Draft version April 5, 2015
Preprint typeset using LaTeX style emulateapj v. 04/17/13

MONTE CARLO METHOD FOR CALCULATING UNCERTAINTY IN OXYGEN ABUNDANCE FROM STRONG-LINE FLUX MEASUREMENTS

Author order to be determined: Maryam Modjaz[1], Federica B. Bianco[1], Seung Man Oh[1,2], David Fierroz[1], Yuqian Liu[1], Lisa Kewley[3,4]

Draft version April 5, 2015

ABSTRACT

MODIFY & FINALIZE AT THE END: We present an open-source Python code for the determination of the strong-emission-line estimators of oxygen abundance in the standard scales, based on the original IDL-code in Kewley & Dopita (2002). The standard strong line Metallicity scales and diagnostics IMPROVE have been used to estimate metal abundance through emission line ratios. Here we introduce a Monte Carlo resampling of these methods in order to better characterize an oxygen abundance confidence region. We output median values, 16th and 84th percentile confidence regions, for various standard metallicity diagnostics, and, when possible, for reddening E(B-V). We produce Monte Carlo parameter distributions for the oxygen abundance and when possible for reddening E(B-V). We test our code on emission lines measurements from a sample of galaxies ($z < 0.15$) and compare our metallicity results with those from previous methods. We show that our metallicity estimates are consistent with previous methods but yields smaller uncertainties. The code is open source and can be found at `www.github.com/nyusngroup/` **(add repo and DOI)**.

Subject headings:

1. INTRODUCTION

Small amounts of carbon, oxygen, nitrogen, sulfur and iron and other elements provide a splash of color to the otherwise dominating greyscape of hydrogen and helium in the stars and gas of galaxies. Nevertheless, even this minute presence of heavy elements (all elements heavier than H and He, also called metals or collectively metallicity) is important for many areas of astrophysics. For example, Johnson & Li (2012), amongst others, suggest that if it was not for the relatively high metallicity level in our Solar System, planet formation may not have been possible. With Z representing the mass fraction of metals, for our own Sun the value is measured to be Z=0.0153 (Caffau et al. 2011), though there are others who suggest a lower solar metallicity of $Z = 0.0134$ mostly because of oxygen (Asplund et al. 2009; Grevesse et al. 2010).Furthermore, when properly observed and estimated, metallicity measurements of galaxies can tightly constrain models of galaxy formation and evolution (e.g., Kewley & Ellison 2008 and references therein), as well as shed light on the metallicity dependence and production conditions for different types of Supernovae (SNe) and long-duration Gamma-Ray Bursts (GRBs) (e.g., Modjaz et al. 2008; Levesque et al. 2010; Anderson et al. 2010; Modjaz et al. 2011; Kelly & Kirshner 2012; Sanders et al. 2012; Lunnan et al. 2014; Leloudas et al. 2014; Pan et al. 2014).

Metals are produced in the cores of massive stars during their fusion life cycle but also during the extreme conditions of stellar explosions. For example, the majority of iron is synthesized in thermonuclear explosions (SNe Ia) while nearly all of oxygen and other α-elements are released in core collapse SNe (SNe Ib, Ic, & II). Since new stars are born from the clouds these explosions enrich, metallicity will increase with each passing generation of stars.

However, for almost all astronomical objects, metallicity cannot be measured directly. The oxygen abundance in the gas-phase is the canonical choice of metallicity indicator for interstellar medium (ISM) studies, since oxygen is the most abundant metal and only weakly depleted onto dust grains (in contrast to refractory elements such as Mg, Si, Fe, with Fe, being depleted by more than a factor of 10 in Orion; see Simón-Díaz & Stasińska 2011). The oxygen abundance[5] is expressed as $12 + \log_{10}(\frac{O}{H})$, where O and H represent the number of Oxygen and Hydrogen atoms, respectively. In particular, (Caffau et al. 2011) measure a solar oxygen abundance of $12 + \log_{10}(\frac{O}{H}) = 8.76 \pm 0.07$, while Asplund et al. (2009) suggest $12 + \log_{10}(\frac{O}{H}) = 8.69$.

Importantly, oxygen exhibits very strong nebular lines in the optical wavelength range of HII regions (e.g., Pagel et al. 1979; Osterbrock 1989; Tremonti et al. 2004), which can be measured. Thus, many different diagnostic techniques, relying on different lines of oxygen, hydrogen and other elements, have been developed (e.g., Kewley & Dopita 2002; Pettini & Pagel 2004; Kobulnicky & Kewley 2004; Kewley & Ellison 2008), which are discussed in the next section. Ultimately the purpose of this paper is to support the release of a public code that computes metallicity according to the standard abundance diagnostics as well as the associated uncertainties due to the measured emission line flux uncertainties.

[1] Center for Cosmology and Particle Physics, New York University, 4 Washington Place, New York, NY 10003, USA
[2] NYU Abu Dhabi PO Box 129188 Abu Dhabi, UAE
[3] Australian National University, Research School for Astronomy & Astrophysics, Mount Stromlo Observatory, Cotter Road, Weston, ACT 2611, Australia
[4] Institute of Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA

[5] We note that in many cases in the literature, including here, the terms metallicity and oxygen abundance are used interchangeably.

1.1. *The different oxygen abundance diagnostics*

Here we present a brief overview of the various observational methods for measuring the gas-phase oxygen abundance - however, for a full discussion with all the caveats we encourage the reader to read the reviews by e.g. Stasińska (2002); Kewley & Ellison (2008); Moustakas et al. (2010); Stasińska (2010); Dopita et al. (2013); Blanc et al. (2015). The so-called "classical" way to estimate the oxygen abundance is the electron temperature (T_e) method, which estimates the electron temperature and density of the nebula using a number of oxygen lines with different ionization states, including the auroral [OIII] λ4363 line[6], to then directly estimate the OII and OIII abundances to obtain the total oxygen abundance, after correcting for the unseen stages of ionization. However, the auroral [OIII] λ4363 line is very weak, except in low-metallicity environments, and saturates at higher metallicity (since at higher metallicities the cooling is dominated by the oxygen NIR fine structure lines) − thus, other methods had to be developed that use other, stronger lines in the spectra of HII regions. These are called *strong-line methods* and are the subject of this manuscript. Strong-line methods can be categorized into two types: theoretical methods, that rely on calibrating various observed line ratios using photoionization methods (basically theoretically simulating HII regions, using stellar model atmospheres, stellar population synthesis and photoionization models) and empirical ones, that calibrate various observed strong line ratios using observed T_e-based metallicities. While historically there have been large systematic offsets between the T_e method and the strong line methods, Dopita et al. (2013) demonstrated that the T_e method gives the same results as the strong line methods, if the energy distribution of the electrons in the HII regions is assumed to not be a simple Maxwell-Boltzmann distribution (as assumed in prior works), but a more realistic κ distribution, as observed in solar system astrophysical plasma. They also find that the effect of the κ distribution on the strong-line methods is minor.

For theoretical strong-line method, a ratio of oxygen line fluxes to $H\beta$, referred to as R_{23}, is commonly used to determine the metallicity of galaxies (Pagel et al. 1979):

$$R_{23} = \frac{[\mathrm{OII}]\lambda 3727 \ + \ [\mathrm{OIII}]\lambda 4959, \lambda 5007}{\mathrm{H}_\beta},$$

where [OIII]λ4959,λ5007 stands for the sum of the two [OIII] lines The drawback of this method is that it is double-valued with metallicity, and thus other line ratios need to be used to break the degeneracy between the high values ("upper branch") and the low values ("lower branch") of the R_{23} metallicities (e.g., Kewley & Ellison 2008). Furthermore, Kewley & Dopita (2002) showed the importance of ionization parameter, which can be physically understood as corresponding to the maximum velocity of an ionized front that can be driven by the local radiation field of hot massive stars that are ionizing the ISM gas. This ionization parameter needs to be taken into account in the various strong-line methods, as HII regions at the same metallicity but with different ionization parameters produce different line strengths. Calibrations of R_{23} by McGaugh (1991) (hereafter M91), by Kewley & Dopita (2002) (hereafter KD02), and by Dopita et al. (2013) (hereafter D13) use different theoretical photoionization models and take the ionization parameter into account, while other calibrations such as that of Zaritsky et al. (1994) (hereafter Z94) do not. Thus, Z94 is mostly valid for only metal-rich galaxies. M91 and KD02 use an iterative process to break the R_{23} degeneracy (KD02 uses different ratios [NII]/[OII] and [NII]/Hα) and to also constrain the ionization parameter q in order to arrive at the metallicity estimate.

As to empirical strong-line methods, the most commonly used ones are that by Pettini & Pagel (2004) (hereafter PP04) and Pilyugin & Thuan (2005). PP04 used HII regions with measured T_e-based metallicities to derive empirical fits to strong-line ratios, and introduced the line ratios of ([NII]/Hβ($N2$) and ([OIII]/Hβ)/([NII]/Hα($O3N2$) as metallicity diagnostics. Since PP04_N2 employs two closely spaced lines (H$\alpha$and NII), which are not affected by stellar absorption, nor (uncertain) reddening, and are easily observed in one simple spectroscopic setup, it has become an often-used scale, at least for low-z SN host galaxy studies (see meta-analysis by e.g,. Sanders et al. 2012; Modjaz 2012; Leloudas et al. 2014). However, it is important to remember that this scale has a number of short-comings: it does not take into account the impact of the ionization parameter, it was initially derived based on only 137 extragalactic HII regions, and the nitrogen emission line employed saturates at high metallicity, and thus the PP04_N2 method saturated for high-metallicity galaxies (at $12 + \log_{10}(\frac{O}{H}) > 8.8$, Kewley & Ellison 2008). An updated calibration by Marino et al. (2013) based on many more T_e-based metallicities (almost three time larger that that of PP04) derives a significantly shallower slope between $O3N2$ index and oxygen abundance than the PP04 calibration. In addition, most recently, Berg et al. (2015) suggest that the auroral [OIII] λ4363 line, commonly used for T_e measurements, is the most problematic auroral line to use amongst those of [OII], [OIII] [NII], [SII], [SIII], giving rise to temperature discrepancies.

As it can be seen, each scale has different advantages and disadvantages and should be used in different metallicity regimes (see detailed discussion in e.g., Kewley & Dopita 2002; Stasińska 2002; Kewley & Ellison 2008; Moustakas et al. 2010; Dopita et al. 2013; Blanc et al. 2015). Thus, this open-source code outputs the oxygen abundance in the main 7[7] metallicity scales (with the KD02 diagnostic having four outputs and the PP04 diagnostic having two outputs). While there is a long-standing debate about which diagnostic to use, as there are systematic metallicity offsets between different methods (recombination lines vs. strong-line method vs. "direct" T_e method, see the above sources), **the relative metallicity trends can be considered robust, if the analysis is performed self-consistently in the same scale, and trends are seen across different**

[6] Note however, that most recently Berg et al. (2015) suggest that [OIII] λ4363 line is the most problematic auroral line to use amongst those of [OII], [OIII] [NII], [SII], [SIII], giving rise to temperature discrepancies.

[7] as of version v1.0, Spring 2015

scales (Kewley & Ellison 2008; Moustakas et al. 2010). Thus, it is then necessary to obtain statistical error bars for relative comparisons that are meaningful. Note however, that while there are conversion values between different scales (Kewley & Ellison 2008), they apply for large data sets, since those conversion values were derived based on ten thousands of SDSS galaxies, and thus should be used with caution (or not at all) for smaller samples In addition, one should note that there is a debate about the value of the solar oxygen abundance (Asplund et al. 2009; Caffau et al. 2011), such that the absolute oxygen calibration is still uncertain.

Here we introduce the open-source python code "**FILL IN**". pro .. In § 2 we describe our method, and the input and output values of the code. In § 4, we compare our method of obtaining abundance uncertainties to previous methods in the literature.

2. DESCRIPTION OF METALLICITY CODE

For computing oxygen abundances, we use the iterative code by Kewley & Dopita (2002), which has been updated in Kewley & Ellison (2008) and reflects ... **LISA: YOUR INPUT HERE:what is the update if any or is it exactly as in Kewley & Ellison 08??** which was initially written in IDL. We translated the code into python, and added new features, most importantly the capability of obtaining uncertainties on the metallicity outputs via Monte Carlo resampling, and made it open source on GitHub, as we explain below.

2.1. *Input and Output of code*

The input of the code is a set of spectral emission line fluxes. We assume that the observed emission lines to be used to indicate metallicity originate in HII regions and are not due to non-thermal excitation by e.g., AGN or interstellar shocks from SNe or stellar winds. Tests to exclude data contaminated by such non-thermal sources should be executed using the recommended line ratios by e.g., Baldwin et al. 1981; Kauffmann et al. 2003; Kewley et al. 2006 *prior to running this code*. Furthermore these lines should have all the correct calibration (at least correct relative calibration) and *should have a signal-to-noise ratio (S/N) of at least 3*. The latter is important for the success of the Monte Carlo resampling technique as described below.

Emission line flux values are fed into our Python implementation as in the original IDL code by Kewley & Dopita (2002), hereafter referred to as IDLKD02. The inputs are emission line flux values and their uncertainties for the following lines: Hα, Hβ, [OI] 6300, [OII] 3727, [OIII] 4959, [OIII] 5007, [NII] 6584, [SII] 6717, [SII] 6731. [SIII] 9096 and SIII 9532 can be used to calculate S_{23}, but are not often observed since they are in the NIR and thus, are currently not used to calculate metallicity with any of the diagnostics enabled by this code. The line fluxes are to be stored in an ASCII file, and the measurement errors in a separate ASCII files. (consult the README.md[8] in the GitHub repository for details about the input format, and sample files). **CHECK AT THE END WITH CODE!** If the fluxes for the specified lines are not available, the entry should be set to 'NaN' and the out-putted oxygen abundance will be calculated only for metallicity scales that use the provided line fluxes. In absence of measurement errors the flux errors entry should be set to 0.0 (and the code will not generate confidence intervals, as explained later in this Section).

As part of the code, the inputted line fluxes are corrected for reddening by using the observed Balmer decrement, for which Hα and Hβ flux values need to be provided. We assume case B recombination, and thus the standard value of 2.86 as the intrinsic Hα/Hβ ratio (Osterbrock 1989), and apply the standard Galactic reddening law with $R_V = 3.1$ (Cardelli et al. 1989). However, the user can choose other extinction laws and R_V values, if desired, given the code's open-source nature. If the input measurements are already de-reddened, the user can easily disable the reddening correction. If either Hα or Hβ are not provided the reddening correction cannot be implemented, and thus, the user is notified and has the option to proceed with the calculation with uncorrected line fluxes.

As output, we obtain metallicity values and their uncertainties in various metallicity scales. The user can choose which of the following calibrations to calculate, which have all have been implemented as prescribed in Kewley & Ellison (2008), except where noted.

- **M91** (McGaugh 1991)
- **Z94** (Zaritsky et al. 1994) which is valid for the upper branch of R_{23} only, and we conservatively constrain it to $\log(R_{23}) < 0.9$, i.e., the range that is covered by the photoionization model grids.
- **C01**: a diagnostic based on R_{23}, C01_R_{23}, and one based on [NII]/[SII], C01_N2S2 (Charlot & Longhetti 2001), **CHECK**
- **D02** (Denicoló et al. 2002) for which we include, in addition to the uncertainties in the measurements, the uncertainty on the fit parameters published in D02.
- **PP04**: (Pettini & Pagel 2004) 2 computations based on the [NII]/Hα ratio, called PP04_N2, and another based on ([OIII]/Hβ) / [NII]/Hα), called PP04_O3N2.
- **P05** (Pilyugin & Thuan 2005)
- **KD02 and KK04**: 4 computations using R_{23} (KD02_R_{23}, which is KK04 **CHECK PLEASE - YES RIGHT MM - NEED TO RESTRUCTURE**), the [NII]/[OII] ratio (KD02_N2O2), the [NII]/Hα ratio (KK04_N2Hα), and a combined method that chooses the optimal method given the input line fluxes (KD comb KK04_R_{23}, KD-comb_new) (Kewley & Dopita 2002, Kewley & Ellison 2008). **FINISH CHECK**
- **D13**: Recently, Dopita et al. (2013) has updated the photoionization models used in KD02 and in KK04 by including new atomic data within a modified photoionization code and by no longer assuming that the energy distribution of the electrons in

[8] https://github.com/nyusngroup/MC_Metalicity/blob/master/README.md

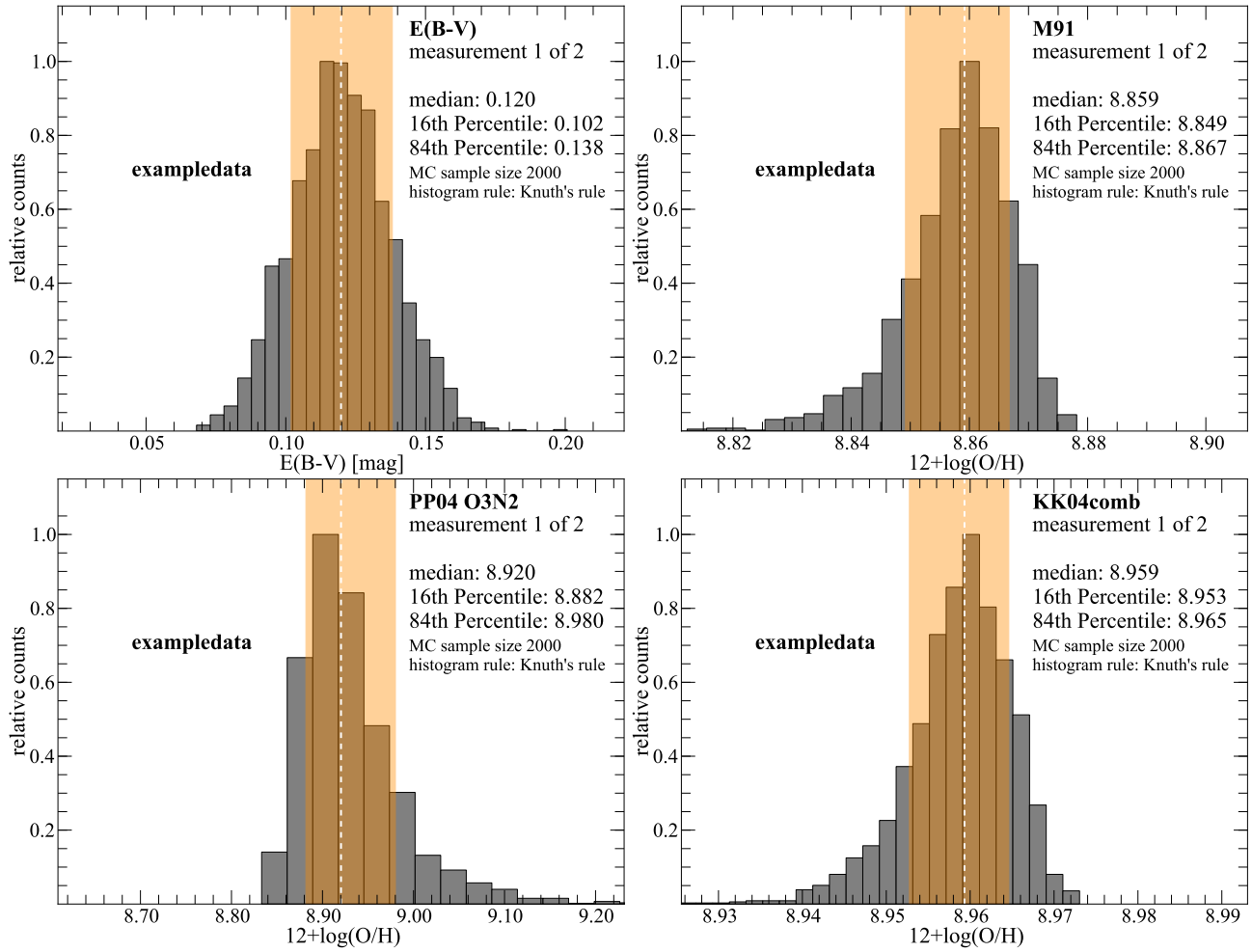

FIG. 1.— Metallicity and reddening E(B-V) parameter distributions based on the example data shown in 1: emission line data of the HII regions at the position of SN 2008D, published in Modjaz et al. (2011). The distribution are generated from N=2,0000 samples. The median values are shown with the dashed lines, while the region between the 16th and the 84th percentile is shaded (orange). We show the metallicity scaled from Kobulnicky & Kewley (2004), updated as described in Kewley & Ellison (2008) (KK04_comb), Pettini & Pagel (2004), using OIII and NII (PP04 O3N2) and McGaugh (1991) (M91). Similar plots are outputted by the code for each metallicity scale calculated. Each plot indicates: the scale, the sequential number of measurement in input, corresponding to a line of the input file, the median, 16th and 84th percentile values, and the method to choose the bin size for the histogram (Kuth's rule in this case, see Section 2.2.1).

the HII regions is a simple Maxwell-Boltzmann distribution (as assumed in prior works), but a more realistic κ distribution, as observed in solar system astrophysical plasma (Nicholls et al. 2012). If the user has installed their publicly available `pyqz`[9] Python module, [NII], [SII], [OIII], Hα, and Hβ lines are fed to the `pyqz` module, which produces up to 8 emission line ratio diagnostics for 12+log(O/H), each using two of the line ratios [NII]/[SII], [NII]/Hα, [OIII]/[SII], and [OIII]/Hβ. Our code sets the κ parameter to 20, which is the value that Dopita et al. (2013) found best resolves the inconsistencies between oxygen abundance values derived from the strong-line methods vs. the "direct" T_e method [10].

- **DP00 & P01 (deprecated)**: (Díaz & Pérez-Montero 2000 and Pilyugin 2001, respectively) are also available, but deprecated: P01 is superseded by P05. Therefore they are not part of the default output; however, they are available upon explicit user request, via the command line options.

All of the above calibrations, except those of D13 (i.e., `pyqz`) and of DP00, are calculated in the original IDLKD02 code, and are discussed in detail in Kewley & Dopita (2002); Kewley & Ellison (2008). DP00 is the only diagnostic that relies on sulfur ratios: $S_{23} = ([\mathrm{SII}]6717 + [\mathrm{SIII}]9069)/\mathrm{H}\beta$. The shortcomings of S_{23} as a tool to measure abundances are discussed in Kewley & Dopita (2002), and include the fact that it is double-valued for all metallicities, that it depends on the ionization parameter, and that the sulfur-to-oxygen ratio is poorly determined and difficult to model due to large ionization correction factors that are needed to account for the presence of unobserved ionization states. The DP00 diagnostic we implement is corrected with the addition of a term $\propto (c + S_{23}^3)^{-1}$ as suggested by (Kewley & Dopita 2002), which corrects the tendency of the scale to systematically underestimates the abundance, with a discrepancy growing larger at higher metallicity. How-

[9] `https://datacommons.anu.edu.au:8443/DataCommons/item/anudc:5037`

[10] The user can modify the value of κ by editing the code, if they wish.

ever, we point out the scatter in the metallicity derived from this diagnostic compared to others remain high.

If the line fluxes necessary for specific scales are not provided, the output metallicities will default to 'NaN'. If the errors in the measurements are not provided, the code will specify that it cannot create a measurement distribution and determine a confidence interval, but it will calculate the nominal metallicity and output it.

The distributions of E(B-V), $\log(R_{23})$ are outputted, together with those of the oxygen abundance in the various scales. While certain parameters, such as the ionization parameter q and the electron density (using the SiII lines) are computed, as long as the necessary lines are provided, they are not outputted in the current version of our code − however, the reader can easily modify the code to suite their needs, given it is an open-source code.

2.2. *Computing Uncertainties*

The novel aspect of our work is that for every set of input line measurements we introduce a Monte Carlo (MC) resampling method to obtain iterations via random sampling within the measurement errors, and thus we obtain a robust result for error estimation (e.g., Efron 1979; Hastie et al. 2009; Andrae 2010).

Given a data set with error bars from which certain parameters are estimated, Monte Carlo resampling generates synthetic data samples drawing from a given distribution. Here we draw synthetic data from a Gaussian distribution centered on each measured line flux value, with standard deviation corresponding to the measurement error. The implicit assumption is made, of course, that the line flux error is Gaussian distributed in nature[11].

For each metallicity scale, for each of N values chosen randomly within the relevant emission line distributions we run the calculations that computes the metallicity. This effectively simulates conducting multiple experiments when repeating observations is impractical or impossible, as in the case of the emission line flux data, and thus generates alternative data sets. The sample size N is set by the user, and one should expect an appropriate value of N to be a few 1000s, depending on the metallicity scale chosen and measurement errors (for example $N = 2,000$ is determined to be sufficient for our example data, as shown in Section 2.3, and we provide tools to assure the sample size is sufficiently large, which we describe in Section 2.3). A distribution of parameter estimates for the oxygen abundance is generated for each scale, from which the median metallicity and its confidence region are calculated, and the results are binned and visualized in a histogram (see Section 2.2.1). This is done for each scale the user chooses to calculate. The fiftieth (50%) percentile, i.e. the median, is reported, as well as the 16th and 84th percentiles of the metallicity estimate distribution as its confidence region. However, the user can choose to also output the full metallicity parameter distribution as ASCII files, in addition to the plots.

This MC resampling approach takes into account the impact of the uncertain reddening (due to the uncertainties in the measurement of the Hα and Hβ fluxes), when the option for de-reddened metallicities is chosen. For each synthetic set of measurements a new reddening value is calculated based on the resampled Hα and Hβ fluxes, and used to compute the de-reddened metallicity value, thus the derived distribution of metallicity values takes into account the uncertain redding. As part of the output, a parameter estimate distribution plot for E(B-V) is provided as well (first panel in figure 1), along with confidence intervals derived using the same method as for the metallicity measurements. If either the Hα or the Hβ flux is not provided, no reddening correction can be applied. The computed metallicity will not be reddening-corrected and the E(B-V) output will be set to zero.

Figure 1 shows the metallicity estimate distribution for three representative scales (M91, PP04_O3N2, and KD02comb), and for the reddening parameter E(B-V) - similar plots that are out-putted by our code for all scales as listed above (not all shown here). Although the input distributions are Gaussian, the metallicity distributions are not, for two reasons: first, since the metallicities are computed based on log values of line flux ratios, symmetric error bars in linear space will translate into asymmetric error bars in log space; and second, some metallicity scale computations are non-linear, and sometimes bimodal, especially those that include R_{23}), since they choose upper vs lower branch to break the degeneracy.

Since the metallicity distributions are not Gaussian, the percentiles we report cannot be expressed in terms of σ values. In determining the confidence region for asymmetric and multi-modal distributions, there are broadly three approaches (e.g., Andrae 2010): choosing a symmetric interval, the shortest interval, or a *central* interval. With the central method we determine the confidence interval by choosing the left and right boundaries such that the region outside the confidence interval on each side contains 16% of the total distribution - in analogy to the one-sigma-interval of a Gaussian distribution. This ensures that the algorithm finds the proper boundaries even for asymmetric, non-Gaussian distributions, and in the case of multiple peaks. In summary, the output for the measured value corresponds to the fiftieth (50%) percentile, while the lower error bar corresponds to the 50^{th}-16^{th} percentile and the upper error bar corresponds to 84^{th}-50^{th} of the metallicity estimate distribution.

The distributions for the D02 scale include the uncertainty in the fit parameters: the oxygen abundance in this scale is generated as $12 + \log_{10}(\frac{O}{H}) = 9.12 \pm 0.05 + (0.73 \pm 0.10)$ NII (Denicoló et al. 2002). The parameters of the fit are generated as the sum of the nominal parameters (9.12 and 0.73) and a Gaussian distributed random value centered on zero, and within a standard deviation of 0.05 and 0.10, respectively, in the above units.

We note that our code does not output the *systematic* uncertainty of each scale, which, for example, is ∼0.15 dex for KD02. However, if all metallicity measurements are in the *same* scale and only *relative* comparisons are made, as recommended by a number of authors, then the systematic error does not have any impact (by definition!). Nevertheless, it is then necessary to obtain sta-

[11] Users may wish to provides their own probability distribution for the emission line uncertainties, and modify the code to suite their needs.

tistical error bars, as computed via the code we are releasing here, for meaningful relative comparisons.

2.2.1. *Visual diagnostics*

In order for the user to check the validity of a measurement, and to better understand the distribution, we provide two visualizations: for each set of input line fluxes, we generate a histogram of the output distribution in all metallicity scales (Figure 1 and 2), and for each set of input line fluxes we generate a *box-and-whiskers* plot (hereafter *boxplot*, for short) summarizing the result of all scales calculated (Figure 3).

Choosing the binning size for a histogram is not a trivial task. Hogg (2008) describes various data analysis recipes for selecting a histogram bin size. Too many bins will result in many empty bins and an "over-fit" histogram, while too few bins may miss features of the distribution. By default, we use *Knuth's Method* to choose the number of bins $N_{\rm bins}$ for each histogram. Knuth's method optimizes a Bayesian fitness function across fixed-width bins (Knuth 2006). Additionally, however, we enable a number of binning options from which the user can choose, including: the square root of the number of bins, *Rice rule* ($N_{\rm bins} = 2\sqrt[3]{N}$, e.g., Hastie et al. 2009), *Doane's formula* ($N_{\rm bins} = 1 + \log_2 N + \log_2\left(1 + {\rm Kurt}\sqrt{(N/6)}\right)$, where Kurt is the third moment of the distribution, Doane 1976[12]), and the full Bayesian solution, known as Bayesian Blocks, which optimizes a fitness function across an arbitrary configuration of bins, such that the bins are of variable size (Scargle et al. 2013). The implementation of the latter method requires the `astroML` python package to be installed on the user's system (Vanderplas et al. 2012[13]). If the `astroML` package is not found, the code will default to Knuth's Rule. As mentioned, Knuth's method implies an optimization. In cases in which the convergence of this minimization takes too long (or if the number of bins after the minimization is $N_{\rm bins}/\sqrt{N} > 5$ or $N_{\rm bins}/\sqrt{N} < 1/3$) the code will revert to Rice rule. Some methods may be computationally prohibitive with a very large sample size, or very little computational power, such as the Bayesian methods that that rely on optimization (Knuth's method and the Bayesian Block method) in which case the user may choose to use Doane's formula, Rice rule, or even the square root of the number of samples, to choose the bin size for the histograms.

Fed: please include more info and the motivation for the Kernel density plot. Lastly, the user can generate and visualize a *Kernel Density* if the `sklearn` package is installed. The Kernel Density of the distribution is then calculated via *KD Tree* with a top-hat function, as explained in the `sklearn` package documentation[14]. The results will then show both a histogram, with $N_{\rm bins}$ chosen via Knuth's method, as well as the distribution Kernel Density, as shown in Figure 2.

[12] Doane (1976) attempted to address the issue of finding the proper number of bins for the histogram of a skewed distribution. Several versions of the so-called Doane's formula can be found in the literature. Our formula can be found, for example, in Bonate 2011

[13] `https://github.com/astroML/astroML`

[14] `http://scikit-learn.org/stable/modules/density.html`

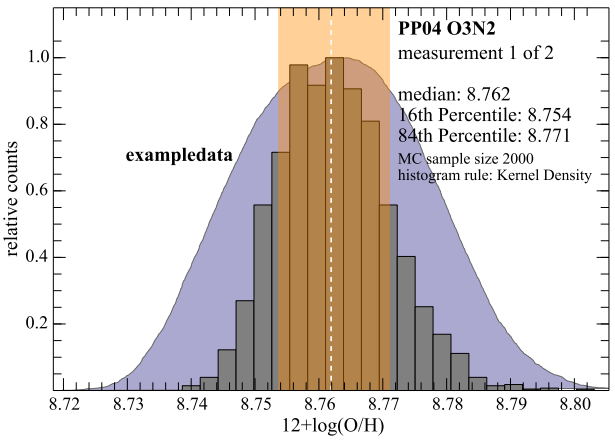

FIG. 2.— The distribution of values for the Pettini & Pagel (2004) scale (PP04 O3N2) for the first measurement of our example data (Table 1) is plotted. The histogram is generated from the $N = 2,000$ samples distribution using Knuth's rule to choose the bin size (Section 2.2.1), and the shaded region represent the Kernel Density estimate for the distribution, calculated via *KD Tree* with a top-hat function.

The boxplot summarizes the result from each scale the user chooses to calculate. For each scale the median of the $12+\log_{10}(\frac{O}{H})$ distribution is plotted as a black horizontal line. The height of the corresponding box represents the 25^{th} percentile of the $12+\log_{10}(\frac{O}{H})$ distribution. The bars represent the maximum and minimum value of the distribution, excluding outliers. The outliers, that are plotted as circles, are defined as all data points farther than $1.5\times$IQR, where IQR is the the *interquartile range* (and the length of the box) from the 25^{th} percentile (i.e. from either end of the box). The Solar oxygen abundance is indicated in this plot for comparison: a gray box shows a range of estimated values for Solar oxygen abundances, from $12+\log_{10}(\frac{O}{H})$=8.69 (Asplund et al. 2009) to $12+\log_{10}(\frac{O}{H})$=8.76 (Caffau et al. 2011). Notice that only the diagnostics requested by the user have a slot in the plot (in the example in Figure 2 the computed scales are M91, the PP04 scales, and the KD02 scales). However these slot exists on the plot whether the diagnostic can be produced or not, i.e. if the set of input lines does not allow a requested scale to be calculated an empty column will be generated in this plot in correspondence of said metallicity scale.

2.3. *Visual diagnostic to assure completeness of the MC simulation*

The user can choose the N, number of samples to be generated. Of course the reliability of the metallicity estimates depends crucially on the sample size being sufficiently large to properly characterize the distribution of metallicity values. It is however not trivial to decide when N is sufficiently large. As soon as N is large enough, and the distribution is well characterized, adding synthetic data will not change its shape. Consider a cumulative distribution for a metallicity scale, D02, for example, which has noise from the measurement errors, as well as from the error in the fit parameters, or KD02, which is a non linear combination of the input line flux values. We plot the cumulative distribution for four randomly selected subsamples of the data points in the dis-

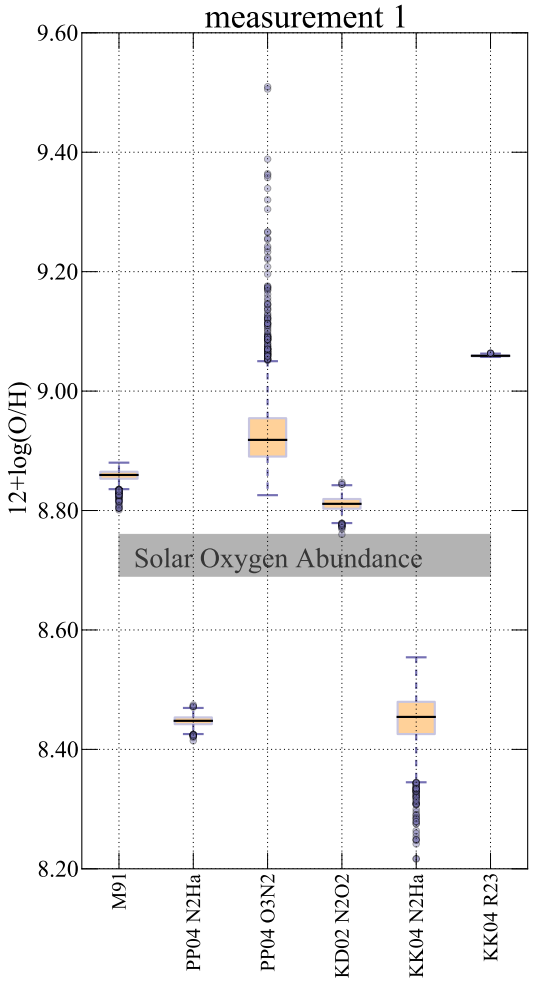

FIG. 3.— A box plot shows the comparison of the results of 6 metallicity scales calculated from the same set of measured lines (Table 1). For each scale, the median of the resulting distribution is plotted as a horizontal line, the inter quartile range (IQR) is represented as an orange box, and the bars, joined to each end of the box by a dashed line, represent the minimum and maximum of the distribution *excluding outliers*, where outliers are defined as any point farther than 1.5× IQR from the edges for the IQR.

tribution, of size 10%, 25%, 50%, 75% of N, as well as for the entire distribution. If 1/10th of the points were a sufficiently large Monte Carlo sample, the cumulative distributions would typically appear smooth and, most importantly, they would overlap. In Figure 4 we show the these cumulative plots for a distribution generated with $N = 200$, $N = 2,000$, and $N = 20,000$, for D02 and KD02comb. Note that the final sample size is 10% larger than the user chosen parameter N. The sample size is in fact automatically increased by 10% at the beginning of the run, in order to assure that even if during the calculations some of the output metallicities were to result in non-valid values ('NaN"s or infinities, for example, if devision by very small numbers is required) the actual sample size is at least as large as the user intended. With reasonable input parameters, the code rarely produces non-valid values, and if the size of the valid output distribution were in the end smaller than the *requested* value N, i.e. if the number of non-valid outputs is larger than 10% of N, the user should in fact worry about the set of input parameters used.

The cumulative distributions for the subsets of the $N = 200$ simulation (left panel in Figure 4) are different, and noisy, indicating that 200 data points is not a sufficiently large dataset. Conversely, at $N = 20,000$ (right panel) the distributions are indistinguishable, indicating that $N = 2,000$ (the smaller subset plotted) is large enough. The $N = 2,000$ cumulative distributions rapidly converge, as the subsample size increases, indicating that a value of N between 200 and 2,000 will characterize the distribution appropriately for our example data set. In the light of this we choose, conservatively, $N = 2,000$ to run our simulations for this dataset.

We remind the user that the appropriate number of N samples will depend on the diagnostic, and on the input data. The errors on the measurements, and the calculations performed to derive the metallicity from line ratios, which are for many scales non-linear, will determine the shape of the distribution, and thus the number of datapoints that are needed to fully characterize it.

2.4. *Performance and Benchmarks*

We ran a bench mark calculation on a MacBook Air with an Intel Core i7 (1.7 GHz) and 8GB of 1600 MHz DDR3 memory. The dataset we used for the calculation includes 2 sets of line measurements (measurement 1, or M1, and measurement 2, M2). The flux values and their associated errors are shown in Table 1. The code performs simple algebraic operations on large data arrays. It is vectorized along the sample-size dimension, so that the code loops over the smaller dimension corresponding to the number of sets of line measurements in input (2 in our example data), while operations are generally performed on the N-dimensional vectors storing the synthetic measurements, and the N-dimensional variable hence derived (certain scales, such as the KD02 and KK04 ones, derive the ionization parameter in an iterative fashion requiring further loops).

With full graphic output (all histograms are being plotted) and performing all default calculations, except those of D13 *pyqz*, for our example data sets and a sample size $N = 2,000$, the time required by the code is ~ 43 seconds (wall clock time), less than 35 seconds of actual CPU time, and less then a second of CPU time spent in the kernel within the process. Including the scales of D13 *pyqz* for the same datasets, the run time becomes ~ 78 seconds, with ~ 65 seconds of actual CPU time. The code running time and the computational time scale roughly linearly with the number of measurements in input (with of course dependence on which lines are available in each measurement). The code time was tested on sets of 1, 2, 5, 10, 25, 50 and 100 identical measuremente (M1 of our example dataset) and the clock is found to scales with a steep slope of ~ 18 with the number of measurements, while for the CPU time spent in the kernel we find a very shallow dependency of of ~ 0.12 (this is the actualy computational time: most clock time is in fact spent in input-puout and plotting activities).

The time spent on plotting functions, which includes

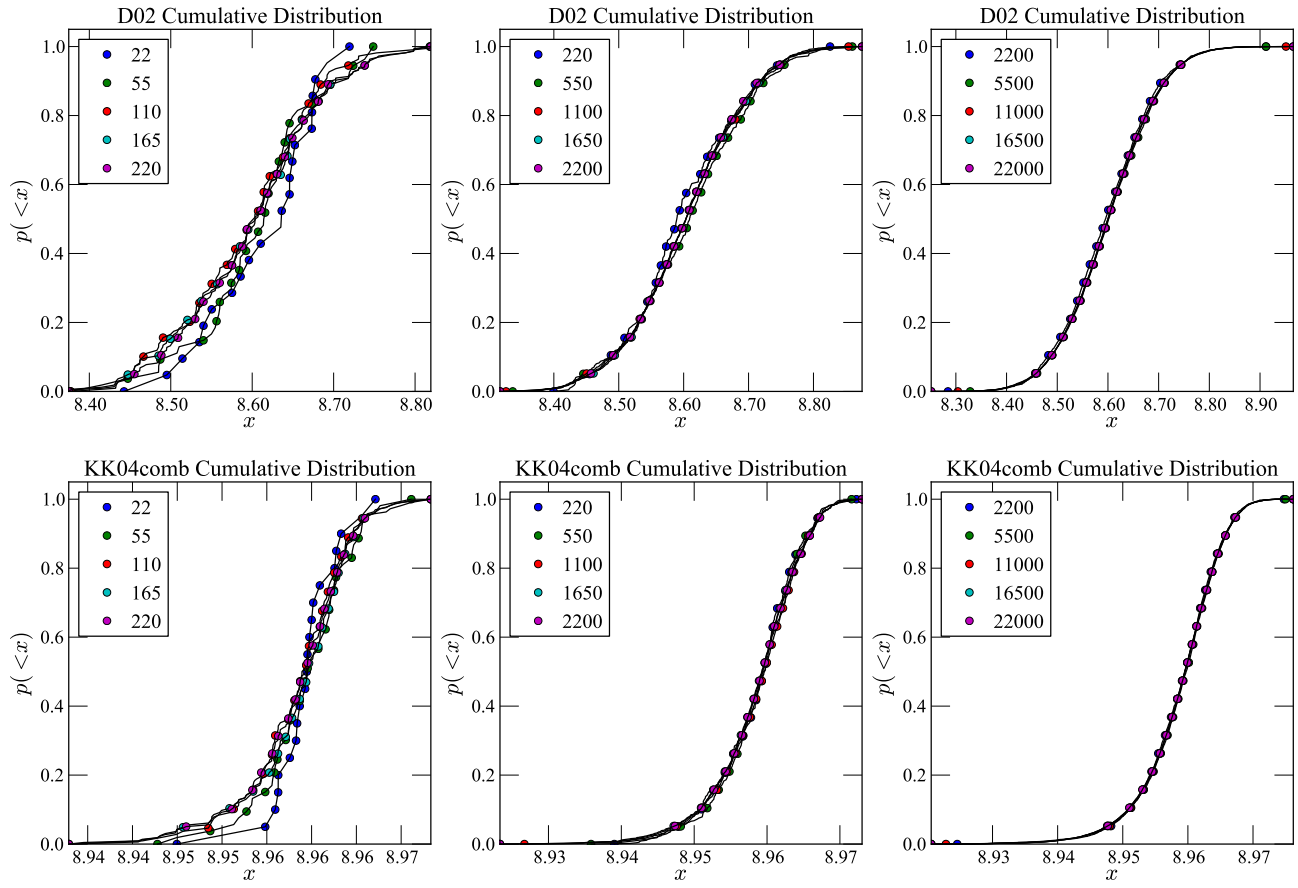

FIG. 4.— Cumulative plots of the distribution of metallicity values for the D02 (Denicoló et al. 2002) and KD02 scale (Kewley & Dopita 2002, updated by Kewley & Ellison 2008), chosen here just as examples, where x refers to $12 + \log_{10}(\frac{O}{H})$. The input used is example data 1. In each plot the cumulative distribution is shown for a randomly chosen subsample of 10% of the data in the sample, as well as a subsample of 25%, 50%, and 75%, of the data, and for all data in the distribution. In the left plots the distributions are generated from an $N = 200$ samples, in the center top and bottom plots from an $N = 2,000$ and in the right-most column from a $N = 20,000$ sample. The increasing overlap of the distributions informs us about completeness. In the left plots the distributions do not fully overlap, indicating that completeness is not achieved with the $N = 200$ sample. On the other end, since all subsamples are indistinguishable in the top and bottom plots on the right, which are generated with $N = 20,000$ samples, we conclude that completeness is already achieved at 10% of the $N = 20,000$ sample (the smallest sample in the plots on the right) for our example data.

TABLE 1
EXAMPLE DATA AND THEIR UNCERTAINTIES BASED ON DATA IN MODJAZ ET AL. (2011)

	[OII]3727	Hβ	[OIII]4959	[OIII]5007	[OI]6300	Hα	[NII]6584	[SII]6717	[SII]6731
M1	1.842 (.053)	0.958 (.032)	NaN	0.302 (.029)	0.127 (.021)	4.746 (.026)	1.642 (.022)	0.941 (.021)	0.543 (.019)
M2	2.875 (.101)	1.251 (.044)	0.168 (.028)	0.064 (.025)	NaN	4.026 (.069)	0.781 (.033)	0.821 (.035)	0.573 (.032)

the calculation of the appropriate number of bins for each scale, is substantial: 1.67 seconds per distribution on average with 14 calls for this data set (one for each metallicity scale, E(B-V), and R_{23}). In Table 2 we summarize the time spent on each metallicity scale for M1, the first measurement of the sample dataset, for each metallicity scale, and we visualize the run time and memory usage in Figure 5. While the CPU usage is modest, the memory usage can be large, depending on the size of the sample. The memory usage ramps up quickly, as soon as the line samples are created, and remains fairly constant throughout the run. Figure 5 shows the memory used by each function in the code as a function of time for a single set of input lines of our example data (M1). The scales that take longer time require the root finding (e.g. M08), or optimizations which are done iteratively (e.g. KD combined).

2.5. *Availability*

The source code is published under MIT-licensed on GitHub[15]. At this time the code is released under DOI NUMBER HERE as version 1.0: NAME HERE. Project details, step-wise tutorials, and further information can be found in the module README [16]. Development is done in Linux and OS X. The package requires standard python packages, such as `numpy`, `scipy`, `pylab`, and additional features are enabled if the packages `astroml`,

[15] `repoURLhere`
[16] `README.mdURL`

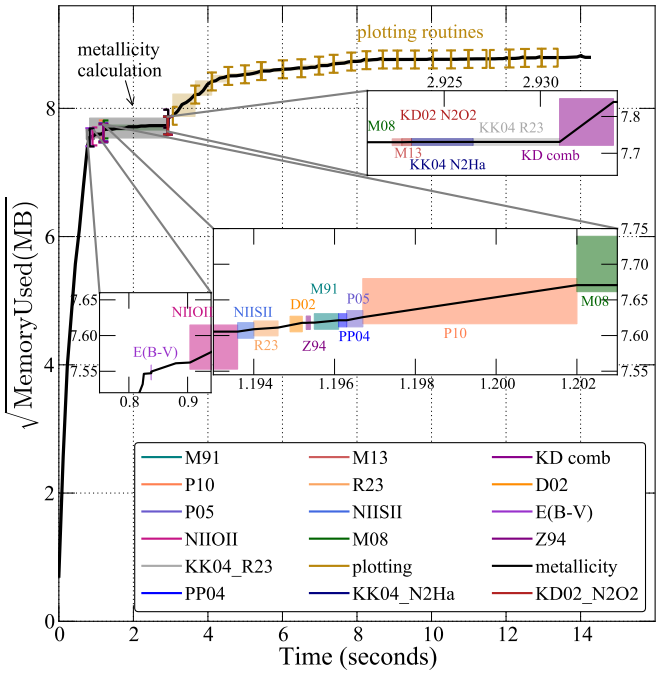

FIG. 5.— Memory usage: we plot the square root of the memory usage in Megabites as a function of time for running our code (using N=2,000 and all metallicity scales except the D13 *pyqz* ones) on a single set of measured emission lines (Table 1, M1). The square root is plotted, instead of the natural value, to enhance visibility. Two inserts show zoom-ins of the region where most of the metallicity scales are calculated, since the run time of the code is dominated by the calculation for the KD combined scale, and, after all scales are computed, by plotting routines, including the calculation of the bin size with Knuth's rule. Each function call is represented by an opening and closing bracket in the main plot, and by a shaded rectangle in the zoom-in inserts.The calculation of NIIOII, which requires 0.32 seconds, is split between the two inserts. Altogether, the calculation of metallicity in all scales, except those in the D13 *pyqz* scales, takes 5.96 seconds. For technical details about the D13 *pyqz* scales performance we refer the reader to the *pyqz* package.

TABLE 2
CPU USAGE BY SCALE

Scale	Time ($\times 10^{-6}$ sec)
E(B-V)	100.1
R_{23}	599.9
M91	600.1
Z94	100.1
P05	399.8
D02	300.2
PP04	200.0
M08	1.720×10^{6}
P10	5300
M13	500.0
pyqz	1.127×10^{6}
KK04 R_{23}	4500
KD02 N2O2	500.2
KD02 N2Ha	3200
KD combined	2800

`skitlearn`, and `pyqz` are installed, but the packages are *not* required. Contact the authors to be included in a mailing list and be notified about critical changes.

3. COMPARISON TO PRIOR UNCERTAINTY COMPUTATION AND OTHER WORKS

A previous method for determining the uncertainty in the oxygen abundance (as used in Modjaz et al. 2008; Kewley et al. 2010; Rupke et al. 2010; Modjaz et al. 2011) was an *analytic* approach of propagating the emission-line flux uncertainties: it found the maximum and minimum abundances via maximizing and minimizing, respectively, the various line ratios by adding/subtracting to the measured line values their uncertainties. For comparison we computed the metallicities and their errors in both ways (both analytic and using our current MC resampling method) for 3 representative scales. We plot our results and the residuals in Fig.6, which shows a number of important points: i) The metallicity reported as the 50th percentile of the metallicity parameter distribution from the MC resampling method is completely consistent with the analytically derived metallicity - well within the respective error bars - and thus, the prior published results still stand (unsurprisingly, since our code, aside for the calculation of the confidence interval, uses the same algorithms developed for IDLKD02). ii) The MC resampling method has smaller error bars than the analytic method, especially for the scales of M91 and KD02. This is easily understandable, since the analytic method assumes the worst-case-scenarios, as it basically yields 2 metallicity parameter draws (the "minimum" and "maximum") which are in the tail of the full metallicity probability distribution. However, the MC resampling method is the more appropriate method as it empirically characterizes the full parameter estimation distribution.

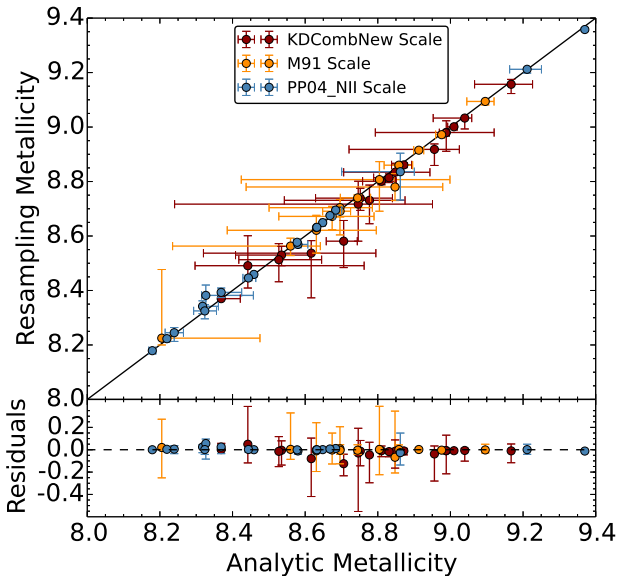

FIG. 6.— Comparison of metallicity estimation between the analytic method and our Monte Carlo resampling method (top) and their residuals (bottom) for three different metallicity scales. Flux measurements come from 19 galaxies previously measured in Modjaz et al. (2011). To add asymmetric errors in quadrature we use $\text{residual}_{\text{min}} = \sqrt{x^2_{\text{max}} + y^2_{\text{min}}}$ and $\text{residual}_{\text{max}} = \sqrt{x^2_{\text{min}} + y^2_{\text{max}}}$. Note that for the scales of KD02-comb of M91, the MC resampling error bars are smaller than those of the analytic propagation, which assumes worst-case scenarios, while metallicity reported as the 50th percentile of the metallicity parameter distribution from the MC resampling method is completely consistent with the analytically derived metallicity in all scales.

3.1. *Comparison with other works*

The field of SN host metallicity studies has been rapidly developing as these kinds of studies may be crucial avenues for constraining the progenitor systems of different kinds of explosions - however, a few of the works do not compute errors and others not show how they compute their statistical errors (e.g., Anderson et al. 2010; Leloudas et al. 2011; Sanders et al. 2012; Leloudas et al. 2014).

In contrast, the general metallicity field has considered in detail how to estimate the uncertainties in measured metallicities- however, none of those codes are open-source and many of them are for specific scales which were chosen by the authors: Moustakas et al. (2010) also use MC resampling to estimate the metallicity uncertainties (in their case using N=500 trials and assuming a Gaussian distribution) but only do this for two scales, KK04 and P05. For computing the metallicities of the SDSS star forming galaxies, Tremonti et al. (2004) fit a combination of stellar population synthesis models and photoionization models to the observed strong emission lines [OII], $H\beta$, [OIII], $H\alpha$, NII and SII and report the median of the metallicity likelihood distribution as the metallicity estimate, with the width of the distribution giving the 1 σ (Gaussian) error. However, this constitutes their own scale (the T04 scale).

In the last stages of preparing this manuscript Blanc et al. (2015) was published. Blanc et al. (2015) employ Bayesian inference for doing something similar to Tremonti et al. (2004) - they use Bayesian inference to derive the joint and marginalized posterior probability density functions for metallicity Z and ionization parameter q given a set of observed line fluxes and an input photoionization model. They provide a publicly available IDL implementation of their method named IZI (inferring metallicities (Z) and ionization parameters) on the author's web site.

4. CONCLUSIONS

FINISH. We hope that this open-access code will be helpful for the many different fields where gas-phase metallicities are important, including in the emerging field of SN and GRB host galaxies, where either it is not described how they got uncertainties or no error bars are computed. Given its public-access nature, the users are free to include any new metallicity diagnostics and modify any parts and assumptions (e.g. that the line fluxes are Gaussian distributed).

The Modjaz SNYU group at NYU is supported in parts by the NSF CAREER award AST-1352405 and by NSF award AST-1413260. F. B. Bianco is supported by a *James Arthur Fellowship* at the NYU-Center for Cosmology and Particle Physics and Y. Liu by a *James Arthur Graduate Award*. This code made use of several Python Modules, including `Matplotlib` (Hunter 2007). Some plots are produced with public code DOI:10.5281/zenodo.15419 available at `https://github.com/fedhere/residuals_pylab`. This research made use of NASA Astrophysics Data System; the NASA/IPAC Extragalactic Database (NED), which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

Anderson, J. P., Covarrubias, R. A., James, P. A., Hamuy, M., & Habergham, S. M. 2010, MNRAS, 407, 2660
Andrae, R. 2010, ArXiv e-prints, arXiv:1009.2755
Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, ARA&A, 47, 481
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Berg, D. A., Croxall, K. V., Skillman, E. D., et al. 2015, ArXiv e-prints, arXiv:1501.02270
Blanc, G. A., Kewley, L., Vogt, F. P. A., & Dopita, M. A. 2015, ApJ, 798, 99
Bonate, P. 2011, Pharmacokinetic-Pharmacodynamic Modeling and Simulation, SpringerLink : Bücher (Springer)
Caffau, E., Ludwig, H.-G., Steffen, M., Freytag, B., & Bonifacio, P. 2011, Sol. Phys., 268, 255
Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245
Charlot, S., & Longhetti, M. 2001, MNRAS, 323, 887
Denicoló, G., Terlevich, R., & Terlevich, E. 2002, MNRAS, 330, 69
Díaz, A. I., & Pérez-Montero, E. 2000, MNRAS, 312, 130
Doane, D. P. 1976, The American Statistician, 30, 181
Dopita, M. A., Sutherland, R. S., Nicholls, D. C., Kewley, L. J., & Vogt, F. P. A. 2013, ApJS, 208, 10
Efron, R. 1979, Ann. Stat., 7, 1
Grevesse, N., Asplund, M., Sauval, A. J., & Scott, P. 2010, Ap&SS, 328, 179
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Science+Business Media, New York)
Hogg, D. W. 2008, ArXiv e-prints, arXiv:0807.4820
Hunter, J. D. 2007, Computing In Science & Engineering, 9, 90
Johnson, J. L., & Li, H. 2012, ApJ, 751, 81
Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, MNRAS, 346, 1055
Kelly, P. L., & Kirshner, R. P. 2012, ApJ, 759, 107
Kewley, L. J., & Dopita, M. A. 2002, ApJS, 142, 35
Kewley, L. J., & Ellison, S. L. 2008, ApJ, 681, 1183
Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961
Kewley, L. J., Rupke, D., Zahid, H. J., Geller, M. J., & Barton, E. J. 2010, ApJ, 721, L48
Knuth, K. H. 2006, ArXiv Physics e-prints, physics/0605197
Kobulnicky, H. A., & Kewley, L. J. 2004, ApJ, 617, 240
Leloudas, G., Gallazzi, A., Sollerman, J., et al. 2011, A&A, 530, A95
Leloudas, G., Schulze, S., Kruehler, T., et al. 2014, ArXiv e-prints, arXiv:1409.8331
Levesque, E. M., Berger, E., Kewley, L. J., & Bagley, M. M. 2010, AJ, 139, 694
Lunnan, R., Chornock, R., Berger, E., et al. 2014, ApJ, 787, 138
Marino, R. A., Rosales-Ortega, F. F., Sánchez, S. F., et al. 2013, A&A, 559, A114
McGaugh, S. S. 1991, ApJ, 380, 140
Modjaz, M. 2012, in IAU Symposium, Vol. 279, IAU Symposium, 207–211
Modjaz, M., Kewley, L., Bloom, J. S., et al. 2011, ApJ, 731, L4
Modjaz, M., Kewley, L., Kirshner, R. P., et al. 2008, AJ, 135, 1136
Moustakas, J., Kennicutt, Jr., R. C., Tremonti, C. A., et al. 2010, ApJS, 190, 233
Nicholls, D. C., Dopita, M. A., & Sutherland, R. S. 2012, ApJ, 752, 148
Osterbrock, D. E. 1989, Astrophysics of Gaseous Nebulae and Active Galaxies (Mill Vallery: University Science Books)
Pagel, B. E. J., Edmunds, M. G., Blackwell, D. E., Chun, M. S., & Smith, G. 1979, MNRAS, 189, 95
Pan, Y.-C., Sullivan, M., Maguire, K., et al. 2014, MNRAS, 438, 1391

Pettini, M., & Pagel, B. E. J. 2004, MNRAS, 348, L59
Pilyugin, L. S. 2001, A&A, 369, 594
Pilyugin, L. S., & Thuan, T. X. 2005, ApJ, 631, 231
Rupke, D. S. N., Kewley, L. J., & Chien, L.-H. 2010, ApJ, 723, 1255
Sanders, N. E., Soderberg, A. M., Levesque, E. M., et al. 2012, ApJ, 758, 132
Scargle, J. D., Norris, J. P., Jackson, B., & Chiang, J. 2013, ApJ, 764, 167
Simón-Díaz, S., & Stasińska, G. 2011, A&A, 526, A48+
Stasińska, G. 2002, ArXiv Astrophysics e-prints, astro-ph/0207500
Stasińska, G. 2010, in IAU Symposium, Vol. 262, IAU Symposium, ed. G. R. Bruzual & S. Charlot, 93–96
Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
Vanderplas, J., Connolly, A., Ivezić, Ž., & Gray, A. 2012, in Conference on Intelligent Data Understanding (CIDU), 47 –54
Zaritsky, D., Kennicutt, Jr., R. C., & Huchra, J. P. 1994, ApJ, 420, 87