Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **527,** 5484–5502 (2024) https://doi.org/10.1093/mnras/stad3464
Advance Access publication 2023 November 14

Optimal metallicity diagnostics for MUSE observations of low-z galaxies

Bethan Easeman [1]★ Patricia Schady [1] Stijn Wuyts[1] and Robert M. Yates [2,3]

[1]*Department of Physics, University of Bath, Bath BA2 7AY, UK*
[2]*Centre for Astrophysics Research, University of Hertfordshire, Hatfield AL10 9AB, UK*
[3]*Astrophysics Research Group, University of Surrey, Stag Hill, Guildford GU2 7XH, UK*

Accepted 2023 November 2. Received 2023 August 31; in original form 2023 February 15

ABSTRACT

The relatively red wavelength range (4800–9300 Å) of the VLT Multi Unit Spectroscopic Explorer (MUSE) limits which metallicity diagnostics can be used; in particular excluding those requiring the [O II]λλ3726,29 doublet. We assess various strong line diagnostics by comparing to sulphur T_e-based metallicity measurements for a sample of 671 H II regions from 36 nearby galaxies from the MUSE Atlas of Disks (MAD) survey. We find that the O3N2 and N2 diagnostics return a narrower range of metallicities that lie up to ∼0.3 dex below T_e-based measurements, with a clear dependence on both metallicity and ionization parameter. The N2S2Hα diagnostic shows a near-linear relation with the T_e-based metallicities, although with a systematic downward offset of ∼0.2 dex, but no clear dependence on ionization parameter. These results imply that the N2S2Hα diagnostic produces the most reliable results when studying the distribution of metals within galaxies with MUSE. On sub-H II region scales, the O3N2 and N2 diagnostics measure metallicity decreasing towards the centres of HII regions, contrary to expectations. The S-calibration and N2S2Hα diagnostics show no evidence of this, and show a positive relationship between ionization parameter and metallicity at $12 + \log(O/H) > 8.4$, implying the relationship between ionization parameter and metallicity differs on local and global scales. We also present HIIDENTIFY, a PYTHON tool developed to identify H II regions within galaxies from Hα emission maps. All segmentation maps and measured emission line strengths for the 4408 H II regions identified within the MAD sample are available to download.

Key words: ISM: abundances – H II regions – galaxies: abundances.

1 INTRODUCTION

Gas-phase metallicity is a key indicator of how star formation has progressed through a galaxy's history, as metals are formed and later expelled into the interstellar medium (ISM) by stars over cosmic time. A principal pursuit in galaxy evolution studies is therefore to trace the build up of heavy elements through cosmic time, and within galaxies.

The oxygen abundance, $12 + \log(O/H)$, that is typically used as a proxy for the gas-phase metallicity of the ISM, is measured via diagnostics which rely on the relative strengths of emission lines (see e.g. Maiolino & Mannucci 2019). Metal recombination lines provide the most direct measure of metallicity, and have the advantage that their line strength is weakly dependent on gas properties such as temperature and density (e.g. Peimbert, Peimbert & Delgado-Inglada 2017). However, the lines are very weak, being around 10^3–10^4 times fainter than the Balmer lines (Maiolino & Mannucci 2019), and can therefore only be detected in the very nearby Universe. Alternatively, the slightly stronger auroral lines also provide a relatively direct tracer of metallicity, whereby the relative strength of auroral to nebular lines originating from the same species provide a measure of the electron temperature (T_e) of the gas. Due to the increased cooling effect provided by the metal lines in more metal-rich gas, the metallicity and electron temperature are linked, so from this measure of T_e, it is possible to fairly accurately determine the metallicity of the gas (e.g. Izotov et al. 2006; Peimbert, Peimbert & Delgado-Inglada 2017; Kewley, Nicholls & Sutherland 2019; Yates et al. 2020). Nevertheless, while stronger than the metal recombination lines, auroral lines are still 10–100 times fainter than hydrogen Balmer lines, becoming increasingly faint in more metal rich systems. This means they are also only detectable in a limited number of systems. While auroral line diagnostics can be considered a direct measurement of the physical conditions within the gas, it must be noted that they rely on a number of assumptions and simplifications (e.g. Pérez-Montero 2017; Cameron et al. 2020; Yates et al. 2020). For example, the gas temperature is assumed to remain constant within a series of concentric shells, rather than taking into account variations on smaller scales (Bresolin 2006; Osterbrock & Ferland 2006), which may lead to diagnostics underestimating the metallicity.

Due to the faintness of these lines, strong line diagnostics have been developed, offering an essential tool to explore the gas-phase metallicity in galaxies too metal-rich or too distant for the recombination and auroral lines to be detected. These diagnostics are developed either by finding empirical relations between a combination of strong line ratios and T_e-based metallicity in H II regions or galaxies, or equivalently, between metallicities and strong line ratios predicted with photoionization models.

★ E-mail: be329@bath.ac.uk

These strong line diagnostics, first developed by Alloin et al. (1979) and Pagel et al. (1979), frequently rely on the [O II] and [O III] nebular lines, such as the R23 (log(([O II] λ3727 + [O III] λλ4959,5007)/H β)) diagnostic, which uses the two principal oxygen states to account for the ionization structure in the H II region. This diagnostic, however, has a large dependence on the ionization parameter, as well as being double-branched, requiring a second, less sensitive metallicity diagnostic to determine which branch applies (Kewley & Dopita 2002; Kobulnicky & Kewley 2004; Maiolino & Mannucci 2019). The N2O2 diagnostic (log([N II] λ6584/[O II] λ3727)) has very little dependence on the ionization parameter, but primarily traces N/O, and is therefore sensitive to the assumed relation between N/O and O/H (Maiolino & Mannucci 2019). The O3N2 (log(([O III] λ5007/H β)/([N II] λ6584/H α))) diagnostic is popular due to all relevant lines being generally accessible in a single grating setting, and its small dependence on dust reddening due to the proximity of the lines in the ratios. However, O3N2 is primarily a tracer of the ionization parameter, log (U), thus requiring an understanding of how metallicity and ionization parameter are related, which can vary on spatial scales and with redshift. An alternative to the O3N2 diagnostic that is similarly insensitive to dust reddening, but apparently additionally independent of log (U) and gas pressure, is the Dopita et al. (2016) N2S2Hα diagnostic, which the authors thus claim is a useful diagnostic to use on high-redshift galaxies.

Thus, while strong line diagnostics are an important tool, it is important to remain mindful of their associated shortfalls, largely related to their dependence on properties other than metallicity, such as ionization parameter, ISM pressure, and electron density (Kewley & Dopita 2002; Dopita et al. 2016). These limitations are manifested in the large discrepancies in metallicity that can be observed between different diagnostics (e.g. Kewley & Ellison 2008) and are not necessarily systematic. Such relative discrepancies can be seen, for example, in the shape of observed radial metallicity profiles (e.g. Belfiore et al. 2017, 2019; Schaefer et al. 2019; Boardman et al. 2020; Mingozzi et al. 2020; Poetrodjojo et al. 2021; Yates et al. 2021). These discrepancies are not yet well understood (Kewley & Dopita 2002; Stasinska 2019).

To this end, in Easeman et al. (2022), we investigated the radial metallicity profiles of galaxies, and found large differences in the prevalence of certain features such as central dips in the radial profiles when different diagnostics were used. The prevalence of these dips in metallicity profiles measured using the O3N2 diagnostic also appeared linked to global properties of the galaxy such as stellar mass and star formation rate, whereas links to global properties were much weaker when the Dopita et al. (2016) N2S2Hα diagnostic was used.

Possible discrepancies between different strong line diagnostics may arise from biases present in the calibration samples used when deriving these diagnostics (e.g. Curti et al. 2017; Kewley, Nicholls & Sutherland 2019; Stasinska 2019). Often these biases are unavoidable, for example in the need for auroral line detections in empirically derived diagnostics, which are primarily detected in low metallicity, high excitation gas (Hoyos & Díaz 2006). Using diagnostics on observations of different spatial scales to the calibration sample can also be problematic. Yates et al. (2020) found that diagnostics calibrated on H II region scales agreed well with T_e-based measurements for observations on similar scales, but less well for observations on global scales. Similarly, diagnostics calibrated on galaxy-integrated observations appeared less reliable for observations on H II region scales.

With higher resolution observations using integral-field unit (IFU) spectrographs such as the Multi Unit Spectrographic Explorer (MUSE; Bacon et al. 2010) and, more recently, the NIRSpec integral field spectrograph on the *JWST* (Gardner et al. 2006), the variation in metallicity within galaxies can be studied on much smaller scales than was previously possible. The combination of high spatial resolution and large field of view has made MUSE an especially powerful instrument for studying the ISM conditions within SF regions, and variations across a galaxy (e.g. Emsellem et al. 2022). However, a drawback of MUSE is its relatively short wavelength range (4650–9300 Å; Bacon et al. 2010), which means that certain key emission lines, such as [O II] λ3727,29, are not visible for low-z galaxies, limiting the metallicity diagnostics that can be applied.

In this paper, we therefore investigate the reliability of a number of strong line diagnostics when applied to MUSE data, using an empirical approach. In Section 2, we detail the observations used, and present the steps taken in our analysis in Section 3. The metallicity diagnostics used are described in Section 4, and our results in Section 5, with a discussion in Section 6. Finally, our conclusions are presented in Section 7. The appendices contain further information about the sample, as well as flux measurements for the relevant emission lines from 4408 H II regions we identify within our sample of 36 galaxies using HIIDENTIFY, our newly developed PYTHON tool.

2 DATA

For our analysis, we use MUSE data taken as part of the MUSE Atlas of Disks (MAD)[1] survey, which covers 38 galaxies on the SF main sequence, selected as a sample of nearby ($z < 0.013$), relatively face-on (inclination $<70°$) galaxies, with a range of masses from $10^{8.5}$ to $10^{11.2}$ $M_\odot$ (Erroz-Ferrer et al. 2019).

The typical spatial resolution is ∼100 pc (Erroz-Ferrer et al. 2019), allowing for the study of individual H II regions. This sample therefore provides a large number of individual H II regions that can be identified and used in our analysis. The range of masses allows us to probe a range of metallicities, as both global and local gas-phase metallicity have been shown to correlate with stellar mass (Tremonti et al. 2004; Sánchez et al. 2013). We give details on the MAD galaxy sample in Table B1 of the appendix.

3 ANALYSIS

The IFU data returned by MUSE are 3D cubes, with two spatial dimensions, and one spectral dimension, meaning each pixel of the image has an associated spectrum. A number of data products have been made available from the MAD survey,[2] including 2D maps of dust-corrected emission line fluxes for all strong lines within the observed wavelength range. However, for our analysis we require flux maps for additional weak, auroral emission lines, as well as associated line flux uncertainties, which are not readily available. We therefore produce our own emission line maps from the reduced MUSE data cubes, which we download from the ESO archive science portal.[3]

In order to measure accurate emission line fluxes, we first need to separate the stellar and gas emission components in order to correct for Balmer absorption from old stellar populations. A failure to correct for such stellar absorption features can result in the Balmer

[1] https://www.mad.astro.ethz.ch
[2] https://www.mad.astro.ethz.ch/data-products
[3] http://archive.eso.org/scienceportal/home

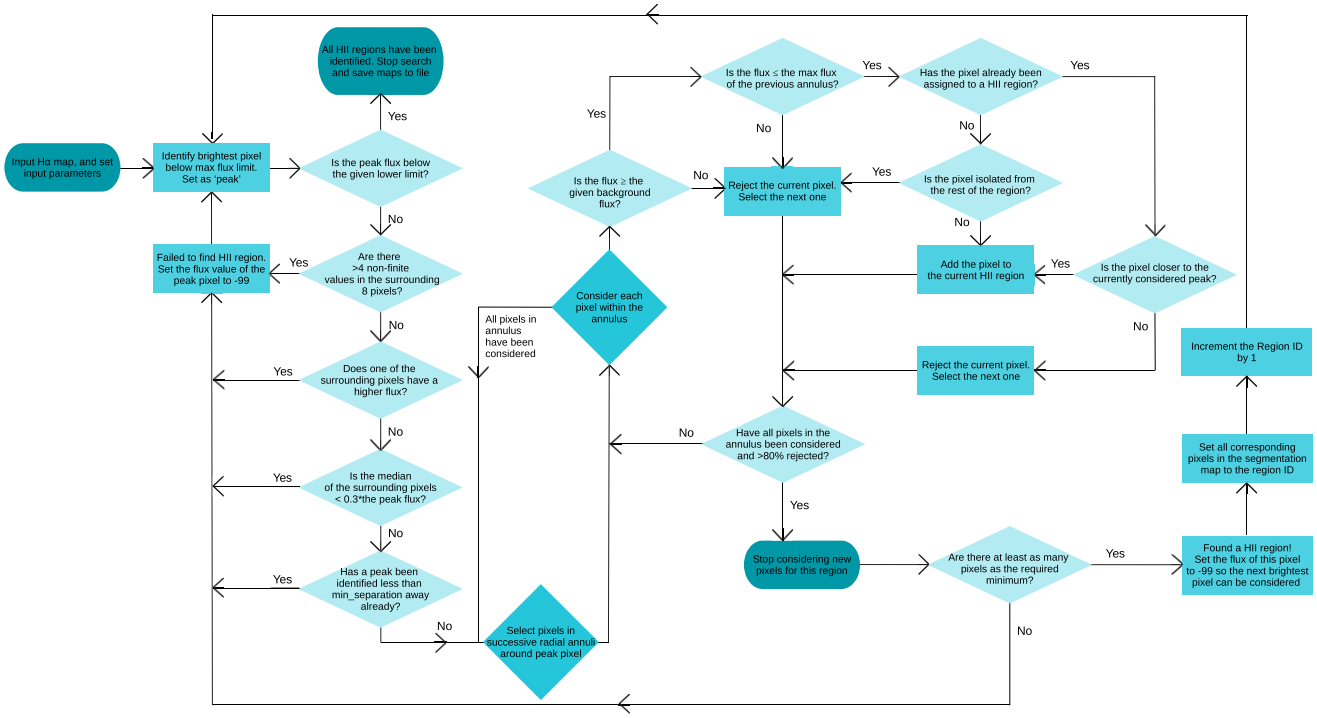

Figure 1. Flowchart illustrating the steps taken by HIIDENTIFY to identify pixels as being peaks of H II regions, removing any spaxels determined to be noise within the image, and to then assign pixels to the regions, returning segmentation maps as shown in Fig. 2.

line fluxes being underestimated, increasingly so for bluer Balmer lines, thus affecting the measured Balmer decrement that we need to produce galaxy dust reddening maps. We use the STARLIGHT software package to separate the stellar and gas emission components (Cid Fernandes et al. 2005, 2009), following a similar procedure as described in Krühler et al. (2017). In summary, we use STARLIGHT to fit a linear superposition of template spectra to each 2 × 2 binned MUSE spectrum, using the stellar population models from Bruzual & Charlot (2003). At the typical redshift of our galaxy sample ($z \sim 0.005$), the 2 × 2 spatial binning corresponds to a physical size of ∼40 pc, reaching up to 100 pc for the most distant galaxy in our sample (NGC3393). We then linearly scale the best-fitting stellar template to the intensity of each of the four spaxels in our bin, and subtract this weighted stellar component from the original data to produce a gas-phase only cube.

We removed NGC3521 from the sample, as it has very weak emission lines, and low S/N of the [O III] and auroral lines. NGC4593 was also removed, due to the large contribution of active galactic nucleus in the centre, leaving us with a final sample of 36 galaxies.

3.1 H II region identification with HIIDENTIFY

To identify H II regions within each galaxy, we use maps of the H α flux, masking out all spaxels with equivalent width of H αEW < 6 Å, which are associated with diffuse ionized gas (DIGs) rather than SF regions. DIGs have different physical properties to H II regions, with lower gas densities, and lower ionization parameters. The ionizing source for DIGs has not been conclusively determined, meaning the metallicity diagnostics that have been calibrated against observations and models of the gas within H II regions are not expected to remain valid when used on DIGs (e.g. Sanders et al. 2017; Zhang et al. 2017; Lacerda et al. 2018). Our choice of a threshold of 6 Å is a compromise between the recommendations of Sánchez et al. (2015), Belfiore et al. (2016), and Lacerda et al. (2018; see also Belfiore et al. 2017). However, our results are not very sensitive to the choice of the H αEW threshold that we use, and increasing this cut to 14 Å, as suggested by Lacerda et al. (2018) to identify purely SF regions, was found to have little effect on the H II regions identified using the process described below.

We developed a PYTHON tool, which we have named HIIDENTIFY, to automatically identify H II regions within a galaxy based on the H α emission, following a similar methodology to codes such as HIIPHOT (Thilker, Braun & Walterbos 2000) and PYHIIEXPLORER (Espinosa-Ponce et al. 2020). To identify H II regions, HIIDENTIFY iterates through the pixels from the brightest to the dimmest, terminating at a user-defined lower limit on the flux. Each of these pixels are considered as a possible peak of a H II region, using several criteria to exclude noise within the image, as shown in Fig. 1. For example, the peak is rejected if > 50 per cent of the surrounding pixels have non-finite flux values, or if the median flux of the surrounding pixels is < 30 per cent of that of the peak. Other criteria must also be met for the pixel to be confirmed as the peak of an H II region – all of the immediately surrounding pixels must have lower fluxes than the currently considered peak, and a minimum required separation between regions can be specified. For this analysis, we used a minimum separation of 50 pc.

Once a pixel is successfully identified as the peak of an H II region, surrounding pixels are considered in circular annuli, and are added to the region if the flux is greater than the user-defined background flux, and the pixel has not already been added to another region. If a pixel is selected to belong to multiple regions, it is assigned to the region with the closest peak. The radius of the regions is not constrained, and instead the growth of the region stops when > 80 per cent of the pixels in the annulus have been rejected.

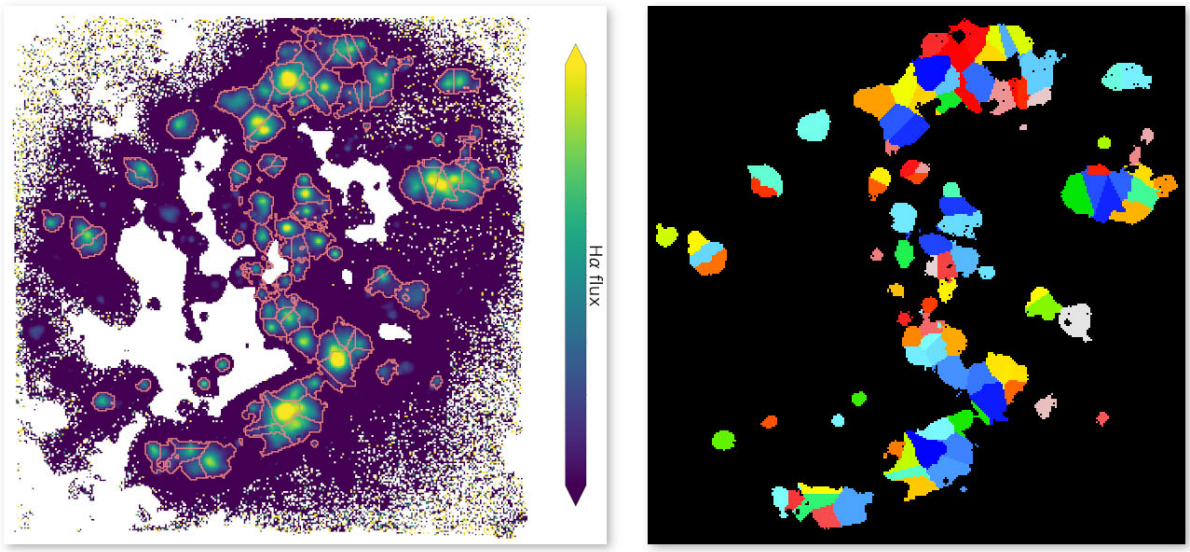

Figure 2. *Left*: H α flux map of NGC1483, with outlines of the regions identified with HIIDENTIFY shown in red. *Right*: Segmentation map, where the pixels within each region are set to the region ID number, shown here by different colours.

Finally, a minimum required number of pixels can be specified, which sets a lower limit on the number of pixels any identified H II region must have. Once all possible H II regions have been identified according to the criteria described above, our code creates a number of maps, including a segmentation map, which depicts which region each pixel belongs to, if any. HIIDENTIFY is publicly available for download, with information on how to install and use it provided in the online documentation.[4] An example MUSE H α map from the MAD sample can be seen in Fig. 2, with the outlines of HIIDENTIFY identified regions overlaid. The associated segmentation map is also shown, indicating the spaxels belonging to each of the identified H II regions. As there are no constraints on the shape or size of the identified regions, it can be seen that HIIDENTIFY identifies all regions with peak flux above a user-defined level, and encapsulates all of the surrounding region out to a given background level, rather than making assumptions about the geometry of the regions.

The results from applying HIIDENTIFY to our sample are shown in Table 1, and the returned segmentation maps have been made publicly available.[5] Input parameters were set so as to ensure that the entire H II region was encapsulated, with any low S/N spaxels removed at a later stage. The background flux was determined using spaxels with S/N $>$ 3, and H αEW$<$ 14 Å, selecting the 75th percentile of the flux to represent the background level at which to set the edge of the H II regions. Using the above input parameters, a total of 4408 H II regions were identified in our sample of 36 galaxies, and they generally had a radius of a few hundred parsecs, which is consistent with observed sizes of H II regions.

Following the release of the MUSE-PHANGS (Physics at High Angular resolution in Nearby GalaxieS; Emsellem et al. 2022) catalogue in Groves et al. (2023), which included segmentation maps from the HIIPHOT code, in Appendix D we compare the results from the two codes, finding very good agreement in the resulting metallicity measurements.

[4]https://hiidentify.readthedocs.io/en/latest/
[5]https://doi.org/10.6084/m9.figshare.22041263

Table 1. The stellar mass (M_*), star formation rate (SFR), and number of H II regions identified by HIIDENTIFY in the 36 MAD galaxies in our sample. We also list the number of H II regions in each galaxy that have [S III] λ6312 S/N $>$ 5 and nebular S/N $>$ 3, and thus were included in our final sample.

Galaxy	log(M_*) ($M_\odot$)	SFR ($M_\odot$ yr^{-1})	Num. H II regions	S/N sel. regions	
NGC4030	11.18	11.08	418	43	(10 per cent)
NGC3256	11.14	3.10	132	39	(30 per cent)
NGC4603	11.10	0.65	248	0	(0 per cent)
NGC3393	11.09	7.06	17	0	(0 per cent)
NGC1097	11.07	4.66	58	8	(14 per cent)
NGC289	11.00	3.58	133	14	(11 per cent)
IC2560	10.89	3.76	167	30	(18 per cent)
NGC5643	10.84	1.46	95	10	(11 per cent)
NGC3081	10.83	1.47	48	0	(0 per cent)
NGC4941	10.80	3.01	18	2	(11 per cent)
NGC5806	10.70	3.61	141	16	(11 per cent)
NGC3783	10.61	6.93	102	35	(34 per cent)
NGC5334	10.55	2.45	122	12	(10 per cent)
NGC7162	10.42	1.73	62	0	(0 per cent)
NGC1084	10.40	3.69	201	38	(19 per cent)
NGC1309	10.37	2.41	257	48	(19 per cent)
NGC5584	10.34	1.29	78	13	(17 per cent)
NGC4900	10.24	1.00	352	43	(12 per cent)
NGC7496	10.19	1.80	140	16	(11 per cent)
NGC7552	10.19	0.59	139	20	(14 per cent)
NGC1512	10.18	1.67	30	1	(3 per cent)
NGC7421	10.09	2.03	70	9	(13 per cent)
ESO498-G5	10.02	0.56	16	0	(0 per cent)
NGC1042	9.83	2.41	44	5	(11 per cent)
IC5273	9.82	0.83	94	17	(18 per cent)
NGC1483	9.81	0.43	97	34	(35 per cent)
NGC2835	9.80	0.38	100	14	(14 per cent)
PGC3853	9.78	0.35	47	6	(13 per cent)
NGC337	9.77	0.57	119	1	(1 per cent)
NGC4592	9.68	0.31	325	79	(24 per cent)
NGC4790	9.60	0.39	208	38	(18 per cent)
NGC3513	9.37	0.21	108	17	(16 per cent)
NGC2104	9.21	0.24	90	11	(12 per cent)
NGC4980	9.00	0.18	89	34	(38 per cent)
NGC4517A	8.50	0.10	19	10	(53 per cent)
ESO499-G37	8.47	0.14	24	8	(33 per cent)

3.2 Spectral line fits

To measure the line fluxes within each spaxel, we fitted the emission lines of interest with Gaussian functions using the SPECUTILS package in PYTHON.

The lines were grouped into chunks with nearby lines (e.g. the Hα and [N II] $\lambda\lambda$6549,84 lines) when fitting, so that the positions could be tied to that of the brightest line to help constrain the fits, and in the case of doublets, the widths could be tied and the ratio of the amplitudes fixed to known theoretical ratios when fitting. The continuum extending $\sim$50 Å either side of the lines was included, allowing the continuum level to be fitted when determining line fluxes. The [O I] $\lambda\lambda$6302,65 sky lines were masked out before fitting the [S III] λ6312 lines. The cube slice was dust-corrected using dust reddening maps that were produced from the measured Hα-to-Hβ Balmer decrement, and assuming a Cardelli, Clayton & Mathis (1989) attenuation law. We assumed an intrinsic Balmer decrement of 2.87 (Osterbrock & Ferland 2006), suitable for SF galaxies with electron density $\sim$100 cm^{-3} and temperature $\sim$10 000 K.

For the [S III] λ9531 line, required for our T_e-based measurements, we use the known theoretical flux ratio between [S III] λ9070 and [S III] λ9531 of 2.47 (Luridiana, Morisset & Shaw 2015), along with our measured flux for the [S III] λ9070 line.

4 METALLICITY AND IONIZATION PARAMETER DIAGNOSTICS

T_e-based methods rely on the detection of auroral lines, which are very faint compared to the strength of nebular lines – for example, the [O III] λ4363 line is around $\sim$100 times fainter than the [O III] λ5007 line (Maiolino & Mannucci 2019). Despite this difficulty, in regions where the auroral lines can be detected, T_e-based diagnostics give a more reliable measure of metallicity. We therefore focus our analysis on the subset of H II regions where we can derive a T_e-based metallicity using the [S III] λ6312 auroral line (see Section 4.1), and use this as our reference metallicity to which we compare the results from a number of strong line diagnostics. Of the full sample of 4408 H II regions, 671 had S/N $>$ 5 detections of [S III] λ6312, as well as S/N $>$ 3 in all nebular lines used in the metallicity diagnostics considered in this paper.

4.1 T_e-based metallicity diagnostics

As the spectral range of MUSE has a lower limit of 4650 Å, the [O III] λ4363 line required for calculating the electron temperature of the O^{++} gas ($T_{e,[O\,III]}$) is not visible for the redshift range of our sample, nor are the [O II] λ3727,29 nebular lines for calculating $T_{e,[O\,II]}$. We are therefore unable to use a T_e-based diagnostic based on the oxygen lines.

Sulphur has been proposed as a useful alternative tracer (e.g. Berg et al. 2015, 2020; Díaz & Zamora 2022), as both sulphur and oxygen are produced in massive stars, and the yield of the two elements are expected to be linked, although with a slight time difference in their ejection from different types of supernovae (Kobayashi, Karakas & Lugaro 2020). The required [S III] λ6312 auroral and [S II] λ6717,31 and [S III] $\lambda\lambda$9070,9531 nebular lines also experience less dust reddening due to the longer wavelengths of these lines.

In this paper, we adopted the recent sulphur-based method presented in Díaz & Zamora (2022). The $R_{S3} = I(9070\,\text{Å} + 9531\,\text{Å})/I(6312\,\text{Å})$ line ratio is used to determine $T_{e,[S\,III]}$, and due to the ionization structure of the H II region, where the overlap between S^{++} and S^{+} appears to cover most of the region (Garnett 1992; Díaz & Zamora 2022), it is assumed that $T_{e,[S\,II]} \approx T_{e,[S\,III]}$.

To account for the contribution of S^{3+} when the required [S IV] λ10540 is not visible, Díaz & Zamora (2022) provide a method relying on the argon lines. Neither the [S IV] nor argon lines are present within the wavelength range of our data, but the abundance of S^{3+} is not expected to be significant in H II regions within SF galaxies such as those present within our sample (Díaz & Zamora 2022).

Using the measure of $T_{e,[S\,III]}$, and the [S II] λ6317,31 and [S III] $\lambda\lambda$9070,9531 nebular lines, respective values of 12 + log(S^+/H) and 12 + log(S^{++}/H) are determined, and combined to give 12 + log(S/H). We then convert from the 12 + log(S/H) returned from this diagnostic, to 12 + log(O/H), using a fixed log(S/O) of $-$1.57 (Asplund et al. 2009). There have been suggestions of the S/O ratio being dependent on metallicity (e.g. Dors et al. 2016; Díaz & Zamora 2022), although other works have found no clear dependence (e.g. Berg et al. 2020). We therefore decided to use a fixed value of $-$1.57 in our analysis. A discussion on the implications of this choice is provided in Section 6.1.

4.2 Strong line metallicity diagnostics

The relatively high lower wavelength cut-off in the MUSE wavelength range restricts the number of diagnostics that we can use, so it was not possible to extend this analysis to commonly used diagnostics such as R23. However, this limitation makes it all the more important that an analysis is carried out to understand the robustness of metallicity diagnostics within the wavelength range and spatial scales covered by MUSE.

In this paper, we test the most commonly used strong line metallicity diagnostics that include emission lines available with MUSE for nearby galaxies. Since we are interested in leveraging the high spatial resolution of MUSE, we focus on those diagnostics that have been calibrated on samples of observed or modelled H II regions, rather than on galaxy-integrated spectra. These are the N2S2Hα diagnostic from Dopita et al. (2016), the N2 and O3N2 diagnostics based on the calibrations from Pettini & Pagel (2004) and Marino et al. (2013), and finally the recent S-calibration diagnostic from Pilyugin & Grebel (2016).

The Dopita et al. (2016) N2S2Hα diagnostic, given in equation (1), was derived using the MAPPINGS 5.0 photoionization model code, using line ratios that are accessible from the ground out to high-redshift within a single configuration, and which have a low dependence on dust attenuation due to the proximity of the lines.

$$\begin{aligned} 12 + \log(\text{O/H}) = {} & 8.77 + \log([\text{N II}]\lambda6584/[\text{S II}]\lambda6717,31) \\ & + 0.264 \log([\text{N II}]\lambda6584/\text{H}\alpha) \end{aligned} \quad (1)$$

The O3N2 and N2 diagnostics are more widely used and there are a number of calibrations available. In this paper, we include the Pettini & Pagel (2004) (PP04) calibration of O3N2 (equation 2) and N2 (equation 3), which were calibrated on a sample of 137 nearby extragalactic H II regions with metallicity measured using predominantly T_e-based methods, but also using detailed photoionization models at the high metallicity end.

$$12 + \log(\text{O/H}) = 8.73 - 0.32 \times \text{O3N2} \quad (2)$$

$$12 + \log(\text{O/H}) = 8.90 + 0.57 \times \text{N2} \quad (3)$$

We also consider the more recent O3N2 and N2 re-calibrations from Marino et al. (2013) (M13), shown in equations (4) and (5),

which used a larger sample of 603 H II regions from the literature, along with 16 H II complexes from the Calar Alto Legacy Integral Field Area Survey (CALIFA; Sánchez et al. 2012) with available T_e-based metallicities.

$$12 + \log(\mathrm{O/H}) = 8.533 - 0.214 \times \mathrm{O3N2} \quad (4)$$

$$12 + \log(\mathrm{O/H}) = 8.743 + 0.462 \times \mathrm{N2} \quad (5)$$

Finally, we also use the S-calibration from Pilyugin & Grebel (2016), which uses a sample of 313 H II regions with spectra from single-slit observations to derive their relationships, with measured T_e-based metallicities spanning $7.0 < 12 + \log(\mathrm{O/H}) < 8.8$. The S-calibration is derived for use when the [O II] λλ3727,29 lines are unavailable, as is the case for our MUSE data, and instead uses the [S II] λλ6717,31 lines. They find very good agreement (within ∼0.05 dex) between their S-calibration and the R-calibration, which uses the [O II] lines for single-slit spectra from a test sample of over 3000 H II regions. Of note is that Pilyugin et al. (2022) find that when comparing spectra of H II regions from IFU to single-slit observations, the S-calibration underestimates metallicities by ∼0.06 dex on average at $12 + \log(\mathrm{O/H}) \gtrsim 8.55$, and by ∼0.02 dex on average at lower metallicities. We discuss the potential impact of this on our results in Section 6. Both the S- and R-calibration are double-branched, and Pilyugin & Grebel (2016) thus recommend using the line ratio $\log(N_2)$ (where N_2 = [N II] λλ6548,84/H β) to separate the upper and lower branch. For the upper branch ($\log(N_2) \geq -0.6$), the relationship in equation (6) is used; for the lower branch, the relationship in equation (7). Here S_2 = [S II] λλ 6717,31/H β, and R_3 = [O III] λλ4959,5007/H β.

$$12 + \log(\mathrm{O/H}) = 8.424 + 0.030\log(R_3/S_2) + 0.751\log(N_2) + (-0.349 + 0.182\log(R_3/S_2) + 0.508\log(N_2)) \times \log(S_2) \quad (6)$$

$$12 + \log(\mathrm{O/H}) = 8.072 + 0.789\log(R_3/S_2) + 0.726\log(N_2) + (1.069 - 0.170\log(R_3/S_2) + 0.022\log(N_2)) \times \log(S_2) \quad (7)$$

4.3 Ionization parameter diagnostic

The O3N2 and N2 diagnostics are strongly dependent on the ionization parameter ($\log(U)$), a measure of the central ionizing source's ability to ionize the surrounding gas. Some works have suggested that these diagnostics may become unreliable under certain $\log(U)$ conditions if, for example, these conditions are not represented in the calibration samples (e.g. Krühler et al. 2017; Mao, Lin & Kong 2018). Therefore, we explore whether there is any evidence of this in our data.

To measure $\log(U)$, we used the Mingozzi et al. (2020) sulphur-based diagnostic, calibrated on spatially resolved IFU data. Sulphur-based ionization parameter diagnostics have been found to have a lower dependence on the metallicity compared to oxygen-based diagnostics (Kewley & Dopita 2002; Dors et al. 2011), making the sulphur-based diagnostic in principle more reliable, but also suggesting that any observed trends between metallicity and $\log(U)$ should not be a consequence of the $\log(U)$ diagnostic itself.

This diagnostic relies on the ratio of the [S III] λλ9070,9531 and [S II] λλ6717,31 line doublets. The Mingozzi et al. (2020) re-calibration accounts for an overestimate of the strength of the [S III] lines, suggested by Kewley & Dopita (2002) to be the cause of sulphur-based diagnostics underpredicting the ionization parameter compared to oxygen-based diagnostics.

5 RESULTS

To investigate the diagnostics on H II region scales, we identified H II regions using HIIDENTIFY, and stacked the spectra of all selected spaxels within each region. We then identified the subsample of regions with S/N of the measured [S III] λ6312 auroral line >5, to allow for a reliable measurement of the T_e-based metallicity. We imposed an additional criteria to ensure that the nebular lines needed in the strong line metallicity diagnostics that we used had an S/N > 3 within each H II stacked spectrum. This left us with a sample of 671 H II regions within a total of 31 galaxies, with logarithmic stellar masses ranging between $\log(M_\star/M_\odot) = 8.5$ and $\log(M_\star/M_\odot) = 11.2$, and SFRs in the range 0.1–11.1 $M_\odot\ yr^{-1}$. The galaxies NGC4603, NGC3393, NGC3081, NGC7162, and ESO498-G5 had no H II regions that met the S/N criteria (see Table 1).

5.1 Sulphur-based T_e metallicity diagnostics

As all of our empirically derived strong line metallicity diagnostics were calibrated against T_e-based diagnostics using the [O III] λ4363 line, we first investigate the agreement between T_e-based diagnostics that rely on the [O III] λ4363 and the [S III] λ6312 lines, before comparing our [S III] λ6312 T_e-based metallicities to the strong line diagnostics.

Since the [O III] λ4363 line is not visible in our MUSE data, to make these comparisons we used the published data of H II regions taken with the Multi-Object Double Spectrographs on the Large Binocular Telescope (MODS; Pogge et al. 2010) as part of the CHemical Abundances Of Spirals (CHAOS) project (Berg et al. 2015). These observations provide a broad wavelength coverage, extending from 3200 to 10 000 Å, enabling the detection of multiple auroral lines, including [O III] λ4363, out to $z \sim 0.36$. Thus far the multislit data for four SF galaxies observed as part of CHAOS have been published, amounting to published spectra for 190 H II regions (Berg et al. 2015, 2020; Croxall et al. 2015, 2016).

For this comparison, we apply the methodology described in Section 4.1 to the CHAOS data to measure the metallicities using the sulphur T_e-based diagnostic from Díaz & Zamora (2022), as well as applying the Izotov et al. (2006) [O III] λ4363-based method. In addition, we compare these sulphur- and oxygen-based T_e values to the metallicities presented in Berg et al. (2020), who use a multiple auroral line method to measure the temperature in a three-zone temperature H II model that incorporates low, intermediate, and high ionization zones. They probe the temperatures (and thus metallicities) of each of these zones using a combination of the [N II], [S III], and [O III] auroral lines. These metallicities are then combined to obtain the total metallicity of the region.

In order to compare the published Berg et al. (2020) HII region metallicities to the respective Izotov et al. (2006) and Díaz & Zamora (2022) oxygen- and sulphur T_e-based metallicities, we selected all H II regions from the CHAOS sample with [O III] λ4363 and [O II] λ7320,30 lines detected with S/N > 3 (as was done by Berg et al. 2020), and with [S III] λ6312 lines detected with S/N > 5. This more stringent criteria on the S/N of the [S III] λ6312 line was used to remain compatible with the selection criteria used in our MAD sample, although we find it makes no clear systematic difference to use a cut of [S III] λ6312 S/N > 3 instead. NGC5194 had no regions with detected [O III] λ4363, leaving a total of 43 regions within the remaining three CHAOS galaxies that meet our S/N criteria.

Fig. 3 (left panel) shows the comparison between the sulphur-based (Díaz & Zamora 2022) method and the oxygen-based method (Izotov

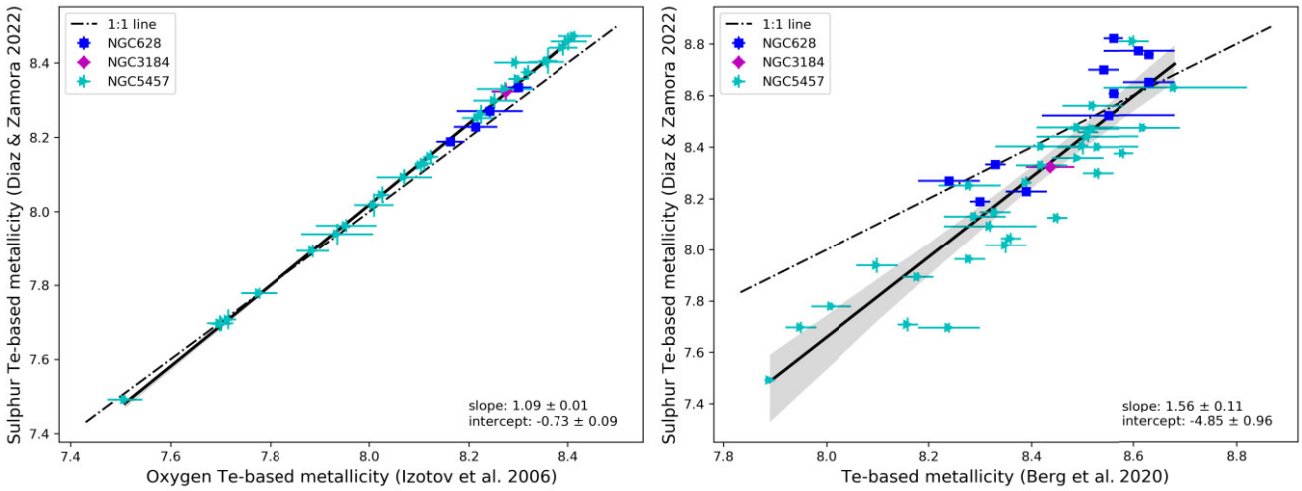

Figure 3. Comparison of the Díaz & Zamora (2022) sulphur-based T_e diagnostic, to other T_e-based diagnostics. In both panels, the dot–dashed black line shows the 1:1 relationship, and the solid black line the best fit to the data, with the shaded region representing the uncertainty on the fit. Comparison to the results from the Izotov et al. (2006) diagnostic shows a good agreement (*left*), but when comparing to the published Berg et al. (2020) values (*right*), a systematic offset at lower metallicities is seen. The slope and intercept of the best-fitting line are given in the bottom right of each panel. Note that not all H II regions within the CHAOS sample have the necessary oxygen auroral line detections required to apply the Izotov et al. (2006) diagnostic, hence the range in *y*-axis values differ between the two panels.

et al. 2006). We find very good agreement between these diagnostics, suggesting that the sulphur-based diagnostic is appropriate as a tracer of the oxygen abundance for verifying the accuracy of strong line diagnostics in our MAD sample. The line of best fit (black line in Fig. 3, left panel) and small scatter suggest a tight positive relationship between the two methods, with the sulphur-based diagnostic returning slightly higher metallicity values for $12 + \log(\mathrm{O/H}) \gtrsim 8.0$, leading to an offset of $\sim$0.02 dex at $12 + \log(\mathrm{O/H}) = 8.4$. For the CHAOS galaxies, Berg et al. (2020) compare T_e[S III] and T_e[O III], and find that at low temperatures (corresponding to high metallicity), T_e[S III] is shifted to lower values than T_e[O III]. This may explain the offset of the sulphur-based diagnostic to slightly higher values at high metallicities.

However, when comparing the [S III] λ6312 based metallicity to the multizone metallicities from Berg et al. (2020), the data no longer fall along the one-to-one line (dot–dashed line in right panel, Fig. 3). There is a systematic offset at low metallicities where the Berg et al. (2020) method returns higher metallicity values, with the sulphur-based measurements offset by -0.37 dex at $12 + \log(\mathrm{O/H}) = 8.0$. Nevertheless, although there is not a one-to-one agreement between the results of the two diagnostics, there is a clear relation, making it possible to convert between the results from the two diagnostics using the parameters of our best-fitting line, with slope $= 1.56 \pm 0.11$, and y-intercept $= -4.85 \pm 0.96$. The implications of applying such a re-scaling to our sulphur-based T_e-based metallicities are discussed in Section 5.2.

5.2 Comparing strong line and T_e-based diagnostics

Having found the Díaz & Zamora (2022) sulphur-based T_e method to agree well with the oxygen-based T_e method, we can now explore the reliability of applying strong-line diagnostics to our MUSE observations, by comparing the results from each strong-line diagnostic to that from the Díaz & Zamora (2022) T_e-based method. These comparisons are shown in Fig. 4, with the points coloured by $\log(U)$. The dot–dashed line shows the 1:1 relationship between the diagnostics, and the dotted line represents the best fit, obtained using the seaborn `regplot` method, with consistent fits obtained using a bootstrap method.

According to the best-fitting relationships, a number of the diagnostics show significant disagreement with the T_e-based measurements, particularly at high metallicity. The PP04 O3N2, PP04 N2, and Pilyugin & Grebel (2016) S-calibration all show some agreement with the T_e-based measurements at low metallicities, but show increasing discrepancies at higher metallicities. For these diagnostics, the best-fitting line to the data returns a slope of $\sim$0.55, leading to the strong line diagnostics returning results $\sim$0.3 dex lower than the T_e-based measurements at $12 + \log(\mathrm{O/H}) = 8.8$. One suggestion for the offsets observed with the PP04 diagnostics could be due to the relatively small calibration sample used. The auroral line becomes increasingly faint at higher metallicities, which results in this region being less well sampled. This is evident in the PP04 sample, where the low O3N2, high oxygen abundance region of the parameter space contains just six data points, four of which have metallicities based on strong line diagnostics. The authors themselves note the need to increase the sample size of their calibration sample at the high metallicity end.

The M13 O3N2 and N2 diagnostics show slightly reduced scatter, but the least agreement with the T_e-based measurements, with an offset of $\sim$0.4 dex below the T_e-based metallicities at $12 + \log(\mathrm{O/H}) = 8.8$. The N2S2H$\alpha$ diagnostic returns an average offset of $\sim$0.2 dex below the Díaz & Zamora (2022) T_e-based metallicities, but with a slope consistent with unity (see Table 2).

As expected, the highly $\log(U)$-sensitive O3N2 and N2 diagnostics show a clear variation of $\log(U)$ along the *y*-axis, with the lowest metallicity regions as measured by the strong line diagnostics having higher $\log(U)$. This leads to the diagnostics showing an increased discrepancy at higher $\log(U)$. Unfortunately, the line fluxes necessary to determine $\log(U)$ for the PP04 sample were not published in the original paper, and we therefore cannot verify this potential $\log(U)$-bias in the PP04 calibration sample.

Interestingly, considering the variation of $\log(U)$ along the *x*-axis, we see no clear trend between T_e-based metallicity and $\log(U)$, suggesting the O3N2 and N2 diagnostics may become unreliable

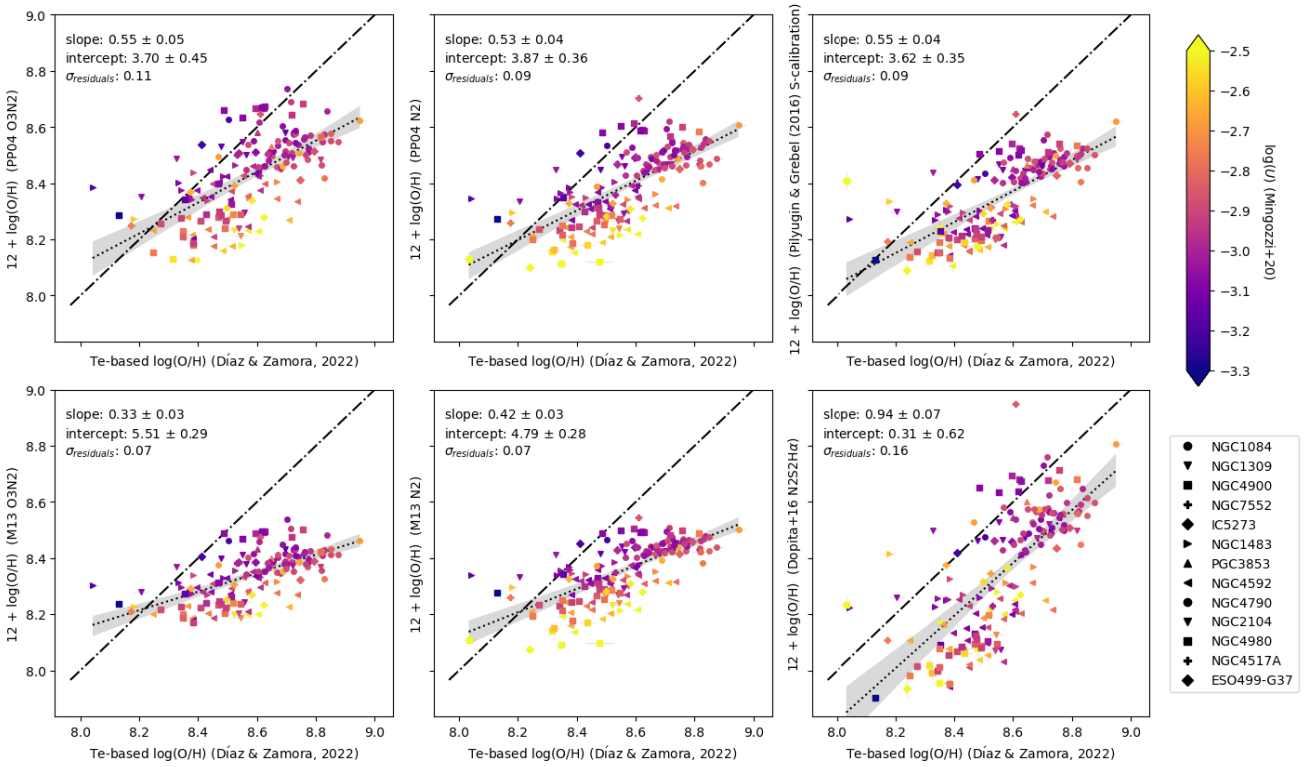

Figure 4. Comparison between strong line and sulphur T_e-based metallicity diagnostics, with the dot–dashed line showing a 1:1 relationship between the diagnostics, and a dotted line representing the best fit. The points are coloured by log (U), measured using the sulphur-based Mingozzi et al. (2020) diagnostic. The scatter about the best-fitting line is denoted by the standard deviation of the residuals ($\sigma_{\rm residuals}$).

Table 2. Parameters for the best-fitting lines shown in Figs 4 and 5, such that $12 + \log(\mathrm{O/H})_X = \text{slope} \times 12 + \log(\mathrm{O/H})_{Te} + \text{intercept}$, where X is the strong line diagnostic.

Strong line diagnostic	Díaz & Zamora (2022) T_e-based slope	intercept	Predicted Berg et al. (2020) T_e-based slope	intercept
PP04 O3N2	0.55 ± 0.05	3.70 ± 0.45	0.86 ± 0.08	1.03 ± 0.71
PP04 N2	0.53 ± 0.04	3.87 ± 0.36	0.82 ± 0.07	1.31 ± 0.56
S-calibration	0.55 ± 0.04	3.62 ± 0.35	0.86 ± 0.06	0.93 ± 0.55
M13 O3N2	0.33 ± 0.03	5.51 ± 0.29	0.52 ± 0.05	3.91 ± 0.46
M13 N2	0.42 ± 0.03	4.79 ± 0.28	0.65 ± 0.05	2.76 ± 0.44
N2S2Hα	0.94 ± 0.07	0.31 ± 0.62	1.47 ± 0.11	−4.24 ± 0.96

under certain log (U) conditions. The Pilyugin & Grebel (2016) diagnostic shows a weaker trend with log (U), and the N2S2Hα diagnostics shows no evidence of a relationship.

If the arguably more sophisticated multizone Berg et al. (2020) metallicity diagnostic is considered the more accurate of the T_e-based diagnostics investigated in this work, this would imply that our sulphur T_e-based metallicities are underestimated at low metallicities. Although this would not improve the discrepancies seen at high metallicity between our T_e-metallicities and the strong line diagnostics shown in Fig. 4, it could, in part, explain the sublinear relation observed in most cases. To explore this further, we use our best-fitting relationship between the sulphur T_e-based metallicities and the Berg et al. (2020) metallicities shown in Fig. 3 to convert our sulphur-based T_e-metallicities to the Berg et al. (2020) metallicity scale, and compare this to the strong line metallicities in Fig. 5. As expected, the PP04 O3N2 and N2, and the S-calibration metallicities show a more linear relation with the T_e-based values (best-fitting slopes increase from ∼0.5 to ∼0.8), but with a systematic offset to lower metallicities of ∼0.1 dex at $12 + \log(\mathrm{O/H}) = 8.2$, and ∼0.2 dex at $12 + \log(\mathrm{O/H}) = 8.8$. Although the relations with the M13 diagnostics also steepen, they still remain fairly flat, with a slope in the range 0.5–0.6 (see Table 2). The N2S2Hα diagnostic, however, shows a much lower level of agreement at low metallicity when converting to the predicted Berg et al. (2020) values, with a best-fitting slope of ∼1.5, and larger offsets of ∼0.4 dex at $12 + \log(\mathrm{O/H}) = 8.2$. The best-fitting parameters are summarized in Table 2.

Although we cannot conclusively state which of the T_e-metallicity diagnostics is the more accurate, it is noteworthy that the PP04 and M13 O3N2 and N2 diagnostics, and the Pilyugin & Grebel (2016) diagnostic were all calibrated against two-zone temperature models. We would therefore expect these diagnostics to be in better agreement with the Díaz & Zamora (2022) sulphur-based T_e-metallicity, which is in far better agreeement with the Izotov et al. (2006) oxygen two-zone temperature model (see Fig. 3) than the Berg et al. (2020) three-zone temperature model. Furthermore, the offsets at high metallicity

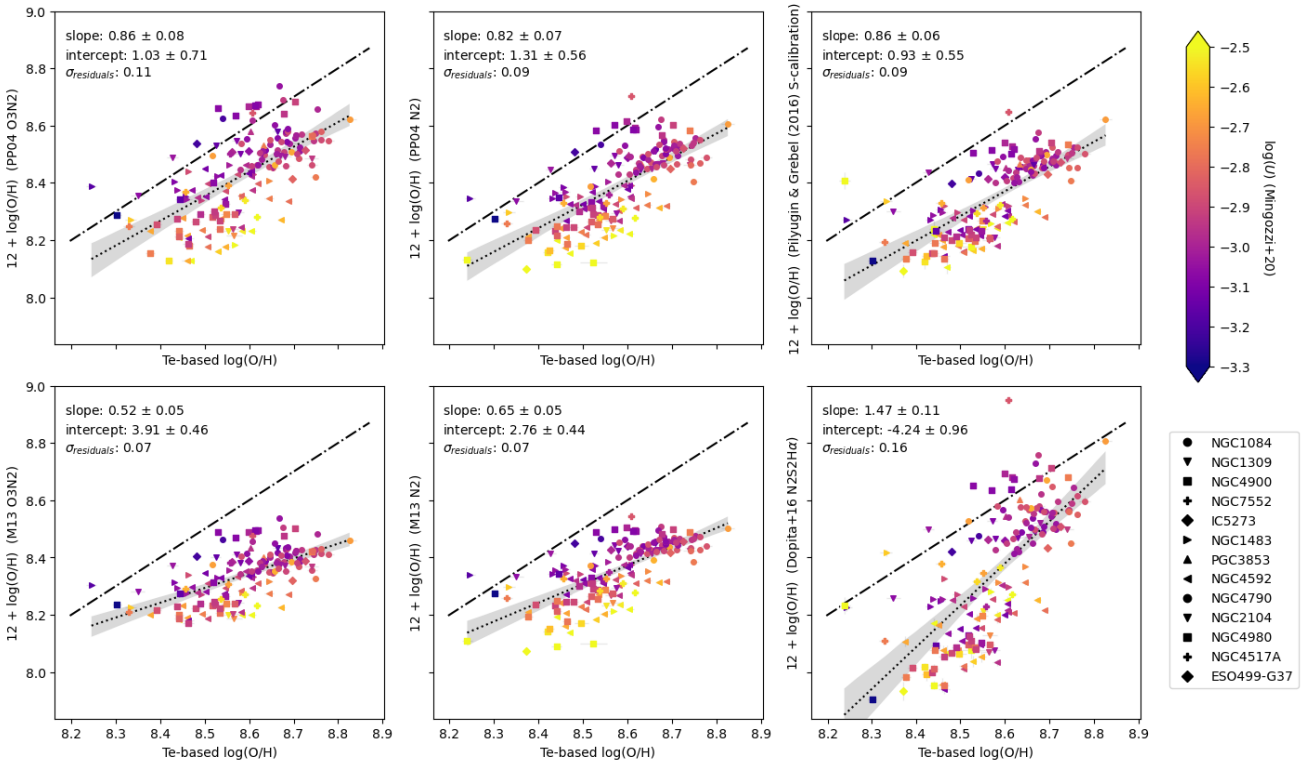

Figure 5. As for Fig. 4, but using the relationship between the Díaz & Zamora (2022) and Berg et al. (2020) T_e-based diagnostics shown in Fig. 3 to convert from the measurements made using the Díaz & Zamora (2022) diagnostic to the predicted value for if the Berg et al. (2020) diagnostic could be used.

apparent in Fig. 4 between T_e-metallicities and the N2, O3N2, and S-calibration can be explained to some degree by selection effects in the calibration samples. We discuss this further in Section 6.2.1.

5.3 Comparing to the Curti et al. O3N2 re-calibration

As the O3N2 diagnostics presented in Fig. 4 show significant discrepancies with the results from the T_e-based diagnostic, and given the potential selection effects in the older calibrations, we investigate the impact of using the more recent Curti et al. (2017) re-calibration of the O3N2 diagnostic. Although this re-calibration was based on stacked galaxy-integrated spectra, and it is preferable to use metallicity diagnostics that have been calibrated on data that probe similar spatial scales to the science data, this re-calibration included a very large, and arguably unbiased calibration sample. Using a consistent underlying calibration sample, Curti et al. (2017) re-calibrated six diagnostics, of which the R_3 ([O III] λ5007/H β), N_2 ([N II] λ6584/H α), and O3N2 all contain nebular emission lines probed by our MUSE data. However, we focus our analysis only on the more commonly used O3N2 diagnostic.

This re-calibration used [O III] λ4363 line fluxes obtained from whole-galaxy observations, with spectra stacked in bins of log([O II]/H β) and log([O III]/H β) to obtain sufficient S/N of the auroral line, in particular at higher metallicities. The sample of galaxies was selected from SDSS data release 7 (DR7; Abazajian et al. 2009) and had a median redshift of 0.072, leading to each fibre corresponding to observations on the scale of ∼3 kpc. Using the oxygen T_e-based metallicities, Curti et al. (2017) then modelled the relation between the strong line ratios and the T_e-based metallicities with a high order polynomial. In the case of the O3N2 diagnostic,

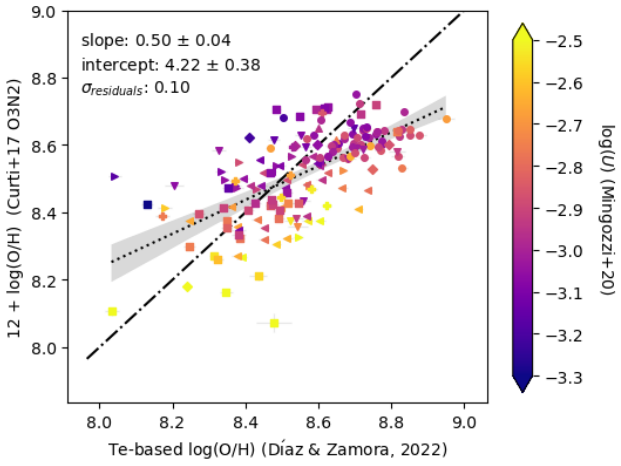

Figure 6. As for Fig. 4, showing the comparison of the Curti et al. (2017) O3N2 against the Díaz & Zamora (2022) T_e-based diagnostic.

their best-fitting polynomial was of the form

$$\mathrm{O3N2} = \sum_{n=0}^{2} c_n x^n, \tag{8}$$

where x is the oxygen abundance normalized to the Solar value ($12 + \log(\mathrm{O/H})_\odot = 8.69$; Asplund et al. 2009), and $c_0 = 0.281$, $c_1 = -4.765$, and $c_2 = -2.268$.

Fig. 6 shows that the average difference between the strong line and T_e-based diagnostics is reduced when the Curti et al. (2017) O3N2 diagnostic is used, compared to the PP04 or M13 calibrations,

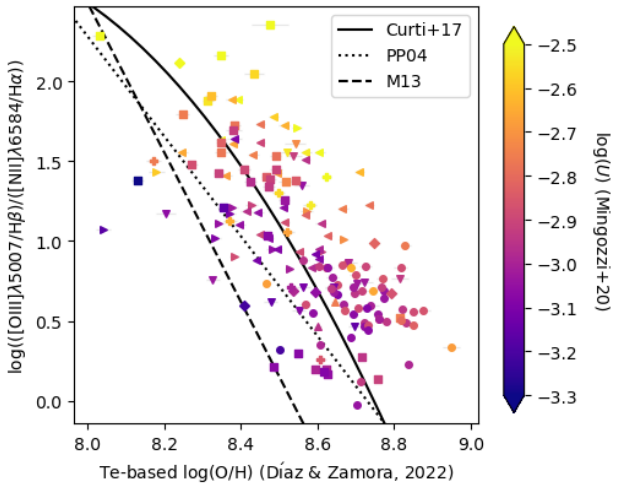

Figure 7. The O3N2 line ratio against sulphur T_e-based metallicity, colour coded by log (U). The best-fitting relations from PP04 (dotted), M13 (dashed), and Curti et al. (2017) (solid) are overplotted. The differences between the O3N2 diagnostics can be clearly seen, with the PP04 and M13 diagnostics shifted to lower metallicities than the Curti et al. (2017) re-calibration. The M13 diagnostic also clearly shows a steeper relationship, hence covering a much smaller range of metallicities.

because the Curti et al. (2017) O3N2 values are generally shifted to higher metallicities. However, the line of best fit still features a sub-unity slope of ∼0.50, similar to the PP04 O3N2, N2, and the S-calibration results. This suggests that the increased discrepancy between the strong line and T_e-based diagnostics at high metallicity cannot simply be explained by biases in the calibration sample.

The Curti et al. (2017) calibration sample, by using observations of galaxy-integrated spectra, averages over a large number of H II regions in each observation, thus averaging over a range of different ionization conditions. Despite this, there remains a clear log (U) dependence in the Curti et al. (2017) metallicities, again suggesting that the log (U) must be taken into account in order to reduce the scatter in O3N2 metallicities.

The relatively flat relation between T_e-based metallicities and O3N2 and N2 diagnostics implies that the strong line ratios used in these diagnostics are not sufficiently sensitive to metallicity, leading to a narrower range in metallicity estimates than T_e-based methods. This can be understood from the unaccounted log (U) dependence in the O3N2 and N2 diagnostics (Kewley & Dopita 2002), which are effectively flattened out in the calibrations. To illustrate this more clearly, in Fig. 7 we show the O3N2 line ratio against T_e-based metallicity for our sample of H II regions, colour-coded by log (U), and we overplot the best-fitting relation between these two parameters from PP04, M13, and Curti et al. (2017). From this figure it can be clearly seen how H II regions with comparable O3N2 line ratios can have very different metallicities, depending on their ionization parameter, as expected from Kewley & Dopita (2002).

5.4 Metallicity diagnostics on sub-H II region scales

The H II regions identified with our HIIDENTIFY code range in size from 22 pixels up to 1740 pixels, corresponding to 0.08–0.98 kpc. Therefore, in addition to investigating the properties of individual H II regions, the MUSE data used in this paper also allow us to explore the range of metallicity and log (U) values *within* H II regions. This may provide further insight into how representative H II-integrated properties are of the range in values within the H II region.

Using the same sample of H II regions with measured T_e-based metallicities as in our previous analysis, we selected the spaxels belonging to these H II regions, as determined using HIIDENTIFY. We repeated our previous analysis using these spaxels, and in Fig. 8 we compare the results from the strong line diagnostics to the Díaz & Zamora (2022) T_e-based diagnostic. For the strong line metallicity diagnostics, we applied an S/N > 3 cut on all relevant emission lines within each spaxel. For the T_e-based, we used an S/N > 3 cut on the nebular lines, and an S/N > 5 cut on the auroral line. We then further required the resulting metallicity to also have S/N > 3. We note that requiring a measure of the T_e-based metallicity means that only 3.5 per cent of the spaxels within these regions are selected, and the impacts of this are discussed further in Section 6.2.1.

Fig. 8 shows that strong-line metallicity estimates are also offset towards lower values compared to T_e-based estimates on sub-H II region scales, covering a similar region of the parameter space as the stacked H II region values.

Overall, these offsets are slightly larger on sub-H II region scales. For example, the best-fitting line to the PP04 O3N2 spaxel measurements suggests an offset of ∼0.35 dex below the T_e-based measurements at 12 + log(O/H) = 8.8, compared to ∼0.26 dex for the H II stack values. On spaxel scales, the M13 O3N2 diagnostic shows an even flatter distribution, with the O3N2 metallicity almost independent of the T_e-based metallicity. The offsets are increased to ∼0.43 dex below the T_e-based values at 12 + log(O/H) = 8.8, compared to ∼0.31 dex in the H II region analysis.

The N2S2Hα diagnostic again shows a close to linear relationship with the T_e-based method, unlike the other diagnostics, with an offset of ∼0.15 dex at 12 + log(O/H) = 8.0 increasing to ∼0.25 dex at 12 + log(O/H) = 8.8.

Despite the similarity in the parameter space occupied by single spaxels and stacked H II regions, of note in Fig. 8 is the almost flat distribution of data points for a given galaxy, especially for the O3N2, N2, and S-calibration diagnostics. A possible explanation for this very flat distribution may therefore be that it reflects large variations within galaxies in log (U) *and* metallicity, but with only a correspondingly small variation in the O3N2 and N2 line ratios. For example, for spaxels in IC5273 with measured T_e-based metallicities, log (U) covers a range of 0.85 dex and the T_e-based metallicities vary by ∼0.7 dex, whereas the line ratios of log(([O III]/Hβ)/([N II]/Hα)) cover only 0.4 dex corresponding to just 0.1 dex range in estimated metallicities from both the PP04 and M13 O3N2 diagnostics. When stacking the spaxels within H II regions, this range in properties is reduced, although there is still evidence of the dependence between O3N2 and N2 line ratios, log (U), and T_e-based metallicity (Fig. 7). This therefore implies that the location of an H II region in Figs 4 and 5 will be sensitive to its average log (U) value, and thus the distribution in log (U) within the H II region.

To investigate this further, for each strong line diagnostic considered in Fig. 4, we plot in Fig. 9 the difference between the stacked and spaxel metallicities (ΔZ), against the region's stacked metallicity, colouring each data point by the spaxel log (U).

Within each region there is a large spread of spaxel metallicities of up to 0.4 dex for the N2S2Hα diagnostic, reduced to mostly within 0.25 dex for the M13 O3N2 and N2 diagnostics. This range of values is in contrast to works such as Peimbert, Peimbert & Delgado-Inglada (2017), which suggest H II regions to be chemically homogeneous. There is additionally a strong anticorrelation between ΔZ and log (U) at fixed stacked metallicity for the O3N2 and N2 diagnostics.

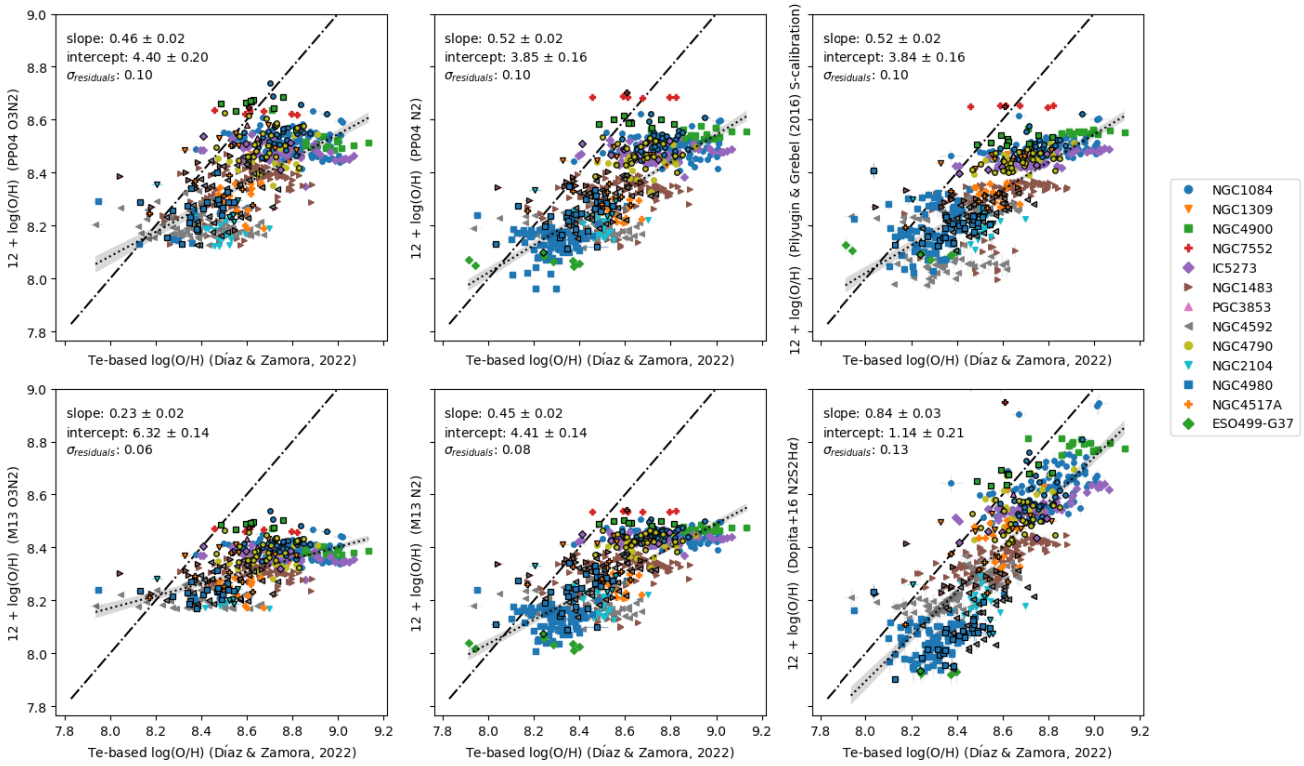

Figure 8. As for Fig. 4, now showing the metallicities of individual spaxels (no outlines), together with the metallicities from whole H II region stacked spectra overlaid with black outlines. The colour and shape of the markers is unique to each galaxy. The best-fitting line, with the uncertainty shown by the shaded region, is fitted only to the spaxel measurements.

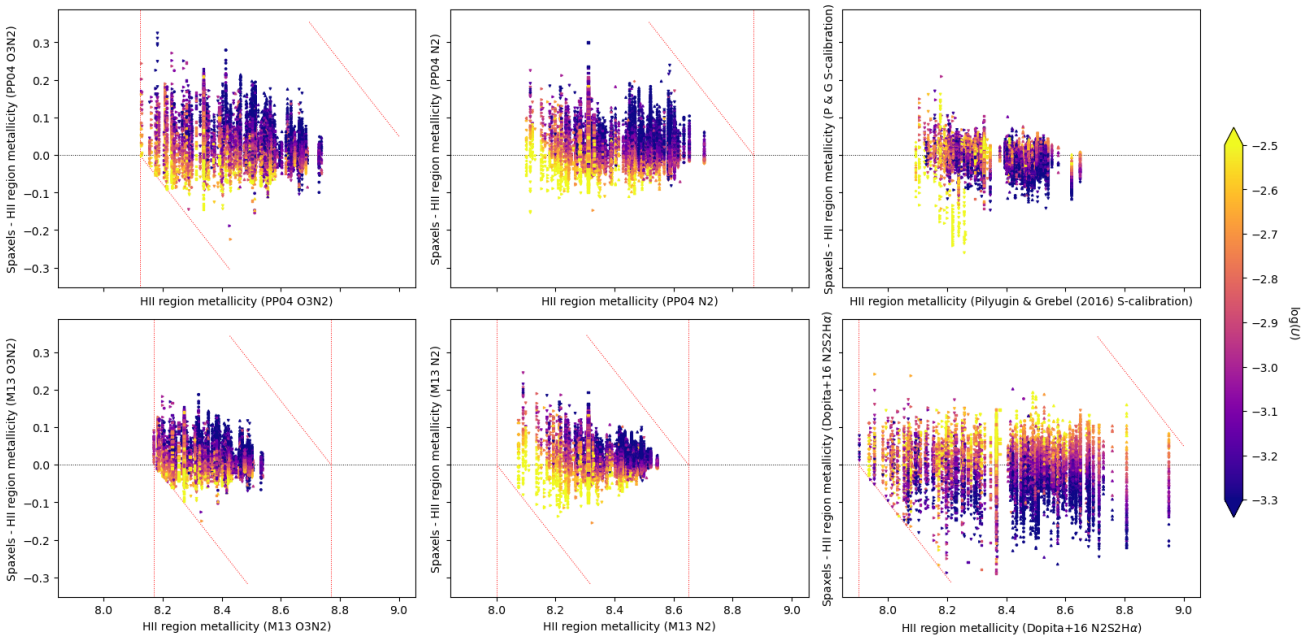

Figure 9. Difference between the metallicity of each spaxel within a given H II region and the metallicity obtained from the corresponding stacked H II spectrum for each H II region in our sample with a T_e-based metallicity. The points are coloured by log (U), as determined using the Mingozzi et al. (2020) diagnostic. Vertical red dotted lines denote the validity limits of the diagnostics, and diagonal red lines show the corresponding limits on the possible differences between the spaxel and region metallicity measurements.

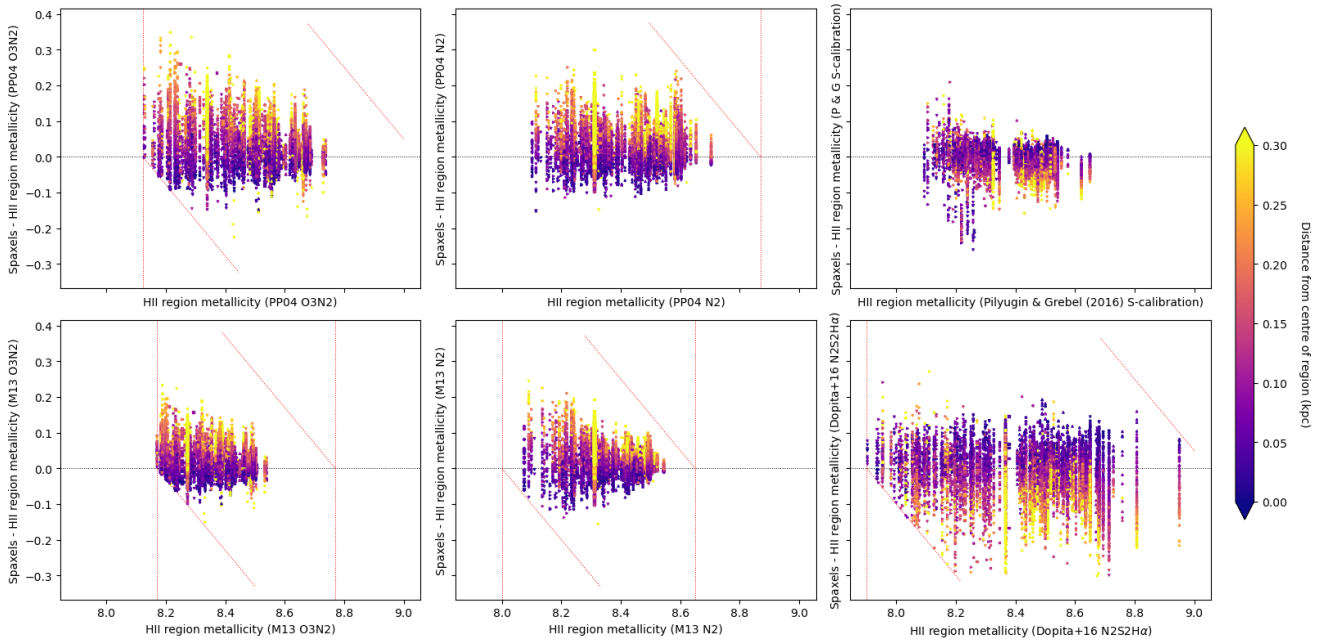

Figure 10. As for Fig. 9, colouring the points instead by the radial distance of each spaxel from the centre of the region, in kpc.

The O3N2 and N2 diagnostics rely on an anticorrelation between the metallicity and log (*U*), which is reflected in the log (*U*) trends seen for these diagnostics in Fig. 9 (left and centre panels, respectively). High values of log (*U*) are seen at the low metallicity end on the *x*-axis in Fig. 9, and along the *y*-axis spaxels with lower metallicities have higher log (*U*) values than spaxels with higher metallicities, leading to a diagonal colour gradient.

Conversely, the N2S2Hα diagnostic should be independent of log (*U*) by using a ratio of [N II]/[S II], and any observed log (*U*)-dependence with metallicity should thus reflect an intrinsic relation. The bottom right panel in Fig. 9 shows that there is no clear trend with log (*U*) at low metallicities, but at higher metallicities (12 + log(O/H) $>$ 8.4) there is a clear positive trend, with spaxels with higher metallicities than that of the overall region also having higher log (*U*). This suggestion of a positive relationship between metallicity and log (*U*) has been previously observed when using the N2S2Hα (Krühler et al. 2017; Easeman et al. 2022), S-calibration (Kreckel et al. 2019; Groves et al. 2023), and the N2O2 ([N II] λ6584/[O II] $\lambda\lambda$3727,29) (Grasha et al. 2022) diagnostics.

The Pilyugin & Grebel (2016) S-calibration shows a much smaller range of metallicities within a single H II region compared to the N2S2Hα values, mostly within $\sim$0.2 dex of the measured stack value, with the exception of a few high log (*U*) points that lie further away. Although the relationship is less clear, the points seem to show a similar relationship with log (*U*) as the N2S2Hα diagnostic, with no clear relationship at low metallicity, but a positive relationship for metallicities $>$8.4.

Aside from the distribution of log (*U*) and metallicity within H II regions in absolute terms, it is also interesting to consider where the spread of data points lie relative to the stacked H II metallicity (dashed horizontal line). For the O3N2 and N2 diagnostics, the data points lie predominantly above the dashed line ($\sim$70 per cent of spaxels lie above). The N2S2Hα and S-calibration instead show the stacked spectra measurements lying closer to the centre of the distribution slightly offset towards higher metallicities than returned from the individual spaxels ($\sim$40 per cent of spaxels lie above, $\sim$60 per cent below). Considering the spaxels with metallicity values within 0.01 dex of the stack value, we found the average log (*U*) to be fairly consistent across the diagnostics, ranging from $-$2.97 to $-$3.00. This implies that for each strong line diagnostic, the H II region metallicity is weighted towards spaxels with comparable environmental properties.

Given this apparent difference between the diagnostics in how the H II stacked metallicity compares to the metallicity distribution within the H II regions, we also investigated the spatial location of the spaxels that carried a higher weight in the stacked H II metallicities. We did this by colour-coding the points in Fig. 9 by the distance of each spaxel from the centre of the region (Fig. 10). In doing so, we noticed that the O3N2 and N2 profiles clearly showed profiles inverted to what would be expected, with lower metallicities measured in the centres compared to the outskirts. This was also observed by Krühler et al. (2017) and Mao, Lin & Kong (2018), and similarly to these papers, we find the N2S2Hα diagnostic to behave as expected, with metallicities increasing towards the centres of H II regions. The S-calibration and N2S2Hα diagnostic show negative radial gradients most clearly for regions with metallicity above 8.4. The inverted profiles seen when using the O3N2 and N2 diagnostics is therefore likely a consequence of the positive relationship observed between metallicity and log (*U*) in more metal rich H II regions with the N2S2Hα and S-calibration diagnostics, contrary to the anticorrelation required by the O3N2 and N2 diagnostics.

For spaxels with metallicities within 0.01 dex of the stack value, the average distance from the centre of the region is also fairly consistent between the diagnostics, ranging from 0.11 to 0.13 kpc.

As the metallicity of stacked spectra appear to trace the same part of the region and gas with similar levels of excitation for all strong line diagnostics, there is no obvious property on the sub-H II region scales that can account for the differences observed in Fig. 4 between the diagnostics considered. Instead, the distribution in H II properties echoes the dependences observed in the stacked spectra between O3N2 and N2 metallicity and log (*U*).

6 DISCUSSION

Strong line metallicity diagnostics are only indirect tracers of metallicity, and they are generally dependent on multiple environmental properties, most notably log (*U*) (Kewley & Dopita 2002). Diagnostics with either just a weak dependence on log (*U*) (e.g. N2O2 at high metallicity; Kewley & Dopita 2002), or ones that include multiple line ratios that can be used to constrain the various environmental properties (e.g. R_{23}; Kobulnicky & Kewley 2004) can therefore be considered more reliable. However, such diagnostics commonly include emission lines at very different wavelength ranges, which are not always accessible with a single spectrograph. In the case of MUSE, the relatively high spectral cut off at the blue end of ∼4800 Å makes the important [O II] line doublet inaccessible for nearby galaxies ($z < 0.3$), for which the spatial resolution is highest. Similarly, the [O III] λ4363 line cannot be detected with MUSE in galaxies at $z < 0.1$, meaning that oxygen T_e-based diagnostics cannot be applied. In this paper, we have therefore used the Díaz & Zamora (2022) sulphur T_e-based metallicity diagnostic to investigate the accuracy of several strong line metallicity diagnostics that are applicable to MUSE observations of nearby galaxies ($z < 0.03$). This includes the Dopita et al. (2016) N2S2Hα diagnostic, which was calibrated against the MAPPINGS photoionization code (Dopita et al. 2013), and which has not yet been compared empirically to T_e-based metallicities on a large sample of H II regions.

In general, we find good agreement between the Díaz & Zamora (2022) sulphur T_e-based metallicities and oxygen T_e-based metallicities (Izotov et al. 2006; see Fig. 3), but there are significant offsets between the T_e-based metallicities and the strong line diagnostics considered in this paper (Fig. 4). In this section, we consider possible causes for the offsets that we observe, and discuss the implications of our findings.

6.1 The sulphur-to-oxygen ratio

The S/O ratio has been suggested to increase with decreasing metallicity (Dors et al. 2016; Díaz & Zamora 2022), and thus by assuming a constant value, as we do in Section 4.1 to convert between 12 + log(S/H) and 12 + log(O/H) we could be overestimating 12 + log(O/H) at low metallicities (relative to high metallicities). For example, using a combination of SF galaxies and H II regions, Dors et al. (2016) found the log(S/O) ratio to vary from −1.68 at 12 + log(O/H) = 8.0, to −1.90 at 12 + log(O/H) = 8.8. Applying such a metallicity-dependent log(S/O) ratio would increase our T_e-based metallicities by ∼0.3 dex at the high metallicity end, further flattening the relation between T_e-based and strong line metallicities, and significantly increasing the offsets we observe between the strong line and T_e-based metallicities in Figs 4 and 5. The log(S/O) values measured for the 'DHR' sample presented in Díaz & Zamora (2022), which best represents our sample of H II regions, similarly imply an inverse relation with metallicity, varying from around −1.3 at 12 + log(O/H) = 8.0 to around −1.6 at 12 + log(O/H) = 8.7.

In contrast to these results, Berg et al. (2020) found no evidence of a metallicity dependence in log(S/O), instead finding a fairly constant ratio with an average value of log(S/O) = −1.34 for the 190 H II regions in their sample of four CHAOS galaxies.

With our choice of log(S/O) = −1.57, we find good agreement between the sulphur- and oxygen-based T_e diagnostics in Section 3. If we were instead to apply a varying log(S/O) such as found by Díaz & Zamora (2022), discrepancies of ∼0.2 dex would be introduced at low metallicities between the Díaz & Zamora (2022) and Izotov et al. (2006) T_e-based diagnostics, and the differences between Díaz & Zamora (2022) and Berg et al. (2020) T_e-based metallicities would also increase. If alternatively we used the best-fitting constant value of log(S/O) = −1.34 from Berg et al. (2020), we would reduce the discrepancies between the strong line and sulphur based T_e metallicities shown in Fig. 4, bringing the N2S2Hα diagnostic in particular in very good agreement. However, unexplained discrepancies of 0.23 dex would then be introduced between the sulphur- and oxygen-based T_e diagnostics (Fig. 3), suggesting that metallicities based on the Díaz & Zamora (2022) method would no longer provide values in alignment with those returned by oxygen-based methods. We therefore find a fixed value of log(S/O) = −1.57 to be the most appropriate choice.

6.2 Possible origin of metallicity offsets

Given that the majority of the strong line metallicity diagnostics in Section 3 are calibrated against T_e-based metallicities, it seems surprising that they show such poor agreement with our sulphur T_e-based metallicities. It is therefore reasonable to consider what differences there may be in the range of environmental properties present in the respective calibration samples and in our sample of H II regions, and how different T_e-based methods compare.

6.2.1 Selection effects in T_e-based samples

The [O III] λ4363 line is stronger in high excitation, low metallicity galaxies (e.g. Hoyos & Díaz 2006), which could bias calibration samples that are based on auroral line detections against low excitation, high metallicity regions. Evidence of such biases can be seen in the M13 H II sample, where for metallicities above 8.5, there are significantly fewer measurements, and for PP04, only 3 of their 131 observed H II regions with [O III] λ4363 line detections have higher than solar metallicities.

To test whether similar selection effects are present in our [S III] λ6312 line sample, we compared the distribution of log (*U*) and strong line metallicities in the full H II region sample selected in Section 3.1 by our HIIDENTIFY code to the subsample of regions with T_e-based measurements. Out of 4408 regions within our sample, 186 (4.2 per cent) have measured T_e-based metallicities. However, we note that our HIIDENTIFY code is capable of separating overlapping H II regions (see Fig. 2), which would result in a larger number of detected H II regions, extending down to lower Hα fluxes, compared to other H II identification codes. For example, from visual inspection of the segmentation maps presented in Grasha et al. (2022, their fig. 3), where the HIIPHOT code was used to identify H II regions, the regions identified appear to be fairly isolated, suggesting that overlapping regions are generally merged together. Evidence of this is present in the galaxy NGC2835, which is present in both our sample and in Grasha et al. (2022). The MUSE pointing covers just the central 1 arcmin × 1 arcmin part of the galaxy, whereas the data in Grasha et al. (2022) cover an area ∼8 times larger (their fig. 1). Nevertheless, the number of identified H II regions in Grasha et al. (2022) is only ∼30 per cent greater. The larger size of our parent sample therefore has the consequence of reducing the fraction of H II regions with [S III] λ6312 detections than if we had used alternative H II detection algorithms.

Similarly to the biases present in [O III] λ4363-based samples, our subset of [S III] λ6312-bright H II regions are indeed shifted towards higher values of log (*U*), covering a range of −3.3 to −2.2, compared to −4.1 to −2.2 for the full sample. The metallicities as measured by the strong line diagnostics are also shifted towards lower values, with

the maximum metallicities in the subsample of T_e-selected regions typically ∼0.2 dex below that for the full sample, despite the parent population peaking towards high metallicities.

Similar trends are seen in our spaxel-scale analysis, where in our sample of H II regions with measured T_e-based metallicities in the stacked spectra, only 3.5 per cent of spaxels within those regions also have measurements of the T_e-based metallicity. Again the range of log (*U*) is reduced, with all spaxels covering a range of −4.0 to −1.8, and those with T_e-based metallicity measurements covering −3.3 to −1.8. The range of metallicity measurements is similarly limited at the high metallicity end, to roughly 0.1 dex below the maximum of that of the parent sample. For example, the PP04 O3N2 values for the parent sample range from 8.15 to 8.8, peaking around 8.6. The subsample of regions with T_e-based measurements cover only from 8.15 to 8.65.

In Figs 4–6, we saw evidence for strong log (*U*) gradients in the O3N2 and N2 comparison plots, whereby H II regions with lower log (*U*) values lay at systematically larger offsets from the line of equality, especially at the higher metallicity end. Given this trend, one could therefore speculate that the large sample of low log (*U*) and high metallicity H II regions omitted by our selection criteria lie closer to the line of equality at the high metallicity end than our T_e-based sample, steepening the slope of the best-fitting line. Nevertheless, given that the [O III] λ4363 T_e-based methods used to calibrate the diagnostics will suffer from similar selection effects, in particular when stacking is not used, as is the case with the PP04 and M13 diagnostics, this bias cannot clearly explain the disagreement that we find in Figs 4–6. Furthermore, the better agreement shown in Fig. 6 between our T_e-based metallicities and the Curti et al. (2017) O3N2 calibration, which uses stacking to reduce selection effects, implies that our results are not significantly affected by selection effects present in our [S III] λ6312-based sample relative to other auroral line-based samples.

To verify that unaccounted selection effects are not inadvertently causing the flat relation that we find between [S III] λ6312 T_e-based metallicities and the O3N2 and N2 metallicities, we investigated the relation between the sulphur-based and PP04 metallicities in the CHAOS sample, which used fixed slit spectroscopy rather than IFU data. We found the O3N2 and N2 relations to be flatter than our MAD galaxy results shown in Fig. 4, with respective best-fitting slopes of 0.30 ± 0.05 and 0.31 ± 0.04 (compared to 0.55 ± 0.05 and 0.53 ± 0.04 when using the MAD galaxy sample). The CHAOS sample of H II regions with [S III] λ6312 detections has predominantly high log (*U*) values (>−3), implying that CHAOS has a larger log (*U*) bias than the MAD sample, and possibly also more biased than the PP04 calibration sample in terms of the range of ionization parameters covered.

6.2.2 *Differences in electron temperatures*

M13 show their relation between T_e-metallicity and the O3N2 and N2 line ratios next to other relations, including PP04, in their figs 2 and 4, where it is clear that the M13 relation is much shallower. Notably, as illustrated in Fig. 7, at $12 + \log(\mathrm{O/H}) \gtrsim 8.4$ the Curti et al. (2017) O3N2 calibration has a similar slope, but is offset to larger metallicities. The sample of H II regions used in M13 is larger than used in PP04, and M13 also use a range of auroral line detections to minimize the impact of temperature-based selection effects. The reason for the systematically lower M13 O3N2 and N2 metallicities at high metallicities, leading to their shallower relation, is not clear. However, it is worth noting that there is scatter in the M13 sample, and in the case of O3N2 (see their fig. 2), the best-fitting relation at high metallicities is clearly driven by a tight distribution of data points from Pilyugin, Grebel & Mattsson (2012), whereas data from Pérez-Montero & Contini (2009), for example, extend to higher metallicities for the same given O3N2 line ratio.

M13 used auroral line detections from various elements to measure the electron temperature of the H II regions in their sample, using the Pilyugin, Vílchez & Thuan (2010) method to re-measure T_e-based metallicities. Which of the auroral lines were detected depended largely on the metallicity of the H II region, with the [N II] auroral line generally used for metallicities >8.4, and metallicities at <8.4 largely measured using the [O III] auroral line. When determining metallicities based on a detection of the [N II] auroral line, their method assumes that $T_{e,[\mathrm{N\,II}]} = T_{e,[\mathrm{O\,II}]}$, and while the two elements do have similar ionization potentials and can be expected to trace similar parts of the H II region, various studies have suggested the two temperatures cannot be considered to be equal, although there is contention in what the relation should be. Yates et al. (2020) found that especially at higher temperatures, $T_{e,[\mathrm{N\,II}]} > T_{e,[\mathrm{O\,II}]}$ in their observations of H II regions and galaxies, whereas Berg et al. (2020) found the opposite using the CHAOS sample of H II regions, suggesting that the relationship between the two temperatures is not well defined, and it is therefore unclear whether it is valid to assume that $T_{e,[\mathrm{N\,II}]} = T_{e,[\mathrm{O\,II}]}$. The mixture of auroral lines used in M13 to trace the electron temperature could therefore contribute to the discrepancies we find between the M13 diagnostics and our sulphur T_e-based metallicities, in particular in the case that $T_{e,[\mathrm{N\,II}]} > T_{e,[\mathrm{O\,II}]}$ (Yates et al. 2020), which would imply that M13 have overestimated $T_{e,[\mathrm{O\,II}]}$, and thus underestimated $12 + \log(\mathrm{O/H})$ at high metallicities. However, the inhomogeneous sample of H II regions used in M13 may also contribute to offsets.

6.2.3 *Fixed slit versus IFU measurements*

The Pilyugin & Grebel (2016) S-calibration has been suggested to return underestimates of the metallicity when used on IFU data, with Pilyugin et al. (2022) finding the results to be underestimated by ∼0.06 dex at $12 + \log(\mathrm{O/H}) \gtrsim 8.55$, and by ∼0.02 below this, on average. They reason that this is due to differences in the intensities of each line in single-slit and IFU observations, due to the limitations of single-slit observations meaning that the entire H II region is not always captured in a single observation. They suggest this introduces systematic effects when their diagnostic, which was calibrated using single-slit observations, is used on IFU data. This could be expected to therefore also affect the PP04 diagnostics, which were calibrated against single slit data, and the M13 diagnostics, which were calibrated using a combination of IFU data and single-slit observations. However, this can only account for some of the offsets observed between the diagnostics in our analysis, as we observed systematic offsets of up to 0.3 dex at high metallicities. Furthermore, our analysis on the spaxel-level data indicate that the H II stacked spectra are predominantly weighted by the central regions of the H II region, implying that emission missed in single slit spectra from the outer regions of the H II region are unlikely to contribute greatly to the measured line fluxes.

6.3 Large- and small-scale metallicity gradients

The strong line O3N2, N2, and S-calibration diagnostics underestimate the metallicity at high metallicities by ∼0.3 dex, which could have implications on measurements of the metallicity gradient,

for example imposing a flattened inner profile for a galaxy with a negative metallicity gradient. The N2S2Hα diagnostic shows no evidence of a metallicity-dependent offset from T_e-based measurements, and no clear log (U) dependence in the discrepancies between the strong line and T_e-based measurements. This could explain why, for a sample of galaxies observed in the MaNGA survey (Mapping Nearby Galaxies at Apache Point Observatory; Bundy et al. 2015), Yates et al. (2021) found evidence of flattened inner profiles in more massive galaxies when O3N2-based diagnostics were used, which were not seen in results from the N2S2Hα diagnostic, as well as finding that the N2 diagnostic also returned much flatter profiles than the N2S2Hα diagnostic.

On sub-H II region scales, we find in Figs 9 and 10 that the O3N2 and N2 diagnostics show clear inverted metallicity profiles, with metallicity appearing to decrease towards the centre of the H II regions. The N2S2Hα diagnostic, however, shows decreasing metallicities from the centres of the regions out to the outskirts, which is in line with expectations. Exploring the relation between metallicity and log (U), we find that at high metallicities, the N2S2Hα diagnostic shows a clear positive relationship between log (U) and metallicity, which would explain the inverted profiles shown by the O3N2 and N2 diagnostics, as they rely on an anticorrelation between metallicity and log (U) observed in H II region and galaxy samples.

6.4 Optimal metallicity diagnostics with MUSE

On both H II region and sub-H II region scales, the N2S2Hα diagnostic results are offset from the T_e-based values, however, this discrepancy has no apparent dependence on either metallicity or ionization parameter, unlike the other diagnostics tested. We therefore find that this diagnostic produces the most reliable results when used on spatially resolved MUSE observations of nearby galaxies.

We note, however, that the H II regions used in our analysis typically account for just ∼10 per cent of the overall Hα luminosity of the galaxy within the MUSE field of view, therefore these results may not be descriptive of trends seen on whole-galaxy scales.

Further study of the diagnostics will be possible with the proposed BlueMUSE instrument (Richard et al. 2019). BlueMUSE will have a spectral range covering much shorter wavelengths (3500–6000 Å), and will provide high spatial resolution IFU data with coverage of the [O II] λλ3727,29 and [O III] λ4363 lines, allowing for T_e-based diagnostics reliant on the oxygen lines to be used, removing the need for any conversion between sulphur and oxygen abundances, and allowing a larger range of strong line diagnostics to be investigated.

7 CONCLUSIONS

We have used a sample of 671 H II regions from 36 galaxies observed as part of the MAD survey, to assess the optimal strong line diagnostic for use with MUSE data. We additionally release a catalogue of 4408 H II regions identified within this sample using our newly developed PYTHON tool HIIDENTIFY,[6] with segmentation maps and tables of emission line strengths made available (see Appendix C). By comparing the results from strong line diagnostics to T_e-based measurements, we find:

(i) The PP04 O3N2, PP04 N2, and Pilyugin & Grebel (2016) S-calibration diagnostics all show consistent results, with a sublinear relationship leading to agreement around 12 + log(O/H) = 8.2, but an offset of ∼0.3 dex below the T_e-based values at 12 + log(O/H) = 8.8. The O3N2 and N2 diagnostics additionally show strong log (U) dependence in the offsets from the T_e-based values. The N2S2Hα diagnostic shows the greatest level of agreement with the T_e-based values, with the slope of the best-fitting line being consistent with a 1:1 relationship. However, this diagnostic has a systematic offset of ∼0.2 dex, which, unlike the O3N2 and N2 diagnostics, has no clear dependence on log (U) or metallicity.

(ii) The O3N2 and N2 diagnostics presented by Marino et al. (2013) show the largest differences with the T_e-based measurements at high metallicities, with these strong line diagnostics returning values over a significantly reduced range compared to the T_e-based values, due to the much steeper relationship between the line ratios and metallicity (see Fig. 7).

(iii) This comparison on H II region scales suggests that when measuring the metallicity of H II regions for galaxies observed with MUSE, the N2S2Hα diagnostic provides the most accurate measurement of the metallicity.

(iv) We used the CHAOS sample (Berg et al. 2020) to assess the validity of using the sulphur T_e-based method presented by Díaz & Zamora (2022) to measure 12 + log(O/H), finding good agreement with the Izotov et al. (2006) oxygen T_e-based method. However, when comparing the results to the published results from applying the multizone T_e-based method presented in Berg et al. (2020), there is a significant offset at lower metallicities (−0.37 dex at 12 + log(O/H) = 8.0). If we use our fitted relationship to 'convert' our measured T_e-based metallicities for the MAD data to predicted values for if the Berg et al. (2020) method could be used on our data, the resulting comparisons between the strong line and T_e-based diagnostics are significantly changed, with reduced metallicity-dependence in the offsets for the O3N2, N2 and S-calibration diagnostics, but increased dependence for the N2S2Hα.

(v) On sub-H II region scales, we find clear evidence of inverted metallicity profiles when using the O3N2 and N2 diagnostics, which are not seen when the S-calibration or N2S2Hα diagnostic is used. For the N2S2Hα diagnostic, we find a clear positive relationship between log (U) and metallicity at high metallicities, in contrast to the negative relationship required by the O3N2 and N2 diagnostics. This could therefore explain the inverted profiles observed within H II regions, as has been suggested by Krühler et al. (2017) and Mao, Lin & Kong (2018).

(vi) Stacked spectra from H II regions appear to generally be weighted towards the inner part of H II regions, where the log (U) is highest. The differing relationships between log (U) and metallicity between the diagnostics may therefore suggest a reason for the disagreement between the N2S2Hα, and O3N2 and N2 diagnostics.

ACKNOWLEDGEMENTS

The authors would like to thank Ángeles Díaz for an interesting and enlightening conversation surrounding the use of sulphur T_e-based metallicity diagnostics, and the S/O ratio. We also thank Dr Julian Stirling for his guidance on publishing HIIDENTIFY, and the referee for their feedback.

DATA AVAILABILITY

Data from the MAD were used, and can be accessed from the ESO archive science portal.[7] The H II region catalogue created and

[6] https://hiidentify.readthedocs.io/en/latest/

[7] http://archive.eso.org/scienceportal/home

analysed as part of this work is provided as an online table (see Appendix C), with the segmentation maps from HIIDENTIFY available at https://doi.org/10.6084/m9.figshare.22041263.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Alloin D., Collin-Souffrin S., Joly M., Vigroux L., 1979, A&A, 78, 200
Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, ARA&A, 47, 481
Bacon R. et al., 2010, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 7735, Ground-Based and Airborne Instrumentation for Astronomy III. SPIE, Bellingham, p. 773508
Belfiore F. et al., 2016, MNRAS, 461, 3111
Belfiore F. et al., 2017, MNRAS, 469, 151
Belfiore F. et al., 2019, AJ, 158, 160
Berg D. A., Skillman E. D., Croxall K. V., Pogge R. W., Moustakas J., Johnson-Groh M., 2015, ApJ, 806, 16
Berg D. A., Pogge R. W., Skillman E. D., Croxall K. V., Moustakas J., Rogers N. S. J., Sun J., 2020, ApJ, 893, 96
Boardman N. F., Zasowski G., Newman J. A., Sanchez S. F., Schaefer A., Lian J., Bizyaev D., Drory N., 2020, MNRAS, 7, 1
Bresolin F., 2006, The Metal-Rich Universe. Cambridge Univ. Press, Cambridge, p. 155
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Bundy K. et al., 2015, ApJ, 798
Cameron A. J., Yuan T., Trenti M., Nicholls D. C., Kewley L. J., 2020, MNRAS, 21, 1
Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245
Cid Fernandes R., Mateus A., Sodré L., Stasińska G., Gomes J. M., 2005, MNRAS, 358, 363
Cid Fernandes R., Schlickmann M., Stasinska G., Asari N. V., Gomes J. M., Schoenell W., Mateus A., Sodré L., Jr, 2009, in Wang W., Yang Z., Luo Z., Chen Z., eds, ASP Conf. Ser. Vol. 408, The Starburst-AGN Connection. Astron. Soc. Pac., San Francisco, p. 122
Croxall K. V., Pogge R. W., Berg D. A., Skillman E. D., Moustakas J., 2015, ApJ, 808, 42
Croxall K. V., Pogge R. W., Berg D. A., Skillman E. D., Moustakas J., 2016, ApJ, 830, 4
Curti M., Cresci G., Mannucci F., Marconi A., Maiolino R., Esposito S., 2017, MNRAS, 465, 1384
Díaz Á. I., Zamora S., 2022, MNRAS, 511, 4377
Dopita M. A., Sutherland R. S., Nicholls D. C., Kewley L. J., Vogt F. P. A., 2013, ApJS, 208, 10
Dopita M. A., Kewley L. J., Sutherland R. S., Nicholls D. C., 2016, Ap&SS, 361, 1
Dors O. L., Jr, Krabbe A., Hägele G. F., Pérez-Montero E., 2011, MNRAS, 415, 3616
Dors O. L., Pérez-Montero E., Hägele G. F., Cardaci M. V., Krabbe A. C., 2016, MNRAS, 456, 4407
Easeman B., Schady P., Wuyts S., Yates R. M., 2022, MNRAS, 511, 371
Emsellem E. et al., 2022, A&A, 659, A191
Erroz-Ferrer S. et al., 2019, MNRAS, 484, 5009
Espinosa-Ponce C., Sánchez S. F., Morisset C., Barrera-Ballesteros J. K., Galbany L., García-Benito R., Lacerda E. A., Mast D., 2020, MNRAS, 494, 1622
Gardner J. P. et al., 2006, Space Sci. Rev., 123, 485
Garnett D. R., 1992, AJ, 103, 1330
Grasha K. et al., 2022, ApJ, 929, 118
Groves B. et al., 2023, MNRAS, 520, 4902
Hoyos C., Díaz A. I., 2006, MNRAS, 365, 454
Izotov Y. I., Stasińska G., Meynet G., Guseva N. G., Thuan T. X., 2006, A&A, 448, 955
Kewley L. J., Dopita M. A., 2002, ApJS, 142, 35
Kewley L. J., Ellison S. L., 2008, ApJ, 681, 1183
Kewley L. J., Nicholls D. C., Sutherland R. S., 2019, ARA&A, 57, 511
Kobayashi C., Karakas A. I., Lugaro M., 2020, ApJ, 900, 179
Kobulnicky H. A., Kewley L. J., 2004, ApJ, 617, 240
Kreckel K. et al., 2019, ApJ, 887, 80
Krühler T., Kuncarayakti H., Schady P., Anderson J. P., Galbany L., Gensior J., 2017, A&A, 602
Lacerda E. A. et al., 2018, MNRAS, 474, 3727
Luridiana V., Morisset C., Shaw R. A., 2015, A&A, 573, A42
Maiolino R., Mannucci F., 2019, A&AR, 27
Mao Y.-W., Lin L., Kong X., 2018, ApJ, 853, 151
Marino R. A. et al., 2013, A&A, 559, 1
Mingozzi M. et al., 2020, A&A, 636, A42
Osterbrock D. E., Ferland G. J., 2006, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei, 2nd edn. University Science Books, Melville
Pagel B. E. J., Edmunds M. G., Blackwell D. E., Chun M. S., Smith G., 1979, MNRAS, 189, 95
Peimbert M., Peimbert A., Delgado-Inglada G., 2017, PASP, 129, 82001
Pérez-Montero E., 2017, PASP, 129, 43001
Pérez-Montero E., Contini T., 2009, MNRAS, 398, 949
Pettini M., Pagel B. E., 2004, MNRAS, 348, 59
Pilyugin L. S., Grebel E. K., 2016, MNRAS, 457, 3678
Pilyugin L. S., Vílchez J. M., Thuan T. X., 2010, ApJ, 720, 1738
Pilyugin L. S., Grebel E. K., Mattsson L., 2012, MNRAS, 424, 2316
Pilyugin L., Lara-Lopez M., Vilchez J., Duarte Puertas S., Zinchenko I., Dors O., 2022, A&A, 5, 1
Poetrodjojo H. et al., 2021, MNRAS, 502, 3357
Pogge R. W. et al., 2010, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 7735, Ground-Based and Airborne Instrumentation for Astronomy III. SPIE, Bellingham, p. 77350A
Richard J. et al., 2019, preprint (arXiv:1906.01657)
Salo H. et al., 2015, ApJS, 219, 4
Sánchez S. F. et al., 2012, A&A, 538, A8
Sánchez S. F. et al., 2013, A&A, 554
Sánchez S. F. et al., 2015, A&A, 574, A47
Sanders R. L., Shapley A. E., Zhang K., Yan R., 2017, ApJ, 850, 136
Schaefer A. L. et al., 2019, ApJ, 884, 156
Stasinska G., 2019, preprint (arXiv:1906.04520)
Thilker D. A., Braun R., Walterbos R. A. M., 2000, AJ, 120, 3070
Tremonti C. A. et al., 2004, ApJ, 613, 898
Yates R. M., Schady P., Chen T.-W., Schweyer T., Wiseman P., 2020, A&A, 634, A107
Yates R. M., Henriques B. M. B., Fu J., Kauffmann G., Thomas P. A., Guo Q., White S. D. M., Schady P., 2021, MNRAS, 503, 4474
Zhang K. et al., 2017, MNRAS, 466, 3217

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Table C1. Example dust-corrected fluxes in 10^{-20} erg s^{-1} cm^{-2} measured for the 4408 H II regions identified within our sample using HIIDENTIFY.

Table C2. Example uncertainties on the measured fluxes for the 4408 H II regions identified within our sample using HIIDENTIFY.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: INVERTED METALLICITY PROFILES OBSERVED WITH O3N2 AND N2 DIAGNOSTICS

Repeating the analysis described in Section 5.4, but instead plotting the log (U) that we measured, we see in Fig. A1 that the stacked H II region value is also weighted towards the value of spaxels near the centre of the region. For spaxels with log (U) values within 0.01 of that of the region stack, the average distance from the centre of the

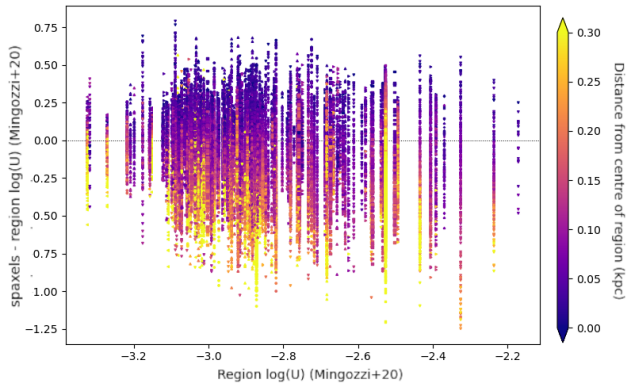

Figure A1. As for Fig. 9, comparing the log (*U*) measured in each spaxel within a region, to the values returned from the region's stacked spectrum, coloured by the distance to the centre of the region.

region is 0.11 kpc, similar to the location of spaxels that contribute most to the stacked metallicities (Fig. 10). 30 per cent of the spaxels have log (*U*) values higher than that of the stacked region value, 70 per cent have lower values, suggesting the stacked value is slightly shifted towards higher log (*U*) than the distribution of values within the region. Smooth gradients in log (*U*) can be seen for all regions, with high values of log (*U*) found in the centres of regions, decreasing with radius.

APPENDIX B: MAD CATALOGUE

Table B1 details the global properties of the MAD galaxies used in this work, with information compiled from Erroz-Ferrer et al. (2019) and Salo et al. (2015).

Table B1. Sample information, compiled from Erroz-Ferrer et al. (2019) and Salo et al. (2015).

Name	z	D (Mpc)	R_e (arcsec)	log($M_*/M_\odot$)	SFR ($M_\odot$ yr^{-1})	PA (deg)	q	RA (deg)	Dec. (deg)
NGC4030	0.004887	29.9	31.8	11.18	11.08	29.6	0.805	180.098510	− 1.099960
NGC3256	0.009354	38.4	26.6	11.14	3.10	60.0	0.680	156.963624	− 43.903748
NGC4603	0.008647	32.8	44.7	11.10	0.65	40.0	0.580	190.230042	− 40.976389
NGC3393	0.012509	55.2	21.1	11.09	7.06	− 20.0	0.720	162.097750	− 25.162056
NGC1097	0.004240	16.0	55.1	11.07	4.66	126.5	0.567	41.579410	− 30.274910
NGC289	0.005434	24.8	27.0	11.00	3.58	123.7	0.658	13.176520	− 31.205830
IC2560	0.009757	32.2	37.8	10.89	3.76	40.0	0.380	154.077985	− 33.563795
NGC5643	0.003999	17.4	60.7	10.84	1.46	87.5	0.670	218.169765	− 44.174406
NGC3081	0.007976	33.4	18.9	10.83	1.47	70.0	0.480	149.873080	− 22.826277
NGC4941	0.003696	15.2	64.7	10.80	3.01	23.5		196.054760	− 5.551620
NGC5806	0.004533	26.8	27.2	10.70	3.61	166.6	0.487	225.001800	1.891270
NGC3783	0.009730	40.0	27.7	10.61	6.93	− 15.0	0.620	174.757342	− 37.738670
NGC5334	0.004623	32.2	51.2	10.55	2.45	15.1	0.735	208.226960	− 1.114760
NGC7162	0.007720	38.5	18.0	10.42	1.73	12.7	0.426	329.912720	− 43.306220
NGC1084	0.004693	20.9	23.8	10.40	3.69	38.4	0.606	41.499690	− 7.578640
NGC1309	0.007125	31.2	20.3	10.37	2.41	78.9	0.891	50.527320	− 15.400070
NGC5584	0.005464	22.5	63.5	10.34	1.29	156.9	0.662	215.599210	− 0.387450
NGC4900	0.003201	21.6	35.4	10.24	1.00	140.8	0.834	195.163120	2.501460
NGC7496	0.005365	11.9	66.6	10.19	1.80	33.2		347.447050	− 43.428020
NGC7552	0.005500	14.8	26.0	10.19	0.59	38.5	0.586	349.044860	− 42.584830
NGC1512	0.002995	12.0	63.3	10.18	1.67	74.5	0.591	60.976170	− 43.348850
NGC7421	0.005979	25.4	29.6	10.09	2.03	64.6	0.739	344.226320	− 37.347200
ESO498-G5	0.008049	32.8	19.8	10.02	0.56	− 25.0	0.740	141.169458	− 25.092694
NGC1042	0.004573	15.0	63.7	9.83	2.41	6.7	0.779	40.099900	− 8.433650
IC5273	0.004312	15.6	33.8	9.82	0.83	49.0	0.638	344.861330	− 37.702880
NGC1483	0.003833	24.4	19.0	9.81	0.43	125.7	0.735	58.198370	− 47.477520
NGC2835	0.002955	8.8	57.4	9.80	0.38	− 22.0	0.870	139.470458	− 22.354667
PGC3853	0.003652	11.3	73.1	9.78	0.35	106.4	0.853	16.270320	− 6.212380
NGC337	0.005490	18.9	24.6	9.77	0.57	119.6	0.633	14.958710	− 7.577960
NGC4592	0.003566	11.7	37.9	9.68	0.31	94.6	0.298	189.828250	− 0.531840
NGC4790	0.004483	16.9	17.7	9.60	0.39	86.2	0.646	193.716380	− 10.247820
NGC3513	0.003983	7.8	55.4	9.37	0.21	67.6	0.722	165.942000	− 23.245480
NGC2104	0.003873	18.0	16.5	9.21	0.24	167.3	0.547	86.769880	− 51.552950
NGC4980	0.004767	16.8	13.0	9.00	0.18	169.6	0.516	197.292010	− 28.641790
NGC4517A	0.005087	8.7	46.8	8.50	0.10	23.8	0.573	188.117170	0.389750
ESO499-G37	0.003186	18.3	18.3	8.47	0.14	40.0	0.640	150.924333	− 27.027806

Table C1. Example dust-corrected fluxes in 10^{-20} erg s^{-1} cm^{-2} measured for the 4408 H II regions identified within our sample using HIIDENTIFY.

Galaxy name	Region ID	H β	[O III] λ4959	[O III] λ5007	[S III] λ6312	[N II] λ6548	H α	[N II] λ6584	...
NGC4030	1	580963.98	24145.05	71701.19	–	218013.01	2145082.91	670310.83	...
NGC4030	2	422113.73	13597.15	40378.11	–	163837.02	1489530.26	503739.32	...
NGC4030	3	174188.67	7093.77	21065.66	1296.39	61948.63	593299.01	190469.54	...
NGC4030	4	612040.38	17424.16	51742.82	–	188234.22	2266623.49	578751.85	...
NGC4030	5	646705.79	15994.38	47496.94	2313.41	227074.56	2291716.1	698171.78	...

Table C2. Example uncertainties on the measured fluxes for the 4408 H II regions identified within our sample using HIIDENTIFY.

Galaxy name	Region ID	H β	[O III] λ4959	[O III] λ5007	[S III] λ6312	[N II] λ6548	H α	[N II] λ6584	...
NGC4030	1	60880.78	6458.79	5789.58	–	30554.0	115493.37	25870.24	...
NGC4030	2	35287.18	3236.24	2931.34	–	16787.9	58892.73	14378.95	...
NGC4030	3	22356.62	2457.38	2191.55	1046.3	9251.78	34267.21	7797.34	...
NGC4030	4	31021.66	3103.83	3058.51	–	19416.09	86717.73	16463.99	...
NGC4030	5	48424.97	3795.56	3499.44	1523.36	27873.0	108669.28	23925.99	...

APPENDIX C: H II REGION CATALOGUE PRODUCED USING HIIDENTIFY FOR THE MAD GALAXIES

Using HIIDENTIFY, the PYTHON tool we developed to identify H II regions within galaxies,[8] we identified a total of 4408 H II regions within 36 galaxies from the MAD sample. For each of these regions, we stacked all spectra within the regions, and fitted the emission lines, as described in Section 3.2. We provide fluxes and associated uncertainties for the H β, [O III] $\lambda\lambda$4959,5007, [O I] $\lambda\lambda$6302,65, [S III] λ6312, [N II] $\lambda\lambda$6548,84, H α, [S II] $\lambda\lambda$6717,31, and [S III] λ9070 lines in an online table, an excerpt of which can be seen in Tables C1 and C2. The segmentation maps denoting which pixels belong to each region can be found at https://doi.org/10.6084/m9.figshare.22041263. These segmentation maps have four extensions, with the first providing a region ID for each pixel, the second a measure of the distance of the pixel from the peak of the region in kpc, the third a map of the peak positions, and the fourth shows the H α flux within the identified regions.

APPENDIX D: COMPARING THE RESULTS FROM HIIDENTIFY TO HIIPHOT

A number of codes exist to identify H II regions within galaxies, and following the release of the MUSE-PHANGS (Physics at High Angular resolution in Nearby GalaxieS; Emsellem et al. 2022) catalogue by Groves et al. (2023), we compare the results from our code to those from HIIPHOT for NGC2835, which has been analysed in both surveys.

The methodologies of the two codes differ slightly, for example HIIPHOT begins with an initial guess of the shape of the region, chosen from a set of six possible models, and grows the regions iteratively based on a slowly declining flux threshold. HIIDENTIFY proceeds by identifying just the brightest pixel within each region, and growing outwards from that point. As the codes iterate outward, different conditions are used to terminate the growth of the region. For the results from using HIIPHOT presented in Groves et al. (2023), the edges of the regions were set using the gradient of the H α surface brightness, with a fixed value used for all galaxies. For our analysis using HIIDENTIFY, we terminate the growth of a region once the flux drops to a defined background value, determined individually for each galaxy.

To compare the results of the two codes, we matched the area observed in our MAD data to the wider field of view of the PHANGS observations, achieved by mosaicking seven MUSE pointings. As shown in Fig. D1, the HIIPHOT code (orange outlines) appears to return very isolated regions compared to HIIDENTIFY (grey regions), suggesting that neighbouring regions may be merged together by HIIPHOT. The HIIPHOT code also identifies a greater number of regions, including a large number within areas removed in our analysis due to having H αEW < 6 Å. Some of the regions appear to match up well between the results from the two codes, but for other regions it appears that HIIDENTIFY has split what was considered as one region by HIIPHOT into multiple regions. Some of the regions identified by HIIDENTIFY are also larger, which may be due to the difference in how the growth of the regions are terminated in the two identification codes, with HIIDENTIFY capturing more of the diffuse emission around the edges of the region.

Due to the differences in the identification of regions between the two codes, we then explored whether this led to any differences in the measured metallicity. Using our MAD data and the HIIPHOT segmentation map, we stacked and fitted the spectra within regions identified by HIIPHOT following the same process as described in Section 3.2. Then for each HIIPHOT region, we selected every region identified by HIIDENTIFY with any overlapping pixels. To compare the metallicities returned, in Fig. D2 we plot the S-calibration metallicity from the HIIPHOT regions on the x-axis, against the metallicity of each overlapping HIIDENTIFY region on the y-axis. The dot–dashed black line shows the 1:1 relation between the two and the points are coloured by the percentage of the HIIPHOT region covered by the HIIDENTIFY region, which is also used to weight the best fit to the data shown as the orange dashed line.

Fig. D2 shows very good agreement between the S-calibration metallicities returned for the regions identified by HIIPHOT and the overlapping HIIDENTIFY regions, especially for the regions that share a large fraction of the region's pixels (purple points). We find a similar amount of consistency for all diagnostics used in our analysis. This implies that the nebular line fluxes that we measure are consistent with the published MUSE-PHANGS measurements, and that our results are robust and show little dependence on the code used to identify the H II regions.

[8] Available at https://hiidentify.readthedocs.io/en/latest/.

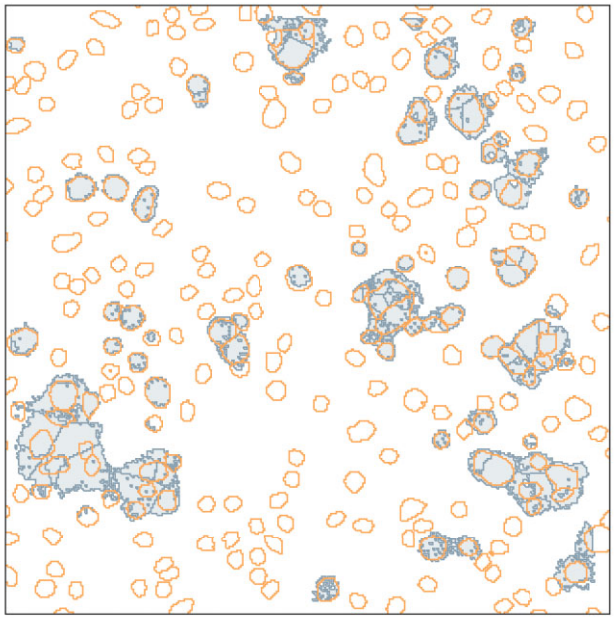

Figure D1. Field of view of the MAD observation of the centre of NGC2835, showing the regions identified with HIIDENTIFY (grey outlines and shaded) and with HIIPHOT (orange outlines) as provided by Groves et al. (2023).

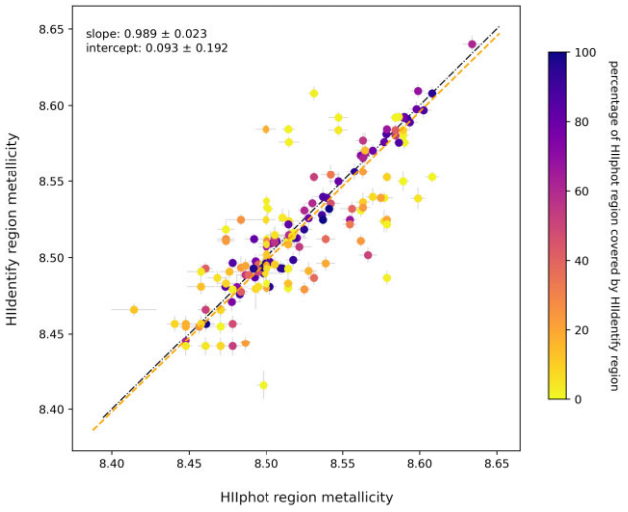

Figure D2. For each region identified by HIIPHOT that overlaps with at least one HIIDENTIFY region, we plot the S-calibration metallicity for the HIIPHOT region on the *x*-axis, and the metallicity for each overlapping region as identified by HIIDENTIFY on the *y*-axis. The points are coloured by the percentage of the HIIPHOT region covered by the HIIDENTIFY region, with the orange line of best fit to the data also weighted by this.

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