Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **465,** 1144–1156 (2017) doi:10.1093/mnras/stw2836
Advance Access publication 2016 November 3

Inferring physical properties of galaxies from their emission-line spectra

G. Ucci,[1⋆] A. Ferrara,[1,2] S. Gallerani[1] and A. Pallottini[1,3,4]

[1] *Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa, Italy*
[2] *Kavli IPMU, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8583, Japan*
[3] *Cavendish Laboratory, University of Cambridge, 19 J. J. Thomson Ave., Cambridge CB3 0HE, UK*
[4] *Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*

Accepted 2016 November 1. Received 2016 October 18; in original form 2016 September 13

ABSTRACT

We present a new approach based on Supervised Machine Learning algorithms to infer key physical properties of galaxies (density, metallicity, column density and ionization parameter) from their emission-line spectra. We introduce a numerical code (called GAME, GAlaxy Machine learning for Emission lines) implementing this method and test it extensively. GAME delivers excellent predictive performances, especially for estimates of metallicity and column densities. We compare GAME with the most widely used diagnostics (e.g. R_{23}, [N II] λ6584/Hα indicators) showing that it provides much better accuracy and wider applicability range. GAME is particularly suitable for use in combination with Integral Field Unit spectroscopy, both for rest-frame optical/UV nebular lines and far-infrared/sub-millimeter lines arising from photodissociation regions. Finally, GAME can also be applied to the analysis of synthetic galaxy maps built from numerical simulations.

Key words: methods: data analysis – ISM: general – ISM: H II regions – ISM: lines and bands – galaxies: ISM.

1 INTRODUCTION

Most of the information on the physical properties of galaxies is encoded in their spectra. These are characterized by a large number of emission lines from which their internal structure can be inferred. Several attempts have been made to recover such physical properties by mean of diagnostics based on small, selected subsets of emission-line ratios. Most of these previous works focused on ionized nebulae and have obtained diagnostics for the physical properties of galaxies based only on the strongest nebular emission lines coming from extragalactic H II regions and star-forming galaxies.

In principle, once calibrated, a diagnostic should be a univocal function of a given physical parameter. For example, each value of the R_{23} diagnostic (see Section 2) should trace a particular value of the metallicity of the gas. However, one of the most limiting aspects of this approach is that it suffers from many systematic errors which may plague the calibration, i.e. non-monotonic behaviour, sizable scatter in the calibration and also that most diagnostics are only limited to strong emission lines (López-Sánchez et al. 2012). Finally, in galaxies, the interstellar medium (ISM) is characterized not only by nebular emission lines typical of H II regions tracing the metallicity, the ionization parameter or the temperature of the gas (Kewley et al. 2001; Kewley & Dopita 2002; Levesque, Kewley & Larson 2010) but also by lines arising from low-ionization species (e.g. [C I], [O I], [N I] and [S II]) typical of neutral and denser regions.

A relatively new interesting field of research in astrophysics is represented by Machine Learning (ML) methods. These have already provided, for example, excellent predictive accuracy in the determination of photometric redshifts or in the star–galaxy classification task (Ball et al. 2006; Kim, Brunner & Carrasco Kind 2015). A number of numerical techniques have been applied such as Artificial Neural Networks (Collister & Lahav 2004) or Decision Trees (Carliles et al. 2010; Carrasco Kind & Brunner 2013; Cavuoti et al. 2015). For a review of the applications of ML to astrophysical problems, we refer the reader to Ball & Brunner (2010) and Ivezić et al. (2014). For other practical applications of these methods, see also Ball, Brunner & Myers (2008), Hoyle et al. (2015a,b), Zitlau et al. (2016), Bellinger et al. (2016) and Kamdar, Turk & Brunner (2016) for an ML framework applied to cosmological simulations and Jensen et al. (2016) for an ML approach to measure the escape fraction from galaxies in the EoR (Epoch of Reionization).

The main purpose of this work is to reconstruct key physical parameters of distant galaxies (metallicity, column density, ionization parameter, density) once their spatially resolved spectra have been acquired. Our aim is to maximize the information that can be extracted from such data by using not only few specific and pre-selected emission lines, but the full information encoded in the spectra. This is now possible thanks to the combination of powerful Supervised Machine Learning (SML) algorithms and large synthetic spectra libraries. As we will see in the following, we covered a very large range of plausible physical properties of ISM clouds to accurately describe the physics beyond not only the ionized

⋆ E-mail: graziano.ucci@sns.it

regions but also of other phases (i.e. neutral, molecular) of the ISM in galaxies.

2 OVERVIEW OF EMISSION-LINES DIAGNOSTICS

In this section, we briefly discuss some popular emission-line diagnostics used to estimate specific physical parameters of the ISM of galaxies. For an extensive review on emission lines, we refer the interested reader to Stasińska (2007).

Let us first consider the classical case of density indicators. This is often done by using two similar energy transitions (but different transition probabilities) of a given ion (Osterbrock 1989). Ions (transitions) typically used are the [O II] (λ3726, 3729) or [S II] (λ6716, 6731). In both cases, the transitions are excited from the ground level to two slightly different upper levels; they correspond to different critical densities. The intensity ratio of the two lines is sensitive to gas density.

One of the most popular indicators for the metallicity of the gas is the R_{23} parameter proposed by Pagel et al. (1979). This is defined as follows:

$$R_{23} = \frac{\mathrm{I([O\,II]\,\lambda3727)} + \mathrm{I([O\,III]\,\lambda4959)} + \mathrm{I([O\,III]\,\lambda5007)}}{\mathrm{I(H}\beta)}, \quad (1)$$

where I denotes the emission-line intensity. A problem is that the indicator is non-monotonic, i.e. for a given value of R_{23}, two different metallicity values are possible solutions. To break this degeneracy, additional diagnostics have been proposed. However, most of these methods rely on the [N II] λ6584 line (Pettini & Pagel 2004) that is usually very weak and tends to become difficult to measure especially at low metallicities, i.e. [N II] λ6584/Hα < 0.1 at $Z \lesssim 0.6\,Z_{\odot}$.

Another alternative to the metallicity calibration is the S_{23} parameter, introduced by Díaz & Pérez-Montero (2000). This is based on the use of the sulphur abundance parameter S_{23} (Vilchez & Esteban 1996):

$$S_{23} = \frac{\mathrm{I([S\,II]\,\lambda\lambda6717, 6731)} + \mathrm{I([S\,III]\,\lambda\lambda9069, 9532)}}{\mathrm{I(H}\,\beta)}. \quad (2)$$

With the recent availability of large data sets with known metallicity, empirical calibrations for metallicity diagnostics over a relatively large range of values have been proposed (Nagao, Maiolino & Marconi 2006; Maiolino et al. 2008; Nagao et al. 2011).

Other useful indicators, along with a critical analysis of different techniques, can be found in Kewley & Ellison (2008). These authors suggest the use of calibrators as [N II] λ6584/Hα (Kewley & Dopita 2002) or [N II] λ6584/[O II] (λλ3726, 3729) (Pettini & Pagel 2004), since these two methods give low residual discrepancies in the estimation of metallicities thus providing also new 'revised' formulae for these calibrators.

All the previous works typically use rest-frame optical diagnostics. Therefore, they are considerably affected by dust extinction and additional assumptions must be made in order to apply a differential correction to line intensitT0.000196649-244.84(such,0019669optical)-2745and ln0.1(eta)19.8(-)inramrd

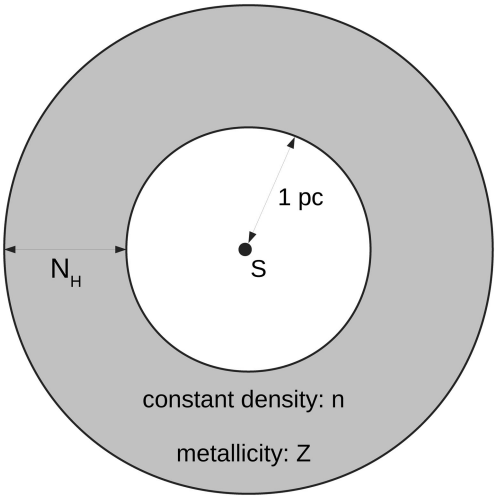

Figure 1. Geometry adopted to build our library. A central source S (with a given ionization parameter U) illuminates the inner part of the cloud located at 1 parsec from the centre. The density (n) and metallicity (Z) of the cloud are assumed to be constant. The outer radius of the cloud is the distance at which our column density ($N_{\rm H}$) reaches the required value for the model under consideration.

3.1 Definitions

The ionization parameter is the number of ionizing photons per hydrogen atoms and we adopt the following definition (e.g. Yeh & Matzner 2012):

$$U = \frac{1}{4\pi R_{\rm S}^2 nc}\int_{\nu_e}^{\infty}\frac{L_\nu}{h\nu}{\rm d}\nu = \frac{Q(H)}{4\pi R_{\rm S}^2 nc}, \quad (6)$$

where $Q(H)$ is the ionizing photon flux, c the speed of light and R_S is the Strömgren radius (Strömgren 1939):

$$R_{\rm S}^3 = \frac{3Q(H)}{4\pi n^2\alpha(T)}, \quad (7)$$

where α is the temperature- (T) dependent recombination rate. By combining equations (6) and (7), we obtain:

$$U = \frac{1}{c}\sqrt[3]{\frac{Q(H)n\alpha^2}{36\pi}}. \quad (8)$$

The minimum ionization parameter inside a H II region generated by a single early-spectral type star can be estimated as follows. If we assume a density $n \sim 100\,{\rm cm}^{-3}$, a recombination rate $\alpha \sim 2.6\cdot10^{-13}\,{\rm cm}^3{\rm s}^{-1}$ and consider the $Q(H)$ from a star with $Z_\star \simeq {\rm Z}_\odot$ and mass $M_\star = 10\,{\rm M}_\odot$, we get $\log U_{\rm min} \sim -4.0$ (Schaerer 2002). Hence, this justifies the minimum value for $\log U$ in Table 1.

For our CLOUDY calculations, we consider a spherical geometry and static conditions. A central ionizing source illuminates the inner part of the cloud situated at distance 1 pc from the centre (see Fig. 1). The outer part of the cloud defines the end of the calculation. For each model (n = cost), the outer cloud distance is set in order to reach the required value for the column density (see Table 1). We have therefore removed the default stopping criterion based on a lower limit on the gas kinetic temperature (4000 K).

Metal abundances for all calculations are assumed to be solar (Grevesse et al. 2010). For dust, we consider contribution from graphite and silicate components that reproduce the observed overall extinction properties of the MW ISM: $R_V = A_V/E(B-V) = 3.1$. The grain size distribution is described by a power-law distribution (Mathis, Rumpl & Nordsieck 1977) resolved by default in CLOUDY into 10 size bins. We did not include the contribution from polycyclic aromatic hydrocarbons (PAHs) in the library used in this work. Observations for local galaxies with lower metallicities (e.g. IZw18 and SBS0335-052) show in fact a suppressed PAH emission (Wu et al. 2007, 2009; Hunt et al. 2010) and these are the prototypes for the high-redshift galaxies ($z \sim 6$) we are interested in (see Section 4.3). The temperature within the cloud is computed by CLOUDY from the balance between heating and cooling.

3.2 Stellar model spectra

The flux illuminating the cloud defines the ionization parameter U (see equation 6) that can be computed by using the STARBURST99 code (Leitherer et al. 1999; Smith, Norris & Crowther 2002; Leitherer et al. 2014). This code computes theoretical spectral energy distributions (SED) combining a stellar atmosphere model with grids of stellar evolutionary tracks that account for different metallicities, star formation histories, IMF and age of stellar populations.

We adopted a Salpeter IMF (Salpeter 1955) in the 0.1–120 $\rm M_\odot$ mass range and a constant SFR of 1 $\rm M_\odot\,yr^{-1}$. The metallicities of the input spectra considered are $Z = 0.001$, 0.008, 0.020 ($\rm Z_\odot$) and 0.040. We calculated models using the set of evolutionary tracks produced by the Geneva group (Schaller et al. 1992). We use the 'Lejeune/Schmutz' option that adopts the extended model atmospheres of Schmutz, Leitherer & Gruenwald (1992) in the case of stars with strong winds and the plane-parallel atmospheric grid of Lejeune, Cuisinier & Buser (1997) otherwise. We fix the stellar population cluster age to 10 Myr. Note that the shape of the EUV spectrum ($100 < \lambda < 1000$ Å) does not change for older stellar populations (Kewley et al. 2001).

STARBURST99 spectra obtained with these prescriptions are given as input to the CLOUDY code. Different examples of the emerging spectra are reported in Fig. 2, for $Z = 0.005\,{\rm Z}_\odot$ (upper panel) and $Z = 0.5\,{\rm Z}_\odot$ (lower panel) and different values of the ionization parameter ($\log U = -2$, 0, 1), at fixed $n = 10^2\,{\rm cm}^{-3}$ and $N_{\rm H} = 10^{20}\,{\rm cm}^{-2}$. The spectrum includes the stellar and dust continuum with the superimposed emission lines of hydrogen, helium and the major elements commonly found in the ISM of galaxies. The figure also shows that the IR/FIR peak due to the dust continuum emission is shifting towards larger wavelengths (i.e. colder dust) at decreasing ionization parameter.

Each CLOUDY output spectrum is then labelled with its corresponding parameters and stored in the library. Then, we implement and train an SML algorithm, described in the following section, which allows us to recover the parameters associated with any given input spectrum.

4 ML METHODS

In this section, we describe the SML approach used in this work and we briefly review its main algorithms, namely Decision Trees and AdaBoost. The main idea of SML is that an observable quantity (i.e. a spectrum) is a set of xs (e.g. spectral lines) that we relate to a set of ys (i.e. the four physical properties $n, Z, N_{\rm H}$ and U). The task is to use a training set in order to find an algorithm $f(x)$ such that for future x in a test set, it will be a good predictor of y.

The SML method tries to infer the physical properties of a given input from labelled data. The training data set consists of a set of input features (i.e. an input vector) and a set of labels (i.e. the desired output values) for each example (see Fig. 3). The SML algorithm analyses the training data set and produces a model that ideally should give as output the same labels required for the training

A common method to produce an ensemble of base learners is the Adaptive Boosting or AdaBoost method (Freund & Schapire 1997; Drucker 1997; Hastie, Tibshirani & Friedman 2009). In the case of Decision Trees as the base learner, the algorithm adds trees sequentially to generate an ensemble of them. AdaBoost iteratively improves the base algorithm by accounting for the incorrectly classified examples in the training set.

First of all, equal weights are assigned to each training examples. At each step of the iteration, the base algorithm is applied to the training set and the weights of the incorrectly classified examples are increased. In each step, the base learner is applied on the training set with updated weights, and after n iterations, the final model is obtained as the weighted sum of the n learners.

4.3 Input features

Starting from a given synthetic SED (see Fig. 2) obtained with the photoionization code CLOUDY, it is possible to obtain the continuum-subtracted intensities of the emission lines. The input vector for the ML algorithms (the input features in Fig. 3) is in our case a collection of intensities associated with a discrete wavelength array and a label containing the four physical properties ($n, Z, N_{\rm H}$ and U) of the model under consideration.

The range of wavelengths used to construct the model spans from 1216 Å (corresponding to the Lyα transition) to 4.0 μm. This particular choice would be equivalent to the rest-frame range in wavelengths obtained by combining the NIRSpec (0.6–5 μm) and MIRI (5–28 μm) instruments on board the James Webb Space Telescope (JWST) and observing a source located at redshift $z \sim 6$. Emission lines from warm/neutral gas are relatively weak but yet observable. For example, [N I] λ5200 and [O I] λ6300,6364 have been observed in the spectra of local galaxies (Moustakas & Kennicutt 2006; Cresci et al. 2015). The strength of our method relies on the fact that the ML algorithm can learn from all the lines present in a spectrum, including the weakest one as those coming from the neutral ISM components. It will then possible to provide at least some constraints on the properties of these phases from observed spectra.

The SML code implementing all the above features will be referred from now on as GAME (GAlaxy Machine learning for Emission lines).

5 RESULTS

5.1 Predictive accuracy

In this section, we present the results of the tests for the AdaBoost SML algorithm in terms of the predicted values of ($n, Z, N_{\rm H}, U$).

The data set used to train GAME consists in a library of 3×10^4 models chosen by randomly selecting the values of the four physical parameters in the ranges reported in Table 1. The data set used for the test (i.e. to predict the labels) consists instead of a [test] sample of 3×10^3 models, also randomly constructed. Thus, although the way of constructing this testing sample is the same used to construct the training data set, the AdaBoost algorithm had never seen these objects before.

The results of the predictive GAME tests performed by using the AdaBoost with Decision Tree as base learner are shown in Fig. 4. The fraction of models for which the actual (i.e. the known values used to generate the testing data set) and predicted values deviate by a factor >2 are 14.8 per cent (n), 1.2 per cent (Z), 1.7 per cent ($N_{\rm H}$) and 23.2 per cent (U). The lower quality predictive performances are somewhat expected. In fact, the determination of the ionization parameter is particularly challenging: it involves the reconstruction from the emerging filtered spectrum of the original U value at the source, which is highly degenerate with the $N_{\rm H}$. GAME delivers top-quality predictions for Z and $N_{\rm H}$, which are almost perfectly recovered by the algorithm.

A different way to appreciate GAME predictive performances is to look at the probability distribution function (PDF, Fig. 5) of the fractional variation between predicted and true physical properties, defined as $\delta_X = \log(X_{\rm PRED}/X_{\rm TRUE})$. In each plot of Fig. 5, the blue shaded area contains 95 per cent of the models. As previously mentioned, the best predictive performances are achieved for $N_{\rm H}$ and Z. For example, for the column density, 95 per cent of the predicted values differ by <60 per cent from the actual ones.

5.2 Weak lines

Up to now, we have considered idealized synthetic spectra. They are idealized because of their 'infinite SNR ratio': they exhibit in fact weak emission lines that would be extremely hard to detect experimentally (e.g. up to 10^5 times fainter than the Hα or [O III] lines). Hence, in our library of synthetic spectra, on average, 500 emission lines are available (i.e. with intensities different from zero). Although GAME does not use all these lines because some of them are unimportant for the construction of the Decision Trees, this idealized situation is unlikely to be reached in a typical observational setup.

We therefore investigated how GAME behaves when a 'detection threshold' is added. We consider null the emission from lines whose intensity is less than a certain fraction $f(\rm H\alpha)$ of the Hα, therefore excluding them from the model. Formally, this approach is equivalent to consider a spectrum whose signal-to-noise ratio (SNR) is SNR $=\sigma/f(\rm H\,\alpha)$, where σ is the rms noise.

In Fig. 6, we report the mean number of available lines in our grid of synthetic spectra as a function of $f^{-1}(\rm H\,\alpha)$. The red dashed line in the figure is the threshold $f(\rm H\,\alpha) = 50$ that we have adopted as a reference in this work. This value is easily reached with state-of-the-art instruments and it is typical of several spectra obtained in observations (Cresci et al. 2015).

We construct two new data sets with spectra that contains only emission lines with an intensity higher than 1/50 of the Hα line. One set is used for the training and the other for the testing phase.

Results of this approach are shown as scatter plots in Fig. 7 and the resulting PDFs are shown in Fig. 8. The fraction of models for which the 'real' (i.e. the known values used to generate the testing data set) and predicted values deviate by a factor >2 are 21.9 per cent (density), 2.6 per cent (metallicity), 2.5 per cent (column density), and 27.3 per cent (ionization parameter). For $f(\rm H\alpha) > 100$, results become similar to the ideal case.

We emphasize that GAME is easy to implement as well as extremely fast even on a laptop computer. Typical run times to train models using our library of 3×10^4 spectra is about 10 min for a single processor run. Therefore, accurately training GAME based on the SNR of the observed input spectra presents no difficulties.

5.3 Sum of different phases

A line of sight (los) passing through a galaxy can cross different ISM phases, e.g. cold neutral medium (CNM), warm neutral medium (WNM), warm ionized medium (WIM) or a dense giant molecular cloud (GMC). The resulting spectrum is then the sum of the spectra

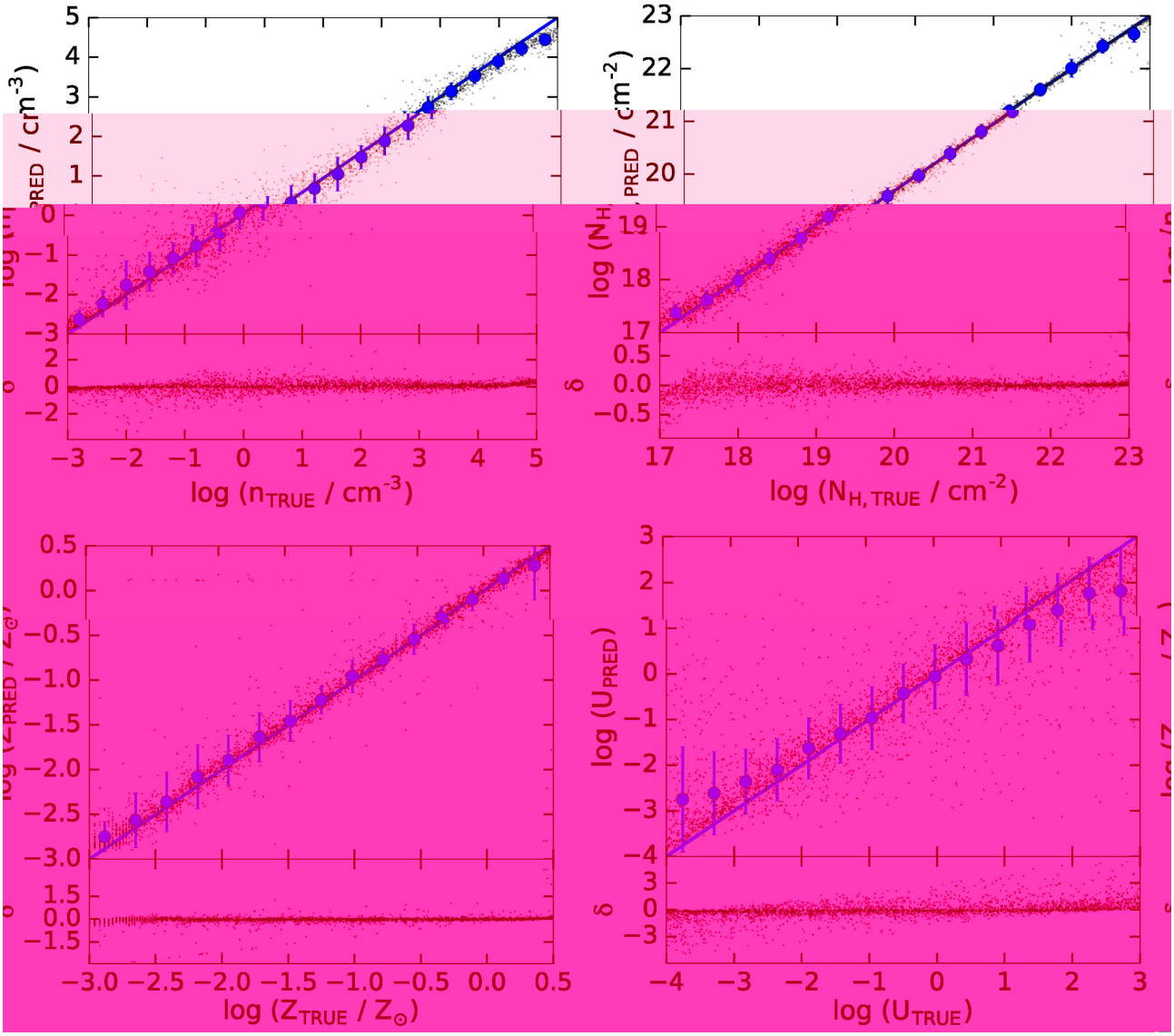

Figure 4. Scatter plots for the predicted (inferred from the model) versus true (used to construct the spectrum) physical properties (density, column density, metallicity and ionization parameter). The blue solid line is the locus of points where true = predicted. The error bars for the binned data (red) show the standard deviation of distribution. δ represents the residuals Δlog (PRED−TRUE) = log (PRED/TRUE).

from individual phases. This is the common case for observations, hence it is important to test GAME in these conditions.

To this aim, we first select spectra from the library labelled with parameters close to the typical values for ISM phases: WNM ($n \sim 1$ cm^{-3}, $U \sim 10^{-4}$), CNM ($n \sim 10^2$ cm^{-3}, $U \sim 10^{-4}$), WIM ($n \sim 10^{-2}$ cm^{-3}, $U \sim 10^{2}$) and GMC ($n \sim 10^3$ cm^{-3}, $U \sim 10^{-4}$). The size of the CNM, WNM and WIM phases is $l \sim 20$ pc, while for GMC, we assumed a size of $l \sim 2$ pc (Larson 1981; Falgarone, Puget & Perault 1992; Ossenkopf & Mac Low 2002; Heyer et al. 2009). Then we sum a variable number of spectra into the final one:

$$S^j(\lambda, Z, N_{\rm H}) = \sum_i S_i^j \left(\lambda, n^j, Z^j, N_H^j, U^j\right), \quad (9)$$

where j labels the phase (j = CNM, WNM, WIM, GMC) and $i = 1, \ldots, N$ is the ith component along the los. We have run cases with $Z \simeq 1$ and $Z \simeq 0.02\,Z_\odot$.

The comparison between the $Z, N_{\rm H}$ values inferred by GAME for the final spectrum and the ones of the individual phases is shown in Fig. 9. GAME performs quite well recovering the emission-weighted values that are intermediate between the inputs of the individual components. This has been verified for all four phases and independently of the assigned mean Z.

As a next step in complexity, we tested a spectrum that is a random combination of different phases along an los (Fig. 10). As GAME returns a single quadruple of $(n, Z, N_{\rm H}, U)$ values, the outcome is biased towards the phases characterized by a larger column density. In other words, GAME is most sensitive to the emitting phase with the largest gas mass.

Although with this limitation, the outcome is very satisfactory as the inferred $(Z, N_{\rm H})$ values are well within the range of the individual phases.

5.4 Comparison with calibrated diagnostics

We compare the calibration of two popular metallicity diagnostics, R_{23} (see equation 1) and [N II] λ6584/Hα, with GAME, in order to

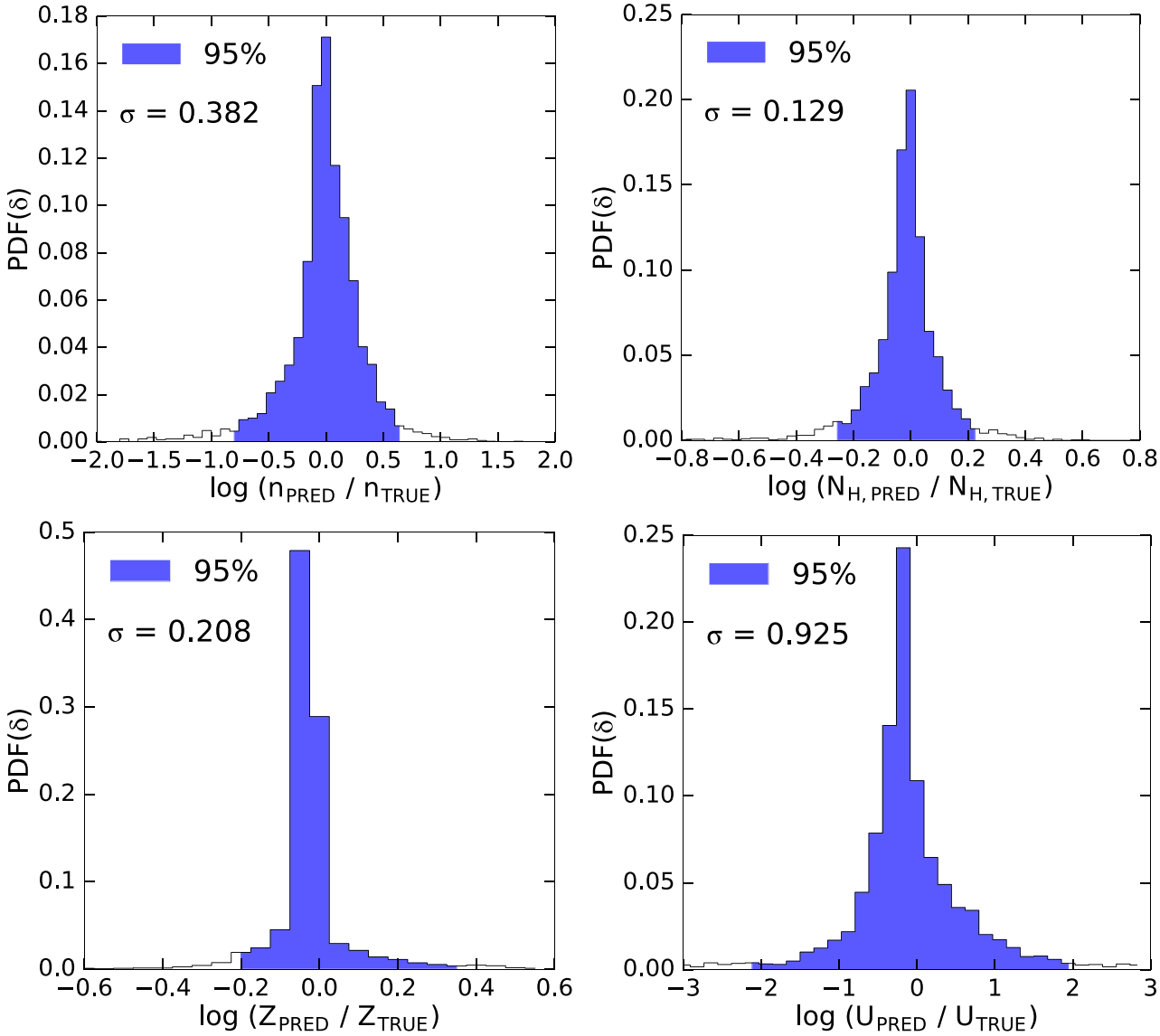

Figure 5. Probability distribution functions (PDF) of the δs reported in Fig. 4, i.e. the logarithm of the ratios between the predicted and true physical properties. σ represents the standard deviation of the distributions. 95 per cent of the models reside within the shaded blue regions.

evaluate the relative performances. As already mentioned, these two diagnostics are based on nebular emission lines and they have also been calibrated on H II regions. For the comparison to be meaningful, we have chosen in our library spectra with physical properties describing standard ionized nebulae: $T \sim 10^4$ K, ionized hydrogen fraction >90 per cent, $N_H \lesssim 10^{20}$ cm^{-2}.

In Fig. 11, we show the values of R_{23} and [N II] λ6584/Hα for generic models (grey dots), and those representing an H II region (blue circles) for $10^{-2} < Z/Z_\odot < 3$. The blue solid lines are the empirical calibrations given by Maiolino et al. (2008) for R_{23}:

$$\log(R_{23}) = 0.7462 - 0.7149x - 0.9401x^2 - 0.6154x^3 - 0.2524x^4 \quad (10)$$

and for [N II] λ6584/Hα:

$$\log\left[\frac{\mathrm{F([N\,II]\,\lambda6584)}}{\mathrm{F(H\,\alpha)}}\right] = -0.7732 + 1.2357x - 0.2811x^2 - 0.7201x^3 - 0.3330x^4, \quad (11)$$

where $x = \log(Z/Z_\odot) = 12 + \log(\mathrm{O/H}) - 8.69$ and F are the reddening-corrected fluxes. As pointed by Maiolino et al. (2008), the previous relations are strictly valid only in the range $7.0 \lesssim 12 + \log(\mathrm{O/H}) \lesssim 9.2$. Outside this metallicity range, the use of these relations relies on extrapolation.

The dashed lines represent theoretical calibrations based on the grid of photoionization models provided by Kewley & Dopita (2002) for an ionization parameter $U = 1.6 \times 10^{-4}$ (red dashed lines in Fig. 11):

$$\log(R_{23}) = -27.0004 + 6.0391y - 0.327\,006y^2, \quad (12)$$

$$\log\left[\frac{\mathrm{F([N\,II]\,\lambda6584)}}{\mathrm{F(H\,\alpha)}}\right] = -2700.08 + 1335.14y - 247.533y^2 + 20.3663y^3 - 0.626\,92y^4; \quad (13)$$

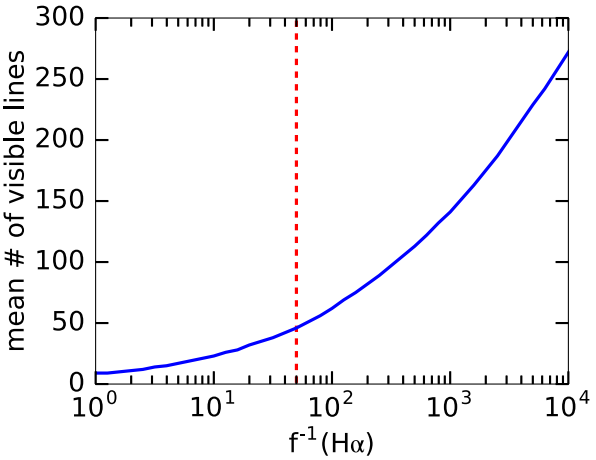

Figure 6. Mean number of available lines in our library of synthetic spectra as a function of $f^{-1}(\mathrm{H}\alpha)$, the fraction of the Hα line intensity used as a threshold. For $f(\mathrm{H}\alpha) = 1/50$ (red dashed line), the mean number of available lines is 50.

and $U = 10^{-2}$ (black dashed lines in Fig. 11):

$$\log(R_{23}) = -45.6075 + 11.2074y - 0.674\,460y^2, \quad (14)$$

$$\log\left[\frac{\mathrm{F([N\,II]}\,\lambda6584)}{\mathrm{F(H}\,\alpha)}\right] = -3100.57 + 1501.77y - 272.883 * y^2 + 22.0132y^3 - 0.6646y^4, \quad (15)$$

where $y = 12 + \log(\mathrm{O/H})$.

Some differences are present between our H II region models (blue circles) and the considered theoretical and empirical calibrations (solid and dotted lines). These are due to the different photoionization code used, the assumption of spherical versus plane-parallel geometry, isochoric versus isobaric assumptions for the gas. However, they are relatively minor.

We compute in this case Z inferred from two different kinds of H II region models: (') spectra for which the R_{23} value is very close to the empirical calibration (blue solid line in Fig. 11, taken as reference) and (2) spectra whose R_{23} strongly differs from the calibration.

For case (a) (red panel of Fig. 11), the 'true' value (the one used to generate the spectrum) and the GAME-predicted one

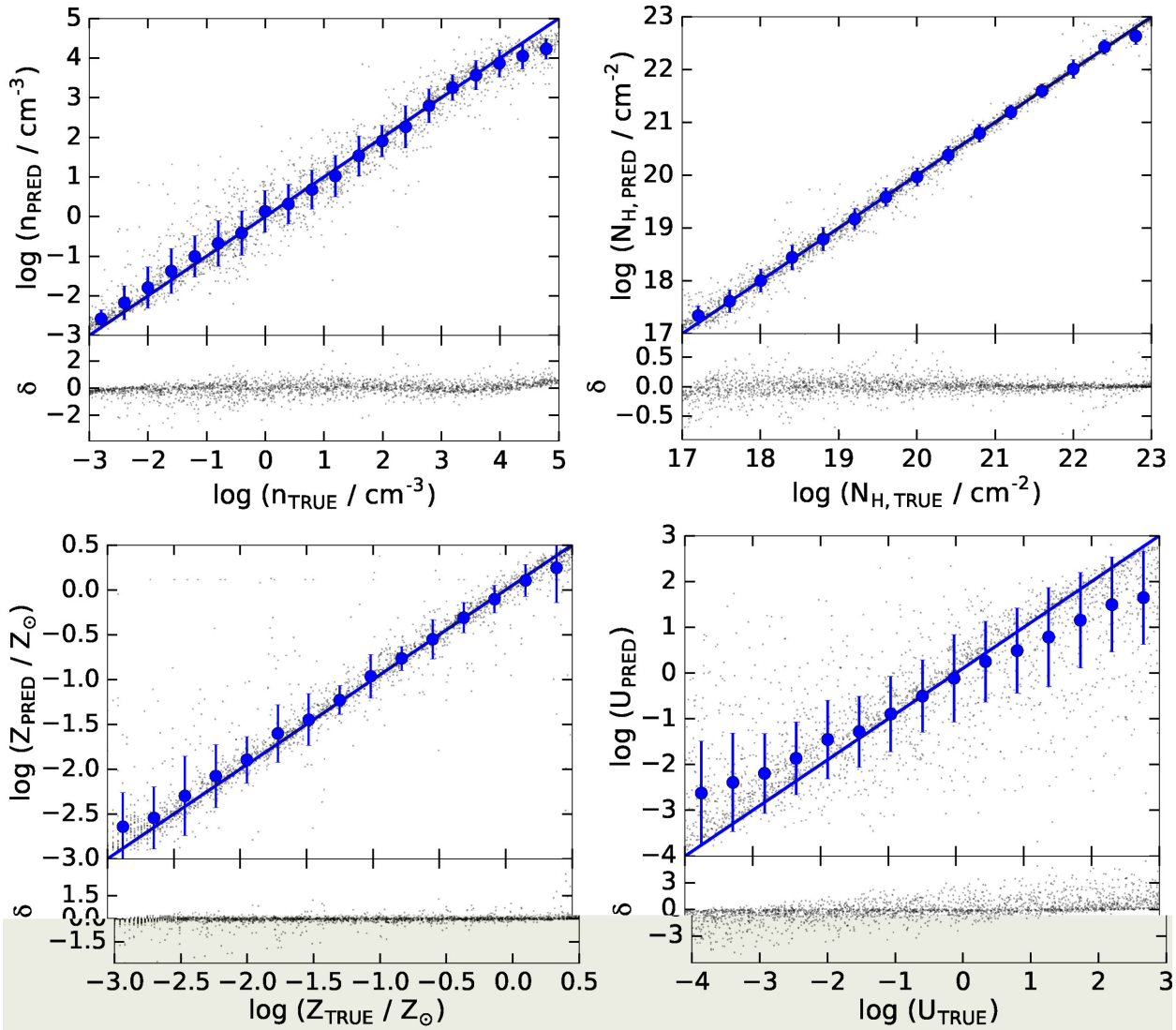

Figure 7. As in Fig. 4, but considering a threshold $f(\mathrm{H}\alpha) = 1/50$ (see the text for the details).

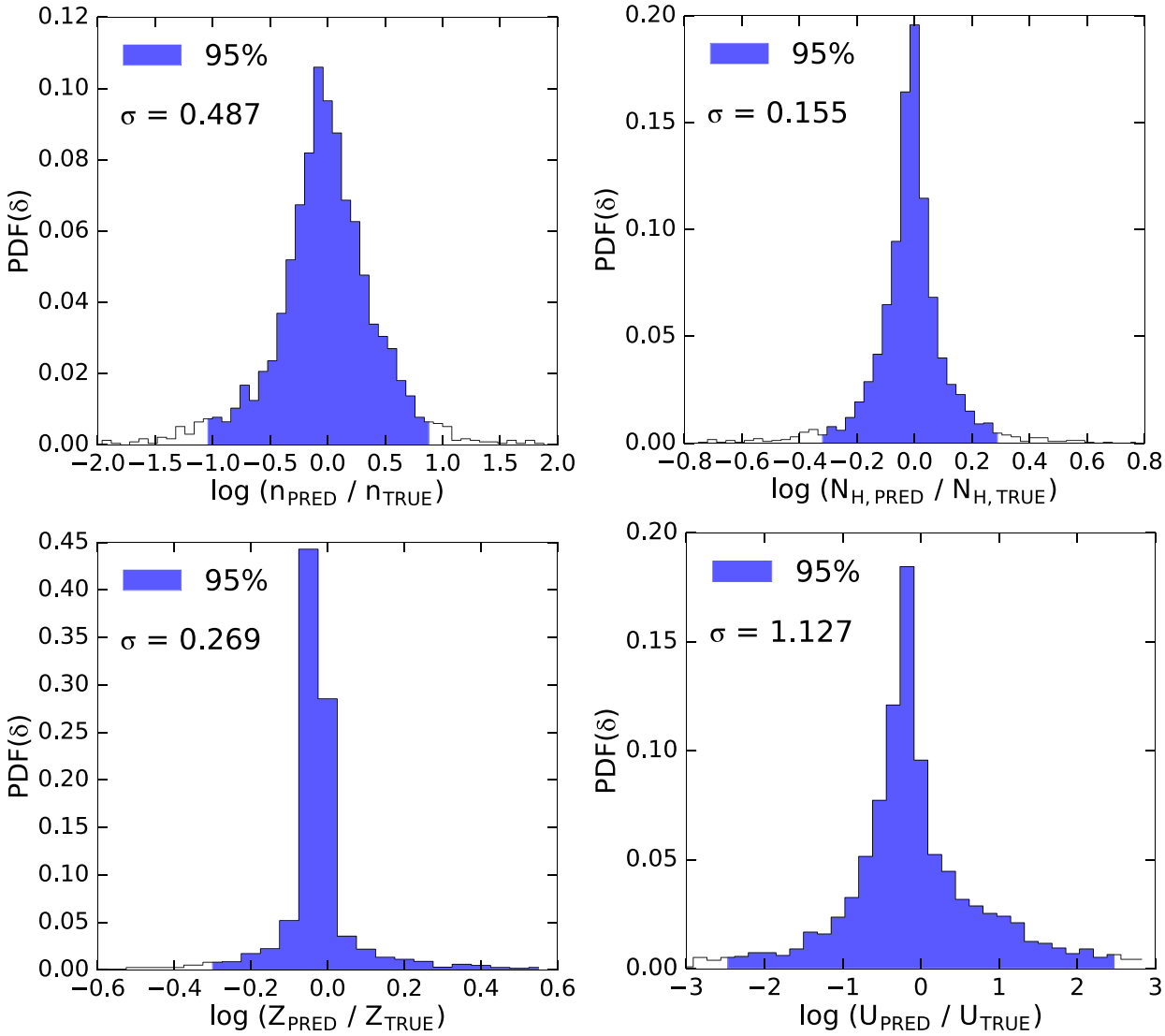

Figure 8. As in Fig. 8, but considering a line intensity threshold $f(\mathrm{H}\,\alpha) = 1/50$ (see the text for the details).

are $Z_{\mathrm{TRUE}} = 0.332$ and $Z_{\mathrm{PRED}} = 0.365\,\mathrm{Z}_{\odot}$; the difference is <10 per cent. For the same case though, the metallicity inferred using calibrations based on the R_{23} and [N II] λ6584/Hα diagnostic are instead $Z_{23} = 0.660$, and $Z_{[\mathrm{N\,II}]} = 0.507\,\mathrm{Z}_{\odot}$, respectively. We see that the performance is much worse than the one delivered by GAME. In fact, the empirical methods predict values in excess by almost a factor of 2 with respect to the actual value.

For case (b) (green panel of Fig. 11), the true and the predicted value are, respectively, $Z_{\mathrm{TRUE}} = 0.0792$ and $Z_{\mathrm{PRED}} = 0.1027\,\mathrm{Z}_{\odot}$, i.e. still in good agreement. However, using the calibrations, we get instead $Z_{23} = 2.384$ and $Z_{[\mathrm{N\,II}]} = 0.850\,\mathrm{Z}_{\odot}$, i.e. they both largely overestimate Z by more than one order of magnitude.

We conclude that GAME appears to be able to extract physical conditions of the gas in a much more precise and reliable way with respect to standard indicators.

Finally, we stress that for the spectra not arising from H II regions (green dots in Fig. 11), the previous calibrators cannot be applied. This happens since it does not longer exist a correlation between the diagnostic value and Z. An SML approach, like the one implemented by GAME, is however capable to extract information from all detected lines. This allows a range of applications that is not only restricted to H II regions, but can extend to a wide variety of ISM phases.

5.5 Connecting theory and observations with GAME

The most natural use of GAME is to extract the physical properties of galaxies from spectroscopic emission-line observations of galaxies. The full power of the code is manifest when it is used in combination with spatially resolved spectroscopy, like the one obtained from IFUs. From such data cubes, it will be possible to readily and robustly obtain physical properties (density, metallicity, ionization parameter) that represent key information to understand galaxy evolution. As new instruments, like MUSE, and telescopes (JWST, ALMA, TMT, E-ELT), will be able to obtain for the first time this type of data also on high-redshift galaxies, a more comprehensive approach to interpret the spectra, as the one presented here, is mandatory to completely exploit their power.

GAME can be also applied to synthetic maps constructed from galaxy simulations. The aim of this procedure is twofold. On one

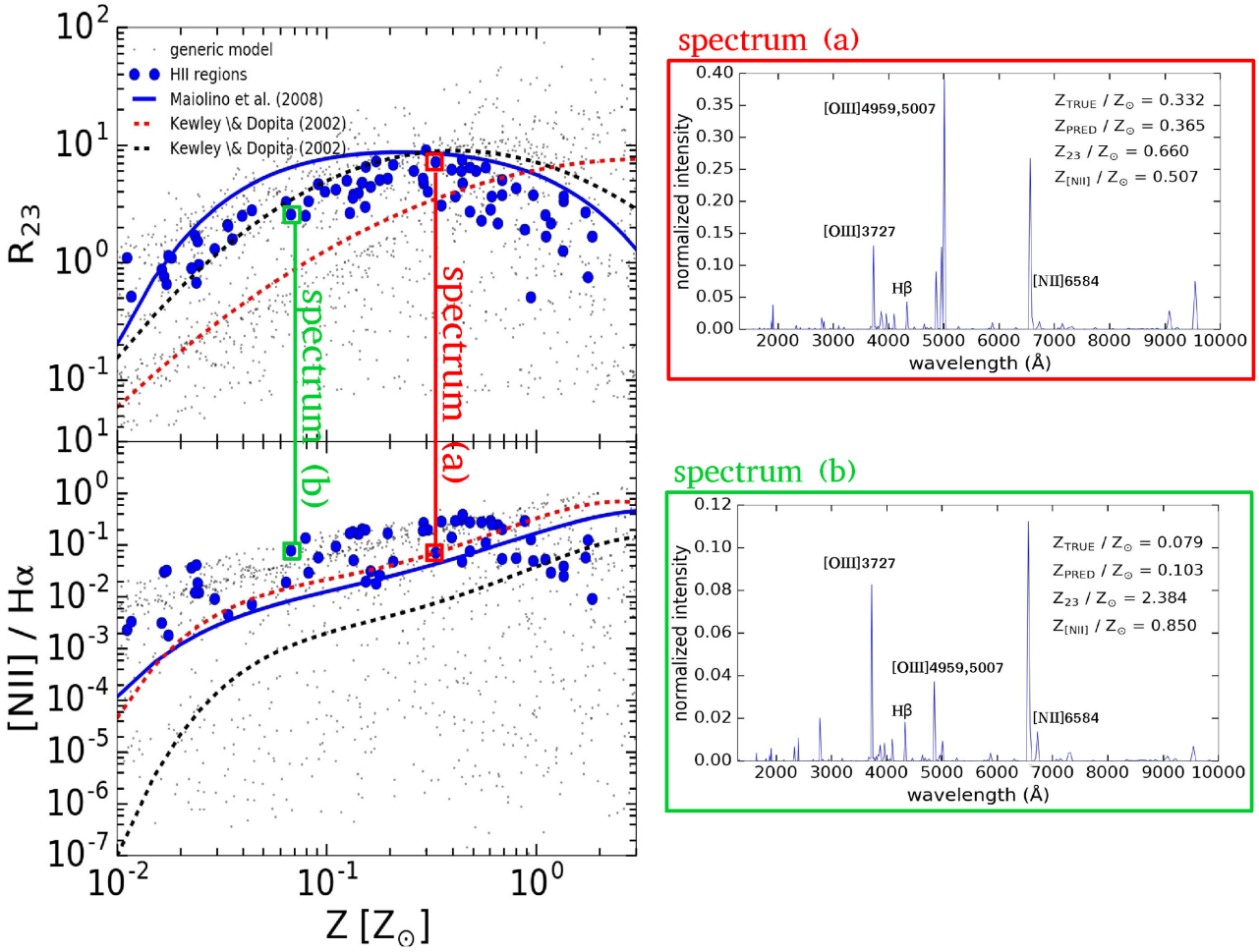

Figure 11. Relations between emission-line ratios and gas metallicity. Green dots are generic models in our library and blue circles are models representing H II regions. The blue solid line shows the empirical calibrations from Maiolino et al. (2008). The red and black dashed lines are the theoretical calibrations reported in Kewley & Dopita (2002), respectively, for a $U = 1.6 \cdot 10^{-4}$ and 10^{-2}, respectively. The red panel (a) shows a spectrum for which the calibrations give a value for the metallicity near to the true value used to generate the model. The green panel (b) shows instead a spectrum for which the calibrations are not good indicators for the metallicity.

to deal with complex, multiphase los, obtaining very satisfactory answers (for details, see Section 5.3).

We emphasize that GAME is easy to implement as well as extremely fast even on a laptop computer. Typical training times using our library of 3×10^4 spectra are ≈10 min and the time required to infer physical parameter values for a given input spectrum is only less than few seconds. Therefore, accurately training GAME to accommodate the specific SNR of the observed input spectra presents no difficulties.

It is worthwhile to add some remarks concerning the comparison with currently used methods, based on emission-line ratios. Different galaxy properties can result in almost equal line ratio for some of the lines. Thus, more line ratios are generally required to break this degeneracy.

The SML approach can overcome these difficulties because (1) it makes use of all the available information present in the spectrum simultaneously, meaning that it is not necessary to choose a priori a subset of lines to use; (2) the training phase is extremely fast and the code can easily adapt to new conditions (e.g. a different SNR ratio of the spectrum, see Section 5.2). The most fundamental aspect is that without any calibration, once trained, the algorithm provides an estimate of the main physical properties with no degeneracy.

Furthermore, we can compare the SML method adopted in this work with the 'traditional' data fitting technique. For an excellent review on the two cited approaches ('algorithmic' versus 'traditional'), we refer the interested reader to the work by Breiman (2001). There are advantages and disadvantages in both these techniques and one should consider the best suited for the science case under study.

Both approaches are based on a strong underlying assumption: the model,[2] the chosen physical properties range and the prescriptions used to generate the library do capture the essential physics governing the ISM. It must be stressed that to get an accurate estimate of the ISM physical properties, one must explore the largest possible range of parameter values when producing the library (see Table 1). This is not always easy and the resulting grid from all the possible combinations of these values can be very large. Although model fitting techniques or Bayesian approaches (Blanc et al. 2015) are very powerful, they suffer from some limitations. The best way

[2] For example, photoionization codes alternative to CLOUDY are MAPPINGS (Sutherland & Dopita 1993; Allen et al. 2008) and MOCASSIN (Ercolano et al. 2003; Ercolano, Barlow & Storey 2005).

to constrain a particular model is in fact to use as many observational constraints as possible. For a Bayesian approach, this can be a problem because using hundreds of features at one time is extremely time consuming. Moreover, adapting a code to deal with an observational spectrum with different wavelength range or with a different SNR ratio can be computationally very expensive.

In this context, an important advantage of the SML method with respect to the Bayesian one is that its performance is not affected by the finite number of models within the library used during the training. In fact, the SML technique, as suggested by the name itself, is capable to 'learn' and explore hidden patterns within the library parameter space. In other words, the SML method is not limited to 'recover' parameter values included in the library, but it can also 'predict' results that are not part of the original one.

Moreover, GAME can be effectively coupled to the analysis of IFU spectroscopic data and synthetic data from numerical galaxy simulations. These applications will be demonstrated in a forthcoming study.

We finally point out that in addition to H II regions, GAME allows us to infer the physical properties of photodissociation regions. These are now being resolved in nearby galaxies. With JWST and ALMA, we will be able to obtain comparable results also for the high-redshift systems.

ACKNOWLEDGEMENTS

We thank D. Cormier, L. Vallini and S. Viti for useful discussions and K. Volk for help with CLOUDY. We thank also the referee G. Ferland for insigthful comments. This research was supported in part by the National Science Foundation under Grant No. NSF PHY11-25915.

REFERENCES

Allen M. G., Groves B. A., Dopita M. A., Sutherland R. S., Kewley L. J., 2008, ApJS, 178, 20
Ball N. M., Brunner R. J., 2010, Int. J. Mod. Phys. D, 19, 1049
Ball N. M., Brunner R. J., Myers A. D., Tcheng D., 2006, ApJ, 650, 497
Ball N. M., Brunner R. J., Myers A. D., 2008, in Argyle R. W., Bunclark P. S., Lewis J. R., eds, ASP Conf. Ser. Vol. 394, Astronomical Data Analysis Software and Systems XVII. Astron. Soc. Pac., San Francisco, p. 201
Bellinger E. P., Angelou G. C., Hekker S., Basu S., Ball W., Guggenberger E., 2016, ApJ, 830, 31
Blanc G. A., Kewley L., Vogt F. P. A., Dopita M. A., 2015, ApJ, 798, 99
Breiman L., 2001, Stat. Sci., 16, 199
Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees. Wadsworth International Group, Belmont, CA
Calzetti D., 2008, in Knapen J. H., Mahoney T. J., Vazdekis A., eds, ASP Conf. Ser. Vol. 390, Pathways Through an Eclectic Universe. Astron. Soc. Pac., San Francisco, p. 121
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Cavuoti S. et al., 2015, MNRAS, 452, 3100
Collister A. A., Lahav O., 2004, Publ. Astr. Soc. Pac., 116, 345
Cresci G. et al., 2015, A&A, 582, A63
De Looze I. et al., 2014, A&A, 568, A62
Díaz A. I., Pérez-Montero E., 2000, MNRAS, 312, 130
Dietterich T. G., 2000, in Kittler J., Roli F., eds, Multiple Classifier Systems, LBCS-1857. Springer-Verlag, Berlin Heidelberg, p. 1
Drucker H., 1997, Improving Regressors using Boosting Techniques. Morgan Kaufmann Publishers Inc., San Francisco, p. 107
Ercolano B., Barlow M. J., Storey P. J., Liu X.-W., 2003, MNRAS, 340, 1136
Ercolano B., Barlow M. J., Storey P. J., 2005, MNRAS, 362, 1038
Falgarone E., Puget J.-L., Perault M., 1992, A&A, 257, 715
Ferland G. J. et al., 2013, Rev. Mex. Astron. Astrophys., 49, 137
Freund Y., Schapire R. E., 1997, J. Comp. Syst. Sci., 55, 119
Grevesse N., Asplund M., Sauval A. J., Scott P., 2010, Ap&SS, 328, 179
Hastie T. J., Tibshirani R. J., Friedman J. H., 2009, The Elements of Statistical Learning : Data mining, Inference, and Prediction. Springer, New York
Heyer M., Krawczyk C., Duval J., Jackson J. M., 2009, ApJ, 699, 1092
Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015a, MNRAS, 450, 305
Hoyle B., Rau M. M., Paech K., Bonnett C., Seitz S., Weller J., 2015b, MNRAS, 452, 4183
Hunt L. K., Thuan T. X., Izotov Y. I., Sauvage M., 2010, ApJ, 712, 164
Ivezić Z., Connelly A. J., VanderPlas J. T., Gray A., 2014, Princeton Series in Modern Observational Astronomy, Statistics, Data Mining, and Machine Learningin Astronomy. Princeton Univ. Press, Princeton
Jensen H., Zackrisson E., Pelckmans K., Binggeli C., Ausmees K., Lundholm U., 2016, ApJ, 827, 5
Kamdar H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 455, 642
Kennicutt R. C., Jr1998, ARA&A, 36, 189
Kewley L. J., Dopita M. A., 2002, ApJS, 142, 35
Kewley L. J., Ellison S. L., 2008, ApJ, 681, 1183
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Kim E. J., Brunner R. J., Carrasco Kind M., 2015, MNRAS, 453, 507
Kroupa P., 2001, MNRAS, 322, 231
Larson R. B., 1981, MNRAS, 194, 809
Leitherer C. et al., 1999, ApJS, 123, 3
Leitherer C., Ekström S., Meynet G., Schaerer D., Agienko K. B., Levesque E. M., 2014, ApJS, 212, 14
Lejeune T., Cuisinier F., Buser R., 1997, A&ASupp., 125
Levesque E. M., Kewley L. J., Larson K. L., 2010, ApJ, 139, 712
López-Sánchez Á. R., Dopita M. A., Kewley L. J., Zahid H. J., Nicholls D. C., Scharwächter J., 2012, MNRAS, 426, 2630
Maiolino R. et al., 2008, A&A, 488, 463
Mathis J. S., Rumpl W., Nordsieck K. H., 1977, ApJ, 217, 425
Mehta M., Rissanen J., Agrawal R., 1995. AAAI Press, pp 216–221.
Mingers J., 1989, Mach. Learn., 4, 227
Moustakas J., Kennicutt R. C., Jr2006, ApJS, 164, 81
Nagao T., Maiolino R., Marconi A., 2006, A&A, 459, 85
Nagao T., Maiolino R., Marconi A., Matsuhara H., 2011, A&A, 526, A149
Nagao T., Maiolino R., De Breuck C., Caselli P., Hatsukade B., Saigo K., 2012, A&A, 542, L34
Ossenkopf V., Mac Low M.-M., 2002, A&A, 390, 307
Osterbrock D. E., 1989, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei. Univ. Science Books, Mill Valley, CA
Pagel B. E. J., Edmunds M. G., Blackwell D. E., Chun M. S., Smith G., 1979, MNRAS, 189, 95
Pallottini A., Gallerani S., Ferrara A., Yue B., Vallini L., Maiolino R., Feruglio C., 2015, MNRAS, 453, 1898
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pettini M., Pagel B. E. J., 2004, MNRAS, 348, L59
Quinlan J. R., 1986, Mach. Learn., 1, 81
Quinlan J. R., 1993, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco
Rubin R. H., Simpson J. P., Lord S. D., Colgan S. W. J., Erickson E. F., Haas M. R., 1994, ApJ, 420, 772
Salpeter E. E., 1955, ApJ, 121, 161
Schaerer D., 2002, A&A, 382, 28
Schaller G., Schaerer D., Meynet G., Maeder A., 1992, A&ASupp., 96, 269
Schmutz W., Leitherer C., Gruenwald R., 1992, PASP, 104, 1164
Smith L. J., Norris R. P. F., Crowther P. A., 2002, MNRAS, 337, 1309

Spinoglio L., Pereira-Santaella M., Dasyra K. M., Calzoletti L., Malkan M. A., Tommasin S., Busquet G., 2015, ApJ, 799, 21
Stasińska G., 2007, preprint (arXiv:0704.0348)
Strömgren B., 1939, ApJ, 89, 526
Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253
Vallini L., Gallerani S., Ferrara A., Pallottini A., Yue B., 2015, ApJ, 813, 36
Vilchez J. M., Esteban C., 1996, MNRAS, 280, 720
Wu Y. et al., 2007, ApJ, 662, 952
Wu Y., Charmandaris V., Houck J. R., Bernard-Salas J., Lebouteiller V., 2009, in Sheth K., Noriega-Crespo A., Ingalls J., Paladini R., eds, The Evolving ISM in the Milky Way and Nearby Galaxies
Yeh S. C. C., Matzner C. D., 2012, ApJ, 757, 108
Zitlau R., Hoyle B., Paech K., Weller J., Rau M. M., Seitz S., 2016, MNRAS, 460, 3152

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