Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST

I. Baronchelli[1] , C. M. Scarlata[2] , G. Rodighiero[1] , L. Rodríguez-Muñoz[1] , M. Bonato[3,4] , M. Bagley[5] , A. Henry[6] ,
M. Rafelski[6,7] , M. Malkan[8] , J. Colbert[9], Y. S. Dai (戴昱)[10] , H. Dickinson[2,11] , C. Mancini[1] , V. Mehta[2] ,
L. Morselli[1], and H. I. Teplitz[9]

[1] Dipartimento di Fisica e Astronomia, Università di Padova, vicolo Osservatorio, 3, I-35122 Padova, Italy; ivano.baronchelli@unipd.it
[2] MN Institute for Astrophysics, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA
[3] INAF-Istituto di Radioastronomia and Italian ALMA Regional Centre, Via Gobetti 101, I-40129, Bologna, Italy
[4] INAF-Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122, Padova, Italy
[5] College of Natural Sciences, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA
[6] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[7] Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
[8] Department of Physics and Astronomy, UCLA, Physics and Astronomy Building, 3-714, LA CA 90095-1547, USA
[9] IPAC, Mail Code 314-6, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA
[10] Chinese Academy of Sciences South America Center for Astronomy (CASSACA)/NAOC, 20A Datun Road, Beijing 100101, People's Republic of China
[11] School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
Received 2019 November 28; revised 2020 May 25; accepted 2020 June 5; published 2020 July 13

Abstract

Future surveys focusing on understanding the nature of dark energy (e.g., Euclid and WFIRST) will cover large fractions of the extragalactic sky in near-IR slitless spectroscopy. These surveys will detect a large number of galaxies that will have only one emission line in the covered spectral range. In order to maximize the scientific return of these missions, it is imperative that single emission lines are correctly identified. Using a supervised machine-learning approach, we classified a sample of single emission lines extracted from the WFC3 IR Spectroscopic Parallel survey, one of the closest existing analogs to future slitless surveys. Our automatic software integrates a spectral energy distribution (SED)-fitting strategy with additional independent sources of information. We calibrated it and tested it on a "gold" sample of securely identified objects with multiple lines detected. The algorithm correctly classifies real emission lines with an accuracy of 82.6%, whereas the accuracy of the SED-fitting technique alone is low (∼50%) due to the limited amount of photometric data available (⩽6 bands). While not specifically designed for the Euclid and WFIRST surveys, the algorithm represents an important precursor of similar algorithms to be used in these future missions.

1. Introduction

A spectroscopic measurement provides the most precise estimate of the redshift of a given source. Even the most precise spectroscopic surveys, however, have a fraction of spectroscopic failures due to incorrectly identified spectral features (see, e.g., Hasinger et al. 2018).

The typical approach adopted in a spectroscopic survey exploits a flagging scheme used to characterize the reliability of a galaxy's redshift measurement. This approach is well exemplified by the VIMOS Very Large Telescope deep survey (Le Fèvre et al. 2005), where the quality of the emission lines' classification is expressed through a four-step scale (or flag). The probability that a classification is not correct can be estimated for each flag value using different methods. In the VIMOS survey, even the best-quality sample shows a residual ∼0.5%–1.0% chance of lines being incorrectly identified. This fraction rises to up ∼50% for the lowest-quality sample. The effect of a misidentified line identification results in what are commonly referred to as catastrophic redshift failures, or "outliers" in the real versus measured redshift plot.

The danger of line misidentification is that the Gaussian uncertainty associated with the λ position is directly translated to the redshift uncertainty. However, given the line misidentification problem, the actual redshift uncertainty is often not Gaussian and it is characterized by the presence of multiple peaks. The Gaussian assumption, therefore, can lead to a large underestimation of the final uncertainty associated with any physical quantity derived from line fluxes and redshifts.

In an alternative to the spectroscopic approach, many color-based criteria allow us to estimate redshifts, albeit with a lower precision than when spectra are available. Photometric redshifts have historically been used to build samples of high-z galaxies by exploiting the shape and features of the UV/optical part of the spectra, in particular the presence of the Balmer and Lyman breaks (at $\lambda \sim 4000\,\text{Å}$ and $1216\,\text{Å}$, respectively) as, for example, in the initial works in the Hubble deep field (e.g., Clements & Couch 1996; Madau et al. 1996; Dickinson 1998; Stevens & Lacy 2001). More refined techniques allow us to extrapolate information from photometric measurements by fitting the spectral energy distributions (SEDs) with theoretical and/or empirical models (for a recent review of this topic, see, e.g., Salvato et al. 2018). Photometric analyses may also result in outliers, as a consequence, e.g., of the misidentification of the Balmer/Lyman break.

Various strategies can be adopted to reduce the number of catastrophic failures in redshift estimates, e.g., by combining spectroscopic and photometric analysis. First of all, instead of a simplistic Gaussian approximation for each source's redshift, one can use the information in the full redshift probability distribution function (PDF). PDFs are typically created by most

common software that perform fitting to SEDs, such as, e.g., Hyperz (Bolzonella et al. 2000). Besides allowing for a more formally correct treatment of the uncertainty and its propagation, the redshift PDF[12] provides a way to compute the probability of a redshift being an outlier. Using the full redshift PDF allows for the selection of samples with different degrees of purity. In any case, the correct computation of PDFs from SED-fitting strategies crucially depends on the availability of template galaxy models able to fit the photometric data.

The outlier problem can also be mitigated using photometric priors. These priors may include galaxy colors (as is the case in photometric redshift estimates via SED fitting) or observed galaxy fluxes. To first approximation, brighter sources are more likely located at lower redshift, so this information can help to disentangle, for example, bright local objects with high metallicity (i.e., steeper optical continuum) from high-redshift sources. In this context, some publicly available SED-fitting software provide the user with the option of applying optical priors (e.g., EAZY; Brammer et al. 2008). Additional empirical techniques can be exploited to improve the precision of photometric redshift estimates and to correct outliers in specific cases. For example, Baronchelli et al. (2018) demonstrated how additional optical priors, in combination with far-IR detections, can help improve the accuracy.

The approaches described above can be used to identify single emission lines in galaxy spectra. However, while the photometric redshift estimate can be refined for every source using these techniques, the wavelength position of a spectral line remains the primary source of information in the case of a spectroscopic determination. In other words, these techniques can help in identifying an emission line, but they do not affect the spectroscopic redshift of a source when the emission lines are already unequivocally identified.

In this paper, we address the problem of correctly identifying single emission lines detected in grism spectra of the Hubble Space Telescope-WFC3 Infra-red Spectroscopic Parallel survey (HST-WISP survey; Atek et al. 2010; I. Baronchelli et al. 2020, in preparation), by combining different sources of information.

Another goal of this paper is to provide a testing ground for the definition of similar algorithms to be used in the context of the future ESA's Euclid (Laureijs et al. 2011) and NASA's WFIRST (Green et al. 2011) missions, in order to maximize the scientific return of their near-IR spectroscopic surveys. It is worth noting that the spectroscopic coverage of the grisms and the photometric bands available in the WISP survey are both very similar to those that are planned to be employed by Euclid and WFIRST. These similarities, together with the wide sky area covered by WISP, make this survey one of the most important proxies for future space-based spectroscopic missions. The Euclid and WFIRST surveys will probably benefit from a large amount of ancillary data. In this sense, focusing on the WISP survey, our analyses represent a pilot study of these future missions. In any case, the modular structure of our algorithm is specifically designed to easily include and remove additional modules and sources of information.

The paper is organized as follows: in Section 2, we present the WISP survey and the samples we use to calibrate and test the algorithm. Section 3 describes the algorithm and its modular structure. In Section 4, we compute the precision of the algorithm, also in terms of completeness and contamination of differently selected samples. In Section 5, we report the new classification of WISP sources obtained using our software, while in Section 6, we discuss the implications of our work in the context of future dark energy missions. The main results and future perspectives are finally summarized in Section 7.

2. Data

We tested our algorithm on the second data release of the WISP survey (WFC3 Infra-red Spectroscopic parallel survey; Atek et al. 2010; I. Baronchelli et al. 2020, in preparation).[13]

2.1. The WISP Survey

WISP is a pure-parallel, near-IR slitless grism spectroscopic survey that efficiently collects WFC3 data while other HST instruments are in use (P.I. M. Malkan, Atek et al. 2010). The spectral coverage of the WFC3's grisms, G102 (0.8–1.1 μm, $R \sim 210$) and G141 (1.07–1.7 μm, $R \sim 130$), enables the detection of the Hα line from $z \sim 0.3$ to $z \sim 1.5$ and the [O III] emission lines from $z \sim 0.7$ to $z \sim 2.3$. To aid in extracting the 1D spectra from the dispersed images and to enable the wavelength calibration, the WISP fields were also observed in direct imaging mode with filters chosen to match the grisms' spectral coverage: F110W for G102 and either F140W or F160W for G141. A relevant fraction of the WISP fields ($\sim$1/3) are also observed with the WFC3 UVIS camera with a subset of the available filters (F475X, F600LP, F606W, and F814W). Finally, about half of the fields are also covered by *Spitzer*/IRAC observations in channel 1 and/or 2 (3.6 μm and 4.5 μm, respectively). To date, the WISP survey has observed 483 fields, collectively covering an area of more than 2000 arcmin^2 (the actual available data in the WISP spectroscopic catalog refer to a total area of 1520 arcmin^2).

Being a parallel survey, the observing strategy of WISP depends on the details of the primary observations. This means that the depth of the coverage varies from field to field. Accordingly, we divided visit opportunities into two categories: "short" and "long" visits, corresponding to fewer than four and four or more orbits, respectively. During the short opportunities, only the G141 grism and the F140W (or F160W) filter are used, while observations in the G102 grism and in the F110W filter are added during long opportunities. In the latter case, the relative integration time between the G102 and G141 grisms is chosen to balance the sensitivity reached by the two grisms. Given these premises, the median 5σ depth reached in both grisms is 5×10^{-17} erg s^{-1} cm^{-2}, with a factor of approximately 2 field-to-field variation.

For the line detection, we apply a wavelet convolution and a SExtractor-type threshold through a custom line-finding software. This approach dramatically reduces the number of false-positive detections. For the detection, we require at least three contiguous pixels with signal-to-noise ratio (S/N) $\geqslant 2.3$. This translates into S/N $\geqslant 4$ integrated over the full emission line (i.e., the sum over the pixels involved). However, this value does not directly correspond to the S/N reported in the final catalog. In that case, the flux and flux uncertainty are measured after the lines and the full spectrum are fit (meaning that the

---

[12] Multiple peaks are commonly observed in a typical PDF. These peaks are due to the degeneracy existing among different SED models when fitted to the available photometric data.

[13] https://archive.stsci.edu/prepds/wisp/

signal becomes the area under the Gaussian curve rather than the sum of the individual pixel values).

Once detected, each emission line is identified through a visual inspection streamlined with a Python-based interactive line-fitting and measuring code. Every object is analyzed by at least two separate reviewers, and a total of 10 reviewers were involved in this process. During this phase, false identifications due to contamination by overlapping spectra or to specific noise features are excluded from the sample. All identified real emission lines are then fitted with a Gaussian profile + continuum, using a line-fitting algorithm that required little input from the reviewer. When an emission line is detected in a spectrum, additional emission lines, initially undetected, can also be fit even if their flux was below the original pixel-based detection threshold.

The quantities measured for each line (total flux, equivalent width (EW), FWHM) are then merged into a catalog that has a unique entry for each galaxy with all the spectroscopic information. Differences among reviewers' classifications and line fits are used to obtain average solutions, when possible, and to define the quality flag associated with each spectrum. In particular, when no agreement is found between the identifications of two reviewers, this information is included in the quality flag. In this case, in the final catalog, only the solution associated with the line fit that minimizes the χ^2 parameter is included. More details on the creation of the WISP emission-line catalog are presented in M. Bagley et al. (2020, in preparation).

2.2. Calibration and Test Samples

The algorithm is calibrated (Section 3.2) and tested (Section 4) on two not-independent (almost completely overlapping) subsamples extracted from the WISP spectroscopic catalog of secure identifications. The choice of using two overlapping samples is justified in Section 4.1. For these samples, the standard method used to identify the spectral lines can be applied, i.e., two or more spectral lines are detected above a 2σ threshold. Additionally, we require that the reviewers agree on the identification of the lines. We only include sources detected in fields covered by both WFC3 grisms. This selection reduces the WISP area usable for calibration and testing to ~900 arcmin². Finally, we exclude all of the emission lines located at the low-sensitivity ends of the grisms' wavelength range ($\lambda < 8500$ Å, $\lambda > 16700$ Å, and 11000 Å $< \lambda < 11400$ Å).[14] We call these two subsamples the calibration (2128 sources) and the test (2283 sources) "gold" sample. The only difference between the two samples is that in the calibration sample we consider only sources detected in the F110W photometric band, while the same requirement is relaxed when testing the algorithm (measuring accuracy, completeness, and contamination). This selection is due to the fact that the F110W magnitude is required to calibrate one of the modules of the algorithm (the magnitude prior described in Section 3.2.2), but the same measurement is not necessary when running the software on unidentified spectral lines. Consequently, the two samples almost fully overlap, with the entire calibration sample included in the test sample (only ~7% of the sources used for the test are not used for the calibration).

---

[14] The high level of noise makes it difficult to clearly identify the strongest line in this spectral region.

After calibrating and testing the algorithm, we run the software on a WISP subsample of sources covered by both grisms but with only one emission line detected above a 2σ threshold (Section 5). We highlight the fact that the algorithm is designed to identify the brightest line detected in a spectrum while it is blind to the possible presence of additional lines besides the strongest one. Consequently, their presence in a spectrum does not influence the calibration and test of the algorithm itself.

3. The Algorithm

3.1. Rationale

The purpose of the algorithm is to provide a probabilistic identification of the strongest emission line observed in a spectrum, by combining various sources of information.

In principle, when only one line is detected in a spectrum, an effective method to identify the line is to compare the spectroscopic redshift expected from different species/ions (e.g., $z_{H\alpha}$, $z_{[O\,III]}$, $z_{[O\,II]}$, etc.), with the independent redshift solution suggested by an SED fit of the available photometric data. The reliability of this approach, however, strongly depends on the number of available photometric measurements and their wavelength coverage. The precision of such a method rapidly declines when only a few photometric measurements are available. In this case, the photometric redshift PDF will not clearly distinguish between, e.g., the Hα and the [O III] redshift solutions. The presence of multiple peaks in the PDF can further complicate the issue, generating photometric redshift outliers.

Figure 1 (top panel) shows the schematic representation of a WISP grism spectrum (G102 and G141) with only one line detected. The same observation can be equally well described as being due to Hα, [O III], or [O II] (cases A, B, and C in the figure), if the emission due to the other additional lines is below the S/N detection threshold. In the same figure, the bottom panel illustrates the effects of a poorly constraining multipeak PDF on the redshift determination.

For a Gaussian PDF, the width of the curve (its value of σ) univocally indicates the uncertainty associated with the value (on the x-axes) corresponding to the peak of the PDF itself. In the more general case, when the shape of the PDF is not Gaussian (e.g., bottom panel of Figure 1), it is difficult to unequivocally describe the width of the PDF with one number. In this case, the redshift range $\Delta z(p)$ corresponding to a given total probability (e.g., $p = 95\%$) is a more correct way to represent the uncertainty associated with the most likely value. Alternatively, one could also consider the integral of PDF*, defined as

$$\int \text{PDF*}(z)\ dz = \int \frac{\text{PDF}(z)}{\max(\text{PDF}(z))}\ dz. \qquad (1)$$

Lower PDF* integrals or smaller $\Delta z(p)$, however, are not always associated with more precise solutions, such as when a PDF shows multiple peaks. This is true even if each of the peaks can singularly be described by a Gaussian function with a small value of σ. This case represents the typical outlier problem. While each Gaussian peak is characterized by a small σ (high precision), the real uncertainty could be badly underestimated (low accuracy).

On the other hand, some correlations such as the magnitude (or size) versus redshift relation are not particularly tight. Using
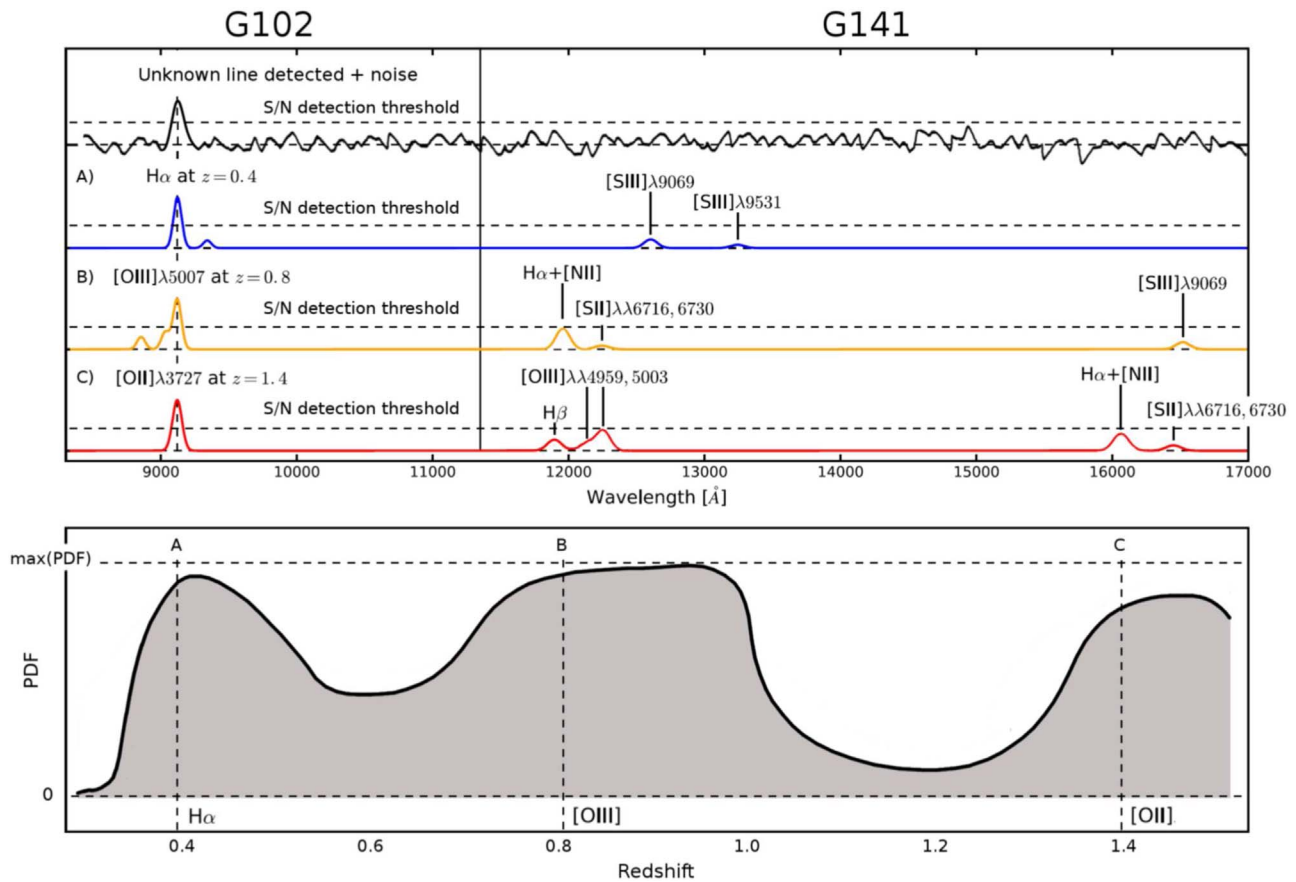
Figure 1. Top panel: schematic representation of a WISP grism spectrum with only one line detected above the S/N threshold (uppermost plot). The lower three plots of the top panel represent three different possible solutions: Hα (A), [O III] (B), and [O II] (C). Bottom panel: a poorly constrained PDF with multiple probability peaks. This kind of PDF would not be able to discriminate among the three possibilities shown in the top panel (A, B, C). This figure is a modified version of Figure 3 of Bagley et al. (2017).

these quantities as priors to determine the redshift itself thus generates wide PDFs (low precision). Because of this characteristic, however, these methods are less prone to the problem of outliers. In other words, while the uncertainty is high, the measure of the uncertainty is more accurate.

Thus, the most effective way to preserve the precision while maximizing the accuracy (limiting the number of outliers) is to exploit the broader PDFs to identify the most probable peaks in the narrower PDFs. This can be obtained by simply multiplying all the PDFs to each other. Besides allowing the PDFs to be computed, the same parameters can provide estimates of the expected flux ratios. These ratios can be used by the algorithm to obtain an independent prediction of the most prominent line observed in a spectrum.

In this section, we analyze in detail all the different methods that we eventually combine in our algorithm. Every method is treated as an independent module of the algorithm. In particular, each module represents one single method to obtain information on one only type from one or more input parameters. This approach allows us to more clearly describe the different types of information that the same input parameter can provide, the relations existing between input parameters and outputs, and the limits of each method.

The output of the algorithm we describe below is, for the strongest emission line in each spectrum, a set of indices $P_{\text{transition}}$. These indices are directly related to the probability

that the observed emission line is correctly identified with each of the transitions considered (Hα, Hβ, Hγ, [O III], [O II], [S II]).[15] The transition with the highest value of $P_{\text{transition}}$ will provide the (probabilistically) best redshift solution. We note that, during this process, the algorithm always assumes that the observed line is a real spectral feature and not a spurious detection due to noise or contamination.

3.2. Structure of the Algorithm

The modular structure of the algorithm is organized as shown in Figure 2. In brief, the modules can be divided in three main categories, or blocks, depending on the kind of information provided. Using PDFs, the algorithm can estimate the probability associated with each value of z (regression) in the redshift range considered ($0 < z < 3.3$). The flux ratios allow us to forecast which species/transition is most probably responsible for the strongest emission (classification). Finally, comparing the results with the input test sample, the algorithm can automatically fine-tune the final function, in order to

---

[15] These indices do not correspond to the actual probabilities for the following reason: first, the algorithm considers only the listed species, without taking into account different possibilities. Moreover, the algorithm does not compute the probability that a detected line is not a real emission, but the result of noise or contamination. For these reasons, the indices are normalized so that the highest probability index $P_{\text{transition}}$ is always equal to 1.
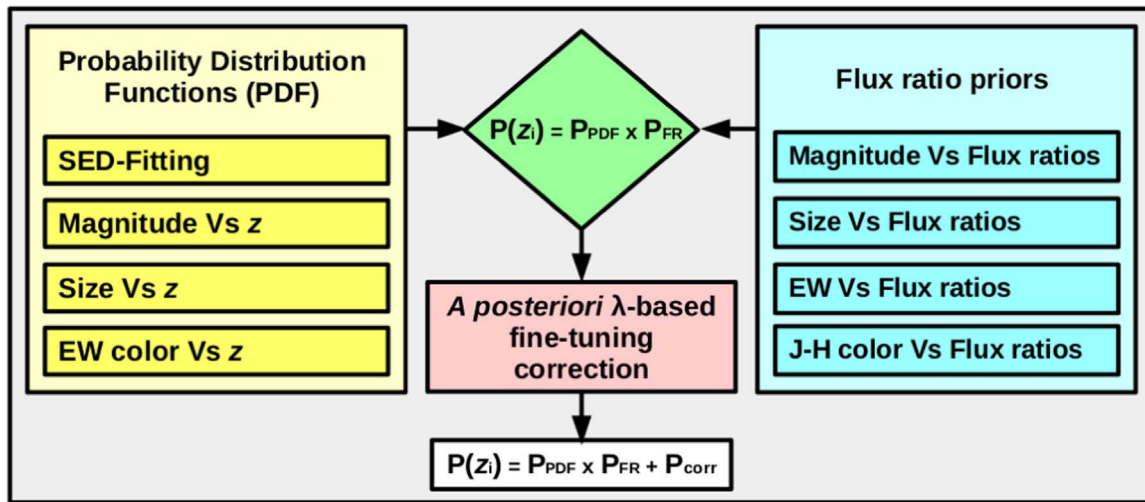
Figure 2. Organization of the modules in the algorithm. The modules can be divided into three categories, or blocks, depending on the kind of information supplied. For each spectrum (for each source), the probability distribution functions PDF(z) provide probability estimation (regression), between $z = 0$ and $z = 3.3$. The line flux ratios can be used to predict which, among the considered species/transitions, is more likely responsible for the strongest line observed (classification). Finally, the probability ratios can be fine-tuned using a λ-dependent a posteriori correction (optimization).

increase the accuracy of the entire process (optimization). In a scheme:

1. REGRESSION BLOCK: photo-z PDFs estimated from SED fitting (Section 3.2.1), J-band apparent magnitude (Section 3.2.2), apparent size (Section 3.2.3), and EW of the strongest line (Section 3.2.4);
2. CLASSIFICATION BLOCK: a set of probability ratios estimated using existing relations between flux ratios and J-band apparent magnitude, size, line EW, and $J - H$ color index (Section 3.2.5);
3. OPTIMIZATION BLOCK: a posteriori fine-tuning of the probability ratios based on the observed wavelength (Section 3.2.6).

In this section, we describe the single modules and how we calibrated them, while in Section 3.3, we detail how the modules are combined with each other to output the final probability estimates.

3.2.1. Photo-z PDF from SED Fitting

We compute photometric redshifts using the Hyperz software (Bolzonella et al. 2000). For each WISP field and for each source, we considered all measurements in the available photometric bands among the WFC3/IR F110W, F140W, F160W, WFC3/UVIS F606W, F600LP, F814W, and IRAC 3.6 and 4.5 μm filters. Figure 3 shows the number of sources covered by every combination of photometric bands. Given the limited number of bands used in the photometric redshift calculation (larger than 2 only for approximately 55% of the sample), the resulting photometric redshift estimate will not be precise. However, we are not interested in the absolute value of the photometric redshift, typically assumed to correspond to the highest peak of the PDF, but rather in the full shape of the PDF itself. This PDF will be combined with additional probability functions as described in Section 3.3. Therefore, even a photo-z estimation derived from two bands only (corresponding to one single color index) adds information on the redshift of a source. For example, the poorly constrained PDF shown in the bottom panel of Figure 1 does not allow for a reliable estimate of the photometric redshift. However, it does allow us to safely say
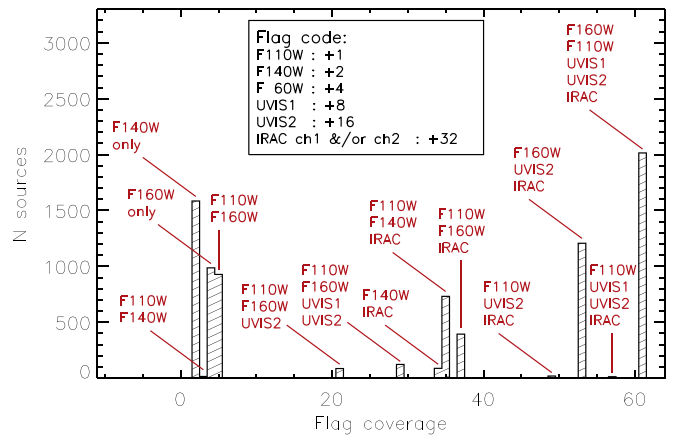
Figure 3. Number of sources vs. coverage in the available photometric bands, for the original WISP spectroscopic catalog (at least one spectral line measured). No selection is applied. The coverage flag (x-axes) is given by the sum of +1 when the source is observed through the F110W filter, +2 for F140W, +4 for F160W, +8 for UVIS1, +16 for UVIS2, and +32 for IRAC channel 1 and/or channel 2 coverage. The naming convention "UVIS1" and "UVIS2" is used to represent the bluer and the redder UVIS filters used in a single observation, while a combination of either F475X or F606W and F600LP or F814W was actually used in the WISP survey.

that the source considered is not located below $z \sim 0.3$ or in the range $1.1 \lesssim z \lesssim 1.3$. This kind of information can be particularly relevant if combined with additional information from other independent sources.[16]

For the Hyperz run, we considered a combination of template models from Bruzual & Charlot (2003), characterized by an exponentially declining star formation rate (SFR) with SFR $\propto \exp(-t/\tau)$, where $\tau = 0.3$, 1, 2, 3, 5, 10, 15, and 30 Gyr. We consider solar metallicity $Z = Z_\odot$ and an extinction law in the Calzetti et al. (2000) form, with A_V ranging from 0.0 to 3.0.

We computed photometric redshifts for all the sources in the sample, regardless of the number of photometric bands available. The typical output PDF rarely shows well-defined

---

[16] For example, the upper panel of Figure 15 shows three poorly constrained PDFs. Their combination is shown in the bottom panel of the same figure.
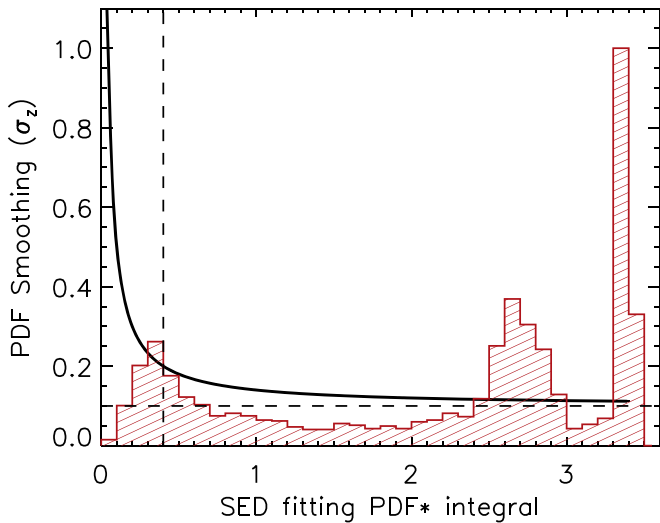
Figure 4. The black curve represents the width of the Gaussian filter used to smooth the photo-z PDF (σ_z in Equation (2)) as a function of the PDF* integral $\int_{0.0}^{3.3}$ PDF*(z) dz. The red histogram shows the distribution of the PDF* integral for the WISP galaxies. Generally speaking, high values of the PDF* integral indicate poorly constrained photometric redshifts (a value of 3.4 corresponds to a PDF* equal to 1.0 everywhere between $z = 0$ and $z = 3.3$). However, given the possible presence of multiple probability peaks (outliers) and the intrinsic discretization of the grid of models used for the SED fit, the precision of the algorithm is not guaranteed by narrower PDFs. The horizontal dashed line represents the best average smoothing factor able to mitigate the effects of the use of a discretized grid of models without limiting the precision of the final PDFs. In Equation (2), this value corresponds to an asymptotic limit for σ_z ($\sigma_z \rightarrow 0.1$ for high values of the PDF* integral). The vertical dashed line represents the best empirically derived value of P_z^*, for which the smoothing factor σ_z is doubled with respect to its asymptotic value.

and unique probability peaks. More commonly, multiple peaks are present. In some cases, their presence is due to the degeneracy of the input parameters, generating solutions at different redshifts with similar probabilities.[17] In other cases, they are a consequence of the discretized grid of models used for the fit, especially when the maxima are close to each other (in redshift). To correct for the effect of model discretization, we smooth the PDFs using a Gaussian filter. Although a good compromise between precision and mitigation of the discretization effect can be obtained with a constant value of σ_z ($\sim$0.1), we decided to vary σ_z as a function of the integral of the PDF* (see Equation (1)), for the reasons explained next.

As we previously discussed, narrow photometric redshift PDFs may still be centered on the wrong redshifts, with resulting uncertainties that are underestimated (resulting in outliers). In general, the overconfidence in the estimation of the z-PDF is a well-known problem, especially for what concerns their low-probability tails (see, for example, Wittman et al. 2016). Because of the small (underestimated) uncertainties, the (possibly wrong) photometric redshift solution would prevail over any other redshift indicator. Thus, in these cases, the emission-line identification would also be compromised. To limit the effects of this problem, we smooth the z-PDFs, a similar solution to that adopted by, e.g., Rodríguez-Muñoz et al. (2019). Differently from that work, where a constant $\sigma = 0.2$ is considered, in our Gaussian smoothing, σ is proportional to the inverse of the PDF* integrals and

---

[17] Often, few photometric optical data can equivalently be well fitted by both the spectrum of an obscured low-redshift source or by the spectrum of a high-redshift galaxy with low/normal extinction.
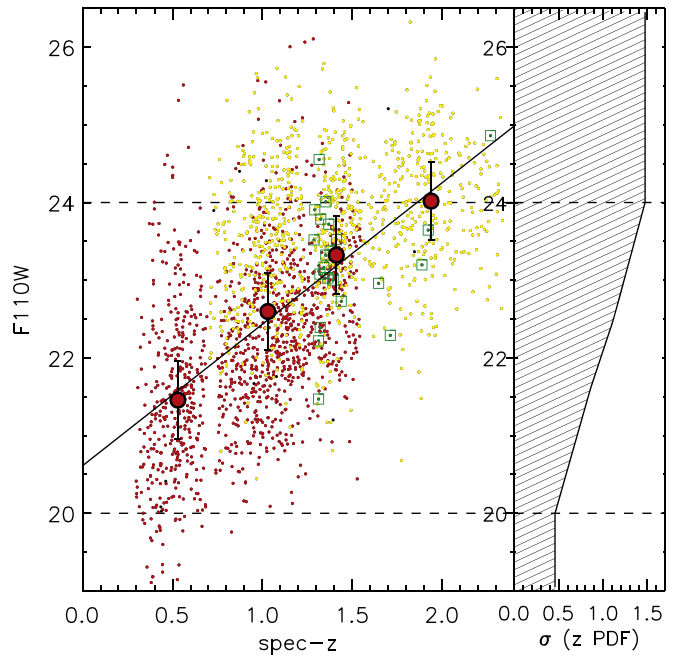
Figure 5. Left panel: relation between the AB magnitude measured in the F110W band (J) and the spectroscopic redshift for the calibration "gold" sample. All of the data are considered for the linear fit, which we use as a prior in the algorithm. However, the prior does not consider values of magnitude $J < J_{min}$ or $J > J_{max}$ (dashed lines): in these cases, the prior assumes $J = J_{min} = 20$ and $J = J_{max} = 24$, respectively. This method limits the contamination of the numerically small sample of sources dominated by [O II] emission from [O III]-dominated sources. The actual nature of the strongest emission line is represented by yellow circles for [O III], red circles for Hα, and green circles surrounded by squares for [O II]. Right panel: given the J magnitude of a source, the PDF is assumed to be a Gaussian function, with σ equivalent to the horizontal dispersion of the data around that specific value of magnitude ($\pm$0.5 mag).

asymptotically converging to $\sigma_z = 0.1$:

$$\sigma_z = 0.1 \times \left(1 + \frac{P_z^*}{\int_{0.0}^{3.3} \text{PDF*}(z) \, dz} \right). \qquad (2)$$

Both the best average smoothing factor ($\sigma_z = 0.1$) and the constant value $P_z^* = 0.4$ are empirically determined by maximizing the accuracy of the algorithm when run on the test "gold" sample. Figure 4 shows the value of the variable smoothing factor σ_z as a function of the original photo-z PDF* integral.

The SED-fitting method relies on the combination of color indices in different photometric bands. As a consequence, if two different sources are characterized by very different apparent magnitudes or sizes, but identical color indices for all of the bands considered, there will be no difference between their output redshifts. Thus, to improve the method, we include the magnitude and size priors described in Sections 3.2.2 and 3.2.3.

3.2.2. Photo-z PDF from F110W (J-band) Magnitude

Galaxies cannot be arbitrarily bright (Schechter 1976). The consequence is that, when surveying a fixed solid angle Ω, galaxies that look very bright are preferentially located at low redshifts, while faint sources more likely correspond to high-redshift galaxies.
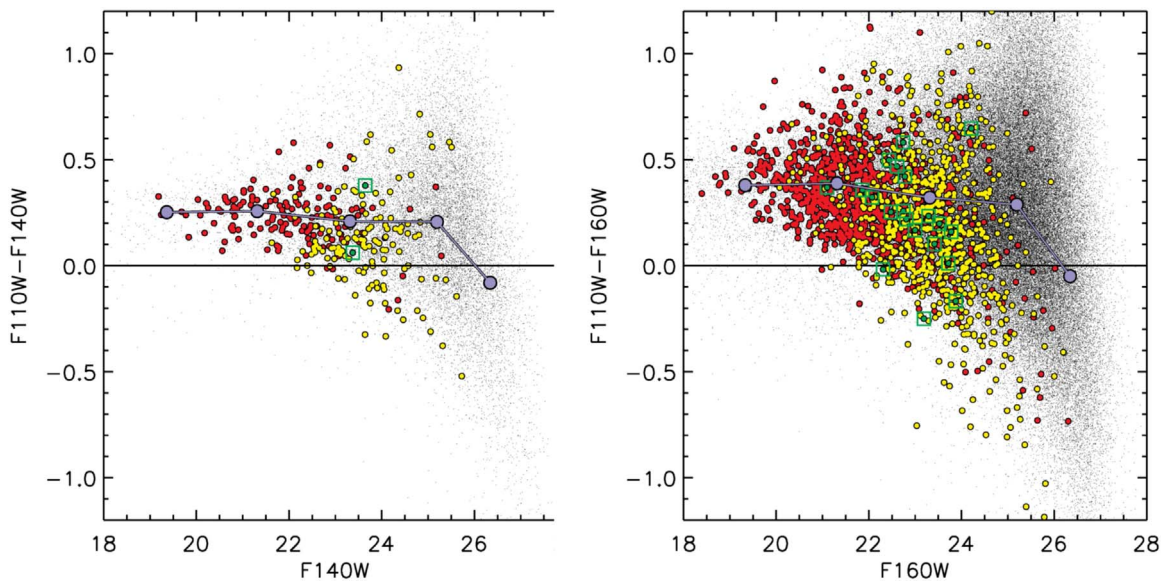
Figure 6. Left panel: $J - H$ (F110W – F140W) color index as a function of the F140W magnitude. We use this color index to infer the expected F110W magnitude from F140W (see Section 3.2.2). The correction is weakly dependent on magnitude and does not depend on the identification of the strongest emission. Red circles show Hα, yellow circles [O III], green circles surrounded by a square [O II], and black dots for no lines detected. The average value of the color index is represented using light purple lines and filled circles, in five bins of magnitude. Right panel: $J - H$ color index computed for the F110W – F160W combination, as a function of F160W.

Figure 5 shows the observed F110W magnitude (hereafter J) as a function of the spectroscopic redshift for WISP sources in the calibration "gold" sample. The algorithm uses the linear fit to these data (solid line in the figure) as the photo-z prior. Specifically, given the J magnitude of a galaxy, the algorithm assumes a Gaussian-shape PDF(z), centered at the redshift indicated by the linear fit, and with a σ equivalent to the horizontal dispersion of the data in the plot of Figure 5 (right panel).

While we computed the linear fit using all the data, when applying the prior we limit the range of validity to $J_{min} < J < J_{max}$, with $J_{min} = 20$ and $J_{max} = 24$. Sources with $J < J_{min}$ or $J > J_{max}$ are set to $J = J_{min}$ and $J = J_{max}$, respectively. This empirical approach allows us to limit the contamination of the numerically small sample of sources dominated by the [O II] emission from sources leaking from the tails of the numerically more consistent [O III] and Hα distributions.

As shown in Figure 3, many sources are not covered by observations in the F110W band. However, to first approximation, we can convert between the H- (F140W or F160W) and the J-band magnitudes, as shown in Figure 6. In this figure, we plot the magnitude-dependent correction that can be applied to the measurements obtained using the F140W and F160W filters to recover the magnitude expected in the F110W band. The same figure also shows that the magnitude correction does not depend on which emission line is the strongest in the spectrum.

3.2.3. Photo-z PDF from Apparent Size

The mass–size relation (e.g., Poggianti et al. 2013; van der Wel et al. 2014) and the shape of the stellar mass function (see, e.g., Lapi et al. 2017 and Figure 4 therein) indicate that the same arguments used for the apparent magnitude can be applied to the apparent size as well: apparently larger sources

are more likely located at lower rather than higher redshifts. Following the Mattig equation (Mattig 1958), the apparent size of a galaxy decreases with distance, up to $z \sim 1.5$.

In Figure 7, we show the galaxies' apparent size (A_IMAGE[18]) computed in the J band (F110W) as a function of the spectroscopic redshift. We include only sources in the WISP calibration "gold" sample. The linear fit to the size–magnitude relation is used as an additional redshift prior in our algorithm. In particular, given the apparent size of a source, we assume a Gaussian PDF in redshift, centered at the redshift indicated by the linear fit, and with a σ equivalent to the local horizontal dispersion of the data in the plot of Figure 7 (right panel).

Also in this case, we computed the linear fit using all of the data, but when applying the prior we limit the range to A_IMAGE$_{min}$ < A_IMAGE < A_IMAGE$_{max}$, with A_IMAGE$_{min}$ = 2.5 pixels and A_IMAGE$_{max}$ = 10 pixels. Sources with sizes larger or smaller than these limits are set to A_IMAGE = A_IMAGE$_{max}$ and A_IMAGE = A_IMAGE$_{min}$, respectively. When no images are available in the J band, the A_IMAGE parameter is computed using the H-band image. Given the negligible differences in the J and H PSFs and sampled stellar populations, no corrections are considered in these cases.

3.2.4. Photo-z PDF from Apparent Equivalent Width

Figure 8 shows the relation existing between the apparent EW of the measured lines and the spectroscopic redshift of the sources. As we do for the apparent J magnitude and size priors (Sections 3.2.2 and 3.2.3 respectively), we compute a linear fit to the observed relation (solid line in the figure). The best-fitting relation is used as a redshift prior by our algorithm. In particular, given the EW of an observed emission line, we

---

[18] The A_IMAGE parameter is an output of the SExtractor software corresponding to the rms of the luminosity distribution of a galaxy, measured in pixels, along the semimajor axes.
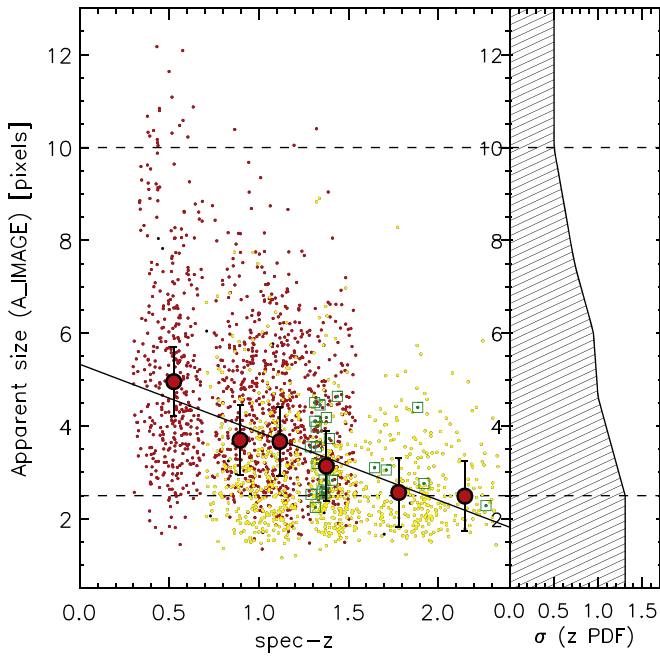
Figure 7. Left panel: relation between apparent size (SExtractor A_IMAGE parameter) measured in the F110W band (J) and spectroscopic redshift, for the calibration "gold" sample. All of the data are used to compute the linear fit, which we use as a prior in the algorithm. However, the prior does not consider values of A_IMAGE $A < A_{min} = 2.5$ pixels or $A > A_{max} = 10$ pixels (dashed lines): in these cases, the prior assumes $A = A_{min}$ and $A = A_{max}$, respectively. The actual nature of the strongest emission line is represented by yellow circles for [O III], red circles for Hα, and green circles surrounded by squares for [O II]. Right panel: given the value of A_IMAGE of a source, the PDF is assumed to be a Gaussian function, with σ equivalent to the horizontal dispersion of the data around that specific value of A_IMAGE (± 0.75).

Figure 8. Left panel: relation between the apparent EW and spectroscopic redshift, for the calibration "gold" sample. All of the data are used to compute the linear fit, which we use as a prior in the algorithm. However, the prior does not consider values of EW $\log(EW) < \log(EW_{min}) = 1.75$ or $\log(EW) > \log(EW_{max}) = 2.75$ (dashed lines): in these cases, the prior assumes $\log(EW) = \log(EW_{min})$ and $\log(EW) = \log(EW_{max})$, respectively. The nature of the strongest emission line is represented by yellow circles for [O III], red circles for Hα, and green circles surrounded by squares for [O II]. Right panel: given the value of EW, the PDF is assumed to be a Gaussian function, with σ equivalent to the horizontal dispersion of the data around that specific value of EW (± 0.125).
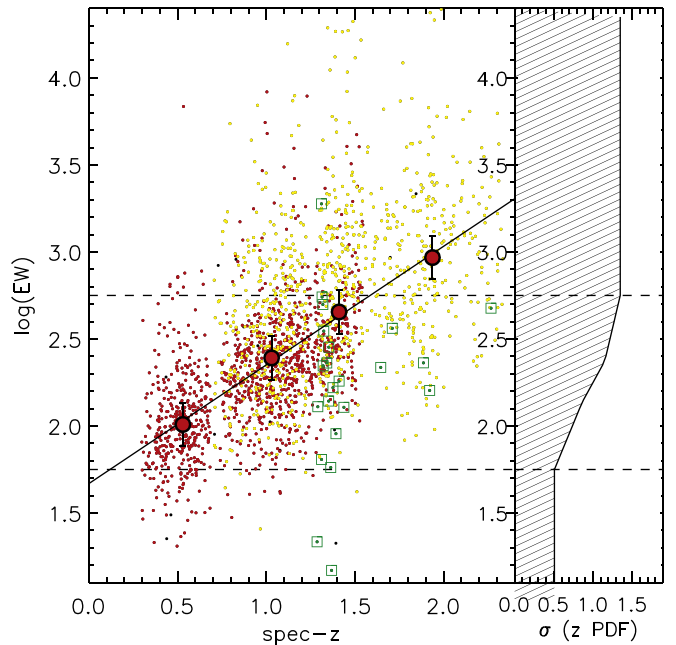
consider a Gaussian PDF centered at the redshift indicated by the linear fit and with a value of σ given by the local horizontal dispersion (right panel of Figure 8).

Also in this case, we limit the range of validity to $\log(EW_{min}) < \log(EW) < \log(EW_{max})$, where $\log(EW_{min}) = 1.75$ and $\log(EW_{max}) = 2.75$. It is possible to observe (Figure 8) that these thresholds allows us to approximately separate sources dominated by the Hα and [O II] emission from those dominated by [O III], helping to limit the contamination of the [O II] sample from the other more numerous samples.

3.2.5. Line Flux Ratio Priors

Unless the z-PDF is particularly narrow, it is difficult to automatically classify an emission line using just the PDF itself. For example, while $\lambda_{H\alpha} \sim \lambda_{[S II]}$, it is very unlikely that [S II] can be the unique single line observed in a spectrum, because Hα is commonly brighter than [S II]. In general, detecting the Hα emission line in a randomly selected spectrum is more common than detecting [O III], and [O III] is more common than [O II] or other lines such as the fainter Hβ, Hγ, or [S II]. Then, in order to obtain a proper classification, the expected flux ratios between different lines can be used as an additional source of information.

The average line flux ratios measured in the calibration "gold" sample allows us to rescale the detection probabilities of the different species/transitions considered. For example, let us assume the case of a spectrum with only one line detected and a similar value of the photo-z PDF measured at $z_{H\alpha}$ and $z_{[S II]}$,

with z-PDF ~ 0 for all the other observable emission lines. Because the Hα emission is always stronger than that due to [S II], the line detected must be classified as Hα. Because the observed wavelength is known (λ_i^{obs}), identifying the nature of the (strongest) emission line measured ($i = H\alpha$, in this example) automatically provides a measure of the spectroscopic redshift: $z = z_{H\alpha} = (\lambda_{H\alpha}^{obs}/\lambda_{H\alpha}) - 1$.

For each of the species/transitions i considered, we computed the average observed flux ratio $F_i/F_{H\alpha}$ as a function of J magnitude, apparent size, EW, and $J - H$ color. This flux ratio is calculated for sources in the calibration "gold" sample and in the wavelength range of observability of the emission lines.[19] As shown in Figures 9 and 10, the line flux ratios show some correlation with all of the parameters considered. Thus, if one or more of the parameters considered are available, it is possible to compute the expected relative strength of the different lines. This information can be directly translated into the relative probability ($P_{FR}(i)$) that a given species/transition, i, corresponds to the strongest line measured in the spectrum. In particular, $P_{FR}(i)$ corresponds to the weighted average of the various $P_{FR}^j(i)$ obtained from each of the different indicators j

---

[19] We do not exploit the ratio between the [O III] and [O II] fluxes, computed in the calibration "gold" sample, above $z \sim 1.50$. This is due to a suspected bias in the original classification of strong [O II] emitters above this redshift. When Hα is redder than the upper λ limit of the G141 grism, and if the [O III] and Hβ lines are weaker than the [O II] emissions, the latter tends to be wrongly misidentified as Hα, especially when the S/N is low.
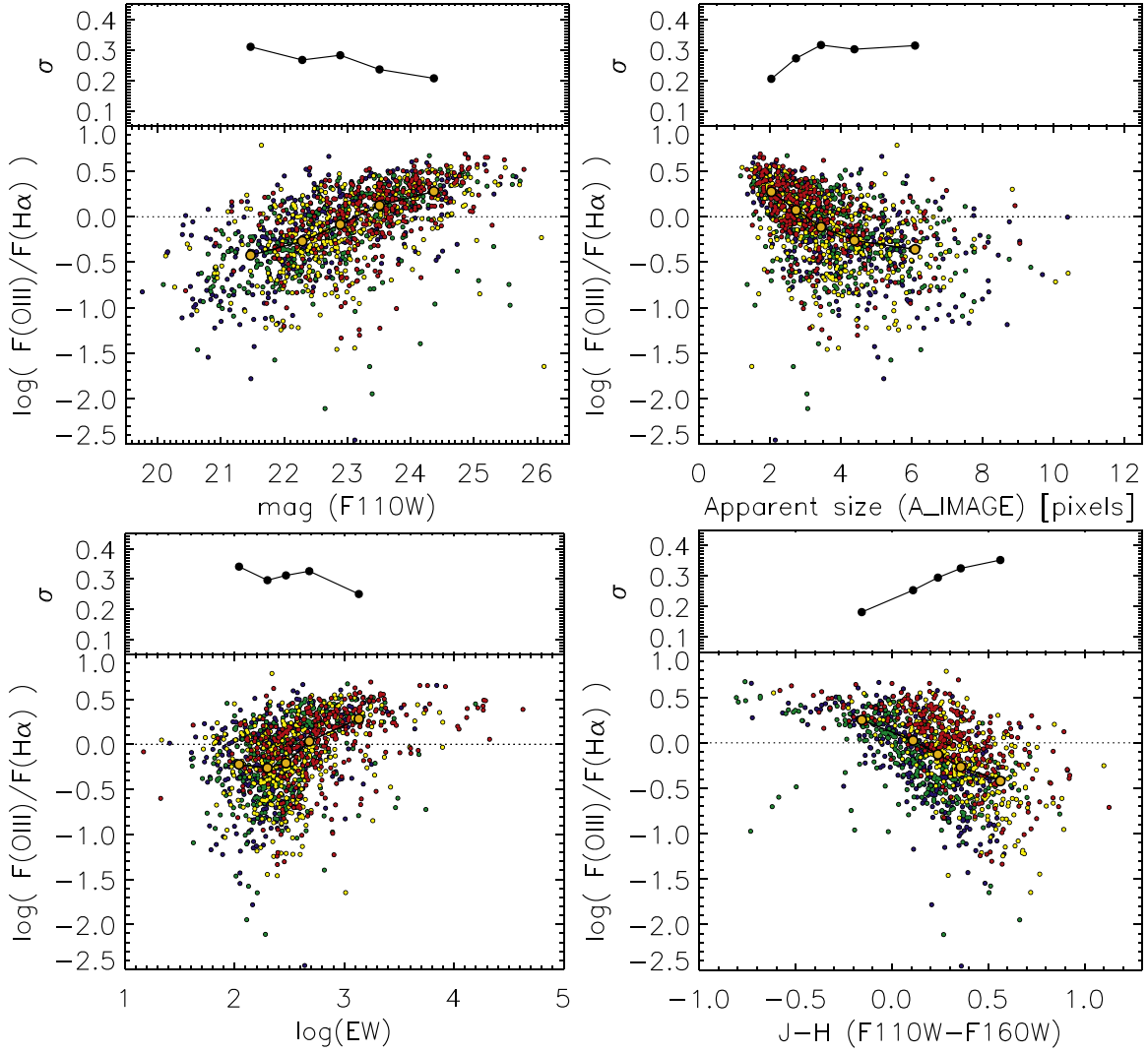
Figure 9. Ratio between Hα and [O III] fluxes, measured as a function of apparent F110W magnitude (top left panel), size (top right), EW (bottom left), and $J - H$ (F110W – F160W) color index (bottom right), for the sources in the calibration "gold" sample. A similar relation is also computed using the F110W – F140W color index, but it is not shown in these plots. For each relation, the dispersion (σ) is shown in the upper parts of each panel. The possible dependence of these relations on redshift cannot be directly taken into account, as the redshift is not known a priori. However, using the observed wavelength prior (Section 3.2.6), it is possible to indirectly correct for such a possible effect. The relations are shown for four different bins of redshifts, from $z \sim 0.87$ to $z \sim 1.42$. In order of increasing redshift, blue, $\langle z \rangle \sim 0.87$; green, $\langle z \rangle \sim 1.05$; yellow, $\langle z \rangle \sim 1.22$; and red, $\langle z \rangle \sim 1.43$.

considered (J magnitude, apparent size, EW, and $J - H$ color):

$$P_{\text{FR}}(i) = [\, P_{\text{FR}}^{J}(i) W_{\text{FR}}^{J}(i) + P_{\text{FR}}^{\text{size}}(i) W_{\text{FR}}^{\text{size}}(i)$$
$$+ P_{\text{FR}}^{\text{EW}}(i) W_{\text{FR}}^{\text{EW}}(i) + P_{\text{FR}}^{J-H}(i) W_{\text{FR}}^{J-H}(i)] / \sum_{j} W_{\text{FR}}^{j}(i), \quad (3)$$

The weights $W_{\text{FR}}^{j}(i)$ are the square inverse of the dispersion of the relations, computed as a function of the reference parameters (see upper panels of Figures 9 and 10):

$$W_{\text{FR}}^{j}(i) = \sigma_{j}^{-2}. \quad (4)$$

Given the observed wavelength of the strongest emission line λ_{obs}, the spectroscopic redshift solution depends on the unknown nature of the species/transition i responsible for such an emission as $z_i = (\lambda_{\text{obs}}/\lambda_i) - 1$, where λ_i is the rest-frame wavelength of the species/transition i. Hence, the algorithm considers both the probability $P_{\text{PDF}}(z_i)$, associated with z_i being the correct redshift solution, and $P_{\text{FR}}(i)$, corresponding to the species/transition i being the responsible for the strongest

emission:

$$P(z_i) = P_{\text{FR}}(i) \times P_{\text{PDF}}(z_i). \quad (5)$$

The previous equation does not provide the complete description of the algorithm, for which we refer to Section 3.3. In fact, the complete computation includes the a posteriori optimization (Section 3.2.6) and the limitations described in Section 3.2.7.

3.2.6. Observed Wavelength Prior

The calibration of all the methods described is performed by using the high-quality data of the calibration "gold" sample. However, because these data correspond to real observations, the phase space used to classify the emission lines cannot be homogeneously calibrated. Additionally, we use a large but simplified (quasi-linear) set of functions describing the correlations between the observational parameters and the outputs. While this approach prevents us from overfitting the calibration data set (because the output does not depend on a
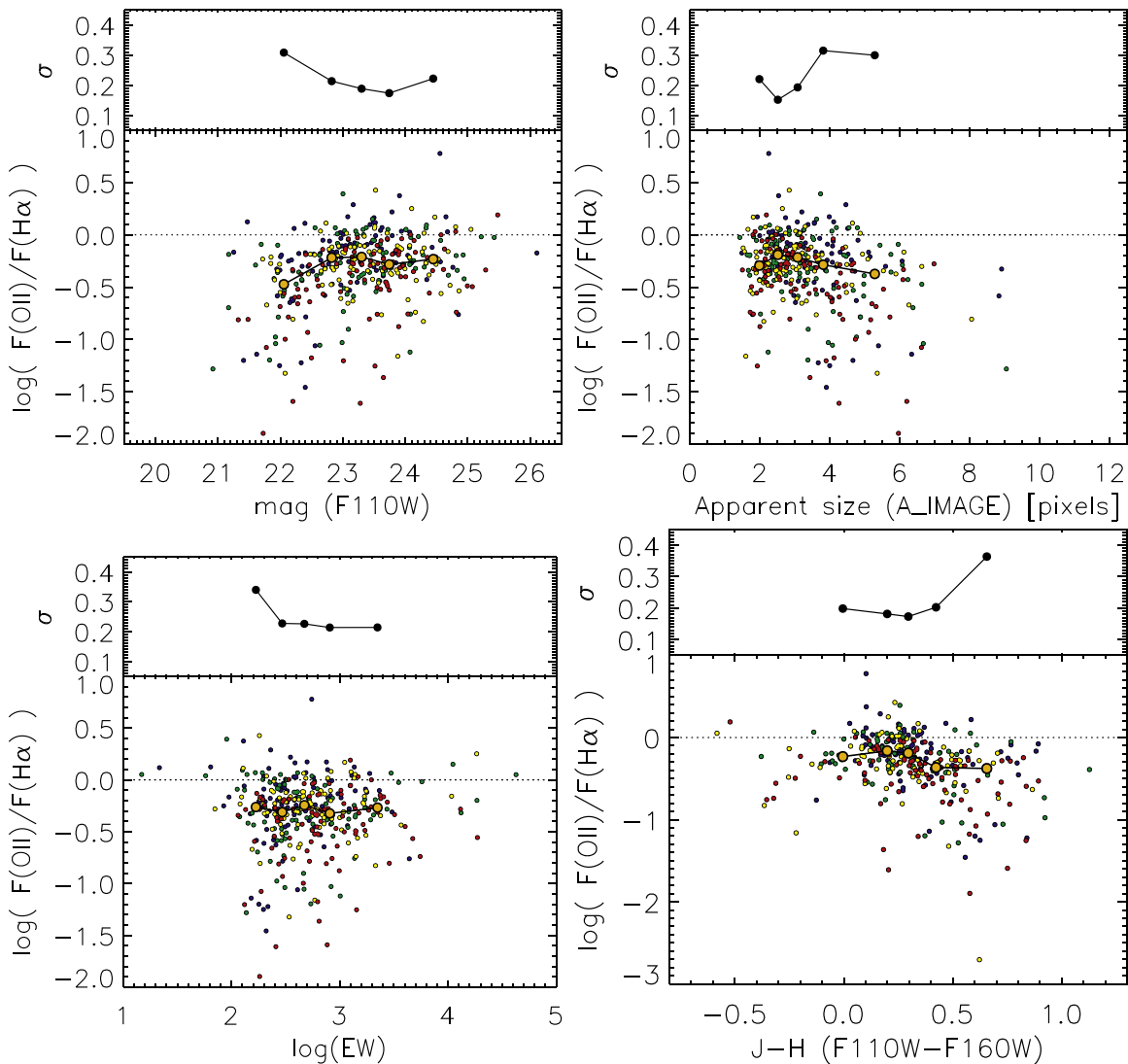
Figure 10. Ratio between Hα and [O II] fluxes, measured as a function of apparent F110W magnitude (top left panel), size (top right), EW (bottom left) and $J − H$ (F110W − F160W) color index (bottom right), for the sources in the calibration "gold" sample. A similar relation is computed also using the F110W − F140W color index, but it is not shown here. For each relation, the dispersion (σ) is shown in the upper parts of each panel. The possible dependence of these relations on the redshift cannot be directly taken into account, as the redshift is not known a priori. However, using the observed wavelength prior (Section 3.2.6), it is possible to indirectly correct for such a possible effect. The relations are shown for four different bins of redshifts, from $z \sim 1.31$ to $z \sim 1.50$. In order of increasing redshift, blue, $\langle z \rangle \sim 1.32$; green, $\langle z \rangle \sim 1.37$; yellow, $\langle z \rangle \sim 1.43$; and red, $\langle z \rangle \sim 1.50$.

discretized configuration of the inputs, but on their weighted average), the same method results in an uneven precision along the dimensions of the parameter space.

In particular, possible biases could be introduced at different λ_{obs} (i.e., at different redshifts). For example, while the average flux ratios are calibrated as a function of magnitude, color, size, and EW, no redshift evolution is taken into account for these relations, as there is no a priori knowledge of the redshift itself. Similar biases are visible in the combined photo-z PDF, showing systematic over- and underestimations of the photometric redshift at different values of the spectroscopic reference.

Some of these redshift-dependent biases can be corrected by using the wavelength of the observed lines as a prior. Combining all the methods described in the previous sections, the algorithm computes probability ratios between the different species (the probability associated with the most probable species is assumed equal to 1.0). We can exploit the wavelength position of the detected lines for a posteriori fine-tuning of these probability ratios. This approach allows for the optimization of the algorithm.

In the WISP survey, the strongest emission lines measured are mostly either Hα or [O III] (1293 and 949 objects, respectively, in the test "gold" sample). For an additional small fraction of sources, [O II] is the most prominent line (24 objects in the test "gold" sample). For this reason, to improve the overall accuracy of the algorithm, the most effective approach consists in the fine-tuning of the ratio between the probabilities computed for Hα and [O III]. Secondarily, another correction can be applied to the ratio between the [O III] and [O II] probabilities to further improve the performance of the algorithm.

In the left panel of Figure 11, we show the original Hα to [O III] probability ratio obtained for all sources of the test "gold" sample before the fine-tuning correction is applied: $P_{\mathrm{H\alpha/[O\,III]}} = P(\mathrm{H\alpha})/P(\mathrm{[O\,III]})$. It is possible to see that the majority of the emission lines are already correctly identified ($P_{\mathrm{H\alpha/[O\,III]}} < 0$ for [O III] and $P_{\mathrm{H\alpha/[O\,III]}} > 0$ for Hα).
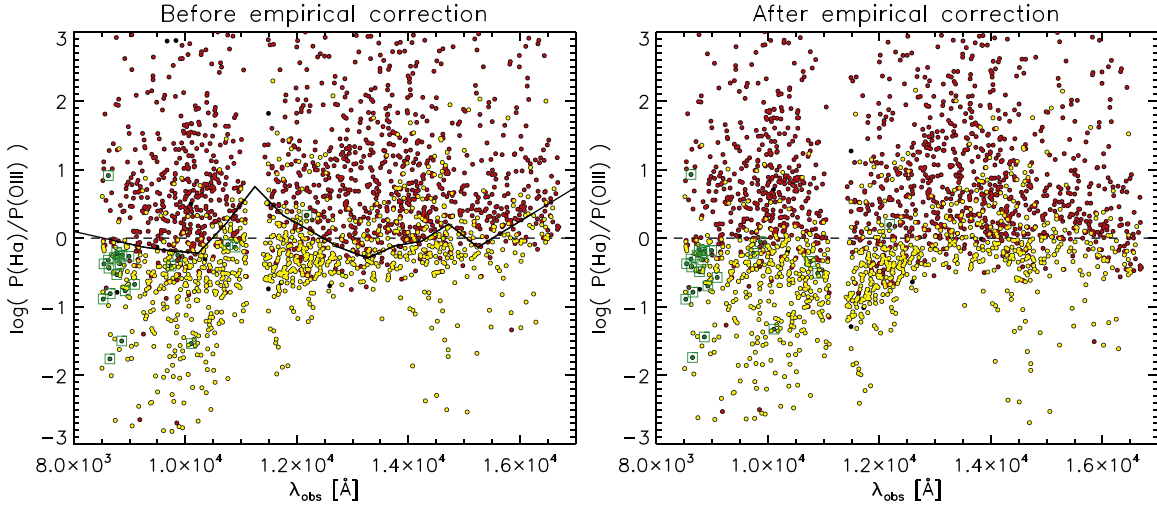
Figure 11. Left panel: Measured $P(\mathrm{H}\alpha)/P(\mathrm{[O\,III]})$ for the sources in the test "gold" sample, before applying the λ_{obs}-based fine-tuning correction ($P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$, black curve, Table 1). The identification of the strongest emission line, detected at λ_{obs}, is represented by yellow circles for [O III], red circles for Hα, and green circles surrounded by squares for [O II]. Right panel: values of $P(\mathrm{H}\alpha)/P(\mathrm{[O\,III]})$ after the application of the $P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$ correction.

<table>
<tr><td colspan="2" align="center">Table 1
Fine-tuning of the $P(\mathrm{H}\alpha)/P(\mathrm{[O\,III]})$</td></tr>
</table>

$\langle \lambda_{\mathrm{obs}}[\text{Å}]\rangle$	$\log P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$
8763.10	−0.042
9259.92	−0.132
9771.42	−0.189
10233.7	−0.218
10730.2	0.211
11248.4	0.750
11755.0	0.331
12234.3	0.111
12755.0	−0.132
13267.0	−0.289
13767.4	−0.111
14242.6	−0.043
14721.9	0.205
15250.4	−0.132
15743.8	0.117
16240.9	0.364

<table>
<tr><td colspan="2" align="center">Table 2
Fine-tuning of the $P(\mathrm{[O\,III]})/P(\mathrm{[O\,II]})$</td></tr>
</table>

$\langle \lambda_{\mathrm{obs}}[\text{Å}]\rangle$	$\log P^{\mathrm{best}}_{[O\,III]/O\,II}(\lambda_{\mathrm{obs}})$
8000	0.4
9400	0.0
9900	0.0
15900	0.7
17000	5.5

prominent [O II] emission above $\lambda_{\mathrm{obs}} \sim 9000$ Å, the probability ratio cannot be precisely fine-tuned. However, above $\lambda_{\mathrm{obs}} \sim 11500$ Å, a systematic deviation of this ratio from $P_{[O\,III]/[O\,II]} = 0$ is immediately evident and easy to correct. In Table 2, we report the values of $P_{[O\,III]/[O\,II]}$ that we applied to $P(\mathrm{[O\,II]})$ to fine-tune the $P(\mathrm{[O\,III]})/P(\mathrm{[O\,II]})$ ratio.
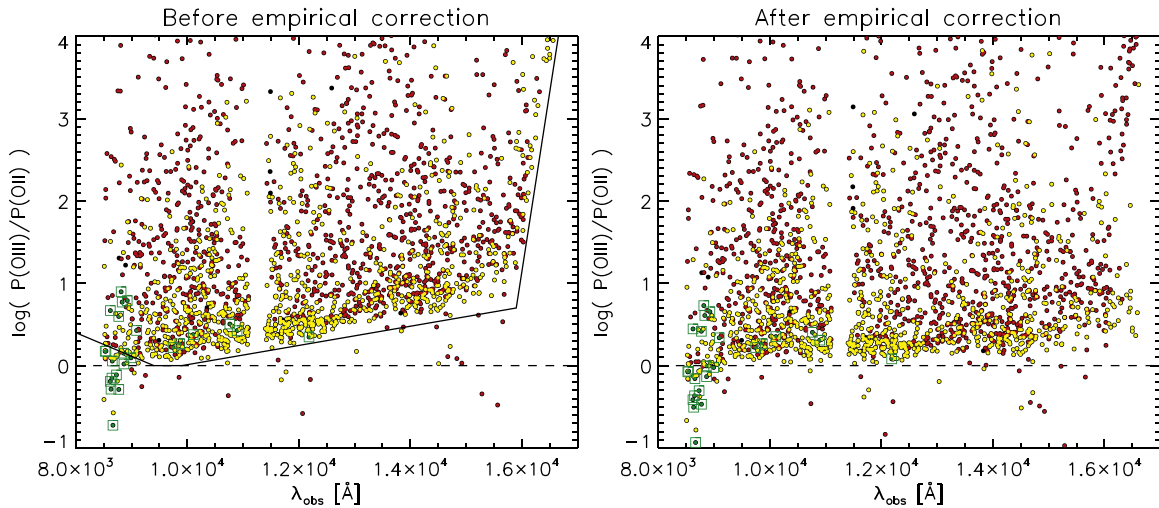
3.2.7. Limitations

The wavelength range covered by WISP is limited. For this reason, the expected flux ratios cannot always be used to identify the strongest measured line. For example, even if the flux ratios indicate that Hα is the expected strongest line, Hα may fall outside of the observable wavelength range. The brightest measured line may be due to other species/transitions (e.g., [O III] or [O II]).

In particular, Hα falls outside the grism at $z > 1.6$, corresponding to $\lambda_{\mathrm{obs}}(\mathrm{H}\alpha) > 17000$ Å and $\lambda_{\mathrm{obs}}(\mathrm{[O\,III]}) > 12900$ Å. Therefore, when (1) the observed wavelength of the strongest (single) observed emission line is $\lambda_{\mathrm{obs}} > 12900$ Å and (2) the expected $F_{\mathrm{H}\alpha}/F_{[O\,III]}$ flux ratio is >1, there is a nonnegligible chance that Hα is indeed stronger than [O III] (and [O II]), but outside the range of observability. Given these circumstances, when conditions (1) and (2) are verified at the same time, we set $P_{\mathrm{FR}}(\mathrm{H}\alpha)/P_{\mathrm{FR}}(\mathrm{[O\,III]}) = 1$, while keeping the values computed for $P_{\mathrm{FR}}(\mathrm{[O\,III]})$ and $P_{\mathrm{FR}}(\mathrm{[O\,II]})$. This is equivalent to not considering the expected flux ratio between Hα and [O III]).

Similarly to the pair Hα–[O III], the same problem is expected to affect the [O III]–[O II] pair when [O III] falls outside the range of observability. However, this mistake is

However, these results can be improved by deriving, for each bin of observed wavelength, the value of $P_{\mathrm{H}\alpha/[O\,III]}$ that better separates the two emission lines: $P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$ (black line in the left panel of Figure 11). In Table 1, we report the tabulated values of $P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$ that we computed in equally spaced wavelength bins ($\Delta\lambda = 500$ Å) ranging from 8500 to 16500 Å (the left column in Table 1 represents the mean λ of the spectral lines in each bin). The right panel of Figure 11 shows the data distribution after the application of the fine-tuning correction $P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$.

While we do not modify the value of $P_{\mathrm{H}\alpha}$, the correction factor $P^{\mathrm{best}}_{\mathrm{H}\alpha/[O\,III]}(\lambda_{\mathrm{obs}})$ is applied to both $P_{[O\,III]}$ and $P_{O\,II}$. This choice allows us to keep unaltered the probability ratio between [O III] and [O II] that is independently corrected.

Similarly to what was done for the Hα and [O III] pair, we applied a correction to the [O II] to [O III] probability ratio. In the left panel of Figure 12, we show the original [O III] to [O II] probability ratio obtained for all sources of the test "gold" sample before the fine-tuning correction is applied: $P_{[O\,III]/[O\,II]} = P(\mathrm{[O\,III]})/P(\mathrm{[O\,II]})$. In this case, given the lack of sources with

Figure 12. Left panel: measured $P([O\ III])/P([O\ II])$ for the sources in the test "gold" sample before applying the λ_{obs}-based fine-tuning correction ($P^{best}_{[O\ III]/O\ II}(\lambda_{obs})$, black curve, Table 2). The identification of the strongest emission line, detected at λ_{obs}, is represented by yellow circles for [O III], red circles for Hα, and green circles surrounded by squares for [O II]. Right panel: values of $P([O\ III])/P([O\ II])$ after the application of the $P^{best}_{[O\ III]/O\ II}(\lambda_{obs})$ correction.

only theoretically possible, as the depth of the WISP survey makes the observability of a $z \gtrsim 2.5$ source unlikely. For this reason, we do not apply any correction to the expected $P_{FR}([O\ III])/P_{FR}([O\ II])$.

On the opposite side of the wavelength range covered by the grisms, the same problem can arise if [O III] is the intrinsic strongest line but it is located below $\lambda \sim 8000$ Å, corresponding to $\lambda_{obs}(H\alpha) < 10536$ Å. While the flux ratios could correctly indicate [O III] as the expected strongest line, the strongest measured line that the algorithm must indicate is Hα. We estimated the effect of this possibility by excluding the flux ratios when $\lambda < 10536$ Å and $F_{H\alpha}/F_{[O\ III]} < 1$. The results that we obtain confirm that the overall effect is negligible on both the test sample (accuracy $\sim 81.7\%$) and the "single-line" sample (for which the recovered distribution in z does not change significantly). For this reason we do consider the flux ratios indicators also when $\lambda < 10536$ Å.

We emphasize the fact that the probability P_{FR}, obtained from the expected flux ratios and corrected as described above, is not used alone. In fact, the outputs of the classification block are combined with the outputs of the regression and optimization blocks, as we will better describe in detail in Section 3.3.

The test "gold" sample includes 632 sources for which the limitations described above apply, out of a total of 2283 galaxies. For this subsample, after applying the correction described, we measure an accuracy (i.e., fraction of spectral lines correctly recovered by the algorithm) of 86.0, which is even higher than the average accuracy obtained testing the complete "gold" sample (82.6%; see Section 4).

Another limitation concerns the availability of the observations needed to compute the $J - H$ color index. We can compute $J - H$ only when one combination of F110W – F140W or F110W – F160W is available. In all other cases, we rely only on the flux ratio priors based on the other indicators.

Finally, when the J magnitude is not available, we consider the H band (F140W or F160W) for the magnitude indicator. In this case, the H magnitude is recalibrated to J as shown in Figure 6.

3.3. Operation Description

Given the strongest line measured in a spectrum (sometimes the only detected line), the main output of the algorithm is the relative probability for each of the species/transition considered to be the responsible for such an emission.

In Section 3.2, we described the modules of the algorithm, their organization in separate blocks. and their independent calibration. In this section, we detail how the different modules are combined with each other, and how the algorithm gets the final probabilities from the outputs of different blocks.

The output of the first block is a combination of redshift PDFs (z-PDF*), i.e., the (peak-normalized) probability associated with each value of z, regardless of the line ID. The final PDF* is obtained by combining the PDF* s independently derived from SED fitting, apparent magnitude, apparent size, and EW (see Sections 3.2.1–3.2.4, and Figures 13 and 14). As such, this function does not provide any information to identify the expected strongest emission line.

The second block statistically associates a set of flux ratios with a given combination of apparent magnitude, size, EW, and $J - H$ color index (see Section 3.2.5 and Figures 9 and 10). This second block does not provide information on the expected redshift, but it predicts which among the considered species/transitions is most likely responsible for the strongest emission in the spectrum considered.

The third block performs the fine-tuning of the relative probabilities computed for the different species/transitions (see Section 3.2.6, Tables 1 and 2, and Figures 11 and 12). This function depends only on the observed wavelength of the strongest emission line (λ_{obs}).

Schematically, for each of the species/transition i, the three functions are combined as follows:

$$P(i) \equiv P(z_i) = P_{PDF}(z_i) \times P_{FR}(i) \times P_{corr}(i, \lambda_{obs}), \quad (6)$$

where $P(i)$ is directly related to the probability that the species/transition i is responsible for the strongest emission line in a spectrum, or equivalently, to the probability that z_i ($\equiv(\lambda_{obs}/\lambda_i) - 1$) is the correct spectroscopic redshift. $P_{PDF}(z_i)$ is the value of the z-PDF* in z_i, while $P_{FR}(i)$ is the
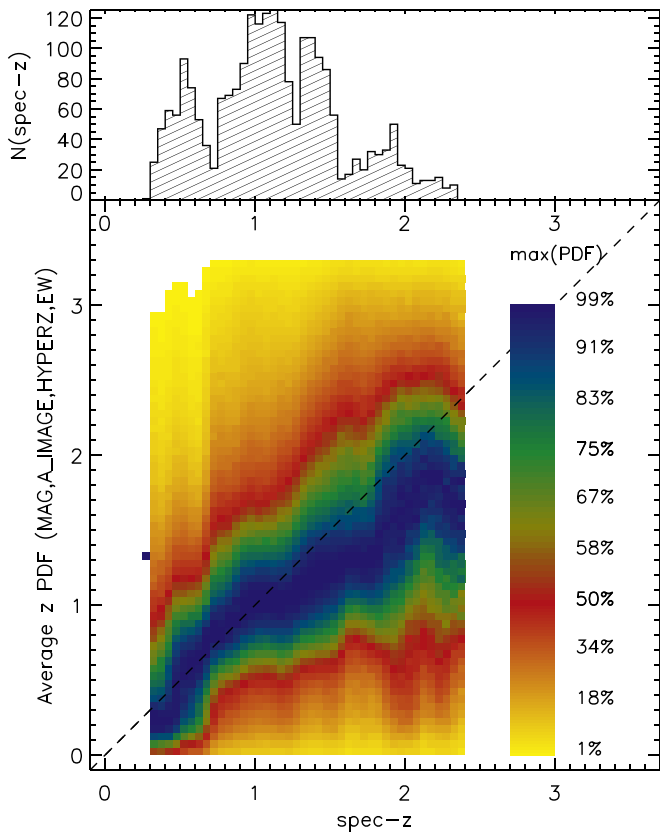
Figure 13. Average PDF* resulting from the combination of the PDF* obtained using all of the redshift indicators described in Sections 3.2.1–3.2.4 (SED fitting, apparent J magnitude, size, and EW, respectively) and singularly shown in Figure 14. The histogram in the upper panel represents the distribution of the spectroscopic sources used in the experiment ("gold" sample).

probability obtained from the flux ratio priors, and $P_{\rm corr}(i, \lambda_{\rm obs})$ represents the $\lambda_{\rm obs}$-dependent fine-tuning correction. The z-PDF* is the combination of the SED-fitting output with the priors obtained from the apparent J magnitude, size, and EW:

$$P_{\rm PDF}(z_i) = P_{\rm PDF}^{\rm sed-fit}(z_i) P_{\rm PDF}^{J}(z_i) P_{\rm PDF}^{\rm size}(z_i) P_{\rm PDF}^{\rm EW}(z_i). \qquad (7)$$

$P_{\rm FR}(i)$, derived from the measured flux ratios, is the weighted average of the values inferred from the apparent magnitude, size, EW, and $J - H$ color, as described by Equations (3) and (4). For each source, the values of $P(i)$ are normalized to the maximum value of $P(i)$ obtained among the different species/ transitions i, so that if, e.g., the transition n is the most probable one responsible for the strongest emission, with the transition m having only half of the probability with respect to n, then the algorithm indicates $P(n) = 1$ and $P(m) = 0.5$, respectively (i.e., the sum over i of the factors $P(i)$ is larger than[20] 1.0).

The output probability ratios are not directly and unequivocally related to the expected flux ratio between two transitions. This can be understood if we consider the hypothetical example of a source characterized by an extremely

---

[20] It is always possible to renormalize the single values of p_i to their sum, obtaining actual values of probability, because all of the terms are known. However, we emphasize that such a convention would be misleading: the algorithm outputs relative probabilities among the options taken into account, but it does not consider all the possibilities, such as additional transitions (e.g., Lyα) or possible spurious detections.

narrow and precise photo-z PDF. In this case, the probability associated with the emission line identified as the strongest would be very high, independent of the flux expected from the other transitions.

Figure 15 shows the typical output of the algorithm. While in the upper panel of this figure the z-PDF* are independently shown, in the lower panel it is possible to see their combination (see Equation (7)). The relative probabilities, associated with the different possible i solutions, are represented using Gaussian functions (where σ corresponds to the redshift uncertainty Δz in the original WISP catalog). Their peak values are proportional to the corresponding values of $P(i)$.

4. Accuracy and Uncertainties

We computed the accuracy of the algorithm by running it on the test "gold" sample and comparing the output predictions with the original WISP classifications. The latter sample includes the identifications performed by WISP reviewers when they agreed on the redshift. Coverage by both grisms is required, and at least two emission lines must be detected. This selection guarantees that the uncertainties estimated can be almost completely attributed to the algorithm itself and not to the original WISP classification. In any case, if a small fraction of incorrect identifications were attributable to the original WISP classification, we would be underestimating the actual accuracy of the algorithm (unless the amount of misidentification is big enough to sensibly modify the calibration of the algorithm itself). We limited the wavelength range of both the calibration and the test phases to two intervals: $8500\,\text{Å} < \lambda_{\rm obs} < 11000\,\text{Å}$ and $11400\,\text{Å} < \lambda_{\rm obs} < 16700\,\text{Å}$. This limitation mitigates the effects of the low grism transmission, including inaccurate flux ratio estimates and possible false detections.

For the test "gold" sample, we know the identity of the strongest emission line in each spectrum. In the test run, we treat these sources as if only one line, the strongest, was measured (i.e., the algorithm is blind to any information concerning the other emission lines).

It is important to notice that the sample that we use to estimate the accuracy of the algorithm is made by real emission lines. This may not be the case in different samples, in particular when only one spectral feature (a possible false detection) is detected. This question is discussed in more detail in Section 5.3.

The test "gold" sample is divided into subgroups for each of which a different species/transitions is responsible for the strongest emission measured. The sources showing emission lines generated by the Hα transition represent the most numerous subgroup ($N_{\rm H\alpha} = 1293$), followed by the [O III] group ($N_{\rm [O\,III]} = 949$), [O II] ($N_{\rm O\,II} = 24$), and other species/ transitions ($N_{\rm other} = 16$ in total for Hβ, Hγ, [S II]).[21]

In our test, the algorithm is run over the complete sample. For each spectrum, the species/transitions that the algorithm indicates as responsible for the most prominent line is the one for which the algorithm itself found the highest value of probability.

Table 3 summarizes the inputs and the results of our test on real lines. For each of the subgroups, we measured the recovery fraction, or completeness (the fraction of emission lines

---

[21] One spectrum showing a strong [S III] 9532 emission is excluded from the test.
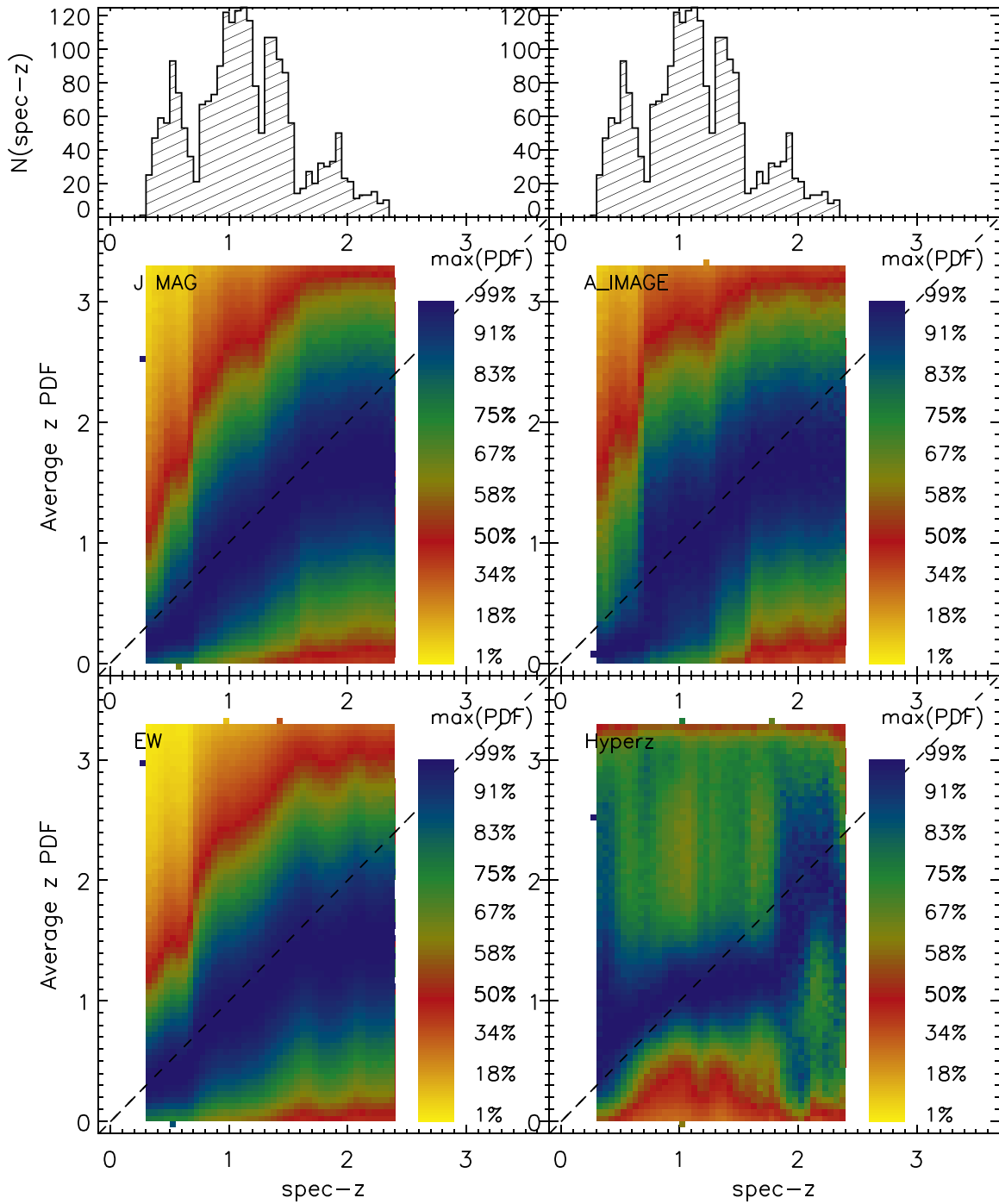
Figure 14. Average PDF* obtained using different methods, as a function of the actual spectroscopic redshift of the sources in the "gold" sample. The PDF* are computed in bins of spectroscopic redshift (*x*-axes) and represented through a color scale. The duplicated histograms in the upper panels represent the distribution of the spectroscopic sources used in the experiment. Top left: average PDF* obtained using the *J* magnitude (F110W band) as a redshift indicator (Section 3.2.2). Top right: average PDF* obtained using the A_IMAGE parameter (apparent size in the F110W band) (Section 3.2.3). Bottom left: average PDF* obtained using the EW (Section 3.2.4). Bottom right: average PDF* obtained using the Hyperz SED-fitting method (Section 3.2.1). While the low-precision SED-fitting technique encounters problems between $z \sim 1$ and $z \sim 2$, the other redshift indicators are better correlated with the redshift in the same range.

correctly identified by the algorithm), and the contamination (fraction of wrong identifications in the recovered sample). All species/transitions summarized in Table 3 are considered in our tests. However, the algorithm is not meant to properly work on the few rare objects characterized by peculiar emissions. Because the [S II], Hβ and Hγ fluxes are almost always weaker than those measured for Hα, [O III], and [O II], only very few sources characterized by a prominent emission from [S II], Hβ,

and Hγ can be used in the calibration process (i.e., our algorithm is not expected to work for these species).[22]

For the most numerous samples, Hα and [O III], the completeness is high (88.3% and 77.0%, respectively), and

---

[22] None of the 16 sources characterized by [S II], Hβ, or Hγ stronger than the other transitions in the test "gold" sample was correctly identified by the algorithm. Only one source was wrongly classified as [S II].
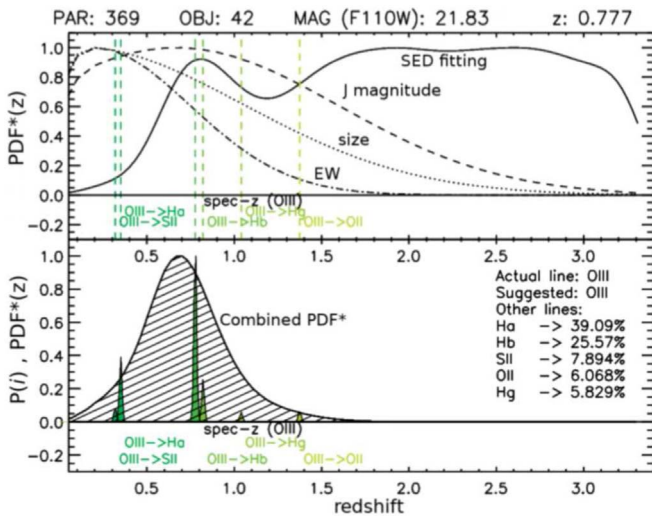
Figure 15. Example output of the algorithm. The upper panel shows the four z-PDF* obtained from SED fitting ($P_{PDF}^{SED-FIT}(z)$, black continuous line), apparent J magnitude ($P_{PDF}^{J}(z)$, black dashed line), apparent size ($P_{PDF}^{size}(z)$, black dotted line), and EW ($P_{PDF}^{EW}(z)$, black dotted–dashed line). The redshift solution originally indicated by the WISP reviewers is indicated by a "spec-z" label, with the corresponding species/transition reported in parentheses (in this example, [O III]). All of the possible redshift solutions z_i are also shown using vertical dashed lines labeled as "original solution $\rightarrow$ other possible solution" (example: [O III] $\rightarrow$ Hα). In the lower panel, the three z-PDF* are combined in a unique PDF* ($P_{PDF}(z)$, continuous black line and shaded area), as described by Equation (7). The possible solutions are represented using Gaussian profiles ($\sigma = \Delta z$) with amplitude equivalent to the normalized value of $P(i)$. For each of the species/transition, the corresponding normalized value of relative probability (Equation (6)) is reported in the legend (the probability is expressed as a percent fraction of the maximum value of $P(i)$ obtained, corresponding to 100%).

the contamination is contained (15.2% and 17.9%, respectively). The [O II] sample still shows an acceptable completeness (50%) but a very high contamination (73.3%). This high level of contamination can be explained by the fact that even a very small fraction of wrong identifications in large samples (such as Hα and [O III]) can cause a strong contamination in a numerically smaller sample (e.g., [O II]). For the other species/transitions considered, we do not expect the algorithm to produce any reliable result, for at least two reason. First, the very limited amount of such emitters in the original WISP calibration "gold" sample does not allow for an opportune calibration of the algorithm itself on these species/transitions. Second, the strong contamination from numerically larger samples is expected to compromise the reliability of the recovered samples (similarly to what happens for the [O II] transition).

The overall accuracy of the algorithm can be measured as the fraction of spectra for which the species/transition responsible for the strongest emission is correctly predicted by the algorithm. Considering all the sources in our test, we found this overall accuracy to be 82.6%. This value demonstrates how the algorithm can strongly improve the accuracy obtained by considering the SED-fitting strategy alone ($\sim$50%, if we consider only Hα, [O III], and [O II]) and the accuracy that we would obtain by always classifying every strongest line as Hα ($\sim$57%), as in the WISP default classification option. We can compute the accuracy also for each single sample (in this case, accuracy = purity = 1 − contamination). As for the completeness and contamination, we report these estimates in Table 3.

The probability indicators can be used to select samples with different levels of purity. This can be achieved by selecting different probability thresholds. Figure 16 shows how increasing values of $P(H\alpha)$ and $P([O III])$, normalized to the sum of the probabilities $\sum_i P(i)$, correspond to samples characterized by increasing levels of purity (accuracy). Given the limited amount of data, we cannot show a similar figure for the [O II] sample. For this case, we refer the reader to the average value reported in the last column of Table 3.

4.1. The Not-independent Calibration and Test Samples

As specified in Section 2.2, the calibration and the test samples are not independent of each other. In particular, all of the data in the calibration sample are also included in the test sample. This decision is justified by the argument that follows. The algorithm is characterized by a modular structure, where each module represents a simple relation (usually linear) between different observational quantities (e.g., apparent size versus redshift, or magnitude versus line flux ratios). These relations are computed using a large amount of data so that each single datum produces a negligible effect on the calibration of a specific module. In other words, removing a datum from the calibration sample would not change the behavior and outputs of the algorithm, even when testing the software on the removed datum. This condition is visually represented in Figure 17. The test that we perform would be unreliable (representing an upper limit to the actual accuracy) only if few data points were available, or if we were considering particularly complex relations prone to overfitting issues.

On the other hand, the accuracy can be assessed by dividing the sample into two independent subsets and calibrating the relations using only one of the two subsamples. This approach guarantees that the accuracy (estimated on the other independent subsample) is not biased by possible overfitting or by the use of too complex functions. However, given the smaller amount of data used for the calibration itself (half of the total), the relations cannot be as accurate as they are when considering the complete sample. In other words, the accuracy estimated using the method just described represents an underestimate of the actual accuracy that can be achieved by calibrating the algorithm using the entire data sample. For the reasons explained, we consider the accuracy computed using the full "gold" sample a more reliable estimation of the actual accuracy of our algorithm.

To convince the skeptical reader, we performed two validation tests by dividing the sample as described above (see Tables A1 and A2 in Appendix A). In these tests, we obtained an overall accuracy of 80.6% and 80.4. These values are lower, but consistent, with the 82.6% accuracy reached using the same sample for both calibrating and testing the algorithm. The lower accuracy obtained in these two tests does not indicate that when the sample is not divided the algorithm is overfitting (i.e., that we are overestimating the accuracy). Vice versa, the small difference is mostly due to the less precise calibration achievable exploiting only half of the available data.

The exception to the arguments reported above is represented by the sample of strong [O II] emitters, for which the small amount of data available, especially at $z > 1.5$, limits our ability to calibrate the algorithm (in particular, regarding the empirical correction described in Section 3.2.6) and to measure the associated uncertainty. For this reason, the precision
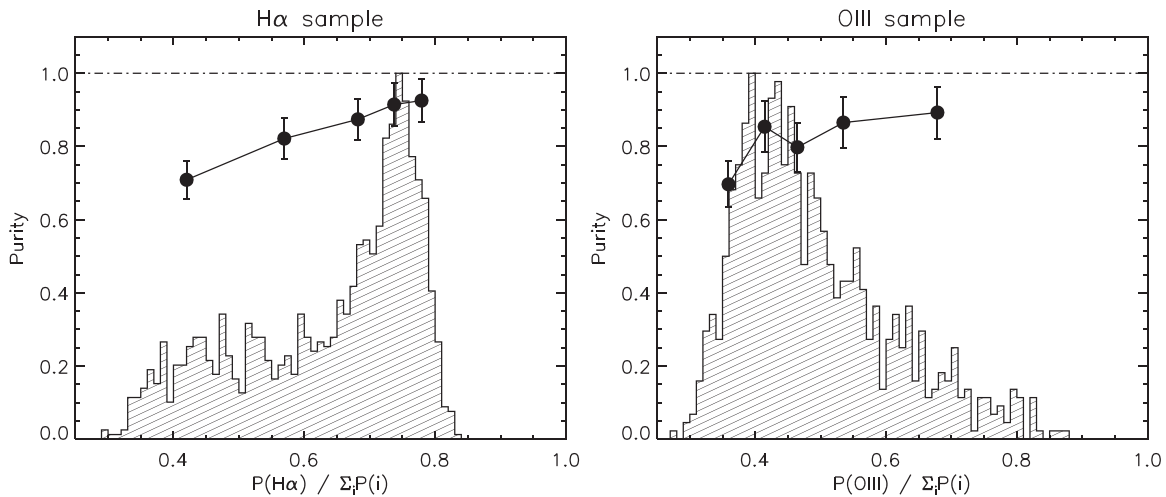
Figure 16. Purity of the recovered Hα and [O III] samples (i.e., fraction of emission lines correctly identified as due to Hα and [O III] by the algorithm) as a function of the probability values $P(H\alpha)$ (left panel) and P[O III] (right panel), normalized to the sum of the probabilities $\sum_i P(i)$. The black filled circles represent the average purity obtained in equally populated bins of probability, each of them containing 20% of the data. Poissonian uncertainties are shown. The background histograms represent the normalized probability distribution of the overall sample (Hα +[O III]).

Table 3
Accuracy Test on the Species/Transitions Considered by the Algorithm, Ordered by Importance (N_{TGS})[a]

Species/ Transitions	λ_{obs} (Å)	N_{TGS}[b]	N_I[c]	N_{CI}[d]	N_{WI}[e]	Completeness (N_{CI}/N_{TGS})	Contamination (N_{WI}/N_I)	Accuracy (Purity) (N_{CI}/N_I)
Hα	6564.5	1293	1346	1142	204	88.3%	15.2%	84.8%
[O III]	4960.3–5008.2	949	890	731	159	77.0%	17.9%	82.1%
[O II][f]	3727.1–3729.9	24	45	12	33	50.0%	73.3%	26.7%
[S II]	6718.3–6732.7	7	1	0 (0%)	1 (100%)	0%	100%	0%
Hβ	4861.4	5	0	0 (0%)	0 (-)	0%	⋯	0%
Hγ	4340.5	4	0	0 (0%)	0 (-)	0%	⋯	0%

Notes. Bold values correspond to the species/transitions on which our analysis focuses.
[a] All the estimates reported refer to samples made of real lines.
[b] Number of sources, in the test "gold" sample (TGS), showing this transition as the most prominent emission line).
[c] Number of sources, in the test "gold" sample, for which the algorithm associates the highest relative probability with the possibility of this transition being responsible for the most prominent emission line measured (Identified, I).
[d] Number of sources, in the test "gold" sample, for which the strongest emission due to this transition is correctly identified (CI).
[e] Number of sources, in the test "gold" sample, for which the strongest emission is mistaken for this transition (wrongly identified, WI).
[f] Due to the small amount of data included in the calibration and test sample, the accuracy estimated should be considered an upper limit to the actual value.

reported for the [O II] sample in Table 3 should be considered an approximate upper limit to the actual value. However, we point out that given the small amount of data, the same uncertainty associated with the precision estimate does affect also any other estimate obtained dividing the complete sample into two independent subsamples (this is confirmed by comparing the precision obtained for the [O II] sample in Tables 3, A1, and A2).

5. Single Lines

5.1. Characterization of the Single-line Sample

After calibrating the algorithm using the secure identifications of the "gold" sample, we run the software on the sample of spectra showing single unidentified emission lines. To this purpose, similarly to what is required for the "gold" sample, we consider only 345 "single-line" sources covered by observations in both the G102 and the G141 grisms at the same time.

It is important to note that for many spectra, after the original identification of the main emission line, the flux of additional

lines could be automatically measured, even if a proper detection of these lines (S/N > 2σ) was not available. In many cases, some of these undetected but visible (and measurable) lines helped the WISP reviewers during the by-eye classification process. However, considering these lines as secure identifications would be misleading (e.g., a bump in the noise can easily be mistaken for Hβ or [S II]). Therefore, these spectra are not included either in the gold or in the single-line sample.

In order to test the homogeneity of the two samples, we compare their main properties. Figure 18 shows the distributions of flux of the strongest line, apparent magnitude, apparent size, and $J - H$ color index. While the $J - H$ color distribution is similar for the two samples, the distributions of apparent magnitude, size, and line flux are not identical. On average, the strongest line of a spectrum is weaker (in absolute terms) when only one line is detected. This behavior is not unexpected, as weaker emission lines will not be detected in a spectrum where the strongest line is already at the level of the S/N detection threshold. On the other hand, because the ratios between the
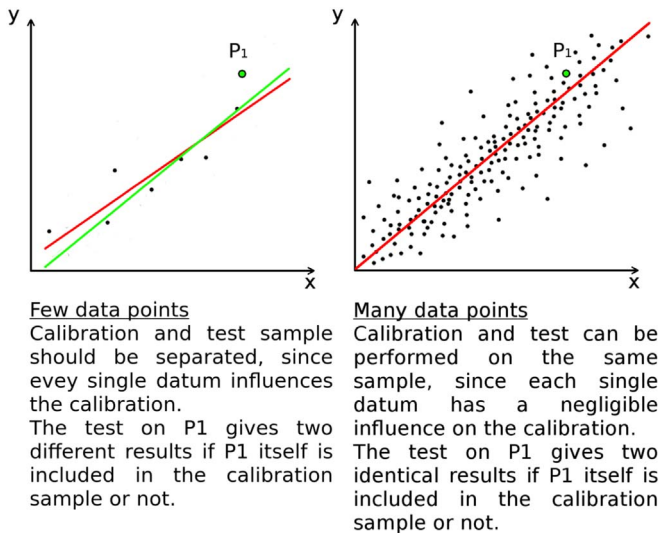
Figure 17. The best approach to validate a relation is to test it on a data sample that is independent of the sample used to compute the relation itself. This is due to the fact that every datum tends to be closer to the relation, if the relation was calibrated using that datum (left panel) or if the function considered is particularly complex (overfitting problem). However, when the function is simple (such as the linear relation in these plots) and the amount of data is sufficiently large, every datum considered singularly has a negligible effect when calibrating the relation (right panel). In this context, when testing the relation on a datum, there is no difference whether or not that specific datum was included in the calibration data set.

fluxes due to different transitions are not completely arbitrary (see plots in Figures 9 and 10), when the strongest line is characterized by a particularly high value of S/N, the weaker lines are also more likely to be detected above the S/N threshold.

Similarly to the flux of the strongest line, the apparent magnitude is usually higher when only one line is detected. Again, this is due to the fact that, given the broad correlation between observed line flux and continuum (see Figure 19), fainter sources are also characterized by weaker emission lines. Hence, in these cases, all emission lines are more likely to be undetected. The opposite argument applies to the strongest emission lines, which are more commonly associated with brighter J magnitudes.

Because the similar $J - H$ color distribution (bottom-right panel of Figure 18) is consistent with a homogeneous redshift distribution, the different apparent size distribution (bottom-left panel of Figure 18) indicates that these effects are probably due to a different average intrinsic size of the two samples.

5.2. Classification of the Single Lines

As suggested by the similar $J - H$ color distribution (bottom-right panel of Figure 18), the automated classification performed by the algorithm confirms that the "single-line" and the "gold" samples are characterized by similar redshift distributions. This can be observed by comparing the red histogram in the right panel of Figure 20 with the black histogram in the left panel of the same figure.

When multiple lines are detected, we can safely consider the human classifications as a reference. Then, the similar distributions in the left panel of Figure 20 represent an additional reliability test of the automatic classification, as it

indicates that no significant biases are introduced by the algorithm.

On the other hand, when only one single line is detected, the software used for the by-eye classification performed by the WISP reviewers is set to initially indicate Hα as a default choice. The human reviewer can modify (or confirm) this default option by considering the presence of additional spectral lines (not detected above 2σ) or features (e.g., the asymmetric shape of the unresolved [O III] doublet). This default option is justified by the fact that indeed Hα is the actual strongest emission for more than half of the WISP "gold" sample ($\sim$56.7%). However, if we were considering the default Hα option as always valid when additional spectral features were not present (black histogram in the right panel), the redshift distribution of the entire "single-line" sample would present strong differences with the distribution of the sources in the "gold" sample (black shaded histograms in the left panel). In particular, the bulk of the sources located between $z \sim 0.7$ and $z \sim 1.5$ in the test "gold" sample would almost completely disappear in the "single-line" sample. Because we can safely accept the original classification of the "gold" sample, it follows that assuming the unidentified single lines is always generated by Hα emission results in a relevant bias.

An independent confirmation of this bias can be found by observing the plots of Figure 19 (flux of the brightest detected line versus apparent magnitude). In these plots, Hα and [O III] lines are located, with some overlap, in two different regions of the bidimensional space. From the bottom panel of the same figure, we can observe that, below $J \sim 24.5$, the "gold" and the "single-line" samples are similarly distributed, indicating that the fraction of Hα and [O III] lines must be comparable for these two samples. The smaller fraction of "single-line" sources above $J \sim 24.5$ is located in a region of the plot where [O III] lines dominate. These observations confirm that the "single lines" cannot be mostly Hα, as assumed in the original WISP default classification. Instead, they are more probably evenly distributed between Hα and [O III], with an excess of [O III] above $J \sim 24.5$.

Summarizing, the distributions of the $J - H$ color index, apparent size, magnitude, and line flux (Figures 18 and 19) further supported by similar redshift distributions of the "gold" and "single-line" samples (as recovered by the algorithm) indicate that the two samples are characterized by galaxies with intrinsic differences in size (and stellar mass), but with similar redshift distributions. Additionally, we point out that the human classification of single lines, driven by the use of a "default option" (such as Hα), biases the sample.

5.3. Contamination from False Detections

Our comparison between the gold and the single-line samples (Section 5.1), and the classification performed by the algorithm (Section 5.2) do not take into account the contamination due to false detections. The identification of multiple lines guarantees that the gold sample is not affected (or marginally affected) by this problem. On the contrary, given the detection of only one line, for the single-line sample, there is no such guarantee.

Most of the contaminants were eliminated from the single-line sample during the original by-eye classification. False detections can be easily identified when the line detected in one of the two grisms is too bright, if compared with the overall
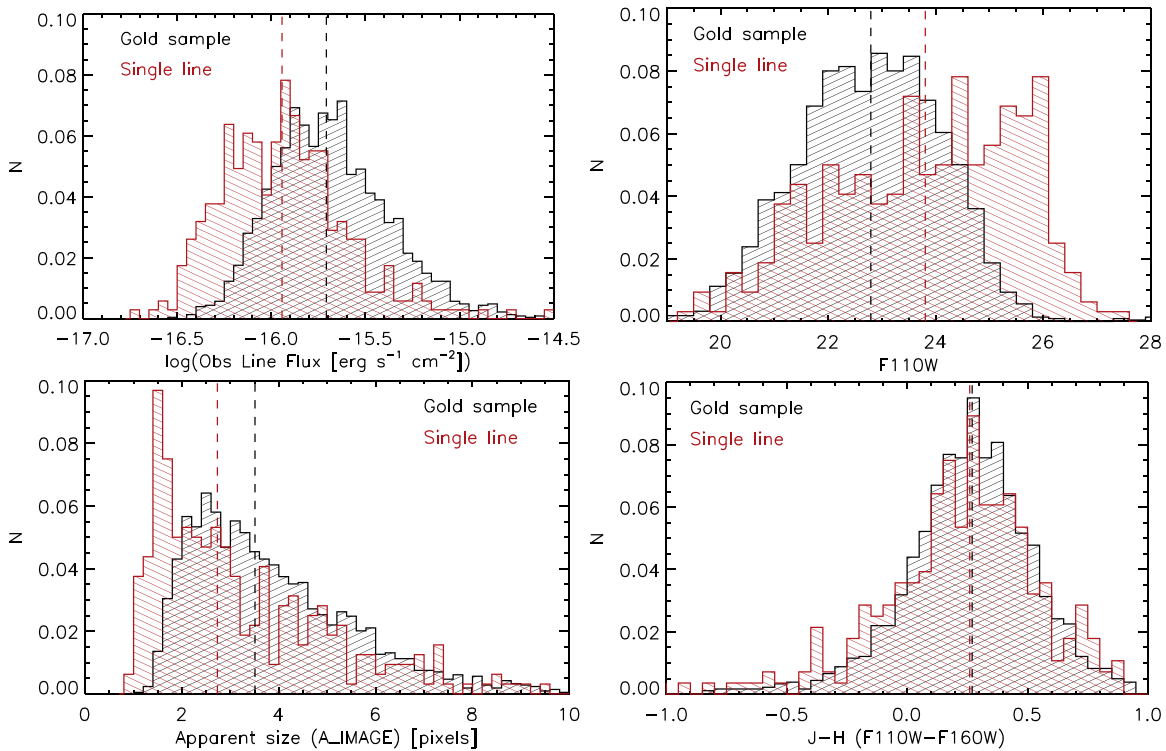
Figure 18. Normalized distribution of the flux of the strongest line (upper-left panel), J magnitude (upper right panel), apparent size (bottom left), and $J − H$ color index (bottom right). The distributions are shown for the calibration sample (in black), characterized by two or more lines detected, and for the sample of spectra with only one line detected (in red). The vertical dashed lines represent the median values for the two samples. The similar $J − H$ color distribution indicates that there are no strong redshift differences between the two samples. Instead, the differences observed in the distributions of line flux, total magnitude, and apparent size suggest an average difference in intrinsic size, stellar mass, or both.

emission of the source in the corresponding photometric band (J or H). In other cases, they can be identified because the shape (or size) of the detected spectral feature is completely at odds with the shape (size) of the emitting source. Most of the time, contaminants are also decentered with respect to the midline of the bidimensional spectrum.

However, in some cases, false detections cannot be easily identified as described above. This happens when both the spectrum and the contaminant are faint and point like. In these cases, it is impossible to compare the apparent shapes of the spectral line with the shape of the source in the image. Additionally, determining the precise position of the source and corresponding midline of the bidimensional spectrum is not as easy as with brighter sources.

Sometimes, false detections can be due to the stochastic nature of noise, but most of the time the contamination is due to the overimposition of a zero-order spectrum of a different source located just outside the field. Because most of the sources in any (randomly located) field are faint and point like, then observing faint, rather than bright, contaminated spectra is much more likely. Consequently, contaminants are usually characterized by high magnitudes, faint line fluxes, and small apparent sizes. Given the depth of the WISP survey, these characteristics are typical of spectra dominated by the [O III] emission (see Figure 9). Therefore, we expect the algorithm to identify almost all of the false detections as [O III], whereas they are classified as Hα in the WISP default classification.

The current version of the algorithm is not designed to identify false detections. However, in Section 5.4, we estimate an upper limit to the fraction of contaminants expected for the

single-line sample in the redshift range $0.8 < z < 1.2$ ($\sim$30%–35%).

Finally, we remind readers that the future Euclid and WFIRST surveys are going to be significantly less contaminated by false detections. This will be achieved by dispersing multiple spectra of the same fields along different directions.

5.4. Single Lines at $0.8 < z < 1.2$

We focus our analysis on a subsample of "single-line" galaxies that our algorithm locates in the redshift range $0.8 < z < 1.2$. Also in this case, we consider only sources covered by observations in both the G102 and the G141 grisms.

As illustrated in Figure 20, the redshift bin considered would be almost empty, if all these sources were classified as Hα when no distinguishable spectral features were visible (default option). On the contrary, the same bin includes the peak of the redshift distribution of the "single-line" sample when we consider the automatic line identification performed by the algorithm. In other words, most of the lines in this range are classified as [O III] by our software, whereas the WISP default classification indicates them as Hα.

This redshift range is particularly relevant also because, given the wavelength coverage of grisms G102 and G141, Hβ, [O III], and Hα are contemporarily located in the visibility range of the WISP survey. We exclude sources located above $z = 1.2$ because, at these higher redshifts (and up to $z \sim 1.5$), the [O III] and Hβ lines fall into the noisier spectral region located between the passband of the two grisms. Additionally, limiting the redshift bin to $z < 1.2$, we can obtain a more homogeneous sample, in terms of redshift distribution.
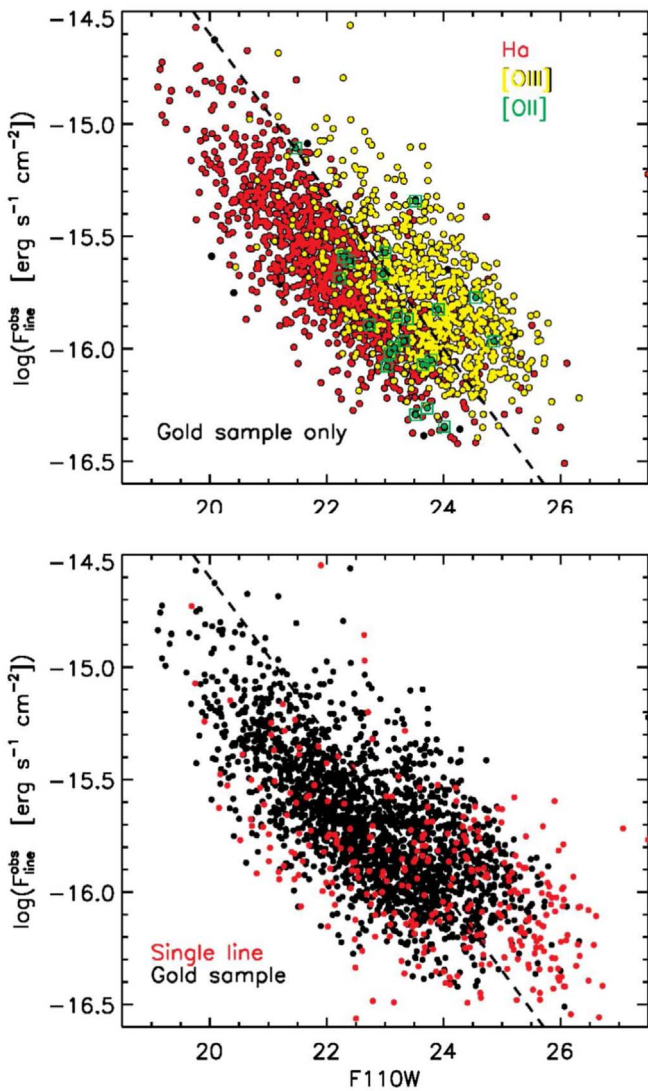
Figure 19. The broad anticorrelation existing between the flux of the strongest spectral line detected and the observed J magnitude. In the upper panel, we consider the "gold" sample, using red filled circles when the brightest line is Hα, yellow when it is [O III], and green circles enclosed in squares when it is [O II]. Two different normalizations are clearly visible for Hα and [O III]. For [O II] lines, the behavior looks more controversial. In the bottom panel, the relation is shown for the "gold" (black filled circles) and for the "single-line" (red filled circles) samples. We can approximately separate Hα from [O III]-dominated sources, by empirically dividing the plots into two regions (black dashed lines). Almost all of the "single-line" sources at high magnitudes ($J \sim 24.5$ corresponds to the average magnitude of the single-line sample at $0.8 < z < 1.2$) are located on the [O III] side of the plot, along the natural continuation of the relation valid for sources dominated by the [O III] emission itself. This evidence independently confirms the prevalence of [O III] among the "single-line" sources at $0.8 < z < 1.2$ (compare with the stacked spectrum in Figure 21).

5.4.1. Stacking

We stacked the "single-line" spectra detected in this redshift range using the method described in Appendix B. Before the stacking, we removed from the sample about 14% of spectra characterized by heavy contamination or other evident problems. The redshift selection and the exclusion of bad spectra reduces the original "single-line" sample to 148 valid spectra that we stacked according to the line identification performed by the algorithm. The result of this process is shown in Figure 21.

While none of the single spectra is characterized by Hβ or [S II] as the strongest (unique) emission line, both these lines are detected well above the local 3σ threshold, in the stacked spectrum. The detection of these lines would not be surprising even if the algorithm's classification was failing in the majority of the cases (provided that it was correct in a significant fraction of cases). However, the detection of these lines at such high intensities (compared with [O III]) indicates that the classification performed by the algorithm is mostly correct.

For example, from the gold sample, we expect log([O III]/Hβ) ~ 0.8 at $J \sim 24.5$ (corresponding to the typical magnitude of the stacked sample). From the stacking, we measure log([O III]/Hβ) ~ 1.07. However, Hβ can contribute to the value measured in the stacked spectrum only when the brightest emission lines in the single spectra are correctly classified. Therefore, we must conclude that the algorithm is correctly classifying lines in at least $\sim 54\%$ of the cases. In reality, due to the presence of some biases affecting this ratio (see Section 5.4.2), the actual accuracy is higher.

In order to test the alternative possibility (i.e., most of the lines due to Hα instead of [O III]), we stacked the same spectra as if the original WISP classification were correct. In this case, [O III] and Hβ would be located at $\lambda < 8000$ Å most of the times, and for this reason they cannot be reliably considered in this test. On the contrary, the [S II] line would always be located inside the wavelength range covered by the two grisms. The stacked spectrum shown in Figure 22, shows that [S II] is not even visible (much less detected). We conclude that the WISP default classification must be wrong in much more than about half of the cases.

Some relevant flux ratios, which we measure from the stacked spectrum, are listed in Table 4 (last column). We can compare these values with the flux ratios expected at similar magnitudes for the gold sample (second column). It is immediately evident that some flux ratios (especially [O III]/Hα and [O III]/Hβ) are particularly high compared with the expected values. The reasons for such discrepancies are explained in Section 5.4.2.

5.4.2. Biases

The line ratios that we measure from the stacked spectrum are affected by three major biases that we estimated by running a series of simulations. In Table 4 we compare the expected flux ratios, which we input into the simulations, with the outputs and with the ratios measured from the stacked spectrum.

The first bias affecting the line ratios (bias 1) is due to the particularly restrictive sample selection. In fact, the single-line sample is made by spectra characterized by only one line detected above 2σ and with no other measurable lines visible (not even below the same 2σ threshold). In fact, when additional lines are not detected but visible (and measurable), the spectra are classified as multiple lines (although they are not included in the gold sample).

Due to the typically high magnitude of the stacked single-line sources ($J \sim 24.5$), most of these spectra are characterized by [O III] stronger than Hα, as shown in Figure 9. Hence, the sample used for the stacking is mostly made by sources characterized by a particularly suppressed Hα emission. In fact, if Hα were not suppressed, then it would simply be measurable, causing the source to be classified as "multiple line," and the spectrum to be excluded from the stacking.
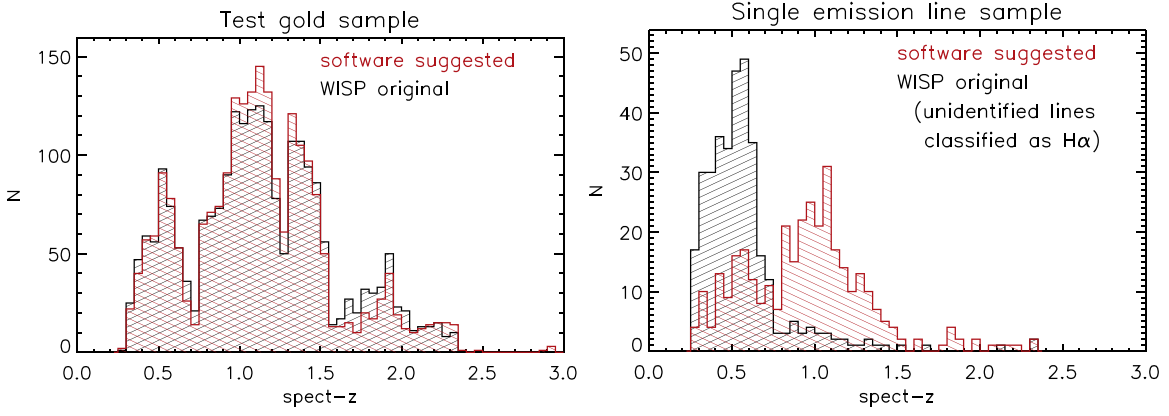
Figure 20. Redshift distribution of the sources resulting from the WISP original (human) and WISP default classification (black shaded histograms in the left and right panels, respectively), compared with the distributions resulting from the automatic classification performed by our algorithm (red shaded histograms). The left panel shows the original and the recovered redshift distributions of the sources in the test "gold" sample. The two distributions agree with each other, indicating that the algorithm does not introduce evident redshift-dependent biases. In the right panel, we consider the sample of sources characterized by only one spectral line detected. In this case, a default identification (Hα) is compared with the automatic classification of our algorithm. The strong differences between the two distributions highlight the limits of using a default option when the emission line is not identified. In fact, while the distribution recovered by the algorithm is consistent with that observed for the "gold" sample, the default option is clearly strongly biased.
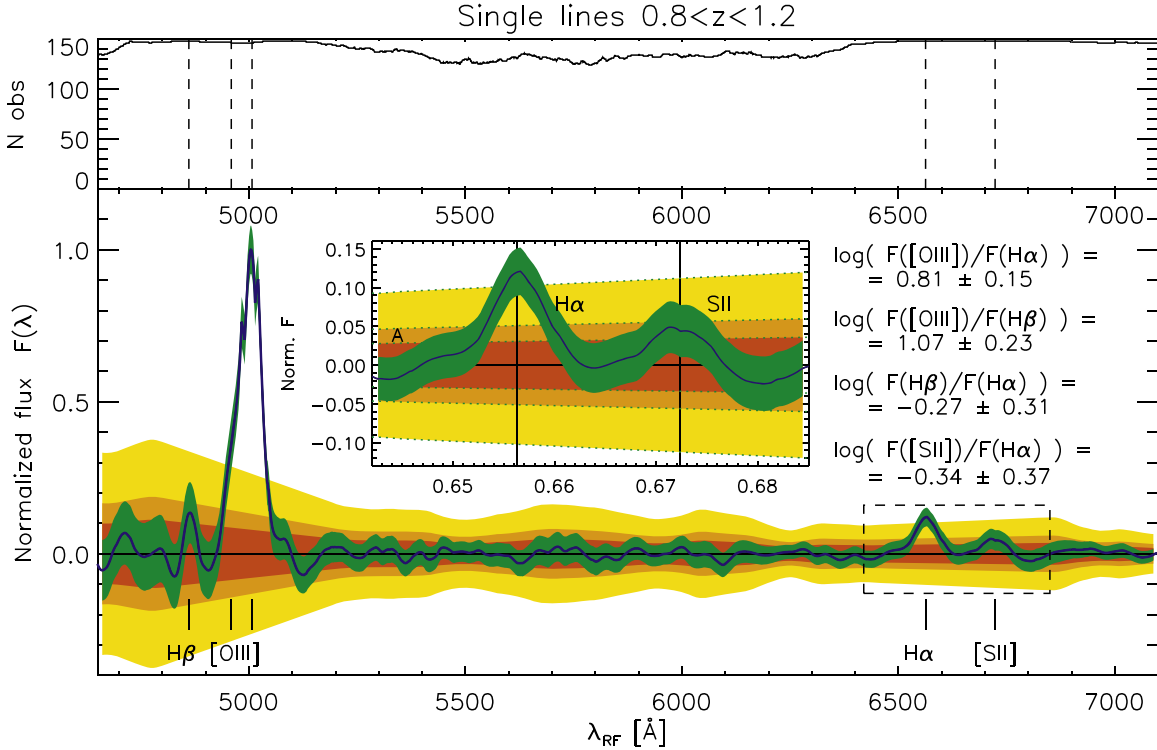
Figure 21. Stacked spectra of "single-line" sources located in the redshift range $0.8 < z < 1.2$. The spectra are stacked consistently with the solution suggested by the algorithm. The position of the main spectral emissions is indicated by black vertical lines (in order of increasing wavelength: Hβ, [O III], Hα, [S II]). The upper panel shows the number of spectra considered for the stacking, at each wavelength. In the lower panel, we show the stacked spectrum (blue line), resulting from the weighted average of the normalized single spectra, after rejecting the values above and below 1σ from the median value (i.e., above the 84th percentile and below the 16th percentile). Single spectra are normalized by the flux of the detected emission line. In order to obtain the average spectrum $F(\lambda)$, at each wavelength every spectrum is weighted using the square inverse of the local noise $(1/\sigma(\lambda)^2)$. In the resulting spectrum, the peak of the strongest emission line ([O III]) is normalized to 1.0. The red, orange, and yellow shaded areas represent the 3σ, 5σ, and 10σ thresholds, respectively, where $\sigma(\lambda)$ is locally computed from the final stacked spectrum (the positions of the main lines are masked in this process). The green shaded area represents the local value of $\pm 3\sigma(\lambda)$ added to $F(\lambda)$. In the inset, we magnify the spectral region surrounding the Hα emission. Some raw line ratios (i.e., not corrected for the biases explained in Section 5.4.2), together with their associated 3σ uncertainties, are indicated in the plot. These ratios are computed considering the Gaussian fits to the detected lines.

In order to measure this selection bias, for each of the 148 stacked single-line spectra, we simulated 21 spectra characterized by a similar noise distribution along λ. Then, we added simulated emission lines to these pure-noise spectra. We place all the simulated sources at $z = 1.0$ (i.e., the average redshift of

the stacked single-line sample). For each spectrum, we set the flux of the strongest emission line ([O III]) so that S/N ∼ 7.67, corresponding to the typical S/N characterizing the single-line spectra. The width $\Delta\lambda(\lambda)$ of the simulated lines is also set in accordance with the median size of the single lines. The flux of
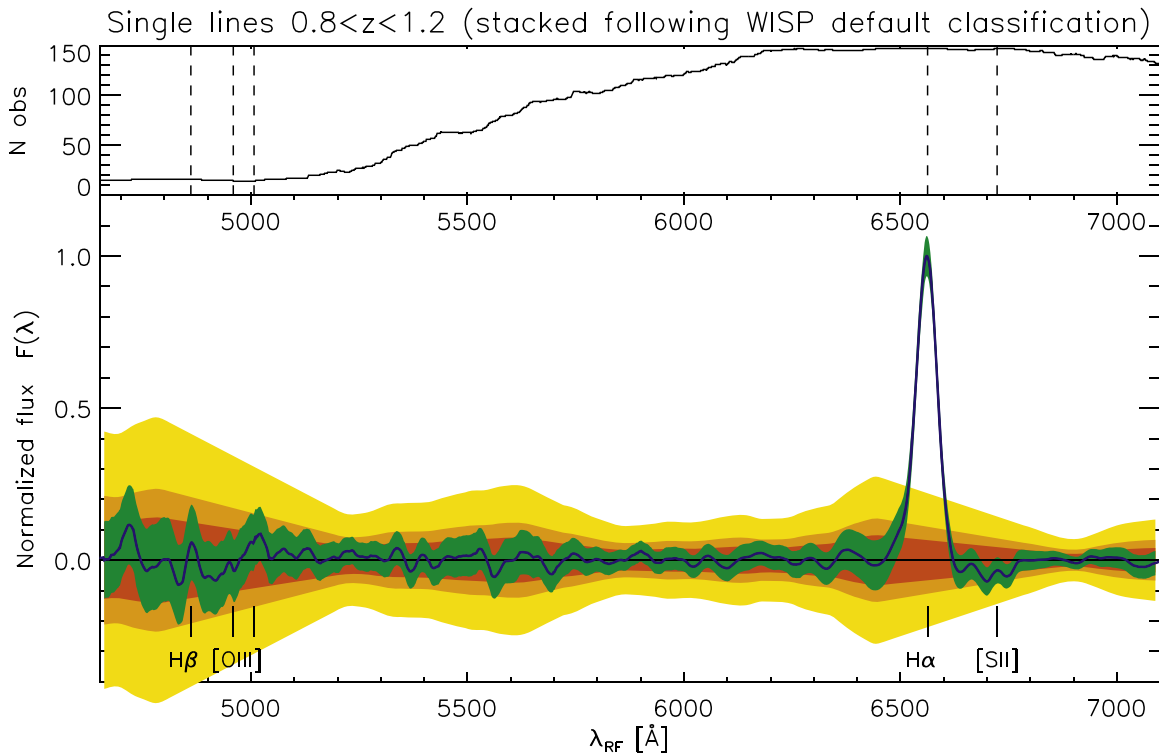
Figure 22. Spectrum obtained by stacking the same "single-line" sources used for Figure 21. In this case, the WISP default classification is assumed. The same fine recentering strategy is used (see Appendix B). In this case, [S II] is not even visible, whereas the same line is detected above 3σ if we follow the classification performed by our algorithm. Instead, following the WISP default classification, Hβ and [O III] would fall outside the wavelength range covered by the grism most of the times, so that a comparison between the two classifications is not possible. This comparison indicates that our algorithm performs better than the WISP default classification. For a complete description of the features visible in these plots, the reader can refer to Figure 21.

all the other lines is set in accordance with the expected line ratios (second column in Table 4) that we measure from the gold sample, at similar magnitudes. Then, from the simulated sample (3108 spectra in total), we selected all of the spectra characterized by $S/N(H\alpha) < 2$, corresponding to the WISP detection threshold.[23]

We repeated the simulation described 100 times. On average, in each of the 100 simulations, 196 ± 6 spectra (6.3%) were selected using the criterion described. For an approximate comparison, the 323 real single lines, located in the combined wavelength range $8500\,\text{Å} < \lambda < 11100\,\text{Å}$ and $11400\,\text{Å} < \lambda < 16700\,\text{Å}$, represent 6.1% of the overall sample of spectra with strongest line detected in the same range (5327 sources). Finally, we stacked the selected simulated spectra by replicating the procedure used for the real ones (we did not apply the recentering procedure because all simulated spectra were already placed at $z = 1.0$). In Table 4 (third column), we report the average flux ratios measured from the 100 simulated stacked spectra.

The flux ratios are affected by an additional bias (bias 2), due to the intrinsic accuracy of the algorithm (see Table 3). Every time the algorithm mistakes Hα for [O III] (or vice versa), this modifies the flux ratios in the stacked spectrum. In fact, while the strongest emission is incorrectly attributed to [O III] (or Hα), there is no contribution to the stacked spectrum from all the other lines.

In order to estimate this bias, we run 100 simulations similar to those described above (for bias1). In this case, we simulated the incorrect identifications by removing all the emission lines from some of the selected spectra. Then, we added only one simulated line (Hα or [O III]), which we set to $S/N = 7.67$. The amount of false identifications is set considering the expected purity of the Hα and [O III] samples (last column in Table 3), and in accordance with the fraction of actual [O III] and Hα identifications in the (real) stacked single-line sample. The combined effects of selection bias and accuracy bias are reported in the fourth column of Table 4.

The third bias is due to the presence of false identifications (see Section 5.3). While the algorithm is not trained to identify false spectral lines, we can simulate their effect on the recovered flux ratios. Similar to bias 2, false identifications are characterized by the absence of additional lines aside from the strongest one. The value of S/N for the false identifications is assumed similar to that characterizing the rest of the single lines (false identifications with peculiar S/N were easily identified and eliminated during the WISP original by-eye classification). For the reasons explained in Section 5.3, we expect almost all the false identifications in this sample to be classified as [O III] by the algorithm.

In Table 4, we report the flux ratios expected from the stacked spectrum by considering the combined effects of bias 1, bias 2, and 30% or 35% contamination from false detections. These values are obtained by running 50 additional simulations for each of the two levels of contamination. The [O III]/Hα ratio is more consistent with a 30% contamination, while [O III]/Hβ agrees better with a 35% contamination. Due to the

---

[23] This criterion should provide a conservative estimate of the actual bias, as some of these lines are visible and measurable (although not detected). Similar spectra would not be included in the real single-line sample.

Table 4
Some Flux Ratios for the Main Emission Lines ($0.8 < z < 1.2$)

| Line Ratio | Expected[a] at $J = 24.5$ | | | | | Measured[b] (Stacking) |
	Unbiased (Gold Sample)[c]	With Bias 1[d]	With Biases 1 and 2[e]	+30% f.d.[f]	+35% f.d.[f]	
$\log([O\ III]/H\alpha)$	0.30	0.56 ± 0.02	0.59 ± 0.02	0.79 ± 0.03	0.85 ± 0.03	0.81 ± 0.15
$\log([O\ III]/H\beta)$	0.79	0.77 ± 0.04	0.84 ± 0.04	0.98 ± 0.06	1.06 ± 0.18	1.07 ± 0.23
$\log(H\beta/H\alpha)$	-0.49	-0.21 ± 0.03	-0.25 ± 0.04	-0.22 ± 0.07	-0.23 ± 0.09	-0.27 ± 0.31
$\log([S\ II]/H\alpha)$	-0.79	-0.49 ± 0.06	-0.52 ± 0.08	-0.57 ± 0.15	-0.75 ± 0.23	-0.34 ± 0.37

Notes.
[a] The associated errors correspond to the dispersion of the values obtained in the different simulations.
[b] The associated errors correspond to the actual uncertainties measured from the stacking.
[c] The unbiased flux ratios expected for the stacked single-line sample are measured from the gold sample, at apparent $J \sim 24.5$. These values represent the input of our simulation.
[d] The spectra in the single-line sample are specifically selected for not showing any additional line besides the brightest one. This particular selection biases the single-line sample (bias 1). At $J \sim 24.5$, in the redshift range $0.8 < z < 1.2$, most of the single lines are identified as [O III]. Therefore, only spectra with a particularly damped Hα can be included in this sample.
[e] The accuracy of the algorithm (Table 3) affects the recovered line ratios (bias 2). When a single line is mistaken for another, the other lines in the same spectrum cannot contribute to the stacked spectrum.
[f] False detections (f.d.) modify the line ratios in the stacked spectrum (bias 3). When a detected line is not due to a real emission, there are no other lines in the same spectrum that can contribute to the stacked spectrum. Almost all the f.d. are classified as [O III] by the algorithm. In this table, we report the results of the simulations for two different levels of contamination from f.d. (30% and 35%). Biases 1 and 2 are taken into account.

Table 5
HST-WISP, Euclid, and WFIRST Grism Main Parameters

| | WISP (HST/WFC3) | | Euclid | | WFIRST |
	G102	G141	Blue	Red	
λ range (nm)	800–1150	1075–1700	920–1250	1250–1850	1000–1930
$\lambda/\Delta\lambda$	210 (at 1000 nm)	130 (at 1400 nm)	380		435–865

higher uncertainty associated with the other flux ratios, these indicators are consistent with both possibilities.

As previously explained, bias 1 could be underestimated due to the conservative criteria adopted in the simulations. Additionally, the accuracy estimated on the gold sample could degrade when classifying single lines. In both these cases, the fraction of false detections that we estimate would represent an upper limit to the actual level of contamination.

6. Discussion

### 6.1. Analogies with the Future Euclid and WFIRST Surveys

The primary purpose of the algorithm that we described in this paper is the automatic classification of the strongest emission lines detected in WISP spectra. However, the modular structure of the algorithm is specifically designed to allow for easy addition, replacement, and removal of each single module (and block). Hence, the algorithm can be easily adapted to different surveys. Given the similarities between WISP and the Euclid and WFIRST surveys, the experiment that we performed represents an important pilot study in the context of these future missions, in particular for what concerns the maximization of their scientific return. Euclid and WFIRST will probe the nature of dark energy by carrying out spectroscopic surveys over an unprecedented sky area (Laureijs et al. 2011, 2012; Green et al. 2012; Spergel et al. 2015), using the Hα and [O III] emission lines as tracers of the large-scale structure over a wide range of redshifts ($0.5 \lesssim z \lesssim 2.9$). In this

context, the correct identification of such lines is going to be crucial.

The WISP survey represents one of the most important proxies for the future Euclid and WFIRST missions. The similar spectral range covered by the grisms of the HST/WFC3, Euclid/NISP, and WFIRST is shown in Table 5 and Figure 23. The presence of two grisms with similar bandwidth (blue and red) makes Euclid/NISP even more similar to HST/WFC3.

Nevertheless, Euclid and WFIRST will probably benefit from a larger amount of ancillary data than WISP. Compared with our analysis, the photometric redshifts (z-PDF) obtained from the SED-fitting approach considered singularly will allow for a more accurate identification of spectral lines, with a smaller fraction of catastrophic failures ($\sim$5%–10% catastrophic failures, $\sigma_z/(1 + z) \leqslant 0.03$–0.05; Laureijs et al. 2011). However, as described in Section 3.2.1, our algorithm does integrate the SED-fitting approach with additional and independent sources of information. Consequently, whatever the actual accuracy obtained using all of the ancillary data will be, we expect a reduced fraction of catastrophic failures when applying an algorithm similar to the one described in this paper, if compared with the SED-fitting strategy used alone.

Our algorithm is not trained to identify false detections (see Section 5.3). In WISP, we estimate these contaminants to represent $\lesssim$35% of the single-line sample. Most of this contamination is due to zero-order spectra corresponding to sources located just outside the WISP fields. Both Euclid and
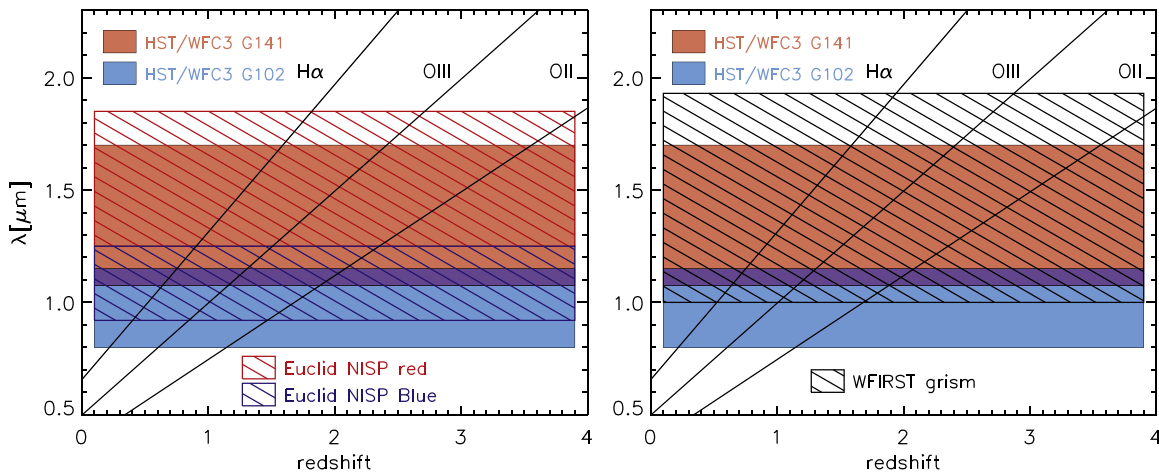
Figure 23. Comparison between the spectral coverages of WISP (HST/WFC3), Euclid NISP, and WFIRST. The wavelength positions of the main spectral lines considered in this paper (Hα, [O III], and [O II]) are represented as a function of the redshift, using thick black lines.

WFIRST will strongly limit such type of contamination by dispersing spectra along different directions.

One of the two primary goals of Euclid is to study the accelerated expansion of the universe through the analysis of baryonic acoustic oscillations.[24] To this purpose, the spectroscopic redshift precision required on single sources is $\sigma_{\Delta z} \leqslant 0.001(1 + z)$. Instead, the constraint on the fraction of catastrophic failures is quite large: $f_c < 20\%$ (Laureijs et al. 2011). However, the uncertainty associated with f_c (σ_{f_c}) must be known to 1%. A more extensive discussion on the effects of line misidentification, for both Euclid and WFIRST, can be found in Addison et al. (2019). With our analysis, we demonstrate that the goal of limiting the fraction of catastrophic failures (f_c) below 20% can be achieved, even considering the poor photometric coverage characterizing the WISP survey. On the other hand, in order to estimate f_c with the precision required, a larger calibration sample would be needed.

Besides WISP, there are currently two other spectroscopic surveys that represent important proxies for the Euclid and WFIRST missions. The 3D-HST (P.I. P. van Dokkum; Brammer et al. 2012; Skelton et al. 2014; Momcheva et al. 2016) and the AGHAST surveys (P.I. B. Weiner; Weiner 2009) obtained spectroscopic observations, through the WFC3/G141 grism, of about 150 pointings located in the CANDELS (Grogin et al. 2011; Koekemoer et al. 2011) fields.

On the one hand, the pointings of the 3D-HST and AGHAST surveys are covered, with some exceptions, only by the G141 grism and by the *H*-band filter (F140W), whereas WISP uses both the G102 and G141 grisms and both the *J* and *H* filters (F110W and F140W/F160W). On the other hand, WISP is a pure-parallel survey, meaning that the pointings are usually randomly located in the sky, while 3D-HST and AGHAST can exploit the richness of ancillary data already available in these major cosmological fields.

For a comparison, the Euclid surveys will use observations in the *Y*, *J*, and *H* bands of the NISP instrument plus the photometric coverage of VIS, between $\lambda \sim 5500\,\text{Å}$ and $\lambda \sim 9000\,\text{Å}$. Additional ground-based observations will be obtained in the *g*, *r*, and *i* bands from ground-based telescopes. Therefore, the SED of the sources covered by the future Euclid

surveys will not be as continuously sampled as those in the CANDELS fields.

Another point favoring WISP as a proxy for the WFIRST and Euclid missions is the total sky area covered, which is more than three times wider that that explored by 3D-HST and AGHAST combined (at the moment, the available WISP data covers a total area of $\sim$1520 arcmin^2). It is worth noting, in any case, that given the analogies among these three surveys, they can be combined to obtain even more accurate forecasts, as has recently been done in Bagley et al. (2020).

Our tests (see Table 3 for a summary) show that, using the current strategy, it is possible to obtain acceptable results concerning the identification of the Hα and [O III] emission lines in WISP. Most of the algorithm's capability is based on the assumption that the strongest emission line in a galaxy spectrum (among the lines considered) falls in the wavelength range covered by the two grisms combined. For the WISP data set, this assumption represents a good approximation of the real situation. However, for a minor part of the WISP data, this assumption is not correct. In these cases, we adopt the less precise strategy described in Section 3.2.7. We highlight the fact that the acceptability of this important assumption depends on the wavelength range covered by the combined grisms. Hence, for the Euclid case, the precision achieved using the methods described could be undermined by the use of only one of the two grisms (as it is planned for the Euclid wide survey).

The accuracy of the algorithm is low when we consider the [O II] sample. This outcome is mostly due to the impossibility of precisely calibrating the algorithm by using the very small number of sources in the "gold" sample ($N_{\text{TGS}}^{[\text{O II}]} = 24$) characterized by an [O II] emission stronger than that of Hα and [O III]. Besides the inaccurate calibration, given the uneven distribution of sources in the Hα, [O II], and [O III] samples, a small fraction of misidentified Hα and [O III] lines can significantly contaminate the much smaller [O II] sample. As a consequence, in order to use similar algorithms in the context of the future Euclid and WFIRST missions, an initial phase of careful characterization of the rarest objects should be considered.

6.2. Possible Improvements and Future Perspectives

The algorithm we discussed can be improved as follows. On the one hand, additional information could be considered by

---

[24] Euclid is optimized also to study the dark matter distribution in the universe through weak gravitational lensing (WL).

the algorithm. For example, we could include the spectral continuum (for the brightest sources) and the rich sets of ancillary photometric and spectroscopic data that are (and that are going to be) available in the fields that will be covered by Euclid and WFIRST. The study of the χ^2 obtained by fitting different spectral models (at different z) could in principle be used as an additional source of information (we propose an alternative to this method in Appendix B). On the other hand, the algorithm could also be improved by taking advantage of the symbiotic use of different strategies, such as supervised and unsupervised machine learning.

Finally, the wide area covered by Euclid and WFIRST will allow us to better characterize the rarest objects (such as those in the [O II] sample described above) from the very initial phases of the surveys, making it possible to limit the contamination affecting these samples of sources. An additional important possibility, which was not explored in this paper, is the ability to quantify their probability, for the single lines identified, of being real emissions or spurious detections.

Even without any further improvement to the actual accuracy (∼82.6%), determining the nature of most of the single emission lines is crucial for many reasons. For example, it makes it possible to determine accurate redshifts of the faintest sources (i.e., the most difficult to explore), or to improve the definition of the faint end of the Hα and [O III] luminosity functions, eliminating particularly pernicious biases on some selected data samples.

In this paper, we run our software on a sample of spectra where only one spectral line is detected. However, this algorithm can also be useful in the more general case, when multiple lines are (barely) visible, or even detected, but there is no consistency among the classification of different reviewers.

7. Summary and Conclusions

We presented an algorithm that can be used to identify the strongest emission lines detected in near-IR grism spectra. The algorithm exploits ancillary information that is usually not considered when determining the spectroscopic redshift. For the algorithm described in this work, we considered low-precision photometric redshifts from SED fitting, priors based on the apparent magnitude, size, EW, color indices, and expected line flux ratios. The classification is optimized by including empirical a posteriori corrections based on the wavelength position of the detections.

Such an approach is very useful to identify a spectral line when it is the only emission detected in a spectrum, especially when the precision of the photometric redshifts from SED fitting is low. However, the same approach can also be helpful in the general case, when the identification of multiple emission

lines is difficult for other reasons (such as contamination, low S/N, etc.).

The WISP survey represents one of the most relevant proxies for the future Euclid and WFIRST missions, especially in terms of wavelength range covered, depth reached, and availability of near-IR photometric bands. In this context, our work is intended as a pilot study toward the maximization of the scientific return of these future surveys.

The approach that we adopted allows us to identify [O III] and Hα emission lines in WISP with good accuracy (∼82.6% of correct identifications on real lines). The ancillary data that will be available for Euclid and WFIRST, together with their observing strategies, should significantly improve this result.

As we illustrated, more or less pure samples (corresponding to less or more complete samples, respectively) can be obtained by using specific probability indicators outputted by the algorithm. However, the accuracy of our algorithm is still low for the rarest objects ([O II]). This problem is mostly due to the limited number of sources that can be used to calibrate the algorithm. In this sense, the large data sets available from the future Euclid and WFIRST surveys should sensibly improve the performance of similar algorithms.

We also expect further improvements of the accuracy from the exploitation of additional observational quantities and independent techniques that we do not consider in this pilot study, such as the combined use of different machine-learning approaches, including unsupervised machine-learning strategies, the analysis of the spectral continuum, the comparison of alternative models fitted to the spectra (χ^2 analysis), or the use of the "contrast factor" that we describe in Appendix B.

Appendix A
Alternative Accuracy Tests

Similarly to Table 3, where the accuracy is computed using the same sample (with few differences) to both calibrate and

Table A1
Alternative Accuracy Test 1[a]

Species/ Transitions	λ_{obs} (Å)	N_{TGS}	N_I	N_{CI}	N_{WI}	Completeness (N_{CI}/N_{TGS})	Contamination (N_{WI}/N_I)	Accuracy (Purity) (N_{CI}/N_I)
Hα	6564.5	639	643	543	100	85.0%	15.6%	84.4%
[O III]	4960.3–5008.2	461	453	354	99	76.8%	21.9%	78.1%
[O II]	3727.1–3729.9	11	19	3	16	27.3%	84.2%	15.8%

Note.
[a] This test (to be compared with Table 3) is performed by dividing the "gold" sample into two independent subsets, using one to calibrate the algorithm and the other to test it.

Table A2
Alternative Accuracy Test 2[a]

Species/ Transitions	λ_{obs} (Å)	N_{TGS}	N_I	N_{CI}	N_{WI}	Completeness (N_{CI}/N_{TGS})	Contamination (N_{WI}/N_I)	Accuracy (Purity) (N_{CI}/N_I)
Hα	6564.5	654	679	565	114	86.4%	16.8%	83.2%
[O III]	4960.3–5008.2	488	453	366	87	75.0%	19.2%	80.8%
[O II]	3727.1–3729.9	13	20	6	14	46.2%	70%	30.0%

Note.

[a] This test (to be compared with Table 3) is performed by dividing the "gold" sample into two independent subsets, using one to calibrate the algorithm and the other to test it. In this second test, the calibration and test samples are inverses of each other, with respect to the first test (Table A1).

test the algorithm, in Tables A1 and A2 we report the accuracy measured by dividing the "gold" sample in two independent subsamples. In the first test (Table A1), we use one of the two subsamples to calibrate the algorithm and the other to assess the accuracy. For the second test (Table A2), we invert the two subsamples. In both cases, all of the relations, including the a posteriori empirical correction (Section 3.2.6), are recomputed considering only the calibration data set. The overall accuracy that we obtain considering all of the species/ transitions is 80.6% and 80.4% for the first and the second tests, respectively. In Tables A1 and A2, we report only the accuracy obtained for the most relevant transitions (Hα, [O III], and [O II]). As explained in Section 4.1, we consider the accuracy obtained with these two tests to be a lower limit to the actual accuracy of the algorithm. For more reliable estimates, we refer the reader to Table 3.

Appendix B
Stacking Technique and Automatic Spectra Recentering

In Section 5.4, we describe the result obtained by stacking the spectra of the "single-line" sources in the range $0.8 < z < 1.2$. Here, we describe how the original spectra are treated and then combined to obtain the final result described.

The three panels of Figure B1 show the steps of the processing that precede the stacking phase. Initially, the contamination is removed from the original spectra.[25] Then, we estimate the continuum at each λ_i (central red curve in the top panel of Figure B1). For this purpose, we linearly fit the observed flux measured in the wavelength intervals immediately preceding ($\Delta\lambda_i^-$) and following ($\Delta\lambda_i^+$) the wavelength considered. We extrapolate the values of the continuum expected at λ_i from both sides, considering their average value. In this process, we mask both the main spectral lines and the noisy central wavelength range, where the sensitivity of the two grisms is lower. After this initial computation, the continuum is further smoothed to obtain a result less dependent on the small-scale variations.

After the continuum subtraction, the local level of noise $\sigma_F(\lambda_i)$ is obtained as the standard deviation of the fluxes $F(\lambda)$ measured in the same λ intervals considered when computing the continuum ($\Delta\lambda_i^-$ and $\Delta\lambda_i^+$). Also, the noise level is subsequently smoothed. Values of $F(\lambda)$ exceeding a local 6σ threshold (highest and lowest red curves in the two upper panels of Figure B1) are masked and not used in the successive steps (unless when the excess is due to the presence of an actual spectral line).

In order to precisely align all of the single spectra to the common rest frame, every "single-line" spectrum is automatically finely recentered following the line identification performed by the algorithm. In principle, the recentering could be automatically performed by fitting Gaussian functions to all of the main emission lines and using the weighted average of the peaks' positions (in λ) to precisely estimate the redshift. This approach works well when the spectral lines considered are detected well above the level of the noise. In practice, the same solution is not well suited for determining the position of weak lines ($F \lesssim \sigma_F$). In these cases, the result is that the alignment to a common rest frame relies only on the (poor) fit of the strongest emission line, detected just above the noise threshold, while the fit to the other lines is completely unreliable and often fails.

To overcome the problems of the technique described, we elaborated a wavelength-dependent contrast factor (CF), which we compute for each of the main emission lines. We use this function to compute the most probable position of each spectral line. For example, $CF_{H\alpha}(\lambda_I)$ quantifies the relative probability associated with the Hα emission being centered at $\lambda_{obs}^{H\alpha} = \lambda_I$. This can be immediately translated into the relative probability for the source to be located at $z = \lambda_I/\lambda_{int}^{H\alpha} - 1$. The sum of the contrast factors computed for each spectral line, weighted for the local noise, can be used to determine the most probable redshift of the source. The bottom panel of Figure B1 shows the behavior of $CF(z, \sigma_\lambda)$ for each of the main lines in the visibility range, and the (not-normalized) PDF resulting from their sum.

Figure B2 schematically describes how the contrast factor is computed for the O III 5007 line that is visible in the spectrum already shown in Figure B1, but the same description can be extended to all emission lines considered. We initially compute two sets of normalized Gaussian functions $G(\lambda)$. In the first set, the value of σ is similar to the width of the emission lines (as expected from the size of each source), while in the second set σ corresponds to $1/3$ of this value. Each of the two sets is made up of three functions, $G_{\lambda_i}^0$, $G_{\lambda_i}^-$, and $G_{\lambda_i}^+$, characterized by identical shapes and normalizations, but centered at λ_i (the wavelength considered) and at $\lambda_i \pm 2\sigma$. Every function $G(\lambda)$ is multiplied, at each λ, for the spectrum $F(\lambda)$. In our computation of CF, we consider the integrals of the six Gaussian functions, normalized for the local level of noise $\sigma_F(\lambda)$ measured from the spectrum. For each of the two sets, we have

$$A(\lambda_i) = \int_0^\infty G_{\lambda_i}^0(\lambda) \frac{F(\lambda)}{\sigma_F(\lambda)} \, d\lambda \qquad (B1)$$

$$B(\lambda_i) = \int_0^\infty G_{\lambda_i}^-(\lambda) \frac{F(\lambda)}{\sigma_F(\lambda)} \, d\lambda \qquad (B2)$$

---

[25] The contamination as a function of λ is obtained directly from the original WISP data.
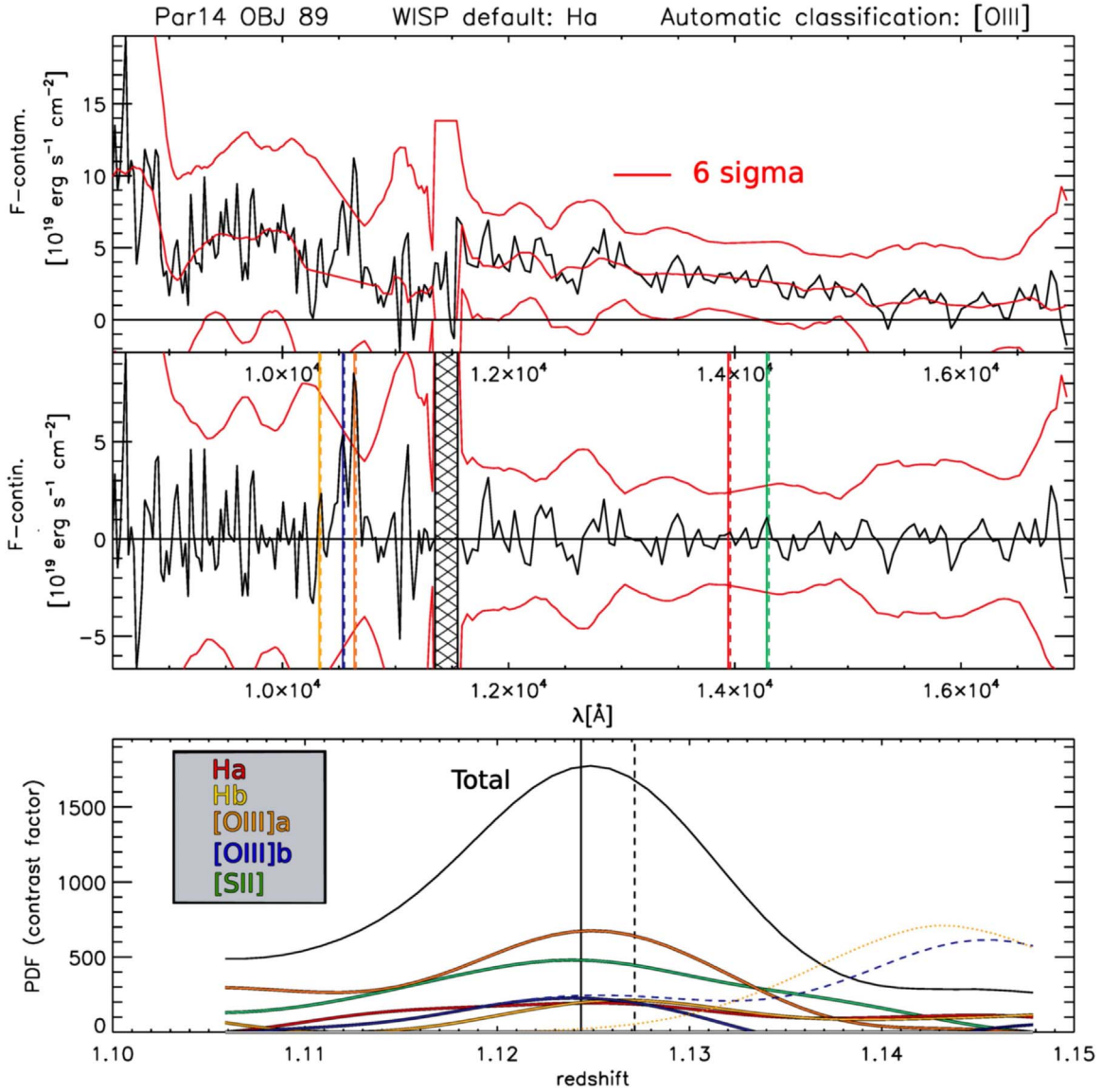
Figure B1. Example of "single-line" spectrum used to compute the stacked spectrum described in Section 5.4 and shown in Figure 21. The contamination-subtracted spectrum is shown in the top panel. We considered locally computed $\pm 6\sigma$ deviations (upper and lower red curves) from the median (central red curve) to exclude unrealistic values, unless when the flux excess was located in close proximity to expected spectral features. The continuum-subtracted spectrum is shown in the middle panel. The position of the only line detected (in this example, it is the only line exceeding our 6σ threshold) is initially obtained through a simple Gaussian fit. Following the identification performed by the algorithm, the automatic recentering of the spectrum is obtained using a "contrast factor" FC (bottom panel), which we compute for each of the lines expected in the range of visibility (see text and Figure B2 for more details). In both the middle and bottom panels, dashed lines indicate the position of the spectral lines determined using a simple Gaussian fit. The continuous vertical lines indicate the positions after the automatic recentering process (in order of increasing λ, Hβ in yellow, O III 4959 in blue, O III 5007 in orange, Hα in red, and [S II] in green).

$$C(\lambda_i) = \int_0^\infty G^+_{\lambda_i}(\lambda) \frac{F(\lambda)}{\sigma_F(\lambda)} \, d\lambda. \qquad (B3)$$

At this point, the six integrals (A_1 and A_2 corresponding with the central Gaussians, B_1 and B_2 corresponding with the Gaussians centered at -2σ, C_1 and C_2 corresponding with the Gaussians centered at $+2\sigma$) are combined to compute the contrast factor:

$$\mathrm{CF}(\lambda_i) = A_1(\lambda_i) - \frac{1}{2}[B_1(\lambda_i) + C_1(\lambda_i)]$$
$$+ A_2(\lambda_i) - \frac{1}{2}[B_2(\lambda_i) + C_2(\lambda_i)]. \qquad (B4)$$

These contrast factors, computed for each spectral line and for each value of λ_i (i.e., for each value of z_i), are finally summed

together, obtaining a unique function $\mathrm{CF}_{\mathrm{TOT}}(z)$. The spectrum is recentered, considering the value of z that maximizes the function $\mathrm{CF}_{\mathrm{TOT}}(z)$.

The contrast factor described works essentially as a combination of two bandpass filters. The band of the first filter (i.e., the σ of the first three Gaussian functions) is tuned on the line width expected from the observed size. This approach allows us to mitigate the effects of the noise at different spatial (λ) scales, focusing on the possible presence of flux excesses at these scales. An additional filter (the other three Gaussians) allows us to refine the estimation of the central position of the line by considering the flux peaks, observable at higher spatial frequencies.

In our analysis, we use the contrast factor described only to recenter the spectra. However, we notice that in future analyses the same approach could be used to improve the accuracy of
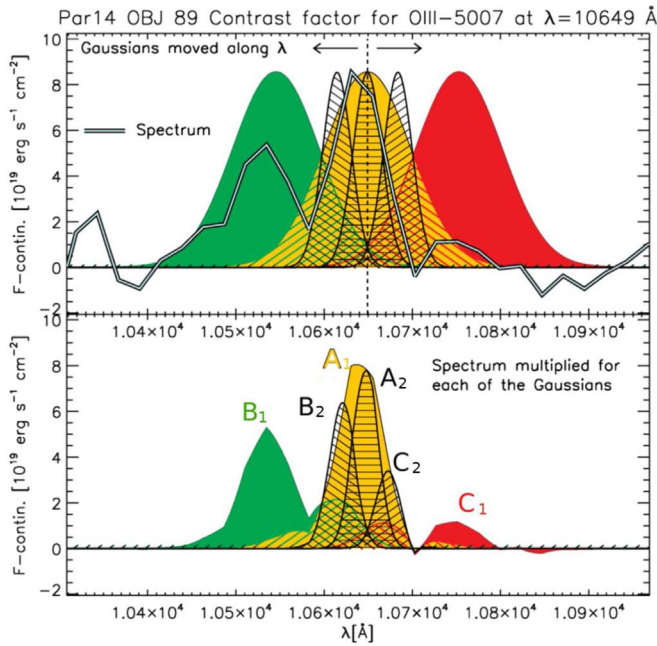
Figure B2. Before the stacking, every spectrum is finely recentered to the common rest frame using a contrast factor $FC(\lambda)$ that we compute for each of the main emission lines. While a simple Gaussian fit is sensitive to the noise at all scales, our contrast factor acts as a combination of two bandpass filters calibrated on the expected width of the spectral line. This approach guarantees the limited influence of the noise at all other scales. The top panel shows the same spectrum of Figure B1, corresponding to the O III 5007 emission (following the identification performed by our algorithm). Two sets of Gaussians functions, each of which is characterized by a specific value of σ, are shown in the same plot. In both sets, three Gaussians are centered at λ_i and at $\lambda_i \pm 2\sigma$, where λ_i is the value of the wavelength considered during one generic step of the process (at each step, the Gaussian functions are shifted along λ, as explained below). For the three Gaussian functions in the first set (green, yellow and red areas), σ is similar to the width of the emission line as expected from the apparent size of the source. In the second set (black shaded areas), σ corresponds to $1/3$ of the same value. At every λ, the spectrum $F(\lambda)$ is multiplied for the value assumed at that wavelength by each of the Gaussian functions (bottom panel). For a given spectral line, and at each wavelength λ_i, the contrast factor $FC(\lambda_i)$ is obtained by combining the integral of these functions as described in Equation (B4). Then, all of the Gaussian functions are coherently shifted and the process is repeated in order to obtain the value of $FC(\lambda)$ at different wavelengths. A similar contrast factor is computed for all the spectral lines considered and the precise redshift of the source is obtained as the value of z maximizing the weighted sum of the resulting functions ($FC_{\mathrm{TOT}}(z)$; black curve in the bottom panel of Figure B1).

the line identification process. In fact, the contrast factor is to all effects a redshift PDF that could be integrated into our computations.[26]

After recentering each single spectrum, we compute, at each λ, the weighted average of $F(\lambda)$, normalized for the total flux of the detected emission line. As a weight, we use the local value of $\sigma_F(\lambda)$. The mean is computed after excluding all values of $F(\lambda)$ exceeding the 16th and 84th percentiles from the median value ($\pm 1\sigma$).

ORCID iDs

I. Baronchelli ⓘ https://orcid.org/0000-0003-0556-2929
C. M. Scarlata ⓘ https://orcid.org/0000-0002-9136-8876
G. Rodighiero ⓘ https://orcid.org/0000-0002-9415-2296
L. Rodríguez-Muñoz ⓘ https://orcid.org/0000-0002-0192-5131
M. Bonato ⓘ https://orcid.org/0000-0001-9139-2342
M. Bagley ⓘ https://orcid.org/0000-0002-9921-9218
A. Henry ⓘ https://orcid.org/0000-0002-6586-4446
M. Rafelski ⓘ https://orcid.org/0000-0002-9946-4731
M. Malkan ⓘ https://orcid.org/0000-0001-6919-1237
Y. S. Dai (戴昱) ⓘ https://orcid.org/0000-0002-7928-416X
H. Dickinson ⓘ https://orcid.org/0000-0003-0475-008X
C. Mancini ⓘ https://orcid.org/0000-0002-4297-0561
V. Mehta ⓘ https://orcid.org/0000-0001-7166-6035
H. I. Teplitz ⓘ https://orcid.org/0000-0002-7064-5424

References

Addison, G. E., Bennett, C. L., Jeong, D., Komatsu, E., & Weiland, J. L. 2019, ApJ, 879, 15
Atek, H., Malkan, M., McCarthy, P., et al. 2010, ApJ, 723, 104
Bagley, M., Scarlata, C., & Mehta, V. 2020, arXiv:2006.03602
Bagley, M. B., Scarlata, C., Henry, A., et al. 2017, ApJ, 837, 11
Baronchelli, I., Rodighiero, G., Teplitz, H. I., et al. 2018, ApJ, 857, 64
Bolzonella, M., Miralles, J., & Pelló, R. 2000, A&A, 363, 476
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brammer, G. B., van Dokkum, P. G., Franx, M., et al. 2012, ApJS, 200, 13
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Clements, D. L., & Couch, W. J. 1996, MNRAS, 280, L43
Dickinson, M. 1998, in The Hubble Deep Field, ed. M. Livio, S. M. Fall, & P. Madau (New York: Cambridge Univ. Press), 219
Green, J., Schechter, P., Baltay, C., et al. 2011, arXiv:1108.1374
Green, J., Schechter, P., Baltay, C., et al. 2012, arXiv:1208.4012
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
Hasinger, G., Capak, P., Salvato, M., et al. 2018, ApJ, 858, 77
Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36
Lapi, A., Mancuso, C., Bressan, A., & Danese, L. 2017, ApJ, 847, 13
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Laureijs, R., Gondoin, P., Duvet, L., et al. 2012, Proc. SPIE, 8442, 84420T
Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845
Madau, P., Ferguson, H. C., Dickinson, M. E., et al. 1996, MNRAS, 283, 1388
Mattig, W. 1958, AN, 284, 109
Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, ApJS, 225, 27
Poggianti, B. M., Calvi, R., Bindoni, D., et al. 2013, ApJ, 762, 77
Rodríguez-Muñoz, L., Rodighiero, G., Mancini, C., et al. 2019, MNRAS, 485, 586
Salvato, M., Ilbert, O., & Hoyle, B. 2018, NatAs, 3, 212
Schechter, P. 1976, ApJ, 203, 297
Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, ApJS, 214, 24
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757
Stevens, R., & Lacy, M. 2001, MNRAS, 325, 897
van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014, ApJ, 788, 28
Weiner, B. 2009, Star formation, extinction and metallicity at 0.7<z<1.5: H-alpha fluxes and sizes from a grism survey of GOODS-N HST Proposal ID 11600, (Baltimore, MD: STScI)
Wittman, D., Bhaskar, R., & Tobin, R. 2016, MNRAS, 457, 4005

---

[26] The same can be said for the χ^2 resulting from the best-fitting Gaussian functions, but with the limitations previously described, concerning line fluxes are at the level of the noise.