split (80-20 per-cent) in training and test galaxies, in order to feed the model with a sufficiently large number of galaxies for training, the *batch size* of the stochastic gradient descent algorithm is set to one-third of the training sample size, and the model is trained for 300 Epochs instead of 100. For the 'multi-parameter' run instead, the batch size is fixed to 8. We have tested that tweaking the hyper-parameters of the network in this way allows us to maintain a reasonable balance between performances and overfitting, the latter becoming of increasing concern especially in small datasets.

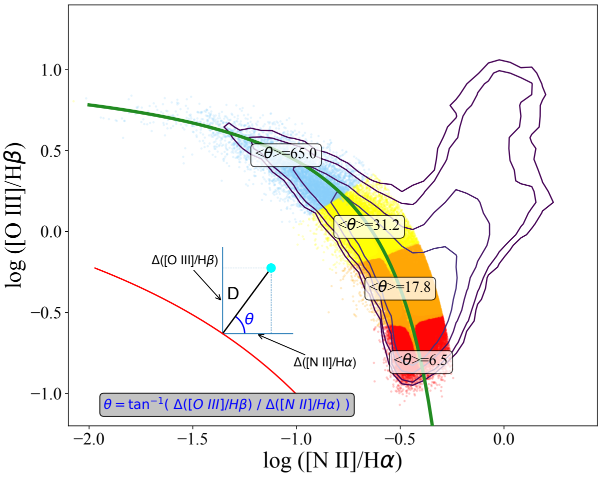

Figure 6. The distribution of star-forming galaxies in the [N ii]-BPT diagram is divided in four sectors, defined by different intervals in the $<\theta>$ angle as formed by the 'offset vector' of each galaxy with the horizontal axis. In this way, we aim to study whether and how the results from the ML analysis presented in Section 4.2 and 4.3 vary as a function of the position of galaxies along the SF sequence.

The results of the ANN analysis for the four different sectors are reported and compared in the upper panels of Fig. 7, where the classification *Accuracy* and the *IoR* in regression are plotted, for both individual features and the 'multi-parameter' run, as a function of the average $< \theta >$ of each of the regions in which the diagram has been divided into. In each sector and for each parameters set, 30 independent ANN runs are performed and the average performances are evaluated.

As a first remarkable result, we find the performances of the network to be quite stable across the entire diagram, scoring $\gtrsim$ 90 per-cent accuracy in classification and $\gtrsim$ 40 per-cent IoR in regression in all sectors. In terms of performances of the individual parameters, those associated to the chemo-dynamical state of the galaxy (M$_\star$, N/O, $\sigma_{\mathrm{H}\alpha}$) score the largest accuracy and IoR in the first three sectors, whereas the performances of parameters associated with star formation and ionisation conditions (like [O iii]/[O ii], SFR, EW[Hα]) increases as moving towards the upper-left part of the diagram, becoming almost dominant in the top-left sector. Nonetheless, Δlog([N ii]/[S ii]) maintains a stable level of performance across the entire diagram, scoring the highest accuracy and IoR everywhere, whereas overall very weak dependency exists, for instance, between the scatter of galaxies in the [N ii]-BPT and variations in electron density (traced by Δ[S ii]λ6717/31) in all sectors but the first one, where this parameter matches the individual performances of M$_\star$ and $\sigma_{\mathrm{H}\alpha}$.

The Random Forest analysis of the four independent sectors is reported instead in the bottom panels of Fig. 7, where the re-

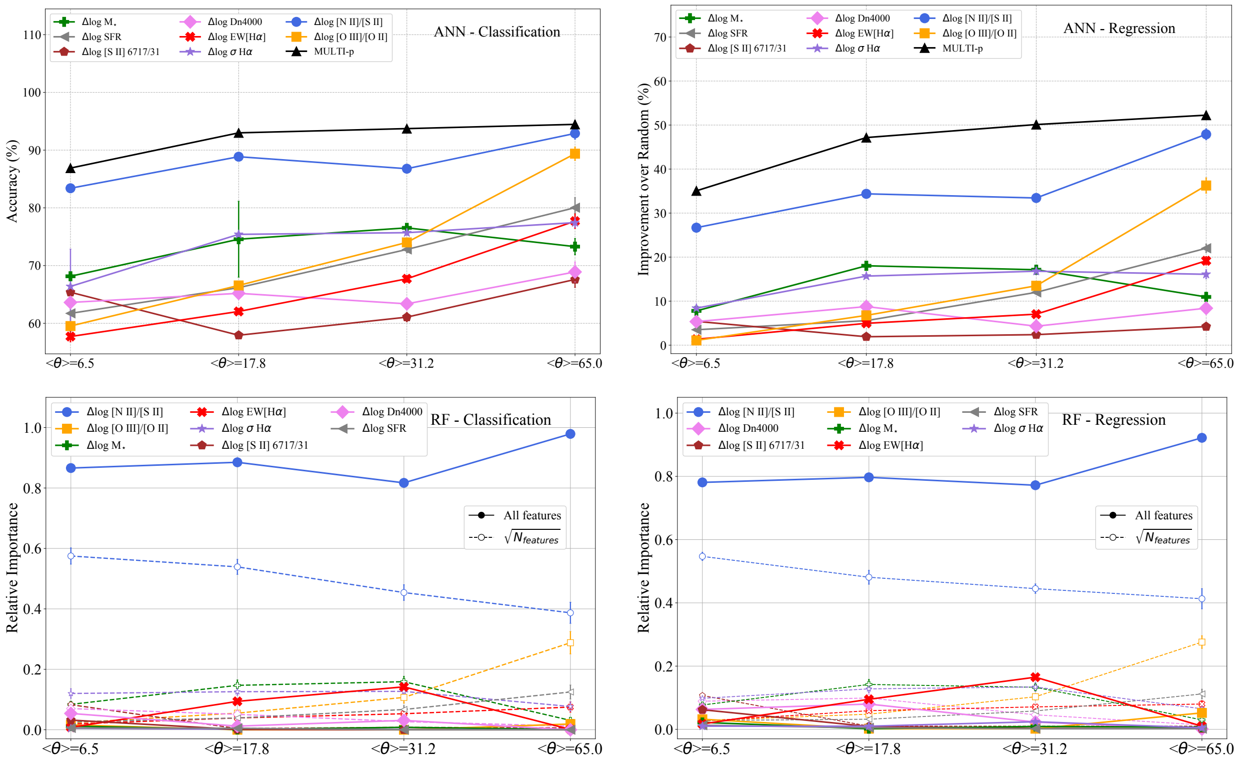

Figure 7. *Upper panels:* Results for the ANN classification (left) and regression (right) analysis for each of the four sectors in which the [N ii]-BPT diagram has been divided (see Fig. 6). The different tracks follow the accuracy and IoR as a function of the average angle $< \theta >$ of the 'offset vector' in each region, and for each of the parameters of interest (colour-coded as in the legend). When trained with the 'multi-parameter' set, the network achieves excellent accuracy and IoR across the entire diagram, with Δlog([N ii]/[S ii]) achieving the best individual performance in all sector. *Bottom panels:* Same as the upper panels, but for the RF analysis. The relative importance of the various features is now reported on the y-axis and plotted as a function of the average $< \theta >$ of each sector: straight lines are representative of the RF analysis with all features allowed at each node, whereas dashed lines are for the $\sqrt{\rm N_{features}}$ case. For both regression and classification problems, Δlog([N ii]/[S ii]) is found as the most relevant parameter across the entire diagram.

lative importance of the various features is plotted as a function of the average <θ> spanned by each region. Deviations in the N/O abundance (traced by Δlog([N ii]/[S ii])) are by far the most relevant quantity to consider (for both the classification and the regression tasks) throughout all the sectors of the diagram, especially when all features are considered at each node (solid lines), and hence the RF truly exposes the parameter which is intrinsically most connected to the target label. Interestingly, variations in EW(Hα) (i.e., in the sSFR) gain a significant ∼ 20% relative importance in the central sectors. In case only $\sqrt{N_{\rm features}}$ are considered at each splitting-node instead (dashed lines), stellar mass, $\sigma_{{\rm H}\alpha}$ and density are found as the most useful alternative parameters to Δlog([N ii]/[S ii])in the first sector, while EW[Hα], SFR and especially Δlog([O iii]/[O ii]) overcome them in the two uppermost regions.

4.5 Discussion

The analysis presented in the previous sections unambiguously suggests that relative variations in the nitrogen-over-oxygen abundance are the primary physical driver of the deviation from the median locus of star-forming galaxies in the [N ii]-BPT. In fact, Δlog([N ii]/[S ii])(or, equivalently, Δlog([N ii]/[O ii])) is robustly identified as the most predictive (individual) parameter and the most relevant feature (among the 'multi-parameter' set) in both classification and regression tasks, for either the global analysis of the sample and within separated regions across the diagram, regardless of the average inclination of the offset vector (and thus, regardless of the strength of its two components along [N ii]λ6584/Hα and [O iii]λ5007/Hβ).

If we recall the tight, monotonic dependence of the position of galaxies along the SF sequence in the diagram with metallicity (as outlined in Section 3.1), we can interpret our global results of Fig.4 and 5 as a manifestation of the existence of an O/H vs N/O relation for SDSS star-forming galaxies, whose intrinsic scatter is reflected and, to some extent, translated into the observed distribution of emission line ratios within the [N ii]-BPT. A tight relationship between O/H and N/O abundances is indeed observed in both HII regions and local galaxies, especially at $M^{\star} \gtrsim 10^{9.5} M_{\odot}$ (Vila Costas & Edmunds 1993; van Zee et al. 1998; Pérez-Montero & Contini 2009; Pilyugin et al. 2012; Andrews & Martini 2013; Hayden-Pawson et al. 2021), and it is set by the predominant nucleosynthetic origin of nitrogen from CNO burning of pre-existing stellar carbon and oxygen in low- and intermediate-mass stars experiencing the AGB phase (i.e., the 'secondary' nitrogen production mechanism, Kobayashi et al. 2011; Ventura et al. 2013; Vincenzo et al. 2016); alternatively, Vincenzo & Kobayashi (2018) reproduced the observed N/O–O/H relation introducing failed supernovae (SNe) in massive stars within their cosmological simulations. Recently, such relationship between O/H and N/O has been suggested as even tighter than the one between $M^{\star}$ and N/O (Hayden-Pawson et al. 2021), in contrast to what claimed by previous studies (e.g., Andrews & Martini 2013; Masters et al. 2016). In light of our results, this would confirm that deviations in N/O at fixed O/H are more likely to be related to the offset from the SF sequence in the [N ii]-BPT than

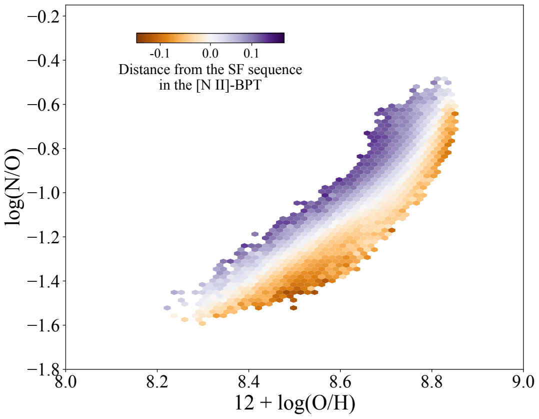

Figure 8. The relationship between N/O and O/H for local SDSS star-forming galaxies is colour-coded by the magnitude **D** of the offset vector from the SF sequence of the [N ii]-BPTdiagram. A clear segregation in **D** is seen in N/O, at fixed metallicity.

relative variations in M*, although the two are clearly physically correlated. The connection between the two diagrams is also readily evident if we look at the distribution of our galaxy sample in the N/O vs O/H diagram, as shown in Fig. 8; here, each hexagonal bin is colour-coded by the average distance **D** of galaxies from the best-fit line of the [N ii]-BPT, almost perfectly tracing the scatter around the median N/O vs O/H relation.

In general, and especially for galaxies located in the bottom-right, high-metallicity region of the diagram (the majority of the sample, with ∼ 70 per-cent of them characterised by $\theta < 22°$), variations in N/O are associated to galaxies of different stellar masses and can be interpreted as age-related effects: galaxies with higher M*, in fact, are more chemically mature than their lower mass counterparts (i.e., those located along iso-O/H lines and with negative **D** values), in the sense that they had more time to enrich the ISM with nitrogen produced by low- and intermediate-mass stars on longer timescales. Hence, to a positive Δlog(M*) corresponds a positive Δlog(N/O) (and viceversa, with relatively lower mass galaxies at fixed O/H which still have nitrogen partly locked in stars), producing the offset in the [N ii]-BPT. Indeed, Δlog(M$_\star$) here acts as a good proxy for Δlog([N ii]/[S ii]) in the ML analysis, reaching high scores in the ANN runs whilst scoring almost zero importance in the RF, but subtracting nonetheless ∼ 10 per-cent of relative importance from Δlog([N ii]/[S ii]) in the $\sqrt{\mathrm{N_{features}}}$ case.

Furthermore, a small but significant (∼13 per-cent) of the global feature importance belong to Δlog(D$_\mathrm{N}$(4000)) (∼8 per-cent) and Δlog(EW(Hα)) (another ∼5 per-cent). We interpret this as an additional contribution to the offset which is still associated to galaxy ageing, but that it is complementary to the information already provided by Δlog([N ii]/[S ii]). In particular, we associate it to a differential impact of older stellar populations (e.g., hot, post-AGB stars, which dominate ionising photon production after ∼ 0.1Gyr), boosting intermediate- and low-ionisation emission lines from a more diffuse ionised gas (DIG) and setting the relative distance of these galaxies from the 'composite' and 'LI(N)ER' area of the diagnostic diagram (e.g., Zhang et al. 2017; Byler et al. 2019). Such an effect is accounted for in particular by Δlog(D$_\mathrm{N}$(4000)) in the 'first' sector of the [N ii]-BPT, whereas it is more prominently seen in Δlog(EW(Hα)) in the 'central' ones: in both cases, galaxies offset above the sequence are characterised by signatures of relatively older stellar populations (higher D$_\mathrm{N}$(4000) and lower EW(Hα)) compared to on-sequence galaxies, and viceversa (see Fig. 2).

Moving upwards along the SF sequence, parameters related to star formation and ionisation state of the gas score progressively higher accuracies in the ANN, although variations in N/O are still identified as the primary driver of the scatter, with ∼ 92 per-cent of relative importance scored in the RF. Although we acknowledge that the small number of galaxies in this sector (< 10,000) might impact the ability of the RF of correctly retrieving the exact relative importance of each of the parameters, nonetheless the overwhelming success of Δlog([N ii]/[S ii]) in a region where the offset vector has a strong component also along [O iii]λ5007/Hβ prompts us some reflections. In particular, one alternative interpretation involve breaking our initial assumption that the offset occur along iso-metallicity lines (i.e., at fixed O/H): as can be seen in Fig. 2 in fact, variations in metallicity across the SF sequence, although very mild (i.e., of the order of < 0.05 dex in Δ log(O/H)), are present in such region of the diagram, and are opposite to variations in N/O. Therefore, the connection between the offset from the median sequence and relative variations in N/O here could just, at least partially, reflect metallicity variations. A decrement in metallicity coupled with an increase in N/O could be explained, in fact, by invoking the presence of differential outflows (i.e., preferentially removing oxygen from the ISM) from relatively younger, low metallicity galaxies with prominent star-formation (e.g., Vincenzo et al. 2016; Magrini et al. 2018). Interestingly then, in the $\sqrt{\mathrm{N_{features}}}$ realisation of the random forest large part of the Δlog([N ii]/[S ii]) importance is taken by both Δlog([O iii]/[O ii])and Δlog(SFR). The former variable traces the ionisation parameter, which has a strong dependence on the stellar metallicity (hence, indirectly on the gas abundances), whereas the latter could be tracing indeed the differential impact of star-formation driven outflows in this galaxy population. Moreover, the [O iii]λ5007/[O ii]λ3727, 29 ratio itself is known to have an intrinsic, although secondary, dependence on metallicity too (Kewley & Dopita 2002). Further analysis based on large samples of galaxies with independent and 'direct' metallicity estimates in such region of the [N ii]-BPTcould certainly help to either confirm or deny such interpretation.

Finally, we note that variations in $\sigma_{\mathrm{H}\alpha}$, although performing overall well in the ANN analysis, picks basically no relative feature importance at all in the RF anywhere across the diagram. Its performance, indeed, closely follows the trend seen for M$_\star$, and this suggests that any information carried by $\sigma_{\mathrm{H}\alpha}$, likely tracing the dynamical mass of the system (Green et al. 2014; Krumholz et al. 2018), is already embedded into the M$_\star$parameter (and/or other parameters, like SFR, Yu et al. 2019; Varidel et al. 2020) within the population of star-forming galaxies. However, we also note that, if Δlog([N ii]/[O ii]) is adopted instead of Δlog([N ii]/[S ii]) to trace variations in N/O abundance, then $\sigma_{\mathrm{H}\alpha}$ retains a non negligible amount of (complementary) relative importance in the RF analysis. We interpret this as a signature of a dependence of [N ii]λ6584/[S ii]λ6717, 31 on $\sigma_{\mathrm{H}\alpha}$, at fixed [N ii]$\lambda$6584/[O ii]$\lambda$3727, 29; these results are presented more in detail in Appendix A. As a final remark, we also note that the relatively small dynamical range of $\sigma_{\mathrm{H}\alpha}$ across the star-forming galaxy populations (whose gas emission lines originates from kinematically 'cold' HII regions with typical velocity dispersions of ∼ 30km s^{-1}), coupled with the intrinsic spectral resolution of the SDSS-II spectrograph of ∼ 70km s^{-1}, also make any inference based on central $\sigma_{\mathrm{H}\alpha}$ measurement more challenging to physically interpret. Exploiting the improved calibration of the instrumental response for the MaNGA spectrograph (Law et al. 2021a), together with the information provided on spatially resolved scales, in a forthcoming paper we aim at revisiting the significance of our ML results on gas velocity dispersion measurements in star-forming galaxies.

5 THE [S ii]-BPT DIAGRAM

5.1 Parameters and metrics

We now shift our focus on the [S ii]-BPT diagram, for which we perform an analysis similar to that previously described for the [N ii]-BPT. Compared to the [N ii]-BPT, in the [S ii]-BPT two separate branches can be seen departing from the star-forming galaxy abundance sequence, one connected to the Seyfert region, while the other one connected to the region where LI(N)ERs are located. Moreover, the distribution of star-forming galaxies in the diagram appears less tight than in the [N ii]-BPT, with a larger scatter, especially in the [S ii]$\lambda\lambda6717, 31$/Hα line ratio at fixed [O iii]$\lambda5007$/Hβ. It can be also be seen, for instance, that galaxies which are located on (or very close to) the best-fit line in the [N ii]-BPT, i.e. which have by construction **D**∼ 0 according to equation 5, are instead more scattered across the best-fit line of the SF sequence in the [S ii]-BPT (with a standard deviation of 0.07 dex in [S ii]$\lambda\lambda6717, 31$/Hα at fixed [O iii]$\lambda5007$/Hβ). This simple observation already suggests that the scatter in this diagram might be primarily associated with different physical mechanisms than in the [N ii]-BPT.

The distribution of our set of parameters among the star-forming galaxy population within the [S ii]-BPT diagram is shown in Fig. 9. Compared to what seen for the [N ii]-BPT in Fig. 1, there are a few remarkable differences. Firstly, the best-fit line of the SF sequence is not monotonic, but it is double-valued in [S ii]$\lambda\lambda6717, 31$/Hα, presenting two distinct branches at high and low [O iii]$\lambda5007$/Hβ, with a turnover point around log([O iii]$\lambda5007$/Hβ) = 0. We perform a fit to the median [O iii]$\lambda5007$/Hβ values in small bins of [S ii]$\lambda\lambda6717, 31$/Hα and provide a fourth-order polynomial representation of the best-fit line of the SF sequence in this diagram, as expressed by the following:

$$\log([\mathrm{S\,II}]/H\alpha) = \sum_{n=0}^{4} p_n \cdot \log([\mathrm{O\,III}]/H\beta)^n \qquad (8)$$

where the coefficients p$_n$ are = [-0.20545, -0.41137, -0.58826, -0.06523, -0.44463] (from 0-th to 4-th order, respectively).

In terms of galaxy properties (and similar to the [N ii]-BPT) the [S ii]-BPT diagram is characterised by a strong sequence in metallicity and ionisation parameter, which can be both clearly visualised in Fig. 9 and quantified by a $\nabla_{||}$ statistics equal to 0.31σ; moreover, we note that, once again, lines of constant metallicity are almost perfectly orthogonal to the best-fit curve everywhere along the sequence. Interestingly, lines of constant star formation rate are almost parallel to the best-fit line of the SF sequence across the entire diagram, whereas significant variations are seen to occur when crossing the line. This suggests a potential strong correlation between the offset from the SF locus and the SFR, which can be also visualised in Fig. 10 (and quantified by a $\Delta_{\perp} = 1.25\sigma$), where, in a similar fashion as for Fig. 2, we plot the logarithmic deviation in each parameter from the average value measured on the closest point along the best-fit line of the sequence (we refer to Section 3.2 for details about how the Δlog(p) metric is computed). In Table 2, we summarise properties and statistics associated to each of the parameters of interest for the [S ii]-BPT.

5.2 Machine Learning Analysis

For the the machine learning analysis, we replicate the framework described in Section 4, with just a few differences as described below. In particular, the set of parameters included in the ML analysis for the [S ii]-BPT is detailed in Table 2. Similarly to the [N ii]-BPT case, only a limited number of parameters is selected for the purposes of assessing the performances of a multi-parameter set in the RF, whereas all the parameters are evaluated in the individual ANN runs, with the exception of metallicity, because its derivation involves exactly the same line ratios of the BPT-axis. As already noted however, the iso-metallicity lines appear orthogonal to the best-fit curve of the SF sequence everywhere across the diagram (as clearly shown by Fig. 9). Therefore, the considerations made in Section 4 for the [N ii]-BPT remain valid for the [S ii]-BPT as well, and removing metallicity from the ML analysis is not expected to bias the final results significantly, as any contribution to an orthogonal offset from variations in log(O/H) can be assumed, in this framework, negligible a priori.

We further note instead that, in contrast to what was done for the [N ii]-BPT, here we select the more ‘direct’ N/O abundance tracer given by the [N ii]/[O ii] ratio: in this way, not only do we rely on a more physically motivated tracer for N/O, but any trivial correlation between the [S ii]/Hα-axis and [N ii]/[S ii] is also removed. The other parameters are selected according to the same criteria outlined in Section 4 for the [N ii]-BPT.

Finally, the distance **D** and angle θ metrics for the [S ii]-BPT are computed in the same way as in equations 5 and 6, as illustrated in Fig. 11. Here we note that, because of the double-branched nature of the SF sequence and the orthogonality of the offset vector, the θ angle can assume also negative values.

5.2.1 Neural Networks

The results of the ANN classification analysis are presented in Fig. 12. As done previously, only galaxies with |**D**|>0.025 are included in the classification analysis, to reduce the noise introduced by the potential misclassification of sources located extremely close to the best-fit line of the SF sequence. However, we note that including all galaxies slightly reduces the performances of the network, but does not impact at all the ranking of the parameters nor affect the interpretation of the results.

Overall, the network achieves a ∼ 87% accuracy (∼ 94% AUC) in the binary classification task when fed with the ‘multi-parameter’ set, only slightly worse than the performance achieved in the [N ii]-BPT. In terms of individual parameters, the distribution of feature performances is quite different from that found for the [N ii]-BPT: here, in fact, Δlog(SFR) is the most predictive variable, followed by deviations in stellar mass, whereas the most predictive feature in the [N ii]-BPT (i.e., relative variations in N/O, here traced by Δlog([N ii]/[O ii])) scores an accuracy of only ∼ 63% in galaxy classification.

Moving to the regression problem, i.e. trying to reproduce the minimum distance **D** of each point from the best-fit line of the star-forming sequence, the network achieves an overall RMSE of 0.045 and a ∼ 37% IoR on the test sample in the ‘multi-parameter’ run, with the comparison between the predicted and the observed target Distance shown in the inset, upper-right panel of Fig. 12. Again, relative variations in star-formation rate score the best performance among the individual parameter runs, followed by $\Delta M_\star$ and Δlog([O iii]/[O ii]). In general, the ranking of individual parameters

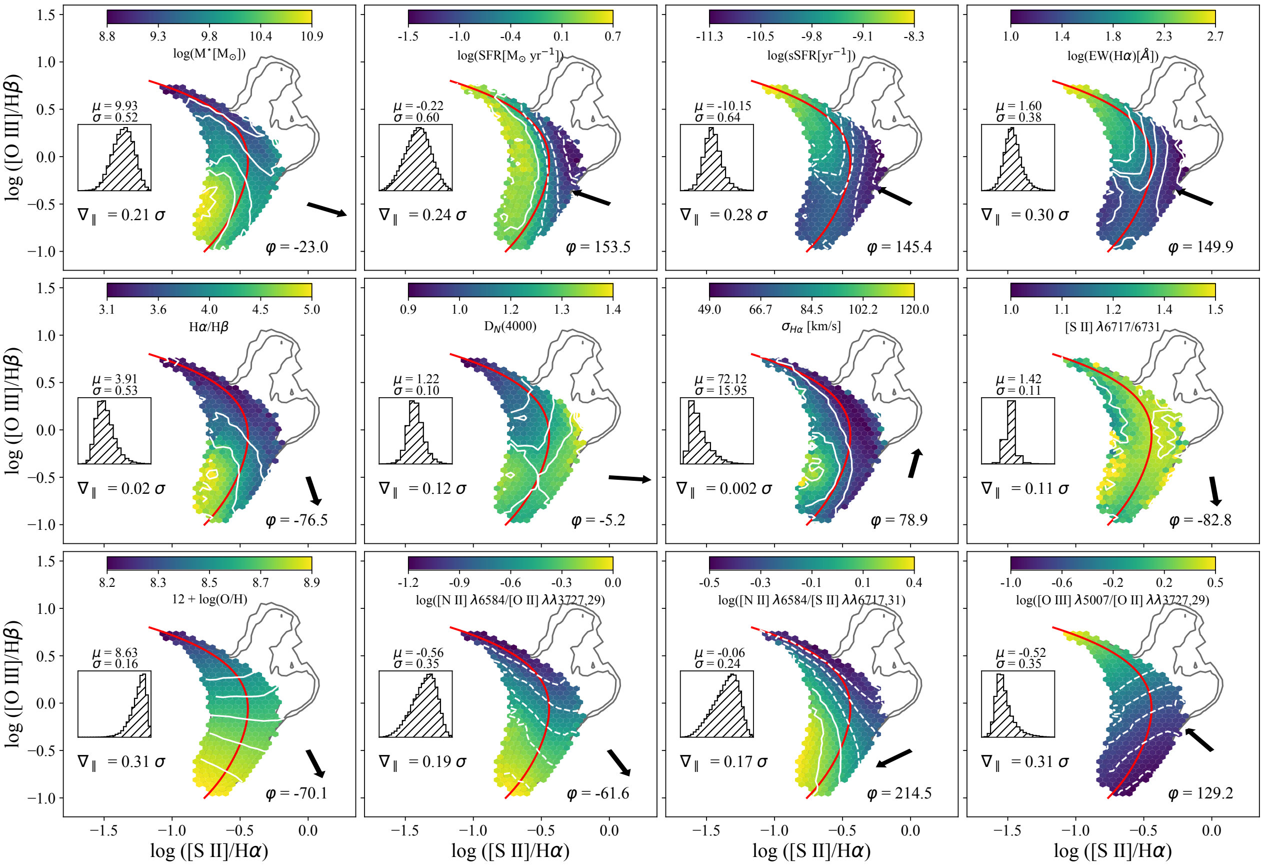

Figure 9. This figure is organised as in Fig. 1, but for the [S ii]-BPT diagram.

Parameter	Physical property	$\nabla_{\parallel}$	φ	$\Delta_{\perp}$	Multi-parameter
$\log(M^{\star}[M_{\odot}])$	Stellar mass	0.21σ	−23.0°	0.41σ	✓
$\log(SFR[M_{\odot}\ yr^{-1}])$	Star formation rate	0.24σ	153.5°	1.25σ	✓
$\log(sSFR[yr^{-1}])$	Specific SFR	0.28σ	145.4°	0.35σ	✗
EW(Hα)[]	Specific SFR	0.3σ	149.9°	0.27σ	✓
Hα/Hβ	Dust extinction	0.02σ	−76.5°	0.42σ	✗
$D_N(4000)$	Age of stellar populations	0.12σ	−5.2°	0.15σ	✓
$\sigma_{H\alpha}$ [km/s]	Gas velocity dispersion	0.002σ	78.9°	1.19σ	✓
[S ii]λ6717/6731	Gas density	0.11σ	−82.8°	0.52σ	✓
12 + log(O/H)	Oxygen abundance	0.31σ	−70.1°	0.01σ	✗
log([N ii] λ6584/[O ii] $\lambda\lambda$3727, 29)	N/O abundance	0.19σ	−61.6°	0.15σ	✓
log([N ii] λ6584/[S ii] $\lambda\lambda$6717, 31)	N/O abundance	0.17σ	214.5°	0.47σ	✗
log([O iii] λ5007/[O ii] $\lambda\lambda$3727, 29)	Ionisation parameter (U)	0.31σ	129.2°	0.09σ	✓

Table 2. List of parameters considered in the analysis of the [S ii]-BPT diagram. The statistics defined in Section 3, and the list of parameters included in the 'multi-parameter' analysis, are also reported.

in the regression task is fully consistent to what was found when solving the classification problem.

5.2.2 *Random Forest*

The random forest analysis on the [S ii]-BPT diagram is presented in Fig. 13, for classification (left-hand panel) and regression (right-hand panel). Deviations in SFR clearly rank as the most important parameter in classifying galaxies within the diagram, whereas Δlog([O iii]/[O ii]) is ranked as the second most important variable to be used in conjunction with Δlog(SFR). Because of the way the RF computes the relative importance of the parameters (fully accounting for their mutual correlations), once again the relative importance of some variables appears here suppressed compared to their absolute performance shown in Fig. 12, demonstrating that part (or all) of their individual predictive power followed purely from correlation with other parameters. In the regression task, the RF achieves similar accuracy as the ANN, and the ranking of relative importance closely follows that of the classification task, with Δ(SFR) and Δlog([O iii]/[O ii]) dominating over the other variables.

When the RF is forced to randomly select only ($\sqrt{N_{\rm features}}$) features at each fork, deviations in $M_{\star}$, EW[Hα] and $\sigma_{\rm H\alpha}$ retain part

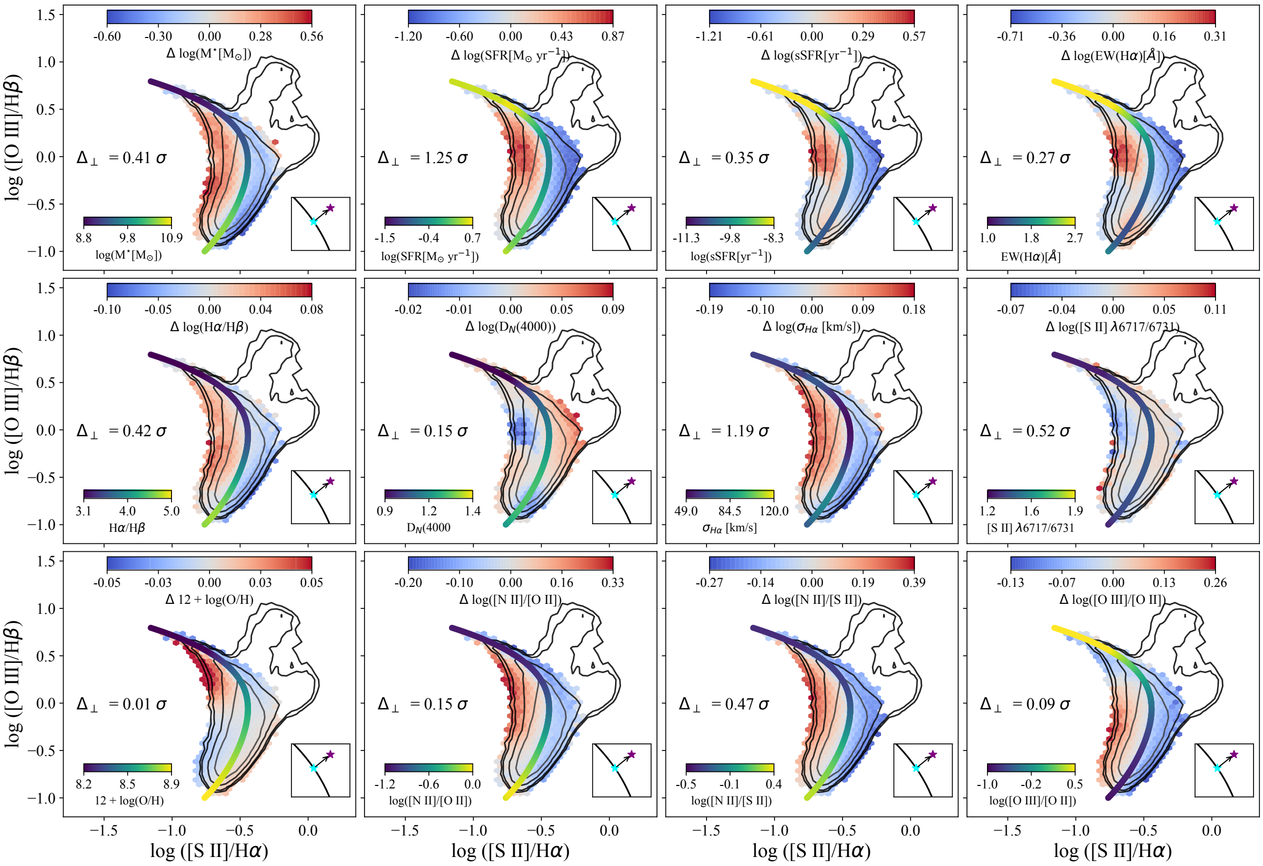

Figure 10. This figure is organised as in Fig. 2, but for the [S II]-BPT diagram.

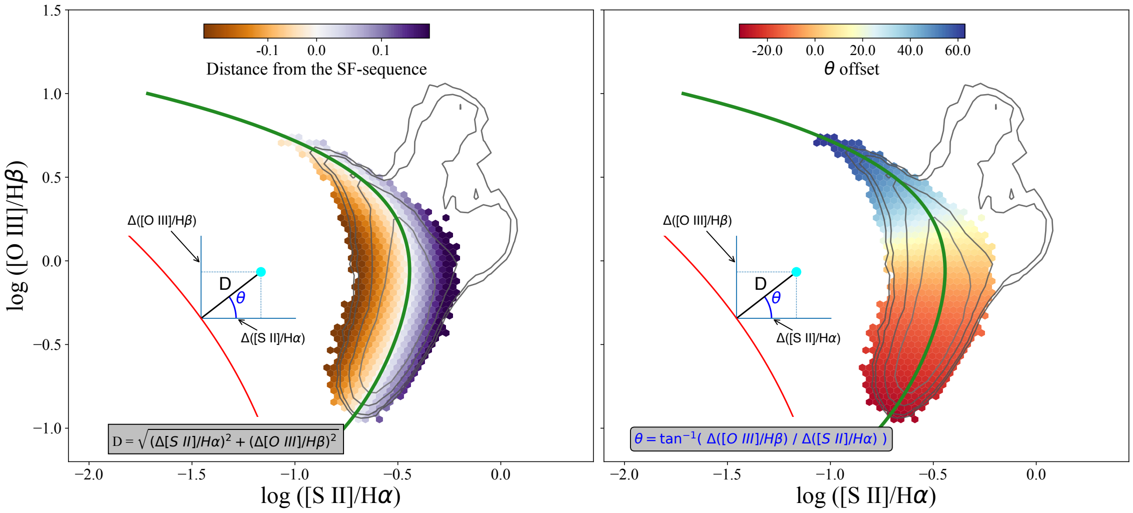

Figure 11. This figure is organised as in Fig. 3, but for the [S II]-BPT diagram.

of the residual importance at the expenses of the two main parameters (as shown by the empty, hatched bars in Fig. 13). Nonetheless, Δlog(SFR) is still strongly identified as the most predictive parameter in the set. Whether such a trend is maintained along the full SF sequence is the subject of the analysis of the following section.

5.2.3 [S II]-BPT sectors

In a similar fashion to what was done for the [N II]-BPT, we here divide the [S II]-BPT into four sectors, in order to assess how the predictivity and relative importance of each parameter changes when considering galaxies in different specific regions across the diagram, parametrised by the inclination of the offset-vector with respect to the horizontal axis. The results are shown in Fig. 15 for ANN (up-

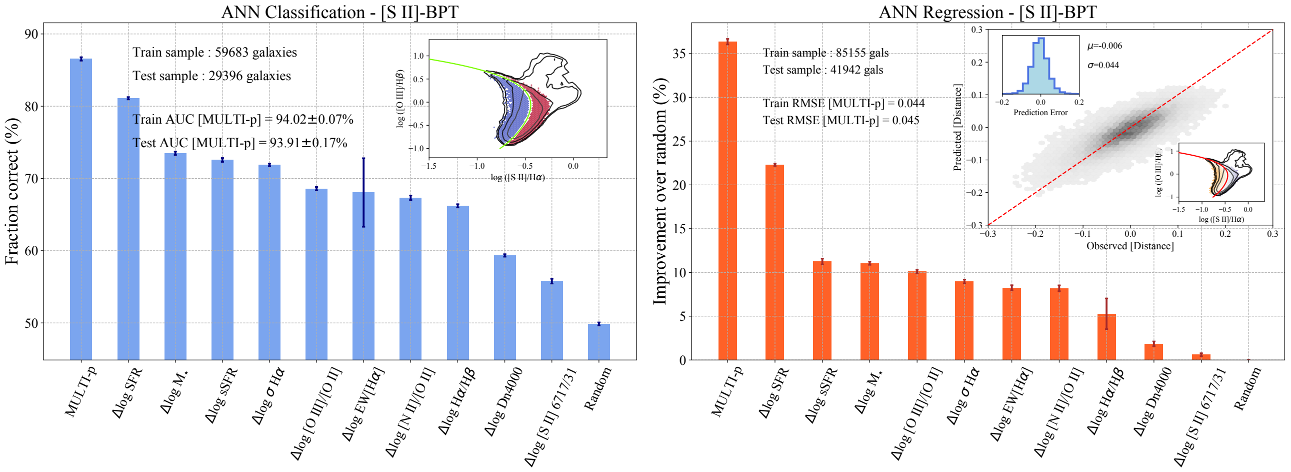

Figure 12. Results of the ANN classification (left-hand panel) and regression (right-hand panel) analysis on star-forming galaxies in the [S II]-BPT diagram. The panels are structured as in Fig. 4. Overall, the network achieves an ∼ 87% accuracy in classification and a ∼ 37% IoR in regression when trained with the 'multi-parameter' set. For both problems, Δlog(SFR) (i.e., relative variation in the star-formation rate) is the parameter that individually performs best.

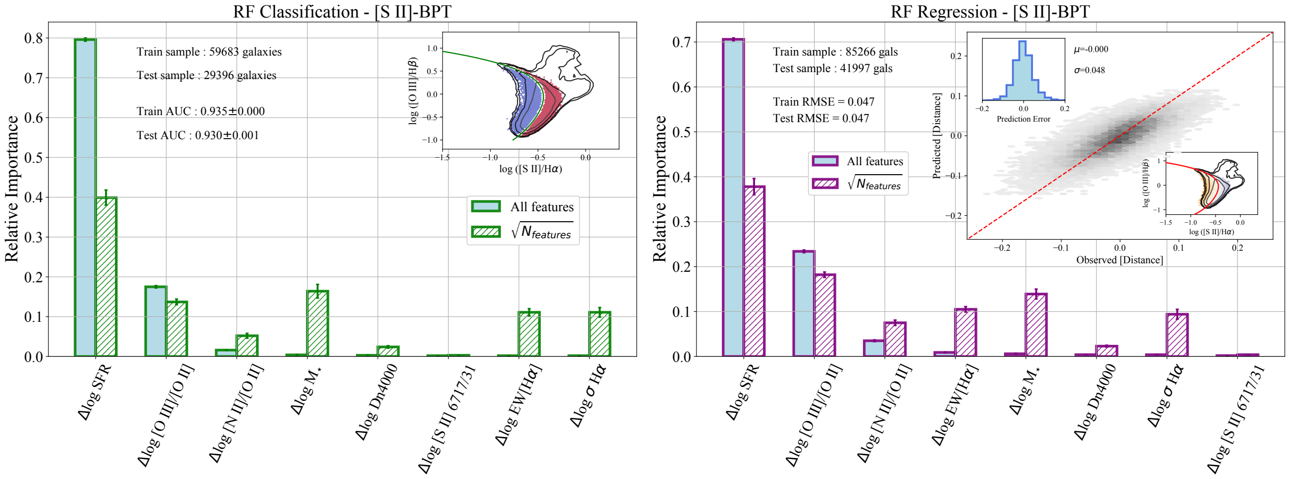

Figure 13. Results of the RF classification (left-hand panel) and regression (right-hand panel) analysis on star-forming galaxies in the [S II]-BPT diagram. The panels are structured as in Fig. 5. Δlog(SFR) is, by far, the most relevant parameter for predicting the offset from the best-fit of the median SF sequence in the [S II]-BPT diagram, and combined with Δlog([O III]/[O II]) account for more than 90% of the total predictive power of the RF. When the trees are randomised in the selection of $\sqrt{N_{\rm features}}$ features at each fork, deviations in $M_\star$, EW[Hα] and $\sigma_{\rm H\alpha}$are the parameters retaining the larger part of the residual importance.

per panels) and RF (bottom panels) respectively, where we compare the performance and relative importance of each parameter in the classification (left panels) and regression (right panels) tasks as a function of <θ>, the median angle (positive counterclockwise from the horizontal axis) formed by the offset vectors of galaxies pertaining to a given sector. Overall, the 'multi-parameter' run in the ANN maintains a constant performance level in both tasks across the entire diagram, with an accuracy close to 90 per cent in classification and an IoR $\gtrsim$ 40% in regression. The ranking of individual parameters is also roughly constant with increasing <θ>, although ΔEW(Hα) does see the sharpest increases in the central regions whilst declining, similar to ΔSFR and Δlog([O III]/[O II]), in the last sector. In contrast, deviations in N/O abundance (here traced by [N II]/[O II]) see a steady but constant increase of its performance moving from the 'bottom' to the 'top' region of the diagram. The dependence of the scatter on relative variations in gas density (traced by the [S II] doublet) remains instead almost negligible across the entire diagram.

For what concerns the RF analysis, we see that in the bottom part of the diagram (i.e., low values of <θ>) deviations in stellar mass and ionisation parameter dominate the relative contribution to the RF predictivity, while parameters associated with star formation like SFR and EW(Hα) gain more importance in the central regions (intermediate <θ> values), where the offset occurs preferentially along [S II]$\lambda\lambda$6717, 31/Hα. Interestingly, in the last sector (i.e., the one including galaxies lying at the top-left of the sequence) the scatter is dominated by deviations in N/O, which hold ∼ 80% of the total information, whereas the other parameters are strongly suppressed.

5.3 Discussion

Overall, both ANN and RF analysis suggest that the scatter in the [S II]-BPT diagram is primarily sensitive to parameters associated to recent star formation activity in galaxies. Assuming that the offset occurs at fixed metallicity (following Fig. 9 and 10 and the

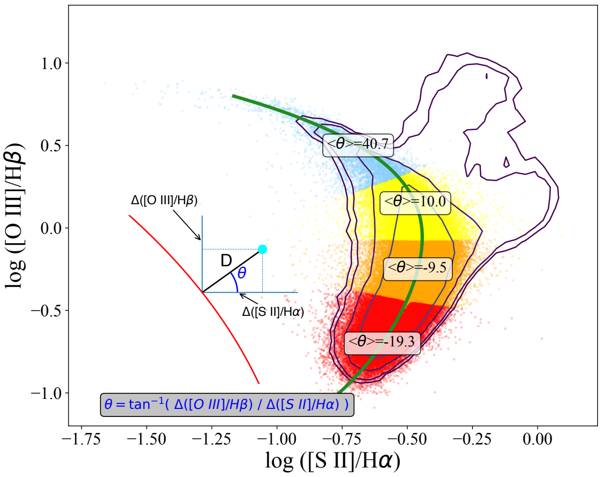

Figure 14. The four sectors in which we divide the [S ii]-BPT in order to assess the variation of parameter performances as a function of the position of galaxies along the sequence.

considerations of Section 3.2), one possible explanation involves the size of HII regions and the relative fraction occupied by the S^+ ions in galaxies with different levels of star formation. HII regions are in fact stratified, with higher ionization species like S^{++} much more common closer to the ionizing source while lower ionization species like S^+ relatively more abundant in the outer parts (see e.g., Levesque et al. 2010; Xiao et al. 2018; Mannucci et al. 2021). Since sulphur has a significantly lower ionisation potential than both oxygen and nitrogen, the S^+ zone is typically much more extended than the O^+ or N^+ ones within the same HII region, and this has an impact also on any line ratio between involving such ions like [S ii]$\lambda\lambda$6717, 31/Hα. It is possible then, that galaxies with increased current star formation (compared to median galaxies on the SF sequence) are characterised by a large number of nearby HII regions which eventually merge, reducing the effective size of the S^+ zone (as also suggested by Masters et al. 2016). We note that, because almost 60% of the total sample of star-forming galaxies are characterised by an offset vector pointing between $\theta = -20°$ and $\theta = 40°$ (hence directed predominantly along [S ii]$\lambda\lambda$6717, 31/Hα), this would likely explain why the contribution from variations in SFR dominates the overall behaviour within the diagram as exposed by the global ANN and RF results of Fig. 12 and 13. Interestingly, the sharp rise of the importance of EW(Ha)(hence, sSFR) at the expenses of SFR in the third sector provides us with additional information on the status of these galaxies, which not only are characterised by higher/lower levels of star-formation compared to 'on-sequence' galaxies, but they are also forming stars at higher/lower pace than in their past history.

The random forest analysis also suggests that coupling Δlog(SFR) with variations in the ionisation parameter (traced by Δlog([O iii]/[O ii]), which picks around 20 per-cent relative importance, hence providing complementary information to that hold by SFR only) maximises the predictivity of the algorithm. An increase in U, in fact, could on the one hand provoke a suppression of the [S ii]$\lambda\lambda$6717, 31/Hα line ratio at fixed [O iii]λ5007/Hβ, as the abundance of doubly ionised sulphur increases at the expense of S^+, while on the other could boost the [O iii]λ5007/Hβ ratio itself.

Interestingly, in the lowest part of the sequence, as probed by the $< \theta >= -19°$ sector, variation in stellar mass is the most predictive quantity of the observed offset from the median sequence; nonetheless, star-formation rate and U tracers still contribute significantly to the total predictivity. This result is probably driven by the presence of a group of high-mass galaxies located at around log([O iii]λ5007/Hβ)= −0.5, log([S ii]$\lambda\lambda$6717, 31/Hα)= −0.75 (see Fig. 9); these sources likely represent a sub-population of relatively older, high-mass, chemically mature galaxies, whose central $\sigma_{\mathrm{H}\alpha}$ (of the order of ∼ 80 − 100km/s) is more typical of bulge structures or other largely pressure-supported systems rather than thin disk structures. In this sense, the large relative importance picked by $M_\star$is likely driven by the information brought by $\sigma_\star$ (which is not represented in our parameter set), plus part of the importance 'borrowed' from SFR, with the relative importance of Δlog(SFR) and Δlog($M_\star$) indeed almost matching in the RF run with $\sqrt{N_{\mathrm{features}}}$ allowed at each node.

Finally, in the uppermost region of the diagram the ML analysis identifies variations in N/O as the most informative parameter for predicting the scatter in the [S ii]-BPT. Being the [S ii]-BPT free of nitrogen lines however, any variation in the N/O abundance cannot have a direct impact on the BPT-line ratios by itself, but should be considered as a reflection of some other underlying physical effect. As already discussed in Section 4.5, one possibility invoke to break the initial assumption of offsets occurring along iso-metallicity lines in this region of the diagram. If this is the case, the connection between the offset from the median sequence and relative variations in N/O in the [S ii]-BPT could just effectively trace metallicity variations. However, in contrast to what is seen for the [N ii]-BPT, these galaxies are observed to deviate in stellar mass, O/H and N/O according to the standard mass-metallicity-N/O relation (i.e., to an increase in $M^\star$ correspond an increase in both O/H and N/O, and viceversa). Hence, this suggest that the high relative importance kept by Δlog([N ii]/[O ii]) in this sector might just be the reflection of the average relationship between these quantities.

6 SUMMARY AND CONCLUSIONS

In this paper, we have presented a novel approach to study the distribution of galaxy properties in the BPT diagnostic diagrams, attempting to link variations in such properties (via different observational tracers) to the variations in the line ratios space observed within the diagrams, following a purely empirical, data-based approach. In particular, artificial neural networks (ANN) and random forest (RF) of decision trees have been trained and tested over a large sample of SDSS galaxies, in order to assess which physical parameters are most connected with the observed offset of local star-forming galaxies from their median sequence in both the [N ii]- and [S ii]-BPT diagrams. Relative variations in a set of physical parameters (in the form of the Δ log(p) metric, equation 4) are linked to the deviation from the sequence itself, with an offset assumed orthogonal to the best-fit line of the sequence at any given point.

The performances of our set of parameters (individually and as a whole), as well as their relative importance, are evaluated in solving both a classification (i.e., predicting whether a galaxy is offset above or below the median sequence) and a regression (i.e., predicting the exact magnitude of the offset) problem. The key points of the paper are summarised below.

- The distribution of star-forming galaxies in both the [N ii] and the [S ii]-BPT diagrams primarily traces a sequence in gas-phase metallicity. A significant gradient in log(O/H) along the best-fit curve of the SF sequence is in fact observed in both diagrams ($\nabla_{\|}$ = 0.31σ), coupled with zero (or very mild) variations assessed orthogonal to it (Δ⊥ ∼0, see also iso-contours of O/H in Fig. 1 and

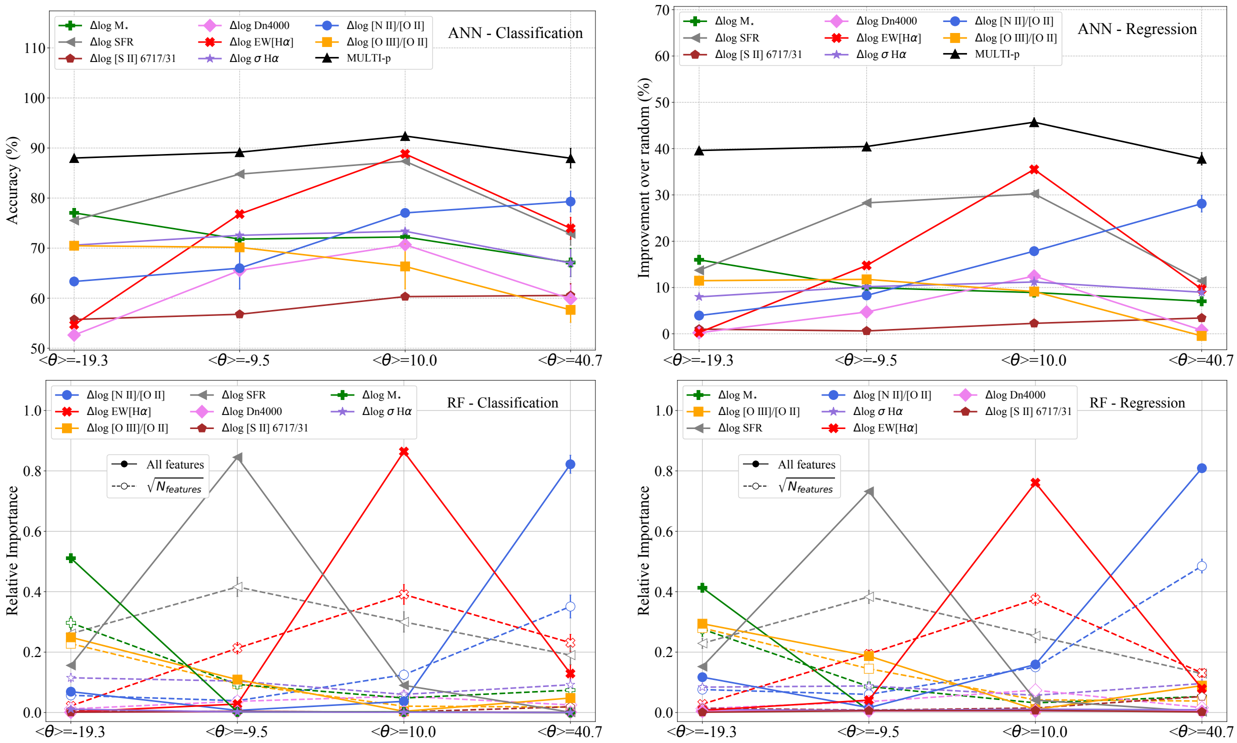

Figure 15. Same as in Fig. 7, for the four sectors the [S II]-BPT diagram has been divided into. The accuracy of the ANN is stable across the entire diagram in both classification (∼ 90%) and regression (≳ 40% IoR). Parameters connected with star formation (Δlog(SFR), Δlog(EW[Hα])) dominates the relative contribution to the observed scatter in the central regions, where the offset vector is mainly directed along the [S II]/Hα-axis (i.e., low <θ> values), whereas at the edges of the SF sequence, parameters related to the chemodynamical properties of galaxies (e.g., Δlog(M$_\star$), Δlog([N II]/[O II])) increase their impact on the predicted offset.

9). Hence, in our framework we assume the gas-phase metallicity to be the main parameter that set the position of galaxies *along* the SF sequence, whereas contributions from relative variations in O/H are not considered to significantly impact the deviations from it, as described by a purely orthogonal ‘offset vector’.

- When trained with multiple parameters, the ANN is capable of classifying whether a galaxy is offset above or below the best-fit of the median SF sequence with > 90 per cent accuracy (AUC= 0.96) in the [N II]-BPT, and to predict the magnitude of the offset from the sequence itself with a RMSE= 0.038 on the test sample. Among individual parameters, Δlog([N II]/[S II]) and Δlog([N II]/[O II]) (tracing relative variations in the N/O abundance compared to ‘on-sequence’ galaxies) are robustly assessed as the most accurate features in both classification and regression tasks for the [N II]-BPTdiagram (Fig. 4).

- From the RF analysis, we find that Δlog([N II]/[S II]) is, by far, the most relevant parameter for predicting the offset from the SF locus in the [N II]-BPT, gathering more than 80% of the total importance among the whole ‘multi-parameter’ set (Fig. 5). Therefore, any offset from the median sequence of star-forming galaxies in this diagram is primarily associated to relative variations in their N/O abundance.

- The impact of the individual parameters on the offset of galaxies from the best-fit curve of the SF median loci in the [N II]-BPT changes as a function of the position along the sequence. In the bottom-right region of the diagram, the offset-vector is almost horizontal, and the deviation from the SF sequence is nicely predicted by properties related to the chemo-dynamical evolution of galaxies (e.g., Δlog([N II]/[S II]), $\Delta\sigma_{\mathrm{H}\alpha}$ and ΔM*), whereas moving along the SF locus, parameters associated to ongoing star-formation in galaxies (e.g., ΔSFR, Δlog([O III]/[O II]), ΔEW[Hα]) increase their predictivity (Fig. 7, upper panels). Nonetheless, Δlog([N II]/[S II]) remains the most relevant parameter of the set for predicting the offset from the SF sequence throughout the entire diagram, regardless of the relative amplitude of the components of the offset vector.

- If we assume the offset to occur at fixed metallicity, these results can be interpreted as a manifestation of the relationship between N/O and O/H (mainly driven by the ‘secondary’ nucleosynthetic production of nitrogen in high mass galaxies), whose median behaviour and scatter is to a large extent reflected in the distribution of galaxies within the [N II]-BPT (Fig. 8). When only $\sqrt{\mathrm{N}_{\mathrm{features}}}$ are considered at each fork of the RF, feature importance is partially shifted from the most important variable (i.e., Δlog([N II]/[S II])) to variables which are strongly correlated with it, like M$_\star$, which acts then as a good proxy for N/O.

- Parameters associated with the age of the stellar populations (D$_\mathrm{N}$(4000)) and specific star formation rate (EW(Hα)) provide complementary information, which is required to maximise the predictivity of the RF (Fig. 5). These are likely associated with the differential contribution of older stellar populations (e.g., hot, post-AGB stars), which impact the strength of intermediate- and low-ionisation emission lines originating from the warm, diffuse ionised gas outside HII regions.

- In the [S II]-BPT diagram, the overall scatter of galaxies around the best-fit SF sequence is primarily associated with relative variations in SFR (Fig. 12). Among the other parameters, Δlog([O III]/[O II])(i.e., variations in the ionisation parameter) retain the most complementary information to ΔSFR that maximises the accuracy of both classification and regression tasks (Fig. 13).

- In particular, parameters associated with recent star formation (Δlog(SFR)and Δlog(EW[Hα])) dominate the relative contribution to the offset in the central part of the diagram, where the majority of galaxies reside (and where the orthogonal offset vector is primarily directed along [S II]$\lambda\lambda$6717, 31/Hα). We primarily interpret this in terms of the relative change in the extension of the low-ionisation, S^{+} zone of HII regions within galaxies with different levels of ongoing star formation. Further contribution from variations in U can either impact the [O III]λ5007/Hβ ratio and affect the relative S^{++}/S^{+} ionic abundance.
- At the edges of the sequence, where the best-fit line of the SF locus in the [S II]-BPT diagram bends, parameters tracing chemo-dynamical properties of galaxies (e.g., Δlog([N II]/[S II]), $\Delta M^{\star}$) increase their scores in the ANN, and gain a higher amount of relative importance in the RF (Fig. 15). This is likely driven by a sub-population of high-mass, bulge dominated galaxies in the bottom-left region, whereas might partially trace residual metallicity variations, which are mainly accounted for by N/O, in the upper-left part.

In conclusion, we have shown how the distribution of star-forming galaxies in the BPT diagnostic diagrams can be well described by a framework in which the offset from the median location of sources along the SF sequence can be ascribed to relative variations in different physical conditions, once the position along the sequence has been set by the knowledge of their gas-phase oxygen abundance. Exploiting a variety of machine learning techniques, we have robustly identified relative variations in the N/O abundance to primarily govern the scatter in the [N II]-BPT diagram, whereas relative variations in parameters associated with star-formation are most relevant to predict the behaviour of galaxies in the [S II]-BPT diagram. Such framework could be tested in the future on both different (and even multi-dimensional) diagnostic diagrams, as well as on high redshift galaxy samples, to provide new and complementary insights for photoionisation models about the evolution of physical conditions in galaxies across cosmic time, as inferred from their observed spectral properties.

ACKNOWLEDGEMENTS

MC, CH-P and RM acknowledge support by the Science and Technology Facilities Council (STFC) and from European Research Council (ERC) Advanced Grant 695671 'QUENCH'. We are also very grateful for several insightful and stimulating discussions on this work, and on machine learning techniques in general, especially to Asa Bluck, Joanna Piotrowska and Sim Brownson.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England.

References

Abazajian K. N., et al., 2009, ApJS, 182, 543
Aller L. H., 1942, ApJ, 95, 52
Andrews B. H., Martini P., 2013, ApJ, 765, 140
Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5
Barchi P. H., et al., 2020, Astronomy and Computing, 30, 100334
Baron D., 2019, arXiv e-prints, p. arXiv:1904.07248
Belfiore F., et al., 2016, MNRAS, 461, 3111
Bluck A. F. L., Maiolino R., Sánchez S. F., Ellison S. L., Thorp M. D., Piotrowska J. M., Teimoorinia H., Bundy K. A., 2019, Monthly Notices of the Royal Astronomical Society, p. stz3264
Bluck A. F. L., et al., 2020, arXiv:2009.05341 [astro-ph]
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Bundy K., et al., 2015, ApJ, 798, 7
Byler N., Dalcanton J. J., Conroy C., Johnson B. D., 2017, ApJ, 840, 44
Byler N., Dalcanton J. J., Conroy C., Johnson B. D., Choi J., Dotter A., Rosenfield P., 2019, AJ, 158, 2
Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245
Chabrier G., 2003, PASP, 115, 763
Cid Fernandes R., Stasińska G., Mateus A., Vale Asari N., 2011, MNRAS, 413, 1687
Curti M., Cresci G., Mannucci F., Marconi A., Maiolino R., Esposito S., 2017, MNRAS, 465, 1384
Curti M., Mannucci F., Cresci G., Maiolino R., 2020, Monthly Notices of the Royal Astronomical Society, 491, 944
D'Agostino J. J., Kewley L. J., Groves B. A., Medling A., Dopita M. A., Thomas A. D., 2019, MNRAS, 485, L38
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, IEEE Proceedings, 97, 1482
Dopita M. A., Evans I. N., 1986, ApJ, 307, 431
Faisst A. L., Masters D., Wang Y., Merson A., Capak P., Malhotra S., Rhoads J. E., 2018, ApJ, 855, 132
Gaia Collaboration et al., 2016, A&A, 595, A1
Green A. W., et al., 2014, MNRAS, 437, 1070
Hayden-Pawson C., et al., 2021, arXiv e-prints, p. arXiv:2110.00033
Hirschmann M., Charlot S., Feltre A., Naab T., Choi E., Ostriker J. P., Somerville R. S., 2017, MNRAS, 472, 2468
Ho I.-T., 2019, MNRAS, 485, 3569
Hsieh B. C., et al., 2017, ApJ, 851, L24
Ivezic Z., et al., 2008, Serbian Astronomical Journal, 176, 1
Ji X., Yan R., Riffel R., Drory N., Zhang K., 2020, MNRAS, 496, 1262
Kashino D., et al., 2017, ApJ, 835, 88
Kauffmann G., et al., 2003a, MNRAS, 341, 33
Kauffmann G., et al., 2003b, MNRAS, 341, 54
Kauffmann G., et al., 2003c, MNRAS, 346, 1055
Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531
Kewley L. J., Dopita M. A., 2002, ApJS, 142, 35
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Kewley L. J., Maier C., Yabe K., Ohta K., Akiyama M., Dopita M. A., Yuan T., 2013, ApJ, 774, L10
Kewley L. J., Nicholls D. C., Sutherland R. S., 2019, Annual Review of Astronomy and Astrophysics, 57, 511
Kobayashi C., Karakas A. I., Umeda H., 2011, MNRAS, 414, 3231
Krumholz M. R., Burkhart B., Forbes J. C., Crocker R. M., 2018, MNRAS, 477, 2716
Lacerda E. A. D., et al., 2018, MNRAS, 474, 3727
Law D. R., et al., 2021a, AJ, 161, 52
Law D. R., et al., 2021b, ApJ, 915, 35
Levesque E. M., Richardson M. L. A., 2014, ApJ, 780, 100
Levesque E. M., Kewley L. J., Larson K. L., 2010, AJ, 139, 712
Levi M., et al., 2013, arXiv e-prints, p. arXiv:1308.0847
Magrini L., et al., 2018, A&A, 618, A102
Maiolino R., Mannucci F., 2019, A&ARv, 27, 3
Maiolino R., et al., 2008, A&A, 488, 463
Mannucci F., et al., 2021, arXiv e-prints, p. arXiv:2109.02684
Masters D., Faisst A., Capak P., 2016, The Astrophysical Journal, 828, 18
McCall M. L., Rybski P. M., Shields G. A., 1985, ApJS, 57, 1
Mingozzi M., et al., 2020, Astronomy & Astrophysics, 636, A42
Osterbrock D. E., Ferland G. J., 2006, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei, 2nd edn. University Science Books
Pettini M., Pagel B. E. J., 2004, MNRAS, 348, L59

Pilyugin L. S., Vílchez J. M., Mattsson L., Thuan T. X., 2012, MNRAS, 421, 1624

Pérez-Montero E., Contini T., 2009, MNRAS, 398, 949

Reza M., 2021, Astronomy and Computing, 37, 100492

Rich J. A., Dopita M. A., Kewley L. J., Rupke D. S. N., 2010, ApJ, 721, 505

Salim S., et al., 2007, ApJS, 173, 267

Shapley A. E., et al., 2015, ApJ, 801, 88

Sooknunan K., et al., 2021, MNRAS, 502, 206

Steidel C. C., et al., 2014, ApJ, 795, 165

Strom A. L., Steidel C. C., Rudie G. C., Trainor R. F., Pettini M., Reddy N. A., 2017, ApJ, 836, 164

Teimoorinia H., Bluck A. F. L., Ellison S. L., 2016, Monthly Notices of the Royal Astronomical Society, 457, 2086

Teimoorinia H., Jalilkhany M., Scudder J. M., Jensen J., Ellison S. L., 2021, arXiv e-prints, 2102, arXiv:2102.07058

Topping M. W., Shapley A. E., Reddy N. A., Sanders R. L., Coil A. L., Kriek M., Mobasher B., Siana B., 2020a, arXiv:2008.02282 [astro-ph]

Topping M. W., Shapley A. E., Reddy N. A., Sanders R. L., Coil A. L., Kriek M., Mobasher B., Siana B., 2020b, MNRAS, 495, 4430

Varidel M. R., et al., 2020, MNRAS, 495, 2265

Vavilova I. B., Dobrycheva D. V., Vasylenko M. Y., Elyiv A. A., Melnyk O. V., Khramtsov V., 2021, A&A, 648, A122

Veilleux S., Osterbrock D. E., 1987, in Lonsdale Persson C. J., ed., NASA Conference Publication Vol. 2466, NASA Conference Publication.

Ventura P., Di Criscienzo M., Carini R., D'Antona F., 2013, MNRAS, 431, 3642

Vila Costas M. B., Edmunds M. G., 1993, MNRAS, 265, 199

Vincenzo F., Kobayashi C., 2018, A&A, 610, L16

Vincenzo F., Belfiore F., Maiolino R., Matteucci F., Ventura P., 2016, MNRAS

Xiao L., Stanway E. R., Eldridge J. J., 2018, MNRAS, 477, 904

Yabe K., et al., 2015, PASJ, 67, 102

Yan R., Blanton M. R., 2012, ApJ, 747, 61

York D. G., et al., 2000, AJ, 120, 1579

Yu X., et al., 2019, MNRAS, 486, 4463

Zhang K., et al., 2017, MNRAS, 466, 3217

de la Calleja J., Fuentes O., 2004, MNRAS, 349, 87

van Zee L., Salzer J. J., Haynes M. P., O'Donoghue A. A., Balonek T. J., 1998, AJ, 116, 2805

APPENDIX A: TESTS ON THE RANDOM FOREST

A1 Different sets of parameters

Throughout the paper, we have analysed the connection between the deviation of star-forming galaxies from the median sequence in the BPT diagrams and a number of physical quantities, traced by rather direct or indirect spectro-phototmetric observables. As discussed already in Section 2.1.2, some of these parameters can be traced by means of different ratios of emission lines: for instance, if we consider the SDSS galaxy sample at the basis of this work, the ionisation parameter can be traced either by the [O iii]λ5007/[O ii]λ3727, 29or the [Ne iii]/[O ii] ratio, whereas the N/O abundance is traced by both [N ii]λ6584/[S ii]λ6717, 31and [N ii]λ6584/[O ii]λ3727, 29(although the latter is a more 'direct' probe than the former). The selection of our fiducial set of parameters to be included in the ML analysis is discussed and motivated in Section 4. However, in this appendix we want to test how much the results and conclusions presented in the main body of the paper are robust to the choice of a different set of parameters, either by changing some of the originally chosen tracers and/or by removing one or more variables. In particular, we assess which impact this might have on the RF analysis in terms of the estimated relative feature importance.

In first instance, we start by considering the [N ii]λ6584/[O ii]λ3727, 29to trace N/O instead of [N ii]λ6584/[S ii]λ6717, 31. As discussed in Section 4, including [N ii]λ6584/[S ii]λ6717, 31in the fiducial RF analysis was required to avoid any trivial correlation between [N ii]λ6584/[O ii]λ3727, 29and [O iii]λ5007/[O ii]λ3727, 29, which would have provided biased relative importance for these two features in the prediction of our target labels (which are based on a combination of the [O iii]λ5007/Hβand [N ii]λ6584/Hαline ratios). In order to include [N ii]λ6584/[O ii]λ3727, 29and maintain at the same time the other parameters in the set independent, we can follow two different approaches, i.e. **i)** change the ionisation parameter tracer accordingly, from [O iii]λ5007/[O ii]λ3727, 29to [Ne iii]/[O ii] or **ii)** remove [O iii]λ5007/[O ii]λ3727, 29at all (hence, any variable related to the ionisation parameter) from the analysis.

Although case **i)** might sound the most obvious and physically motivated choice, we note that, given the low signal-to-noise of the [Ne iii]λ3869 in many individual SDSS spectra, requiring significant detections in such emission line inevitably impact the final selected sample, introducing a bias which is directly correlated with the position of galaxies in the BPT diagram itself (i.e., galaxies would be preferentially removed from the bottom-right, metal-rich region of the diagram). To mitigate this effect, for the purposes of the present test we require only a 2.5σ detection in the [Ne iii]λ3869 emission line (on top of the S/N requirements outlined in Section 2), which is enough to provide a $\gtrsim 3\sigma$ significance in the [Ne iii]/[O ii] ratio; this brings the final selected galaxy sample to 22, 840.

The output from the RF classification analysis for such dataset are shown in the left panel of Fig. A1 (which replicates the structure of Fig. 5): Δlog([N ii]/[O ii])is ranked as the most important parameter in the set, confirming the results obtained with Δlog([N ii]/[S ii]), whereas [Ne iii]/[O ii] retains a similar level of relative importance (in combination with Δlog([N ii]/[O ii])) as that originally scored by Δlog([O iii]/[O ii])in Fig. 5.

In case **ii)**, we decide instead to remove completely any tracer associated to the ionisation parameter (whose importance is, as we have seen, overall minimal when variations in N/O are already accounted for) and perform the RF analysis on the original full sample of galaxies by including [N ii]λ6584/[O ii]λ3727, 29as the N/O tracer. We stress here again that any change in the composition of the set of parameters to be fed to the RF might affect the overall distribution of relative importance among the different quantities. Nonetheless, in this way we can not only assess the performance of [N ii]λ6584/[O ii]λ3727, 29in the RF exploiting our full statistical power, but also test to what extent removing the other emission line-based parameter from the set would impact its final score (i.e., whether a large part of the relative importance of Δlog([N ii]/[O ii])is just driven by trivial correlations with other parameters based on emission line ratios). The results of this second test are shown, for the RF classification task, in the right panel of Fig. A1. Again, Δlog([N ii]/[O ii])is assessed as the most relevant parameter from the RF, whereas the small residual importance associated with the ionisation parameter tracers is now mostly accounted for by $\sigma_{\mathrm{H}\alpha}$and M*.

Both these tests confirms that relative variations in the N/O abundance are most predictive for characterising the position of a galaxy with respect to the median SF sequence in the [N ii]-BPT, regardless of the choice of the N/O tracer and of the inclusion of different emission lines-based features. For the sake of brevity, we do not discuss the RF regression analysis here, which we have verified to give fully comparable results for both case **i)** and **ii)** discussed above.

We just further briefly comment on the amount of relative importance retained by $\sigma_{\mathrm{H}\alpha}$in the RF analysis, when Δlog([N II]/[O II]) is adopted instead of Δlog([N II]/[S II]) to trace variations in N/O. In fact, where [N II]λ6584/[O II]λ3727, 29 traces more directly N/O by virtue of the closest ionisation potential of N^{+} and O^{+}, [N II]λ6584/[S II]λ6717, 31 presents further, secondary dependence on different parameters on top of that on N/O. A partial correlation analysis reveals, in fact, that a strong correlation exists between [N II]/[S II] and SFR, $\mathrm{M}^{\star}$ and $\sigma_{\mathrm{H}\alpha}$, at fixed [N II]/[O II], with partial Spearman ranks equal to 0.9583, 0.9178 and 0.9418, respectively. It is plausible, then, that decoupling the N/O tracer from such dependencies is reflected into the RF 'seeing' the relative importance of these parameters, and in particular $\sigma_{\mathrm{H}\alpha}$, increasing with respect to the others. We also note in fact that central $\sigma_{\mathrm{H}\alpha}$, tracing predominantly the dynamical mass of galaxies, is one of the parameters showing the strongest 'absolute' variation across the SF sequence compared to the variation along it (see Fig. 1 and 2), with a $\nabla_{||} = 0.002\sigma$ and $\Delta_{\perp} = 0.74\sigma$. Therefore, relative variations in $\sigma_{\mathrm{H}\alpha}$are well connected to the deviations from the SF sequence in the [N II]-BPT, and this is now exposed also by the RF, whereas previously this information was already partially embedded in the use of Δlog([N II]/[S II]). The performance of $\sigma_{\mathrm{H}\alpha}$is likely driven by the highest masses galaxies at the edge of the star-formation Kauffmann et al. (2003b) diving line; indeed, the relative importance carried by $\sigma_{\mathrm{H}\alpha}$ and $\mathrm{M}_{\star}$ is almost equivalent in the $\sqrt{N_{\mathrm{features}}}$ case of the RF.

A2 Assessing trivial correlations between the target labels and emission line-based parameters

One potentially critical point of the hereby presented analysis concern the level of trivial correlation which exists between our target label in the ML, which is based on a combination of [O III]λ5007/Hβand [N II]λ6584/Hα, and some of the involved parameters (in particular Δlog([N II]/[S II])and Δlog([O III]/[O II])), which shares one of their emission lines with the target label itself (i.e, [N II]λ6584and [O III]λ5007, respectively). In this section we aim to test if, and to what extent, the final outputs of the ML algorithms (especially the success of Δlog([N II]/[S II])and Δlog([N II]/[O II])) are just trivially recovered from the covariance of target labels and emission lines-based parameters.

In order to perform this test, we keep the flux of the [N II]λ6584emission line fixed for all galaxies in our sample, while randomly shuffling the [S II]λ6717, 31line fluxes (which goes at the denominator of both line ratios) among the full selected star-forming galaxies. In such way, we create a 'hybrid' variable, which we refer to as [N II]/$\overparen{[\mathrm{S\,II}]}$, which share the emission line at the numerator with the target label of the ML, but it does not retain any longer a clear, physical interpretation as an N/O tracer. Therefore, if the predictivity of the fiducial Δlog([N II]/[S II])parameter resided only (or mostly) in having the [N II]λ6584line flux in common with **D**, the RF should still pick [N II]/$\overparen{[\mathrm{S\,II}]}$ as the (or one of the) most relevant variable among the full set of parameters; on the contrary, if the relative importance of [N II]/$\overparen{[\mathrm{S\,II}]}$ is strongly suppressed, that would suggests that the information resides in the full line ratio (hence, in the actual N/O abundance) rather than just being driven by trivial mathematical covariance.

The results of the RF classification and regression analysis are shown in the left and right panel of Fig. A2, respectively. In both cases, the relative importance of [N II]/$\overparen{[\mathrm{S\,II}]}$ appears strongly suppressed (scoring less than 10 per-cent), both compared to its fiducial value of Fig. 5 and to that of the other parameters in the set, whose relative weights in the prediction of the target label are now instead increased. Interestingly, we note that the overall performances of the algorithm are clearly hampered, with a reduction in the AUC for classification and in RMSE for regression of XX and YY per cent, respectively, compared to the fiducial analysis presented in Section 4.3. These two observations, if taken together, confirms that, on the one hand, the relative importance of Δlog([N II]/[S II])does not follow from trivial correlations with our target label as driven by line fluxes in common (if not for less than 10 per-cent), whereas on the other, that the overall performance of our parameters set is significantly affected if we remove observables carrying direct information about the N/O state of galaxies, because even the combination of all the remaining parameters is not capable of providing the same level of predictive power. This further corroborates the interpretation of N/O (and its relative variations compared to median behavior of galaxies) as the primary responsible for the observed scatter in the [N II]-BPT diagram.

APPENDIX B: PARTIAL CORRELATION ANALYSIS

In this appendix we present a rather different but complementary approach to the analysis of the connection between the offset from the SF sequence in the BPT diagrams and the set of physical parameters adopted in the paper, based on the evaluation of (partial) correlation coefficients.

In Fig. B1 we represent the matrix of Spearman correlation coefficients computed among the features in the 'multi-parameter' set (in their 'Δlog' form), and including the distance **D** from the SF sequence too, for both the [N II]-BPT(left panel) and [S II]-BPT(right panel). Each square in the matrix is colour-coded (on a diverging 'blue-to-red' scheme) according to the value of Spearman correlation rank scored by the two parameters representing the 'coordinates' of that element in the matrix.In this way, it is readily immediate to visualise the amount of correlation between all the involved parameters, and between each parameter and our target label: Δlog([N II]/[S II])is the quantity scoring the highest correlation rank with **D** in the [N II]-BPT, whereas Δlog(SFR)is for the [S II]-BPT, in agreement with the findings of the ML analysis presented in the main body of the paper.

Starting from these observations, and following Bluck et al. (2020), in Fig B2 we present a more detailed assessment of the correlation between Δlog([N II]/[S II])and **D** (upper panel), and Δlog(SFR)and **D** (lower panel), but with a rather different way to visualise the results. In the upper panel of Fig B2 for instance, the Spearman correlation strength of Δlog([N II]/[S II])with **D** is presented as light shaded red bars, and is reported adjacent to the Spearman rank correlation strengths of each other parameter in the set, as listed along the x-axis (in light shaded blue bars). The light shaded bars confirms what already shown by the correlation matrix in Fig. B1., i.e. that the correlation strengths of Δlog([N II]/[S II])with **D** are higher than for any other variable. For other few parameters (e.g., Δlog($\mathrm{M}^{\star}$) , Δlog(σ)), the correlations ranks are $\sim 30-40$ per cent lower than Δlog([N II]/[S II]), whereas for the remaining parameters are even more suppressed.

The red, left-hand, solid shaded bars in the upper panel of Fig. B2 represent instead the partial correlation strengths of Δlog([N II]/[S II])with **D**, at fixed values of each other parameter; the partial correlation strengths of each other variable with **D**, at a fixed Δlog([N II]/[S II])are instead represented by the solid, blue, right-hand bars. Therefore, any subgroup of bars (and their relative x-axis labels) should be intended as representative of a pair of parameters,

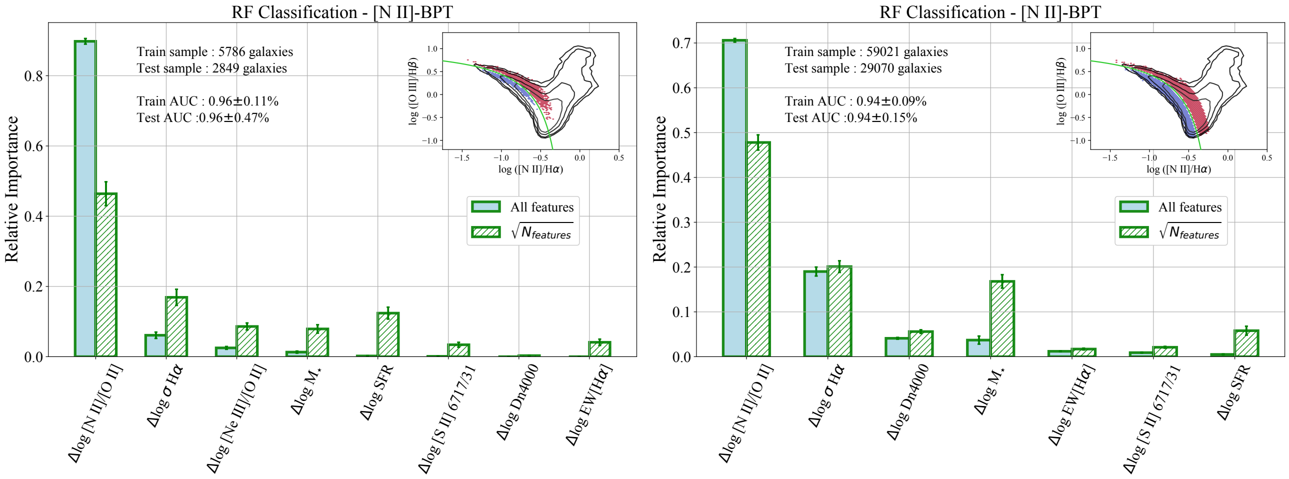

Figure A1. This figure replicates Fig. 5, but with a difference choice of the involved parameters. In particular, [N II]/[O II] is adopted instead of [N II]/[S II] as a tracer of the N/O abundance: in the *left panel*, [Ne III]/[O II] is also included to independently trace the ionisation parameter, causing a strong reduction of the number of selected galaxies, whereas in the *right panel* no tracer of U is adopted at all, and the full star-forming sample is hence considered. In both cases, the RF picks Δlog([N II]/[O II]) (hence, again, deviations in N/O) as the most relevant feature in the classification task, confirming the results presented in the main body of the paper. For sake of brevity, we do not show here the regression analysis, which nonetheless leads to equivalent conclusions.

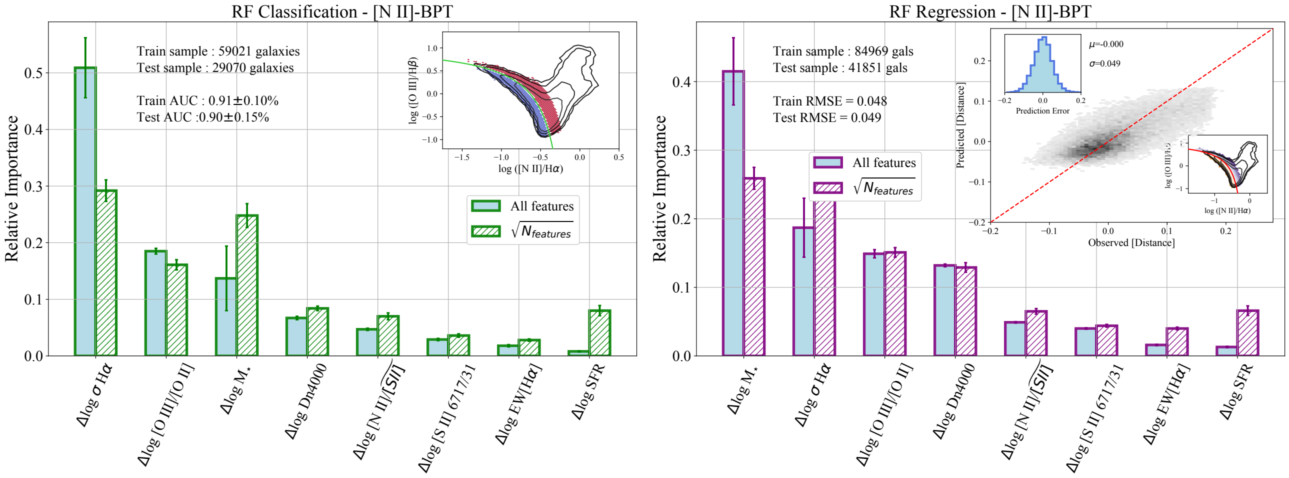

Figure A2. RF analysis of the [N II]-BPT diagram, where the [N II]/[S II] ratio has been replaced by a 'pseudo-hybrid' parameter (i.e., [N II]/$\widetilde{[S II]}$) obtained by randomly shuffling the [S II]λ6717, 31 fluxes among the full star-forming galaxy sample, while keeping the fluxes of the [N II]λ6584 line fixed. In this way, we can test to what extent the connection between our target label and the Δlog([N II]/[S II]) parameter, as recovered by the ML algorithms, is trivially induced by the presence of the [N II] emission line. The results, which see the relative importance of [N II]/$\widetilde{[S II]}$ strongly suppressed compared to both our fiducial analysis and the rest of the features in the set, confirms that the information resides in the full [N II]/[S II] ratio (hence in the N/O abundance), rather than just being driven by a trivial correlation between the nitrogen line fluxes. An equivalent conclusion can be drawn if we perform the same test on the [N II]/[O II] ratio.

each constituted by Δlog([N II]/[S II])and one of the other variables in the set, alternatively. For instance, from the comparison of partial correlation coefficients, we observe that the strength of correlation between Δlog([N II]/[S II])and **D** is only mildly reduced, at a fixed Δlog($M^{\star}$) or Δlog(σ)). However, at a fixed Δlog([N II]/[S II]), the correlations between Δlog($M^{\star}$), Δlog(σ)) (which were the second and third ranked parameters in both RF and in terms of global correlation coefficients) and **D** are strongly affected and reduced in magnitude. Therefore, fixing the (variations in) N/O abundance almost completely removes the correlations of **D** with both Δlog($M^{\star}$) and Δlog(σ)). Moreover, for some parameters like Δlog(SFR), the direction of the correlation is even inverted.

The same relationships between pairs of (partial) correlation coefficients are shown in the lower panel of Fig B2 for the set of parameters adopted in the analysis of the [S II]-BPT; here, Δlog(SFR)is taken as the reference parameter to which all other variables should be compared to. Again, Δlog(SFR)shows both the highest correlation coefficient and partial correlation coefficient with **D** than any other parameter, whose partial correlation ranks with our target label are on the contrary strongly suppressed when evaluated at fixed Δlog(SFR).

In summary, the analysis based on (partial) correlations rank establish Δlog([N II]/[S II])(hence deviations in the N/O abundance) as the most intrinsically connected parameter with the distance **D** from the SF sequence in the [N II]-BPTdiagram, and Δlog(SFR)as the most connected parameter with **D** in the [S II]-BPT, in excellent agreement with the machine learning analysis presented in the main body of the paper.

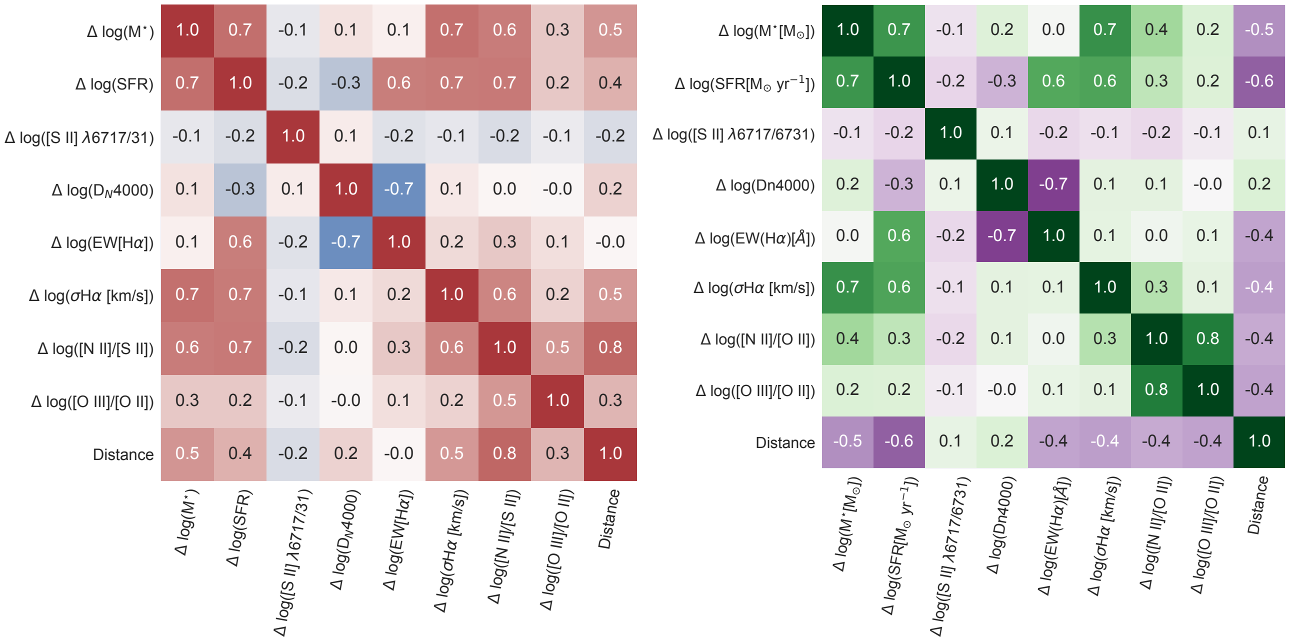

Figure B1. Matrix of Spearman correlation coefficients for the 'multi-parameter' set of the [N II]-BPT (left panel) and [S II]-BPT (right panel), respectively. The target label for the ML regression problem, i.e, the distance **D** from the SF sequence, is also included.

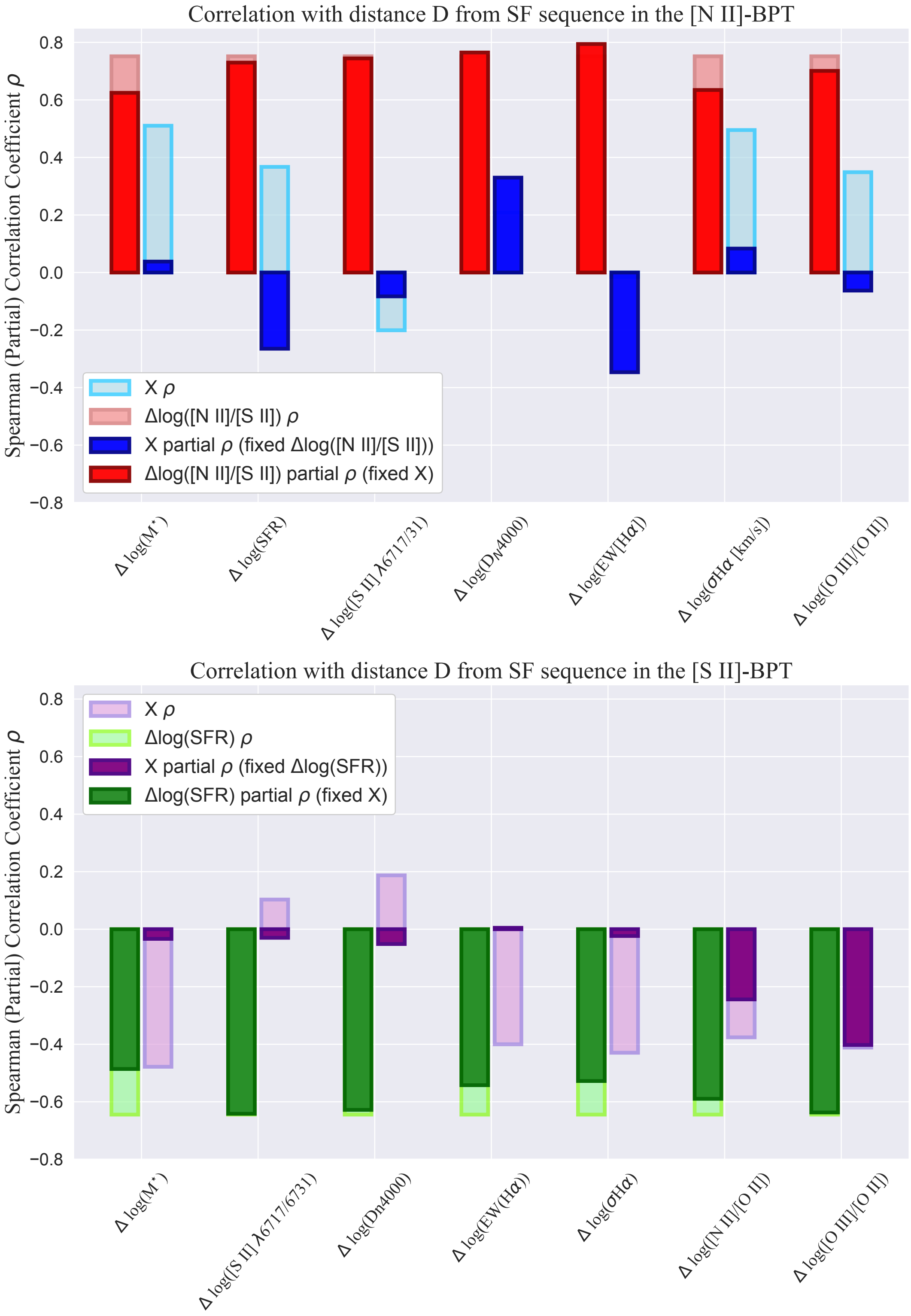

Figure B2. *Upper panel*: Spearman (partial) correlation coefficients for our set of parameters with the distance **D** from the median SF sequence in the [N II]-BPT diagram. The x-axis labels each parameter under consideration, and the bars are grouped into pairs for comparison with Δlog([N II]/[S II]), taken as reference because identified as the most relevant parameter in the ML analysis of Section 4 and in the matrix of Spearman rank coefficients of Fig. B1. Light shaded bars indicate the global Spearman rank correlation strength of each parameter with **D**, whereas solid coloured bars instead indicate the partial correlation strengths. For blue bars, the partial correlation is computed by keeping fixed the value of Δlog([N II]/[S II]), whereas the red bars report the partial correlation strength of Δlog([N II]/[S II]) at a fixed value of each of the other parameters, in turn. These results are fully consistent with our main ML analysis in showing that Δlog([N II]/[S II]) presents both the highest correlation and partial correlation coefficients with **D** than any other parameter.
Lower panel: Same as *upper panel*, for the [S II]-BPT. Here, the reference parameter is Δlog(SFR), which again shows both the highest correlation and partial correlation coefficients with **D** than any other parameter.