Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **477,** 1484–1494 (2018) doi:10.1093/mnras/sty804
Advance Access publication 2018 March 28

GAME: GAlaxy Machine learning for Emission lines

G. Ucci,[1⋆] A. Ferrara,[1,2] A. Pallottini[1,3,4,5] and S. Gallerani[1]

[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]*Centro Fermi, Museo Storico della Fisica e Centro Studi e Ricerche 'Enrico Fermi', Piazza del Viminale 1, Roma, I-00184, Italy*
[4]*Cavendish Laboratory, University of Cambridge, 19 J. J. Thomson Ave., Cambridge CB3 0HE, UK*
[5]*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*

Accepted 2018 March 22. Received 2018 March 22; in original form 2017 June 8

ABSTRACT

We present an updated, optimized version of GAME (GAlaxy Machine learning for Emission lines), a code designed to infer key interstellar medium physical properties from emission line intensities of ultraviolet /optical/far-infrared galaxy spectra. The improvements concern (a) an enlarged spectral library including Pop III stars, (b) the inclusion of spectral noise in the training procedure, and (c) an accurate evaluation of uncertainties. We extensively validate the optimized code and compare its performance against empirical methods and other available emission line codes (PYQZ and HII-CHI-MISTRY) on a sample of 62 SDSS stacked galaxy spectra and 75 observed HII regions. Very good agreement is found for metallicity. However, ionization parameters derived by GAME tend to be higher. We show that this is due to the use of too limited libraries in the other codes. The main advantages of GAME are the simultaneous use of all the measured spectral lines and the extremely short computational times. We finally discuss the code potential and limitations.

Key words: methods: data analysis – galaxies: ISM – ISM: general.

1 INTRODUCTION

Understanding the structure and physical properties of the interstellar medium (ISM) in galaxies, especially at high redshift, is one of the major drivers of galaxy formation studies. Measurements of key properties as gas density, column density, metallicity, ionization parameter, and Habing flux, rely on galaxy spectra obtained through the most advanced telescopes (both earth-based and space-borne) and, in particular, on their emission lines (Osterbrock 1989; Stasińska 2007; Hammer et al. 2017; Pérez-Montero 2017; Stanway 2016). However, finding diagnostics that are free of significant systematic uncertainties remains an unsolved problem (Bresolin 2017, and references therein).

Today, extensive spectroscopic studies of the redshifted ultraviolet (UV) and optical emission lines of distant galaxies ($z \gtrsim 2$) are limited to relatively small (with respect to the local Universe) samples of star-forming galaxies (Shapley et al. 2003; Erb et al. 2010; Stark et al. 2014). The situation is even worse for higher redshifts ($z \gtrsim 6$) where emission line measurements are limited to few very bright galaxies (Sobral et al. 2015; Stark et al. 2017) and gravitationally lensed objects (Stark et al. 2015a,b). One of the main issues is that for such distant galaxies, the number of UV and optical emission lines available is almost always limited to a couple if not to a single extremely bright line. High-quality rest-frame UV and optical spectra, including also fainter lines of high-z galaxies, have to wait for new facilities such as the James Webb Space Telescope and the Extremely Large Telescope. These instruments will revolutionize the current research, in terms of both quality and amount of data (estimated production rate is petabyte yr^{-1}; Garofalo, Botta & Ventre 2017).

This latter aspect, in particular, will bring the astrophysical community into an era where Machine Learning (ML) algorithms and Big Data Analytics architectures will become fundamental tools in the data-mining process. This is already the case for local observations where, e.g. Integral Field Units (IFUs) are already able to provide observations of local galaxies containing tens of thousands of spaxels (Cresci et al. 2017). The development of new methods will be therefore crucial, especially in the analysis of galaxy spectra, which, combining the efforts from different instruments, will include faint lines arising from very extended wavelength ranges [i.e. from UV to far-infrared (FIR) rest frame].

The usual emission line diagnostics will still represent extremely useful tools, albeit they will be likely unable to reach firm conclusions as they rely on a very limited set of lines. These diagnostics are in fact based on UV/optical line ratios (Pagel et al. 1979; McGaugh 1991; Kewley & Dopita 2002; Kobulnicky & Kewley 2004; Pettini & Pagel 2004; Pilyugin & Thuan 2005; Bresolin 2007; Bresolin et al. 2009a; Marino et al. 2013; Brown, Martini & Andrews 2016; Pilyugin & Grebel 2016). This is mostly because UV/optical

⋆ E-mail: graziano.ucci@sns.it

Recombination Lines (RLs) emitted by the lightest elements (hydrogen and helium) are the strongest feature in galaxy spectra. The analogue ones emitted by heavy element can be extremely weak (up to 10^{3-4} times fainter than the H β 4861 Å line) and thus very difficult to measure in very distant objects. The brightest metal lines are instead the Collisionally Excited Lines (CELs), corresponding to forbidden transitions.

Some recent works have started to explore new diagnostics based also on FIR lines (Nagao et al. 2011; Farrah et al. 2013; De Looze et al. 2014; Pallottini et al. 2015; Vallini et al. 2015; Pereira-Santaella et al. 2017; Vallini et al. 2017; Rigopoulou et al. 2018). Facilities such as the Atacama Large Millimeter/submillimeter Array will allow us to construct catalogues containing not only information on the UV/optical/near-infrared (IR) part of the spectrum but also on the FIR lines (i.e. [C II] λ157 μm, [O I] λ63 μm, [N II] λ122 μm, [O III] λ88 μm). Such FIR lines would perfectly complement UV/optical/near-IR lines. However, to fully exploit the information contained in the combined line set, new methods and libraries of synthetic observations are needed.

In this context, several methods in the literature have been developed that widely use ionized gas to infer galaxy physical properties from emission line intensities (Stasińska 2004). A common method to measure abundances is based on the determination of the electron temperature. The electron temperature T_e can be calculated from auroral to nebular emission-line ratios such as R_{O3} = [O III] λ4959+5007 / [O III] λ4363 (Osterbrock 1989; Pérez-Montero 2017). Alternatively, one can use the ratio of RLs of ions, which shows a weak dependence on T_e and the electron density n_e (Peimbert 2003; Tsamis et al. 2003; Peimbert, Peimbert & Ruiz 2005; García-Rojas & Esteban 2007; López-Sánchez et al. 2007; Esteban et al. 2009; Bresolin et al. 2009b; Esteban et al. 2014; Peimbert & Peimbert 2014; Bresolin et al. 2016). Given that RLs and auroral lines can be extremely weak in faint, distant, or high metallicity objects, other methods that use CELs and Balmer lines have been devised to compute the ISM physical properties. These are referred to as the Strong Emission Line (SEL) methods and they are based on the comparison of theoretical spectra from a grid of photoionization models (McGaugh 1991; Zaritsky, Kennicutt & Huchra 1994; Kewley & Dopita 2002; Kobulnicky & Kewley 2004; Tremonti et al. 2004; Kewley & Ellison 2008; Dopita et al. 2016) or on empirical calibrations obtained for samples for which a previously derived metallicity estimate via the electronic temperature method exists (Alloin et al. 1979; Pagel et al. 1979; Pettini & Pagel 2004; Nagao, Maiolino & Marconi 2006; Maiolino et al. 2008; Nagao et al. 2011; Marino et al. 2013; Pilyugin & Grebel 2016; Curti et al. 2017). In addition to these, there are codes capable to infer the abundance and ionization parameter of H II regions: IZI (Blanc et al. 2015, based on a Bayesian approach), PYQZ (Dopita et al. 2013), HII-CHI-MISTRY (Pérez-Montero 2014), and BOND (Vale Asari et al. 2016).

Within our group, we have developed GAME (GAlaxy Machine learning for Emission lines), a new fast method to reconstruct the physical properties of the ISM by using all the information represented by the emission line intensities present in the whole available spectrum (Ucci et al. 2017, hereafter U17). U17 represented a sort of 'feasibility study': our primary objective was to verify whether ML techniques can be used to predict the physical properties of galaxies. Using the AdaBoost method, we verified that the emission line intensities can effectively provide information on the state of the gas (especially for metallicity).

As a next step, we present here the current (updated and optimized) version of the GAME code (Section 2, for the workflow of the code, see also Appendix A[1]) that is based on a new library of photoionization models (50 000 synthetic spectra). In addition to the technical improvements, we present also a strategy to deal with uncertainties for an ML-based code. We further implemented an approach to include noise in the library during the ML training phase, in order to apply GAME to real spectroscopic observations. Another key result concerns the treatment of the emission line degeneracy with physical properties. This is discussed in the framework of the application of GAME to study the ISM of star-forming galaxies (Section 3). In Section 4, we also test the performances of GAME, comparing it to other methods/codes and against a sample of H II regions with available abundance determinations. We finally discuss the potential and limitations of GAME in Section 5.

Table 1. Range of ISM physical properties used to construct the GAME library.

Parameter	Min value	Max value
$\log(Z/Z_\odot)$	−3.0	0.5
$\log(n/\mathrm{cm}^{-3})$	−3.0	5.0
$\log(U)$	−4.0	3.0
$\log(N_H/\mathrm{cm}^{-2})$	17.0	23.0

2 GAME

GAME is a code that, by using as input spectral emission line intensities, infers key galaxy ISM physical properties (see U17 for full details). It is based on a Supervised Machine Learning algorithm called AdaBoost with Decision Trees as base learner trained with a large library of synthetic spectra.

To generate each spectrum, we ran the photoionization code CLOUDY v13.03 (Ferland et al. 2013), using as input the quadruplet (n, N_H, U, Z), where n is the total hydrogen density, N_H the column density, U the ionization parameter, and Z the metallicity. We assume an oxygen abundance $12 + \log(O/H) = 8.69$ and solar abundance ratios for all the elements[2] (Allende Prieto, Lambert & Asplund 2001; Asplund et al. 2004, 2009). The library covers a large range of physical conditions found in the ISM, as reported in Table 1. The total wavelength coverage of the synthetic spectra in the library ranges from the Lyα (1216 Å) wavelength up to 1 mm. However, GAME is able to deal with any subset of emission lines within the input spectra.

A Decision Tree (Breiman et al. 1984) recursively partitions the data with respect to the input feature space in branches first (i.e. the branches correspond to different regions of the input feature space) and then into an increasing number of 'leaves'. Because it is possible to improve the power of many base learners into 'ensemble learning methods' (Dietterich 2000), we can combine many Decision Trees to make a 'forest'. A common way to produce a forest of Decision Trees is the algorithm called Adaptive Boosting or Adaboost (Drucker 1997; Freund & Schapire 1997; Hastie, Tibshirani & Friedman 2009). AdaBoost improves the performance of a base learner by accounting for the elements in the training set that have large prediction errors. We refer the interested reader to

[1] The Appendices are available online on the Monthly Notices of the Royal Astronomical Society website.

[2] Such assumption can be relaxed, but this would require to build a dedicated library. We plan to explore the effect of peculiar abundances in the future work.

Schapire (1990) and Drucker (1997) for a detailed description of the AdaBoost algorithm.[3]

By running GAME, it is possible to infer four 'default labels': (n, $N_{\rm H}$, U, Z). Besides these default labels (i.e. the physical properties used to generate the CLOUDY photoionization models; see Section 2.1.1), in the GAME library the user can find 'additional labels', as for example the radius of the cloud r, the visual extinction in magnitudes $A_{\rm V}$, or the FUV (6–13.6 eV) flux in Habing units G/G_0.[4] In this paper, we have chosen as additional label G/G_0: the GAME output is therefore a set of values for n, $N_{\rm H}$, U, Z, G/G_0.

2.1 Model improvements

In this section, we describe some new important improvements introduced here with respect to the original version of the code presented in U17. These concern (a) the build-up of the spectral library, (b) the inclusion of noise in the training procedure, and (c) the evaluation of the uncertainties; they are described in the following. More technical and detailed materials can be found in the Appendices.

2.1.1 *Library of synthetic spectra*

Concerning the library, the main improvements in the current version are as follows:

(i) The library now contains 50 000 synthetic spectra, i.e. it is ∼65 per cent larger than that used in U17.

(ii) We added SEDs of Population III stars generated via the YGGDRASIL code (Zackrisson et al. 2011). We adopted the Zackrisson et al. (2011) model with a zero-metallicity population and a Kroupa IMF (Kroupa 2001) in the interval 0.1–100 $M_\odot$ based on a rescaled single stellar population from Schaerer (2002). For the star formation history, we have chosen an instantaneous burst with age set to 2 Myr.

(iii) In addition to the graphite and silicate dust grains, we have added polycyclic aromatic hydrocarbons (PAHs) as these particles considerably affect EUV extinction and are an important heating source, especially in neutral regions. These contributions are modelled following Weingartner & Draine (2001). For the effect of stochastic heating, we refer to the work of Guhathakurta & Draine (1989). A power-law distribution of PAH is assumed with 10 size bins (Abel et al. 2008). The dust-to-gas ratio has been linearly scaled with metallicity.

2.1.2 *Training library with noise*

The library discussed so far is made of purely theoretical models. Observed spectra contain noise that must be taken into account. Our approach is to include within the ML training phase a library containing noisy models. To this aim, we have generated a new library (100 000 models) made by two parts: the first is the original library of 500 000 photoionization models and the second is the same library to which Gaussian random noise has been added to the lines with an amplitude equal to 10 per cent of the line intensity.[5] The sum of these two libraries represents therefore our final training data set. We test GAME against synthetic noisy spectra in Section 3.1.

Table 2. Different models used for the GAME validation test (Section 3).

Name	$\log(n/{\rm cm}^{-3})$	$\log(N_{\rm H}/{\rm cm}^{-2})$	$\log(U)$	$Z/Z_\odot$
model (a)	2.861	19.725	−3.166	0.2283
model (b)	2.827	19.643	−3.093	0.2361
model (c)	2.795	19.737	−3.084	0.2498
model (d)	2.758	19.648	−3.103	0.2939
model (e)	2.634	19.606	−2.964	0.2548
model (f)	2.574	19.757	−2.847	0.2932

2.1.3 *Uncertainties on the inferred physical properties*

The final improvement concerns a robust estimate of the uncertainties associated with the inferred physical properties. The method works as follows. For each input spectrum, we construct N modified versions of it by adding to the line intensities Gaussian noise. For each input line intensity I with associated error e, the code extracts N new intensities i from the following Gaussian distribution:

$$P(i) = \frac{1}{\sqrt{2\pi e^2}} \exp\left[-\frac{(i-I)^2}{2e^2}\right] . \quad (1)$$

For upper limits instead, the code generates a new line by taking a random number uniformly distributed between zero and the upper limit.

With this procedure, the code generates multiple individual new observations of each spectrum. Then, for each input spectrum, we obtain N determinations of the quintuplet of physical properties (n, $N_{\rm H}$, G/G_0, U, Z). By default, the code gives as final output for each spectrum the average, the median, and the standard deviation of these N values. Optionally, the code can return all the N determinations of the physical properties, which can be subsequently combined into a probability distribution function.

3 VALIDATION OF THE CODE

3.1 Noise in observed spectra

We now analyse how noise on the line intensities can affect the determination of the physical properties.

We start from the emission lines of a synthetic spectrum for which the true physical properties, i.e. those used to generate it, are known. The effect of noise can be mimicked by perturbing each of the synthetic line intensities around its original value. If we then apply GAME to the perturbed intensities, it is possible to assess how the inferred physical properties vary as a function of the noise amplitude. This can be done by using in the ML training phase both the original library (500 000 models) and the one including noise (Section 2.1.2). To perform this analysis, we apply the following steps:

(i) generate two synthetic spectra: 'model (a)' and 'model (b)' (see Table 2 for details);

(ii) choose a set of emission lines to use for the calculation (reported in Table 3);

[3] Both the implementations of AdaBoost and Decision Trees within GAME are included in the scikit-learn PYTHON package (Pedregosa et al. 2011), http://scikit-learn.org/.

[4] $G_0 = 1.6 \times 10^{-3}$ erg s^{-1} cm^{-2} (Habing 1968)

[5] Higher noise values would lead to completely noise-dominated models: given the large degeneracy in the emission line intensities (see Section 3.2), GAME would then not be able to reliably discriminate among different models.

Table 3. Emission lines used to compute the values of the physical properties in Section 3.1.

Line	Wavelength (Å)
[O II]	3726
[O II]	3729
[Ne III]	3869
Hδ	4102
Hγ	4341
[O III]	4363
Hβ	4861
[O III]	4959
[O III]	5007
He I	5876
[O I]	6300
[N II]	6548
Hα	6563
[N II]	6584
[S II]	6717
[S II]	6731
[Ar III]	7135

(iii) choose a set of 20 values of noise percentages (n_i = 1 per cent, 2 per cent, 3 per cent,..., 20 per cent) and compute 50 different realizations of these two spectra for each n_i value: for each realization, we added to the emission line intensities a Gaussian random value between 0 and n_i percent of the intensity of the line itself;

(iv) run GAME to infer the values of the physical properties on both the original library (500 000 photoionization models) and the noisy library (1000 000 models; see Section 2.1.2).

Results are shown in Fig. 1, where we report the inferred physical properties for models (a) and (b). When the code is trained without noise – first and third columns for models (a) and (b), respectively – adding noise to the spectrum to a level as low as $n_i = 4$ per cent can lead to differences between the true and inferred properties up to 2 orders of magnitude. Noisier spectra yield even larger differences.

Second and fourth columns of Fig. 1 show that including the noise in the training procedure reduces these differences in the determination of the physical properties. Interestingly, we also see that metallicity determinations are quite robust. In fact, even adding noise at 20 per cent level, the inferred metallicity is within a factor of 2 from the true value, confirming previous conclusions in U17.

Although the library has been constructed with a noise percentage up to 10 per cent, training the ML algorithm with this library allows GAME to return outputs consistent with the 'true' values even if applied to spectra with greater noise. In fact, as can be seen from Fig. 1, the mean of the inferred values agrees quite well with the 'true' values up to noise level as large as 20 per cent.

Using noiseless libraries with low SNR spectra can lead to wrong determinations of the intrinsic physical properties (see Fig. 1). This issue is even more severe for weak lines such as [O I] λ6300 or [Ar III] λ7135 (see Table 3), which can be fundamental for the determination of the physical properties (i.e. they could have a very high feature importance; see Appendix B).

3.2 Emission line degeneracy

Measurements of ISM physical properties are generally based on the comparison between observations and empirically calibrated line ratio or synthetic spectra obtained from photoionization models.

Although empirical calibrations using the electronic temperature are preferable (with respect to theoretical calibrations) because they are based on a quantity directly inferred from observables (Curti et al. 2017), they also suffer from some limitations (Pérez-Montero 2017, and references therein). The major among these is that calibrations often use galaxy samples that do not properly cover all the physical properties of space. Hence, empirical calibrations obtained from a sample of low-excitation H II regions could give unreliable results when applied to global galaxy spectra (Curti et al. 2017).

Comparisons with photoionization models are usually performed changing the metallicity and the ionization parameter, but they are limited to a small range of other ISM physical properties (i.e. density) if not only to a single value (Pérez-Montero 2014; Vale Asari et al. 2016; Pérez-Montero & Amorín 2017). This is problematic, as the ISM density distribution has a dynamic range that can easily span several dex (Hughes et al. 2017).

For this reason, we produced an extended library of physically motivated theoretical models (with 500 000 models described in Section 2.1.1), whose purpose is to cover the large range of physical conditions found in the ISM. In this section, we will show that, with such a library, the emission lines arising from different combinations of the input physical properties (density, column density, ionization parameter, and metallicity) are extremely degenerated even if the variation of the physical parameters is at the percent level.

We considered 300 photoionization models with the following physical properties:

$$\begin{aligned} -1.8 &< \log(n/\mathrm{cm}^{-3}) < -1.3 \\ 17.3 &< \log(N_\mathrm{H}/\mathrm{cm}^{-2}) < 17.8 \\ -1.5 &< \log(U) < -1.0 \\ 0.3 &< Z/Z_\odot < 0.6. \end{aligned}$$

We then compute the following line intensities: [O III] λ3727, HeII λ4686, [O III] λ5007, Hα, [N II] λ6584, [S II] λ6717. In Fig. 2, we report the ratio of the line intensities for the i-th model and the first one within this set of 300 models. This figure shows that although the physical parameters considered in the models vary within small ranges, they still result into intensity variations of several orders of magnitudes, especially for [O III] λ3727, [N II] λ6584, and [S II] λ6717.

In Fig. 3, in a similar way, we show the variation for some emission line ratios (based on commonly used emission line diagnostics):

$$\begin{aligned} \mathrm{O_3N_2} &= ([\mathrm{O\,III}]\lambda5007/\mathrm{H}\beta)/([\mathrm{N\,II}]\lambda6584/\mathrm{H}\alpha)\,, \\ \mathrm{R_{23}} &= ([\mathrm{O\,II}]\lambda3727 + [\mathrm{O\,III}]\lambda4959 + [\mathrm{O\,III}]\lambda5007)/\mathrm{H}\beta\,, \\ \mathrm{N_2} &= [\mathrm{N\,II}]\lambda6584/\mathrm{H}\alpha\,, \\ \mathrm{O_{32}} &= [\mathrm{O\,III}]\lambda5007/[\mathrm{O\,II}]\lambda3727. \end{aligned}$$

These emission line ratios suffer also from intensity variations up to 2.5 dex. What emerges is that not only the intensity of the lines but also their ratios seem to be affected by a non-negligible degeneration. Interestingly, this is not true in the case of R_{23}, which remains approximately constant, at least in the small range of physical properties considered in this section. However, it must be noticed that in a larger metallicity range ($-2 \lesssim \log(Z/Z_\odot) \lesssim 0.5$), R_{23} is not monotonically dependent on Z (e.g. Nagao et al. 2006, first panel in fig. 6). Moreover, one of the fundamental advantages of using GAME lies in the fact that instead of inferring one single physical property at one time (i.e. metallicity in the case of R_{23}), it can retrieve simultaneously n, U, N_H, G/G_0, Z for an extended range of physical properties (see Table 1).

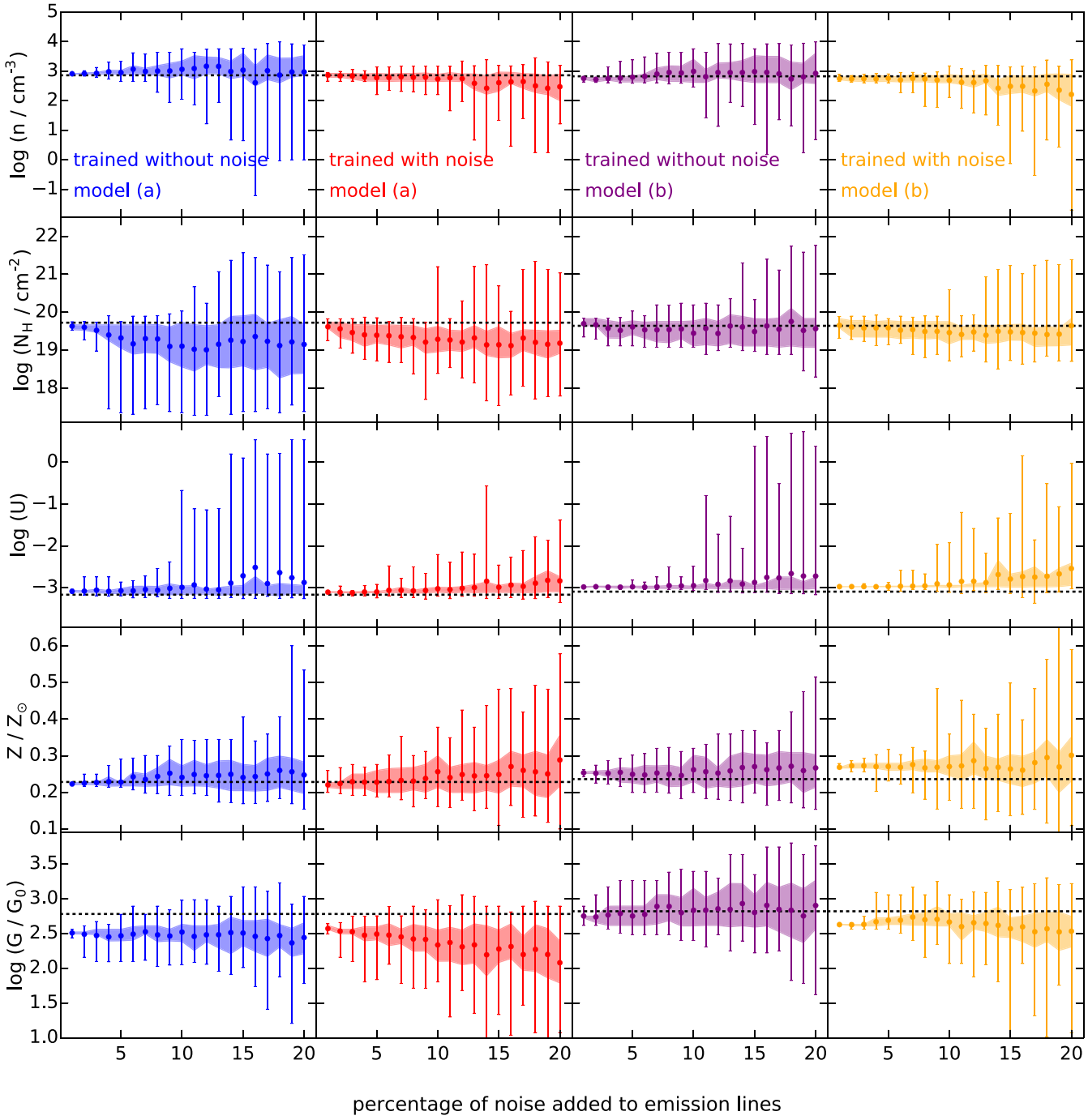

Figure 1. Inferred values of the physical properties for models (a) and (b) in Table 2 with noise added to the emission line intensities. The circles represent the mean of the inferred values for each of the 50 realizations and the error bars denote the minimum and maximum value inferred. The shaded region represents values between the first and third quartile. The dashed horizontal line is the 'true' value used to generate the two models.

To fully appreciate this aspect over the entire range of optical wavelengths, using the set of photoionization models listed in Table 2, we show in Fig. 4 how a small change in the physical properties is mirrored into large line intensity variations. Panel 1 of Fig. 4 shows the spectrum for model (a); the remaining panels show the spectral differences with the other models (b)–(f). Although the variation of metallicity between model (a) and (b) is <0.008 dex, varying simultaneously n, N_H, and U by 0.03, 0.08, and 0.07 dex induces large differences in the resulting spectrum.

GAME is based on a very large library and uses all the information carried by the spectral lines on the physical properties. This approach, allowed by the ML implementation, could overcome the degeneracy better than model fitting techniques, typically based on χ^2 methods, i.e. the minimization of the Euclidean distance between models:

$$D(m, l) = \sqrt{\sum_{i=1}^{n} (m_i - l_i)^2}\,, \qquad (2)$$

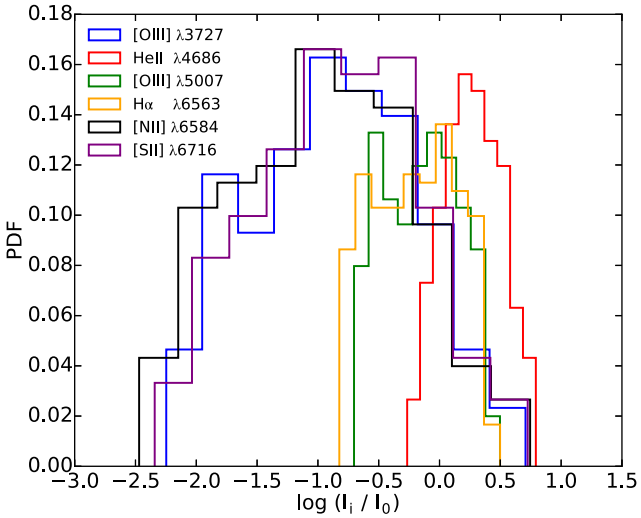

Figure 2. Ratios of the line intensities, I_i, for different emission lines between the i-th model and the first one (intensity I_0) within our set of synthetic models described in Section 3.2.

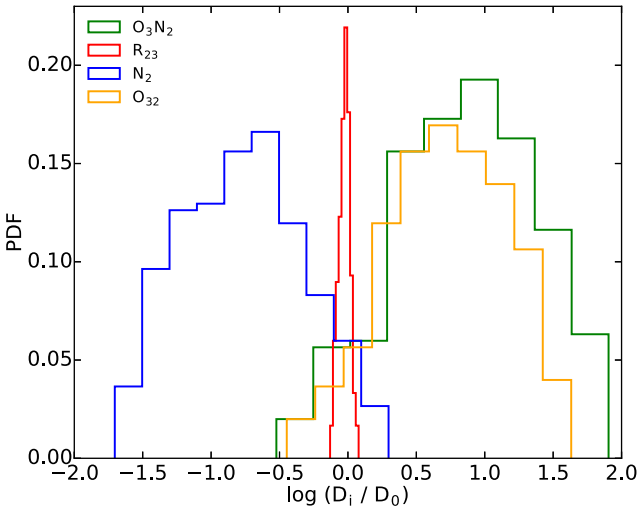

Figure 3. Ratios of commonly used line intensity diagnostics (O_3N_2, R_{23}, N_2, O_{32}) between the i-th model (D_i) and the first one (D_0) within our set of synthetic models described in Section 3.2.

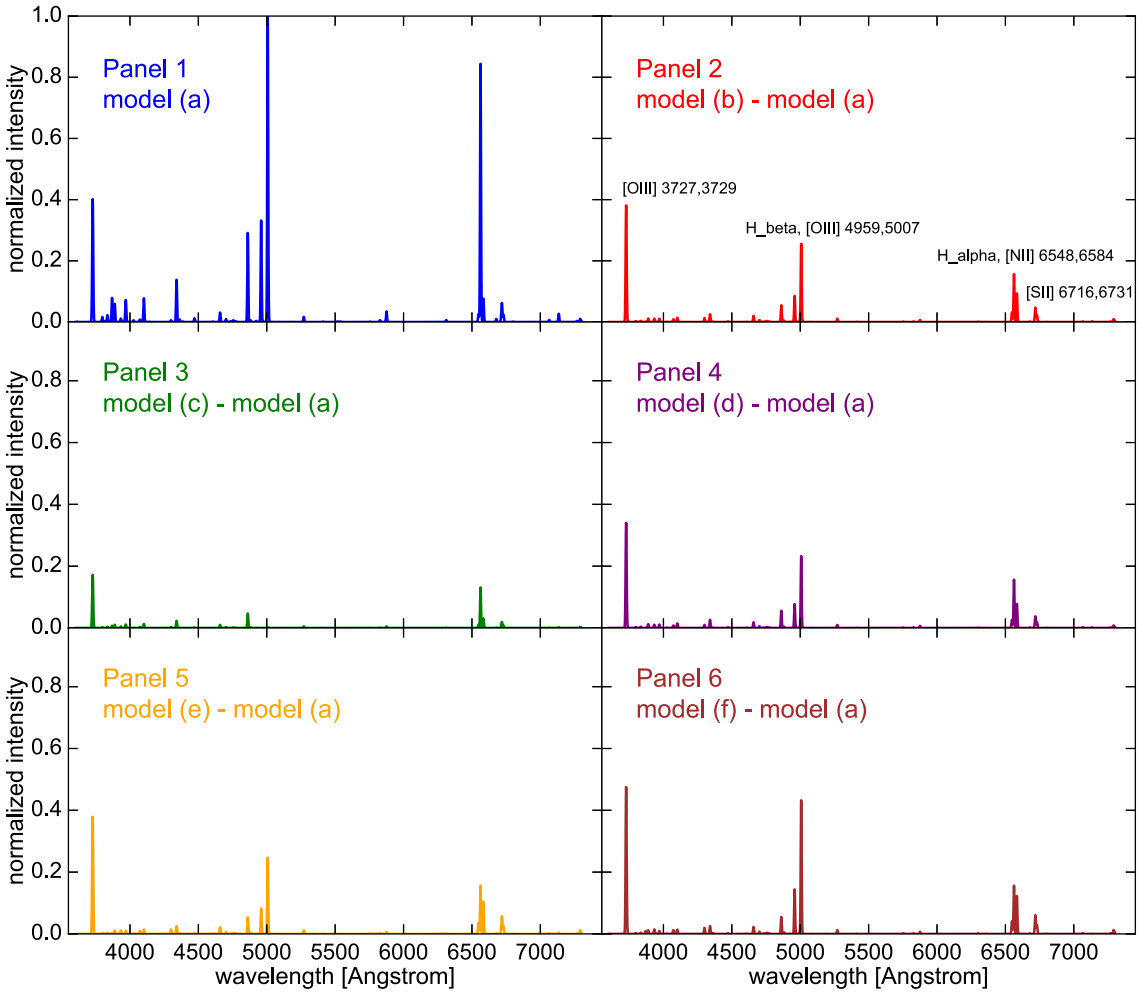

Figure 4. Optical spectra of the models reported in Table 2. Panel 1 shows model (a); other panels show the difference between models (b)–(f) and model (a).

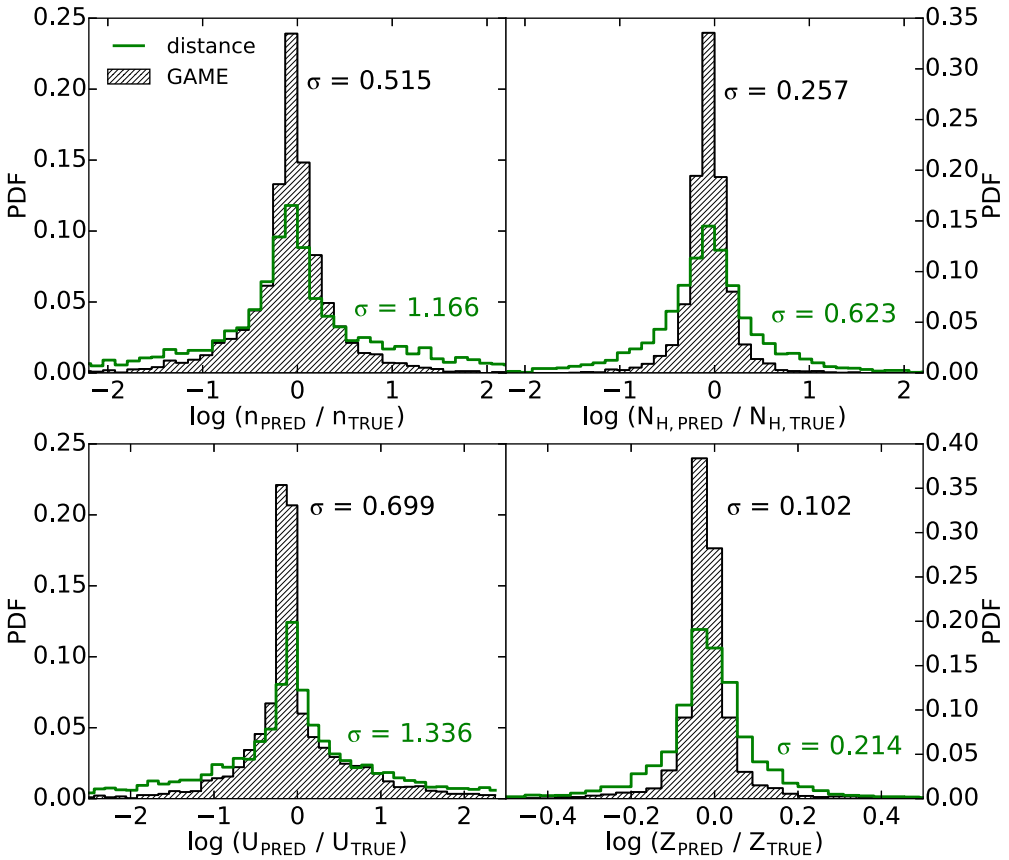

Figure 5. Ratios between the inferred and true values of the physical properties using two approaches: GAME (black shaded region) and the minimization of the Euclidean distance $D(m, l)$ (green line, see equation 2). The test has been performed using 10 per cent of noiseless library. The standard deviations of the distributions (σ) are also shown.

Table 4. Set of emission lines used to assess the predictive performances of GAME (see the text for details).

Line	Wavelength [Å]
Hβ	4861
[O III]	5007
He I	5876
[O I]	6300
Hα	6563
[N II]	6584
He I	6678
[S II]	6717
[S II]	6731

where m_i is the value of the emission line intensities of a given model and l_i the corresponding i-th value contained in the library. In the following, we show that minimizing such function does not necessarily lead to a correct determination of the physical properties.

In Fig. 5, we compare the inferred values of the physical properties using the two approaches, GAME versus $D(m, l)$ minimization (equation 2), using the set of emission lines in Table 4. We perform the test on 10 per cent of the noiseless library (GAME uses the remaining 90 per cent as the training data set), including 50 000 models. On this reduced set, we infer the physical values both with GAME and by distance minimization.

Trying to recover a spectrum that is as similar as possible to the input spectrum (minimizing the distance) leads to worse results with respect to GAME. In Fig. 5, it is evident that the standard deviation of the logarithm of the ratio between the 'predicted' and 'true' values in the case of GAME is a factor of 2 smaller than the $D(m, l)$ minimization approach. The similarity between two spectra, given the extreme degeneration (see Section 3.2), does not necessarily mean a good correlation with their physical properties. This was expected since other ML techniques as k-Nearest Neighbour (with k = 1 in the case of equation 2) work well when the input space has a low dimensionality, which does not apply to our problem.

4 COMPARISON WITH OTHER METHODS

4.1 Overview

The most prominent feature of GAME is that it exploits the full information encoded in a spectrum. Instead of using small, pre-selected subsets of emission line ratios, GAME can use an arbitrary number of lines to infer the ISM physical properties. Additionally, its usage is not limited to a specific ISM phase (i.e. H II regions).

Another key advantage of the ML implementation is that, once trained, it requires a very short computational time (see also Appendix C). This is a crucial feature in view of applications to modern IFU observations with ∼1000 000 spaxels, each one with a substantial number (>10) of observed lines per spectrum.

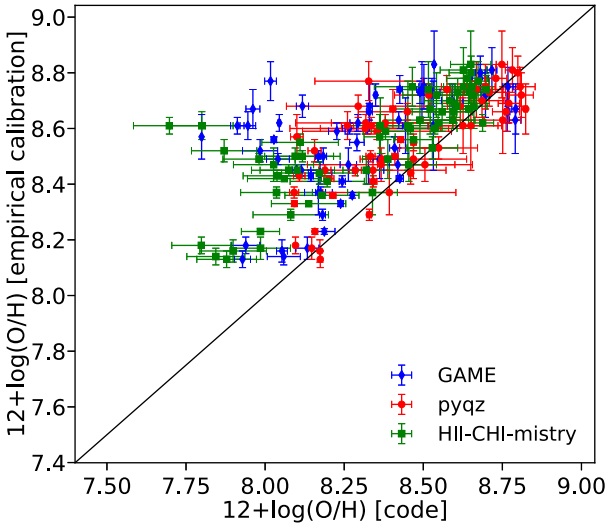

Figure 6. Comparison between the metallicity inferred with the empirical calibrators of Curti et al. (2017) and by GAME (blue diamonds), PYQZ (red circles), and HII-CHI-MISTRY (green squares) on 62 stacked galaxy spectra (see text for details).

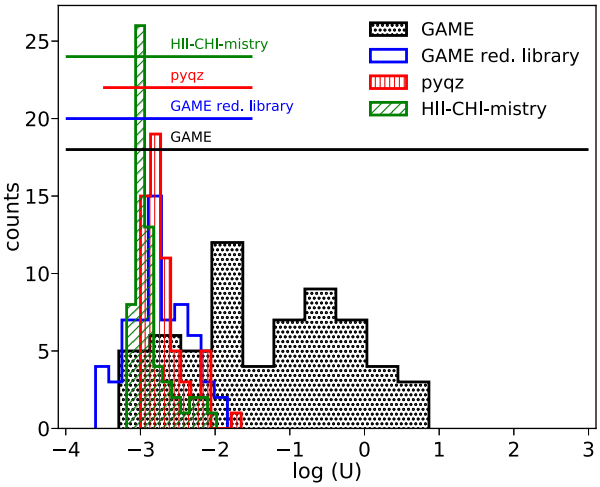

Figure 7. Ionization parameter, U, distribution for 62 stacked galaxy spectra inferred by GAME (grey dotted histogram), GAME with a reduced library (empty), PYQZ (red), and HII-CHI-MISTRY (green). The horizontal lines denote the U range considered in the libraries of the different codes.

4.2 Quantitative comparison: galaxies

Here we quantitatively compare GAME with other empirical calibrations and results from different codes. To perform this comparison, we use a sample of 62 stacked galaxy spectra. These spectra are obtained from SDSS star-forming galaxies in $0.027 < z < 0.25$ and are fully described in Curti et al. (2017). The spectra are stacked in bins of 0.1 dex according to the log values of their [O II] λ3727/H β and [O III] λ5007/H β ratios. For each of the stacked spectra, Curti et al. (2017) derived the corresponding metallicity, using a set of new and self-consistent empirical calibrations. We compare such determination with the results of GAME and the results from two widely used emission line codes, PYQZ and HII-CHI-MISTRY.

PYQZ[6] (Dopita et al. 2013) is a public PYTHON code that uses abundance- and excitation-sensitive line ratios to define a plane in which the oxygen abundance and ionization parameter can be determined by interpolating a grid of photoionization models to match the observed line ratios.

HII-CHI-MISTRY [7] (Pérez-Montero 2014) is a publicly available PYTHON code. This code takes the extinction-corrected emission line fluxes and, based on a χ^2 minimization on a photoionization model grid, determines chemical abundances (O/H, N/O) and ionization parameter. In this work, we use the version 3.0 of the code, dealing with input line uncertainties via a Monte Carlo approach.

As evident from the comparison shown in Fig. 6, the empirical method tends to slightly overestimate metallicity with respect to all codes. Besides, the mean offset and the standard deviation between GAME predictions and those from other methods are small (less than 0.3 dex). Overall, the agreement on metallicity is good.

We extend the comparison among the three codes to the ionization parameter (Fig. 7). Interestingly, the GAME results markedly differ from those obtained with the other two codes. This can be due to the fact that PYQZ and HII-CHI-MISTRY only contain models with a narrower U range (i.e. $-3.5 < \log U_{\rm PYQZ} < -1.5$, and $-4.0 < \log U_{\rm HII\text{-}CHI\text{-}MISTRY} < -1.5$). To verify this point, we ran GAME with a reduced library containing ionization parameter values in the range $-4.0 < \log(U) < -1.5$. Fig. 7 clearly shows that with the reduced library, GAME infers U values that are in agreement with those from the other two codes; e.g. the means of the inferred values are respectively μ(PYQZ) $= -2.67$, μ(GAME) $= -2.76$, and μ(HII-CHI-MISTRY) $= -2.88$. Hence, it is important to consider U values that are larger than usually assumed.

Values of U up to 10 are expected in line-emitting galaxies. Let us consider the relation between the ionizing photon flux and the star formation rate (e.g. Murray & Rahman 2010):

$$Q(\mathrm{H}) = 2.46 \times 10^{53} \left(\frac{\mathrm{SFR}}{\mathrm{M}_\odot\ \mathrm{yr}^{-1}} \right) \mathrm{s}^{-1}. \quad (3)$$

For a compact galaxy with an SFR $= 1\ \mathrm{M}_\odot\ \mathrm{yr}^{-1}$, $n = 1\ \mathrm{cm}^{-3}$, and radius 1 kpc, we obtain $\log(U) \sim 0.6$, consistently with the upper limit found with GAME (see Fig. 7). As discussed in U17, the spectrum emerging from a galaxy is weighted by the column densities along the line of sight. If diffuse phases are dominant in the build-up of the final spectrum, their low density pushes $U \propto n^{-1}$ towards large values.

Note that although the full library contains spectra with $-4.0 < \log(U) < 3$ (see Table 1), GAME does not infer values $\log(U) > 1$. This implies that this bound is set by the physical conditions, and it is independent of the library extension.

We finally note that U is the physical parameter most affected by uncertainties. The bootstrap routine now included in the code should however significantly mitigate the problem: although the inferred values for U cover a large range, their PDF obtained with the bootstrap is highly peaked (see Appendix A4 and Fig. A3).

4.3 Quantitative comparison: H II regions

As an additional comparison, we apply GAME to a sample of 75 observed H II regions with available chemical abundance determinations. We choose H II regions for which a large number ($N > 13$) of high-quality (typical errors <10 per cent) emission lines is available. Our final sample is composed of the following:

[6] http://fpavogt.github.io/pyqz/index.html
[7] http://www.iaa.es/epm/HII-CHI-mistry.html

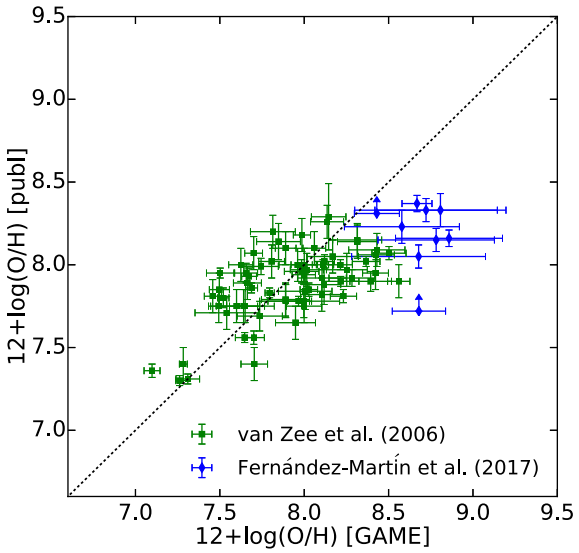

Figure 8. Comparison of the oxygen abundance determinations for H II regions reported in van Zee & Haynes (2006) and Fernández-Martín et al. (2017) with those inferred by GAME.

(i) 66 H II regions located in 21 dwarf irregular galaxies observed by van Zee & Haynes (2006). For 25 H II regions, oxygen abundances are obtained via direct detection of emission lines tracing the electron temperature; for the remaining 41 H II regions, abundances are inferred through SEL methods. The emission lines (where available) used as input for GAME are [O II] λλ3727+3729, [Ne III] λ3869, [O III] λλ4959+5007, [O I] λ6300, [S III] λ6312, Hα, [N II] λλ6548+6584, HeI λ6678, [S II] λλ6717+6731, and [Ar III] λ7135 (table 3 of that paper);

(ii) 9 H II regions located towards the Galactic anticentre observed by Fernández-Martín et al. (2017) for which more than 60 emission line measurements are available (table A.1 of that paper).

Results are shown in Fig. 8, where we compare the abundance determinations reported in the cited works with the GAME results. The metallicity values inferred by GAME ($Z/Z_{\odot}$) are converted into oxygen using the relation 12+log(O/H) = 8.69. The overall agreement between the two methods is good. The scatter does not seem to correlate with metallicity and the dispersion is of ≃0.2–0.3 dex. The mean offset (between the results inferred with GAME and the published results reported in Fig. 8) and its standard deviation are, respectively, 0.08 and 0.29. Differences among the results of different methods can arise because of several reasons. First of all, the oxygen abundance we infer from the metallicity assumes solar abundances. Moreover, while GAME adopts input emission lines that are not de-reddened, the reddening corrections adopted by van Zee & Haynes (2006) and Fernández-Martín et al. (2017) (i.e. n_e = 100 cm^{-3}, Case B recombination, different extinction curves) may have a significant impact on the determination of physical properties.

The main advantage of GAME is that it does not use any assumption on the physical properties of gas (e.g. density, temperature) defining the emission line spectrum we are looking at. The code is able to recover the physical properties (in this case, metallicity) by extracting them from a library containing a vast collection of physical conditions of the ISM. We also note that GAME does not even use the information that the sample refers to H II regions: it simply searches the library, trains itself, and gives the best predicted physical conditions.

5 SUMMARY AND DISCUSSION

We presented an updated and optimized implementation of GAME (U17), a code designed to infer ISM physical properties from emission line spectra. The code is based on an ML algorithm (AdaBoost with Decision Trees as base learner) to calculate density (n), column density (N_H), ionization parameter (U), metallicity (Z), and FUV flux in the Habing band (G/G_0), given an arbitrary set of emission line flux measurements (or upper limits) with their uncertainties.

GAME is extremely reliable, particularly for the metallicity determination: the five-fold cross-validation score with the set of emission lines reported in Table 4 is higher than 0.95. Although some properties, as metallicity, are easily and robustly recovered also from noisy spectra with few emission lines, other physical properties require higher quality data with more lines measurements. We have shown that for a given set of emission lines if the cross-validation score for the metallicity is 0.95, it might happen that lower scores are obtained for n and N_H (both ∼0.8) and U (∼0.7).

Another key point in our analysis is that the emission line intensities are highly degenerate. A small variation of the physical properties leads to large changes in the emission line intensity ratios. The ML approach used here can overcome this issue much better than classical fitting methods based on χ^2 minimization. Noticeably, such important result can be achieved also when the spectra include noise as in real observations.

We have compared GAME with methods based on empirical calibrations (Curti et al. 2017) and other codes (PYQZ and HII-CHI-MISTRY) by considering a sample of 62 stacked spectra from SDSS galaxies (Curti et al. 2017). While a very good agreement has been obtained in terms of the metallicity determination for the considered sample, we find discrepancies in the derived values of the ionization parameter. We discuss possible reasons for such disagreement in Section 4.

Finally, we have tested our code on a sample of 75 H II regions with direct method and SEL abundance determinations (van Zee & Haynes 2006; Fernández-Martín et al. 2017) to study how GAME can recover these values. We found that the oxygen abundance determinations are in good agreement with those inferred with GAME with a typical scatter around 0.2–0.3 dex. The applications of GAME are not only limited to H II regions, but the code can also deal equally well with different phases of the ISM, including the molecular one. This is because the underlying library covers the largest possible range of physical conditions characterizing the ISM. Furthermore, GAME offers the possibility to use an arbitrary set of emission lines, which span a wavelength range from the Lyα one (1216 Å) to 1 mm. These features allow the user to infer the physical properties of phases ranging from the hot ionized medium to dense molecular cores.

One of the main limitations of GAME (and other methods/codes) relies on the possible presence of different ISM phases/gradients along the line of sight contributing to the same spectrum. This introduces a complexity that cannot be managed at the present time. We preliminarily discussed this issue in U17. The main result there was that in such conditions, the returned physical parameters are biased towards the phase with largest gas column density. To make progress, we plan to investigate this issue in more details using emission line spectra generated from high-resolution galaxy simulations. Simulated galaxies and their spectra offer the advantage that the physical conditions of the gas shaping the observed spectra are precisely known. This will allow us to (a) devise more stringent reliability tests for GAME, and (b) understand how to maximize the information retrieval from spectra arising from multiphase lines of sight (for this issue, we also refer the reader to section 5.3 of U17).

It is nowadays possible to obtain spatially resolved spectra of galaxies. The information coming from different regions of a galaxy requires the development of new methods in order to obtain the physical conditions of the different phases of the ISM. Large libraries and robust algorithms will be crucial in the analysis of galaxy spectra that include faint lines arising from extended wavelength ranges. Combining the UV/optical/IR/FIR information from the same object will be the next step towards a better understanding of the internal structure of distant galaxies.

ACKNOWLEDGEMENTS

We thank M. Curti for providing the galaxy spectra used in our analysis and the anonymous referee for constructive insights. We also thank B. Greig, N. Gillet, C. Behrens, and L. Vallini for useful discussions and comments. AF acknowledges support from the ERC Advanced Grant INTERSTELLAR H2020/740120.

REFERENCES

Abel N. P., van Hoof P. A. M., Shaw G., Ferland G. J., Elwert T., 2008, ApJ, 686, 1125
Allende Prieto C., Lambert D. L., Asplund M., 2001, ApJ, 556, L63
Alloin D., Collin-Souffrin S., Joly M., Vigroux L., 1979, A&A, 78, 200
Asplund M., Grevesse N., Sauval A. J., Allende Prieto C., Kiselman D., 2004, A&A, 417, 751
Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, ARA&A, 47, 481
Blanc G. A., Kewley L., Vogt F. P. A., Dopita M. A., 2015, ApJ, 798, 99
Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees. Wadsworth International Group, Belmont, CA
Bresolin F., 2007, ApJ, 656, 186
Bresolin F., 2017, Outskirts of Galaxies, Astrophysics and Space Science Library. Springer International Publishing, Cham, p. 145
Bresolin F., Ryan-Weber E., Kennicutt R. C., Goddard Q., 2009a, ApJ, 695, 580
Bresolin F., Gieren W., Kudritzki R.-P., Pietrzyński G., Urbaneja M. A., Carraro G., 2009b, ApJ, 700, 309
Bresolin F., Kudritzki R.-P., Urbaneja M. A., Gieren W., Ho I.-T., Pietrzyński G., 2016, ApJ, 830, 64
Brown J. S., Martini P., Andrews B. H., 2016, MNRAS, 458, 1529
Cresci G., Vanzi L., Telles E., Lanzuisi G., Brusa M., Mingozzi M., Sauvage M., Johnson K., 2017, A&A, 604, A101
Curti M., Cresci G., Mannucci F., Marconi A., Maiolino R., Esposito S., 2017, MNRAS, 465, 1384
De Looze I. et al., 2014, A&A, 568, A62
Dietterich T. G., 2000, Multiple Classifier Systems, LBCS-1857. Springer, New York, p. 1.
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, 61
Drucker H., 1997, in Proc. Fourteenth Int. Conf. Mach. Learn. ICML '97. Morgan Kaufmann Publishers Inc., San Francisco, CA, p. 107. Available at: http://dl.acm.org/citation.cfm?id=645526.657132
Erb D. K., Pettini M., Shapley A. E., Steidel C. C., Law D. R., Reddy N. A., 2010, ApJ, 719, 1168
Esteban C., Bresolin F., Peimbert M., García-Rojas J., Peimbert A., Mesa-Delgado A., 2009, ApJ, 700, 654
Esteban C., García-Rojas J., Carigi L., Peimbert M., Bresolin F., López-Sánchez A. R., Mesa-Delgado A., 2014, MNRAS, 443, 624
Farrah D. et al., 2013, ApJ, 776, 38
Ferland G. J. et al., 2013, RMxAA, 49, 137
Fernández-Martín A., Pérez-Montero E., Vílchez J. M., Mampaso A., 2017, A&A, 597, A84
Freund Y., Schapire R. E., 1997, J. Comput. Syst. Sci., 55, 119
García-Rojas J., Esteban C., 2007, ApJ, 670, 457
Garofalo M., Botta A., Ventre G., 2017, IAU Symp. 12, Astrophysics and Big Data: Challenges, Methods, and Tools. Cambridge University Press, Cambridge, p. 345
Guhathakurta P., Draine B. T., 1989, ApJ, 345, 230
Habing H. J., 1968, Bull. Astron. Inst. Neth., 19, 421
Hammer F., Puech M., Flores H., Rodrigues M., 2017, Studying distant galaxies: A Handbook of Methods and Analyses (arXiv:1701.03794), http://arxiv.org/abs/1701.03794
Hastie T. J., Tibshirani R. J., Friedman J. H., 2009, The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer series in statistics, Springer, New York, http://opac.inria.fr/record=b1127878
Hughes T. M. et al., 2017, A&A, 602, A49
Kewley L. J., Dopita M. A., 2002, ApJS, 142, 35
Kewley L. J., Ellison S. L., 2008, ApJ, 681, 1183
Kobulnicky H. A., Kewley L. J., 2004, ApJ, 617, 240
Kroupa P., 2001, MNRAS, 322, 231
López-Sánchez Á. R., Esteban C., García-Rojas J., Peimbert M., Rodríguez M., 2007, ApJ, 656, 168
Maiolino R. et al., 2008, A&A, 488, 463
Marino R. A. et al., 2013, A&A, 559, A114
McGaugh S. S., 1991, ApJ, 380, 140
Murray N., Rahman M., 2010, ApJ, 709, 424
Nagao T., Maiolino R., Marconi A., 2006, A&A, 459, 85
Nagao T., Maiolino R., Marconi A., Matsuhara H., 2011, A&A, 526, A149
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
Peimbert A., 2003, ApJ, 584, 735
Peimbert M., Peimbert A., 2014, Rev. Mex. Astron. Astrofis., 44, 137
Peimbert A., Peimbert M., Ruiz M. T., 2005, ApJ, 634, 1056
Pereira-Santaella M., Rigopoulou D., Farrah D., Lebouteiller V., Li J., 2017, MNRAS, 470, 1218
Pérez-Montero E., 2014, MNRAS, 441, 2663
Pérez-Montero E., 2017, Publ. Astron. Soc. Pac., 129, 043001
Pérez-Montero E., Amorín R., 2017, MNRAS, 467, 1287
Pettini M., Pagel B. E. J., 2004, MNRAS, 348, L59
Pilyugin L. S., Grebel E. K., 2016, MNRAS, 457, 3678
Pilyugin L. S., Thuan T. X., 2005, ApJ, 631, 231
Rigopoulou D., Pereira-Santaella M., Magdis G. E., Cooray A., Farrah D., Marques-Chaves R., Perez-Fournon I., Riechers D., 2018, MNRAS, 473, 20
Schaerer D., 2002, A&A, 382, 28
Schapire R. E., 1990, Mach. Learn., 5, 197
Shapley A. E., Steidel C. C., Pettini M., Adelberger K. L., 2003, ApJ, 588, 65
Sobral D., Matthee J., Darvish B., Schaerer D., Mobasher B., Röttgering H. J. A., Santos S., Hemmati S., 2015, ApJ, 808, 139
Stanway E. R., 2016, IAU Symp. 12, What can distant galaxies teach us about massive stars? Cambridge University Press, p. 305
Stark D. P. et al., 2014, MNRAS, 445, 3200
Stark D. P. et al., 2015a, MNRAS, 450, 1846
Stark D. P. et al., 2015b, MNRAS, 454, 1393
Stark D. P. et al., 2017, MNRAS, 464, 469
Stasińska G., 2004, in Esteban C., García López R., Herrero A., Sánchez F., eds, Cosmo Chemistry: The Melting Pot of the Elements. Cambridge University Press, Cambridge, p. 115
Stasińska G., 2007, preprint (arXiv:0704.0348)
Tremonti C. A. et al., 2004, ApJ, 613, 898
Tsamis Y. G., Barlow M. J., Liu X.-W., Danziger I. J., Storey P. J., 2003, MNRAS, 338, 687
Ucci G., Ferrara A., Gallerani S., Pallottini A., 2017, MNRAS, 465, 1144
Vale Asari N., Stasińska G., Morisset C., Cid Fernandes R., 2016, MNRAS, 460, 1739

Vallini L., Gallerani S., Ferrara A., Pallottini A., Yue B., 2015, ApJ, 813, 36
Vallini L., Ferrara A., Pallottini A., Gallerani S., 2017, MNRAS, 467, 1300
van Zee L., Haynes M. P., 2006, ApJ, 636, 214
Weingartner J. C., Draine B. T., 2001, ApJS, 134, 263
Zackrisson E., Rydberg C.-E., Schaerer D., Östlin G., Tuli M., 2011, ApJ, 740, 13
Zaritsky D., Kennicutt Jr R. C., Huchra J. P., 1994, ApJ, 420, 87

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Appendix

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