



Publication Year	2020
Acceptance in OA	2022-01-24T10:43:45Z
Title	Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST
Authors	BARONCHELLI, IVANO, Scarlata, C. M., Rodighiero, G., Rodríguez-Muñoz, L., BONATO, MATTEO, Bagley, M., Henry, A., Rafelski, M., Malkan, M., Colbert, J., Dai, Y. S., Dickinson, H., MANCINI, CHIARA, Mehta, V., Morselli, L., Teplitz, H. I.
Publisher's version (DOI)	10.3847/1538-4365/ab9a3a
Handle	http://hdl.handle.net/20.500.12386/31346
Journal	THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES
Volume	249

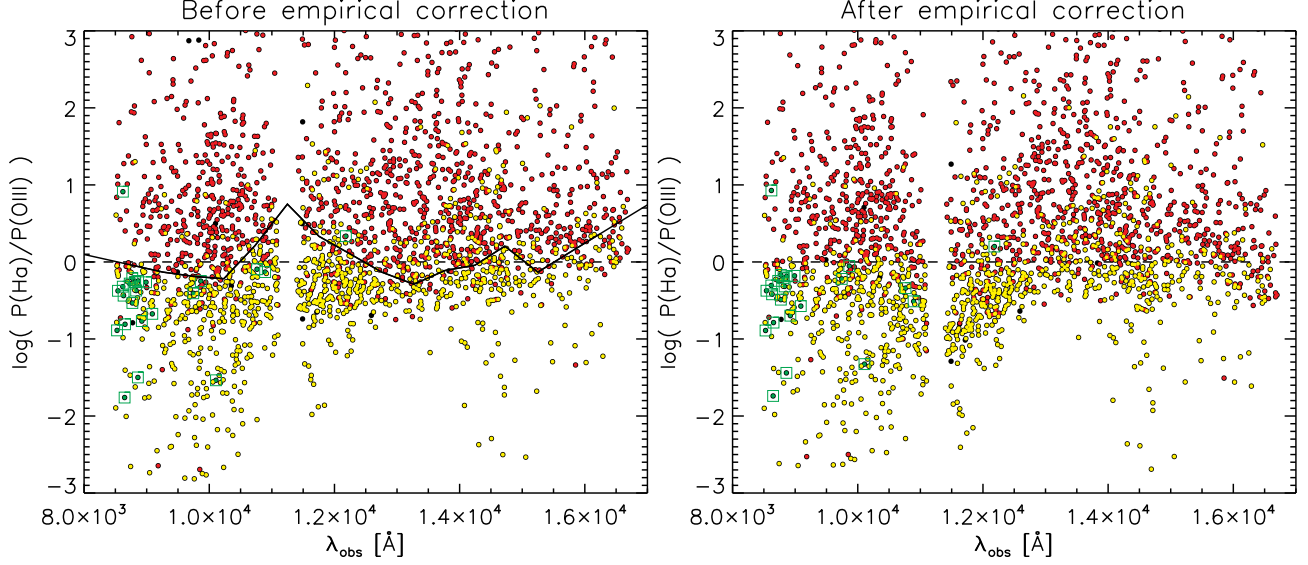


Figure 11. Left panel: Measured $P(\text{H}\alpha)/P([\text{O III}])$ for the sources in the test “gold” sample, before applying the λ_{obs} -based fine-tuning correction ($P_{\text{H}\alpha/[\text{O III}]}^{\text{best}}(\lambda_{\text{obs}})$, black curve, Table 1). The identification of the strongest emission line, detected at λ_{obs} , is represented with yellow circles for [O III], red circles for H α and green circles surrounded by squares for [O II]. **Right panel:** values of $P(\text{H}\alpha)/P([\text{O III}])$ after the application of the $P_{\text{H}\alpha/[\text{O III}]}^{\text{best}}(\lambda_{\text{obs}})$ correction.

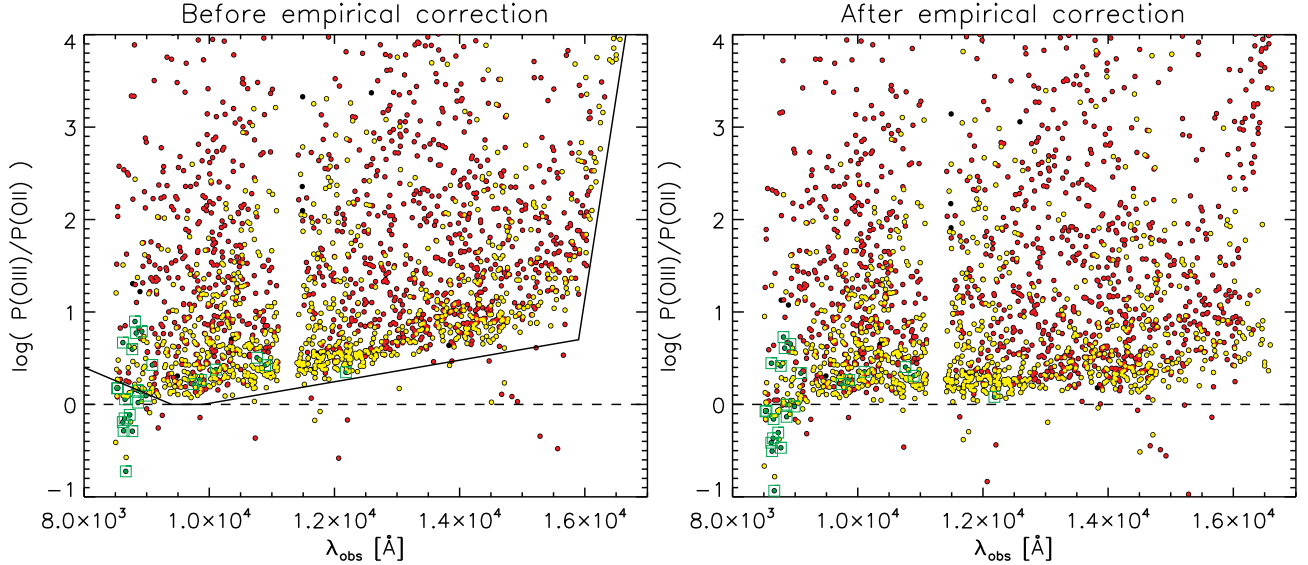


Figure 12. Left panel: Measured $P([\text{O III}])/P([\text{O II}])$ ratio for the sources in the test “gold” sample, before applying the λ_{obs} -based fine-tuning correction ($P_{[\text{O III}]/[\text{O II}]}^{\text{best}}(\lambda_{\text{obs}})$, black curve, Table 2). The identification of the strongest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

lines correctly recovered by the algorithm) of 86.0, that 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 the 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 Figures 6.

3.3. Operation description

Given the strongest line measured in a spectrum (sometimes the only detected line), the main output of the algorithm

Table 1. Fine-tuning of the $P(\text{H}\alpha)/P([\text{O III}])$ ratio

$\langle \lambda_{\text{obs}} [\text{\AA}] \rangle$	$\log P_{\text{H}\alpha}/[\text{O III}]^{\text{best}}(\lambda_{\text{obs}})$
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 2. Fine-tuning of the $P([\text{O III}])/P([\text{O II}])$ ratio

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

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 on 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 probability distribution functions (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 equivalent width (see Sections 3.2.1, 3.2.2, 3.2.3, 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, equivalent width and J-H color index (see Section 3.2.5 and Figures 9 and 10). This second block does not provide infor-

mation on the expected redshift, but it predicts which among the considered species/transitions is more likely the 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}).

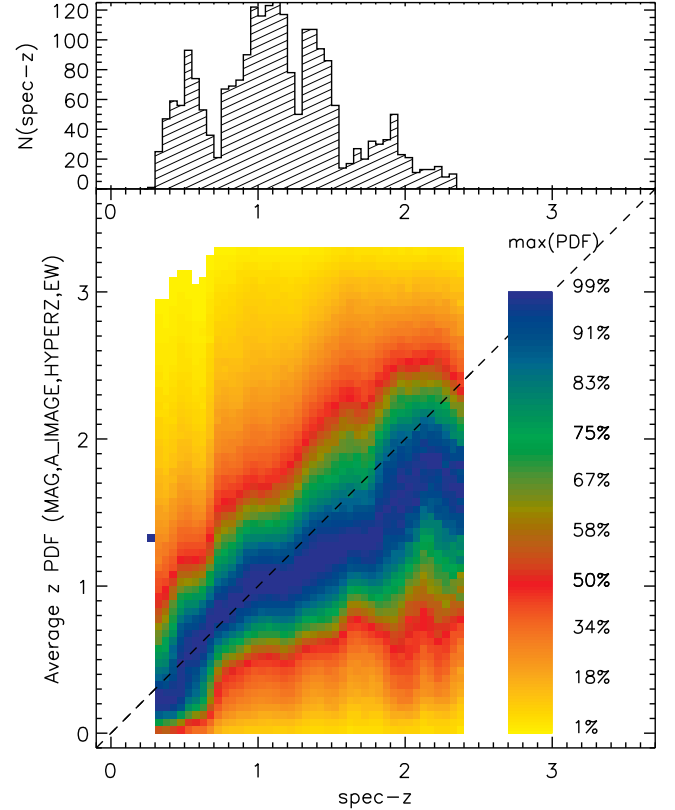


Figure 13. Average PDF* resulting from the combination of the PDF*s obtained using all the redshift indicators described in Sections 3.2.1, 3.2.2, 3.2.3, 3.2.4 (SED fitting, apparent J magnitude, size, and equivalent width respectively), and singularly shown in Figure 14. The histogram in the upper panel represents the distribution of the spectroscopic sources used in the experiment ("gold" sample).

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

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

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 ($\equiv \frac{\lambda_{\text{obs}}}{\lambda_i} - 1$) is the correct spectroscopic redshift. $P_{\text{PDF}}(z_i)$ is the value of the z -PDF* in z_i , while $P_{\text{FR}}(i)$ is the probability obtained from the flux ratio priors, and $P_{\text{corr}}(i, \lambda_{\text{obs}})$ represents the λ_{obs} dependent fine-tuning correction. The z -PDF* is the combination of the SED fitting output with

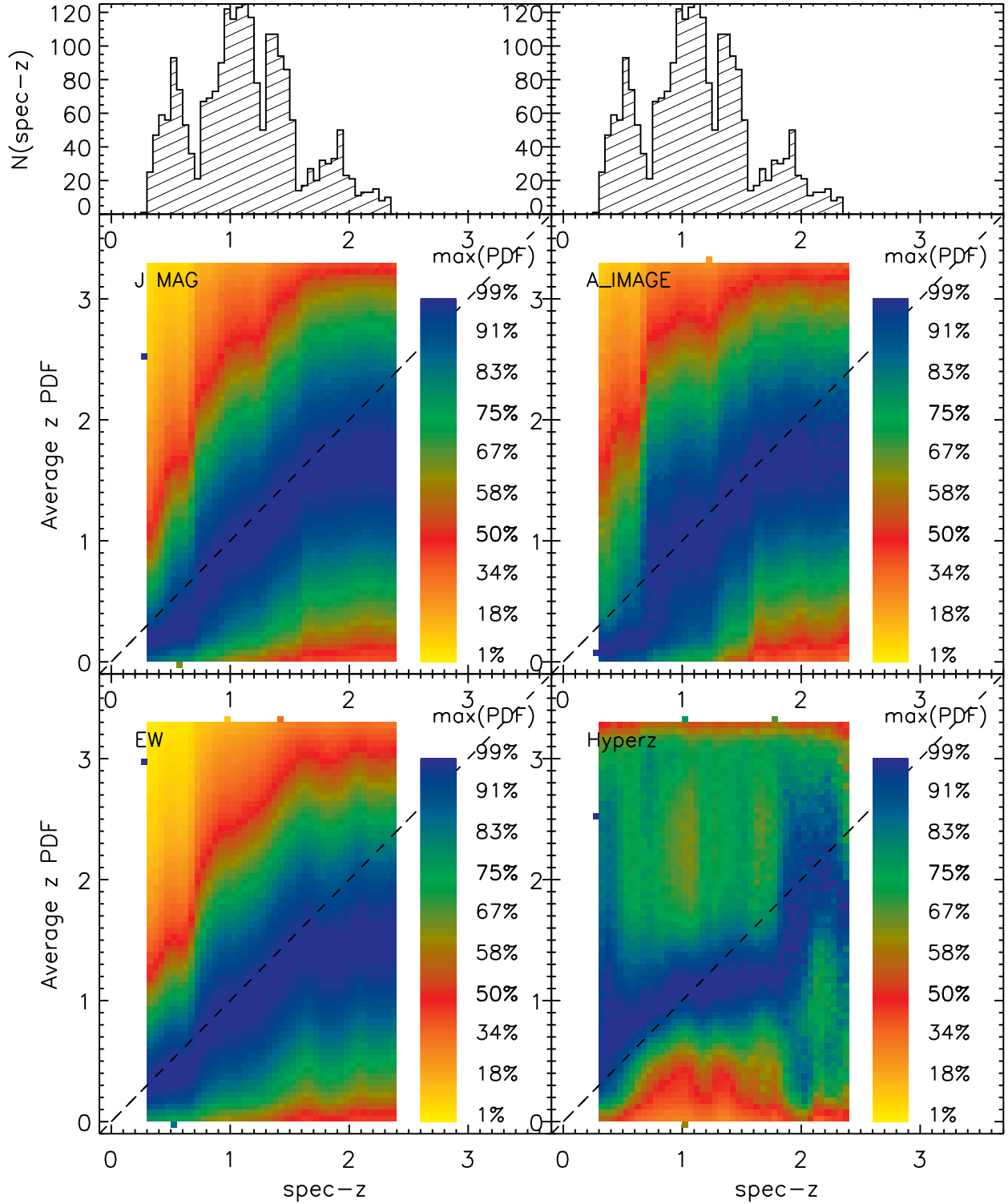


Figure 14. Average PDF* obtained using different methods, as a function of the actual spectroscopic redshift of the sources in the "gold" sample. The PDF*s are computed in bins of spectroscopic redshift (x-axes) and represented through a color scale. The duplicated histograms in the upper panels represent 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 equivalent width (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.

the priors obtained from the apparent J magnitude, size and equivalent width:

$$P_{\text{PDF}}(z_i) = P_{\text{PDF}}^{\text{sed_fit}}(z_i) P_{\text{PDF}}^{\text{J}}(z_i) P_{\text{PDF}}^{\text{size}}(z_i) P_{\text{PDF}}^{\text{EW}}(z_i). \quad (7)$$

$P_{\text{FR}}(i)$, derived from the measured flux ratios, is the weighted average of the values inferred from apparent magnitude, size, equivalent width 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 1.0⁹).

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 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*s 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. The 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 wrong 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

⁹ It is always possible to renormalize the single values of P_i to their sum, obtaining actual values of probability, since all 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. Ly α) or possible spurious detections.

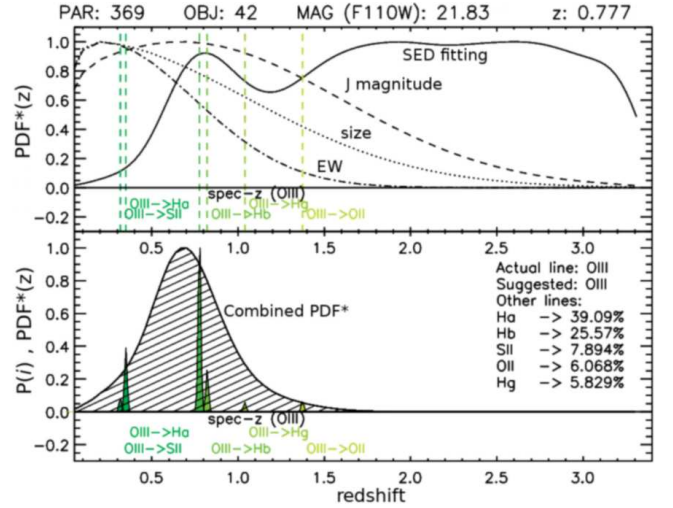


Figure 15. Example output of the algorithm. The **upper panel** shows the four z -PDF*s obtained from SED fitting ($P_{\text{PDF}}^{\text{SED-FIT}}(z)$, black continuous line), apparent J magnitude ($P_{\text{PDF}}^{\text{J}}(z)$, black dashed line), apparent size ($P_{\text{PDF}}^{\text{size}}(z)$, black dotted line), and equivalent width ($P_{\text{PDF}}^{\text{EW}}(z)$, 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 the possible redshift solutions z_i are also shown using vertical dashed lines labelled as: “original solution $\rightarrow$ other possible solution” (example: [O III] $\rightarrow$ H α). In the **lower panel**, the three z -PDF*s are combined in a unique PDF* ($P_{\text{PDF}}(z)$, 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%).

to two intervals: $8500 \text{ \AA} < \lambda_{\text{obs}} < 11000 \text{ \AA}$ and $11400 \text{ \AA} < \lambda_{\text{obs}} < 16700 \text{ \AA}$. 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 note 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 H α transition represent

the most numerous subgroup ($N_{H\alpha}=1293$), followed by the [O III] group ($N_{[\text{O III}]}=949$), [O II] ($N_{\text{O II}}=24$) and other species/transitions ($N_{\text{other}}=16$ in total for $H\beta$, $H\gamma$, [S II])¹⁰.

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 correctly identified by the algorithm) and the contamination (fraction of wrong identifications in the recovered sample). All the 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. Since the [S II], $H\beta$ and $H\gamma$ fluxes are almost always weaker than those measured for $H\alpha$, [O III] and [O II], only very few sources characterized by a prominent emission from [S II], $H\beta$ and $H\gamma$ can be used in the calibration process (i.e. our algorithm is not expected to work for these species)¹¹.

For the most numerous samples, $H\alpha$ and [O III], the completeness is high (88.3% and 77.0% respectively), and 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 $H\alpha$ 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 reasons. First, the very limited amount of such emitters in the original WISP calibration “gold” sample does not allow us 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 $H\alpha$, [O III], and [O II]), and the accuracy that we would obtain by always classifying every strongest line as $H\alpha$ ($\sim 57\%$), 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 $P([\text{O III}])$, 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 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 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 can not 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 6 and 7 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

¹⁰ One spectrum showing a strong [S III]-9532 emission is excluded from the test

¹¹ None of the 16 sources characterized by [S II], $H\beta$ or $H\gamma$ 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].

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 ($N_{\text{CI}}/N_{\text{TGS}}$)	Contamination ($N_{\text{WI}}/N_{\text{I}}$)	Accuracy (Purity) ($N_{\text{CI}}/N_{\text{I}}$)
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%
H β	4861.4	5	0	0 (0%)	0 (-)	0%	-	0%
H γ	4340.5	4	0	0 (0%)	0 (-)	0%	-	0%

^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 to the possibility of this transition to be the responsible for the most prominent emission line measured, the highest relative probability (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.

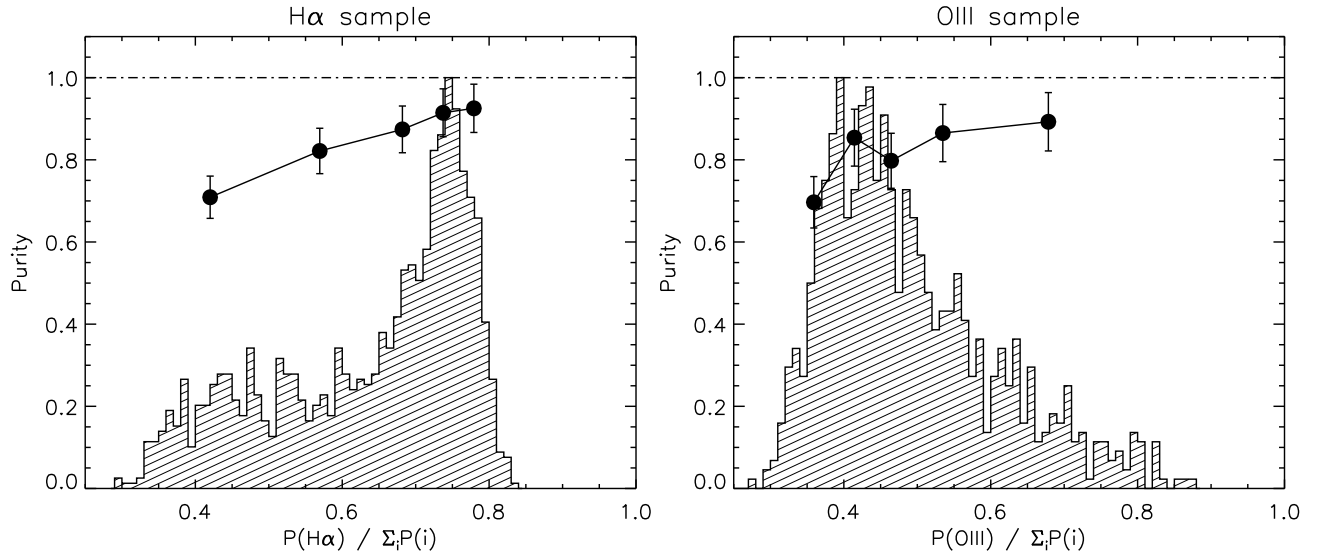


Figure 16. Purity of the recovered H α and [O III] samples (i.e. fraction of emission lines correctly identified as due to H α and [O III] by the algorithm) as a function of the probability values $P(\text{H}\alpha)$ (left panel) and $P(\text{[O III]})$ (right panel), normalized to the sum of the probabilities $\sum_i P(i)$. 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 (H α + [O III]).

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 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 in two independent subsamples (this is confirmed by

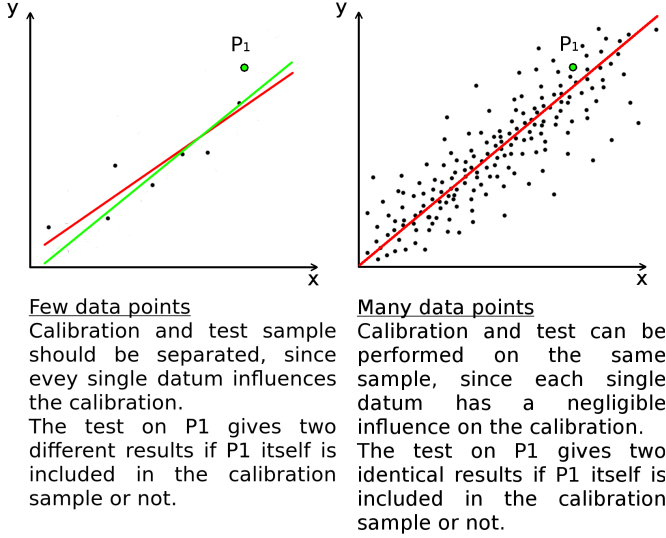


Figure 17. The best approach to validate a relation is to test it on a data sample that is independent from the sample used to compute the relation itself. This is due to the fact that every datum tends to be closer to the relation, if the relation was calibrated using that datum (**left panel**), or 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 if that specific datum was included in the calibration data-set or not.

comparing the precision obtained for the [O II] sample in Tables 3, 6 and 7,).

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 ($S/N > 2\sigma$) 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 $H\beta$ 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, ap-

parent 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 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 $H\alpha$ 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 $H\alpha$ is the actual strongest emission for more than half of the WISP “gold” sample ($\sim 56.7\%$). However, if we were considering the default $H\alpha$ 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

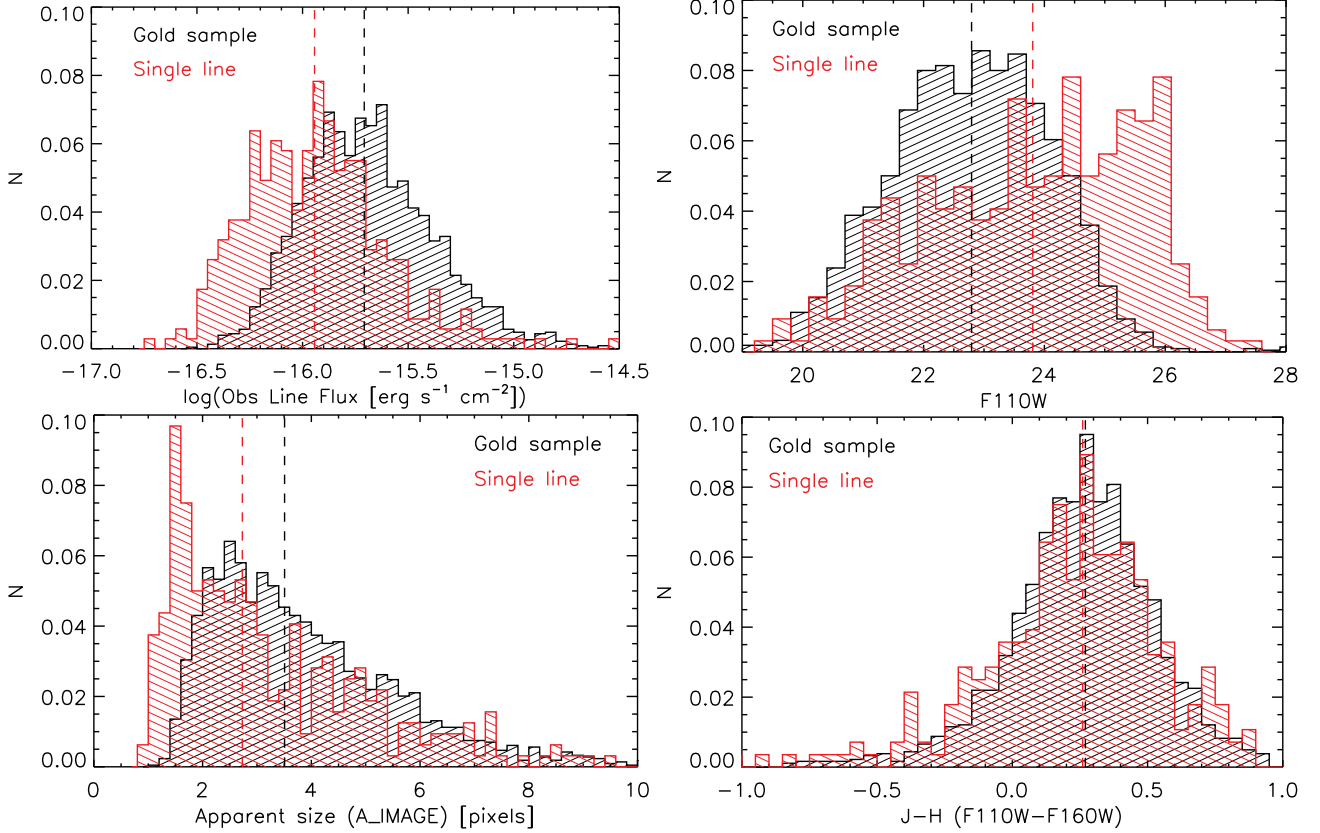


Figure 18. Normalized distribution of the flux of the strongest line (upper left panel), J magnitude (upper right panel), apparent size (bottom left), and J-H color index (bottom right). The distributions are shown for the calibration sample (in black), characterized by two or more lines detected, and 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 not 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.

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 $H\alpha$ 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, $H\alpha$ 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 $H\alpha$ 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” can not be mostly $H\alpha$, as assumed in the original WISP default classification.

Instead, they are more probably evenly distributed between $H\alpha$ 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 the similar redshift distributions of “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 $H\alpha$), 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 a guarantee.

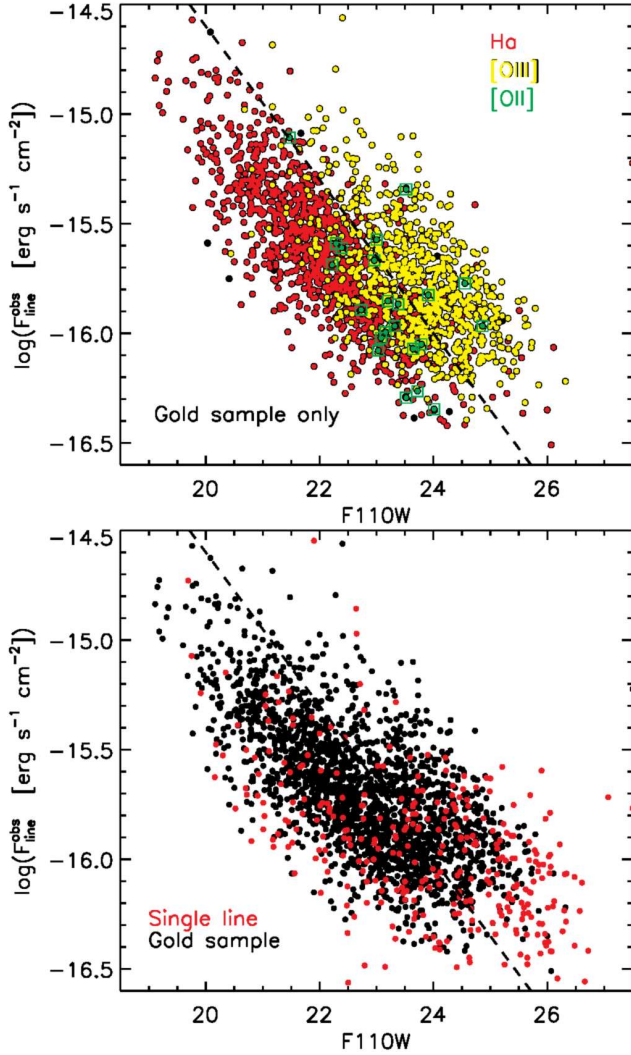


Figure 19. The broad anticorrelation existing between the flux of the strongest spectral line detected and the observed J magnitude. In the **upper panel**, we consider the “gold” sample, using red filled circles when the brightest line is $H\alpha$, yellow when it is $[O III]$, and green circles enclosed in squares when it is $[O II]$. Two different normalizations are clearly visible for $H\alpha$ 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 $H\alpha$ from $[O III]$ dominated sources, by empirically dividing the plots in 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).

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 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 times, contaminants are also decentered with respect to the midline of the bidimensional spectrum.

However, in some cases, false detections can not 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 superimposition 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, contaminants 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 $H\alpha$ 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 $0.8 < z < 1.2$ ($\sim 30\text{--}35\%$).

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 $H\alpha$ 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 $H\alpha$.

This redshift range is particularly relevant also because, given the wavelength coverage of grisms G102 and G141, $H\beta$, $[O III]$, and $H\alpha$ are contemporarily located in the visi-

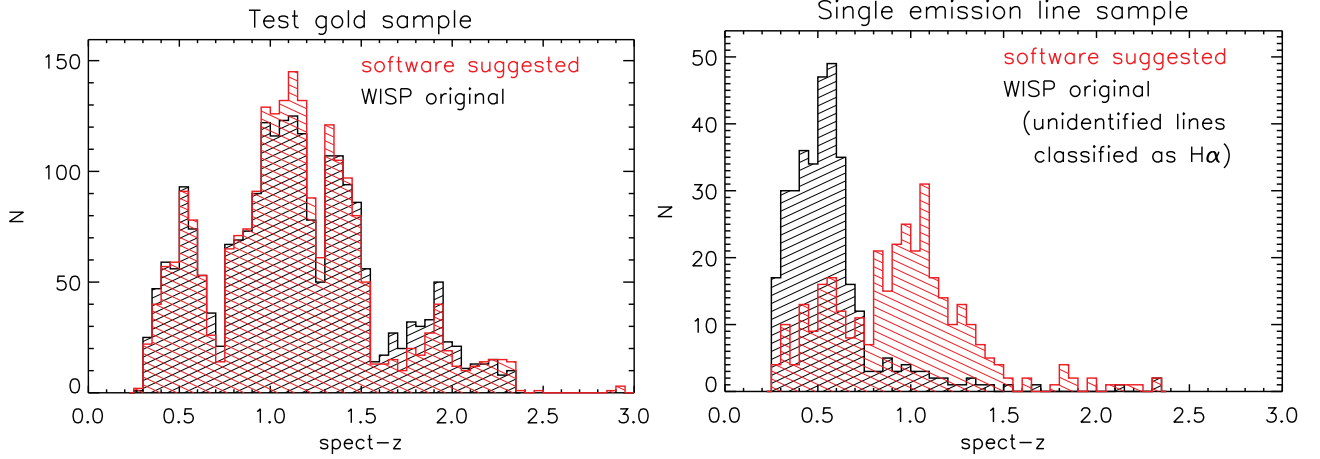


Figure 20. Redshift distribution of the sources resulting from the WISP original (human) and WISP default classification (black shaded histograms in the left and right panels, respectively), compared with the distributions resulting from the automatic classification performed by our algorithm (red shaded 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 ($H\alpha$) 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.

bility 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 $H\beta$ 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.

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 $H\beta$ or $[S II]$ as the strongest (unique) emission line, both these two 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 the 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 $\log([O III]/H\beta) \sim 0.8$ at $J \sim 24.5$ (corresponding to the typical magnitude of the stacked sample). From the stacking, we measure $\log([O III]/H\beta) \sim 1.07$. However, $H\beta$ 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 $H\alpha$ instead of $[O III]$), we stacked the same spectra as if the original WISP classification were correct. In this case, $[O III]$ and $H\beta$ would be located at $\lambda < 8000\text{\AA}$ most of the times, and for this reason they can not 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 $[O III]/H\alpha$ and $[O III]/H\beta$) are particularly high if 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.

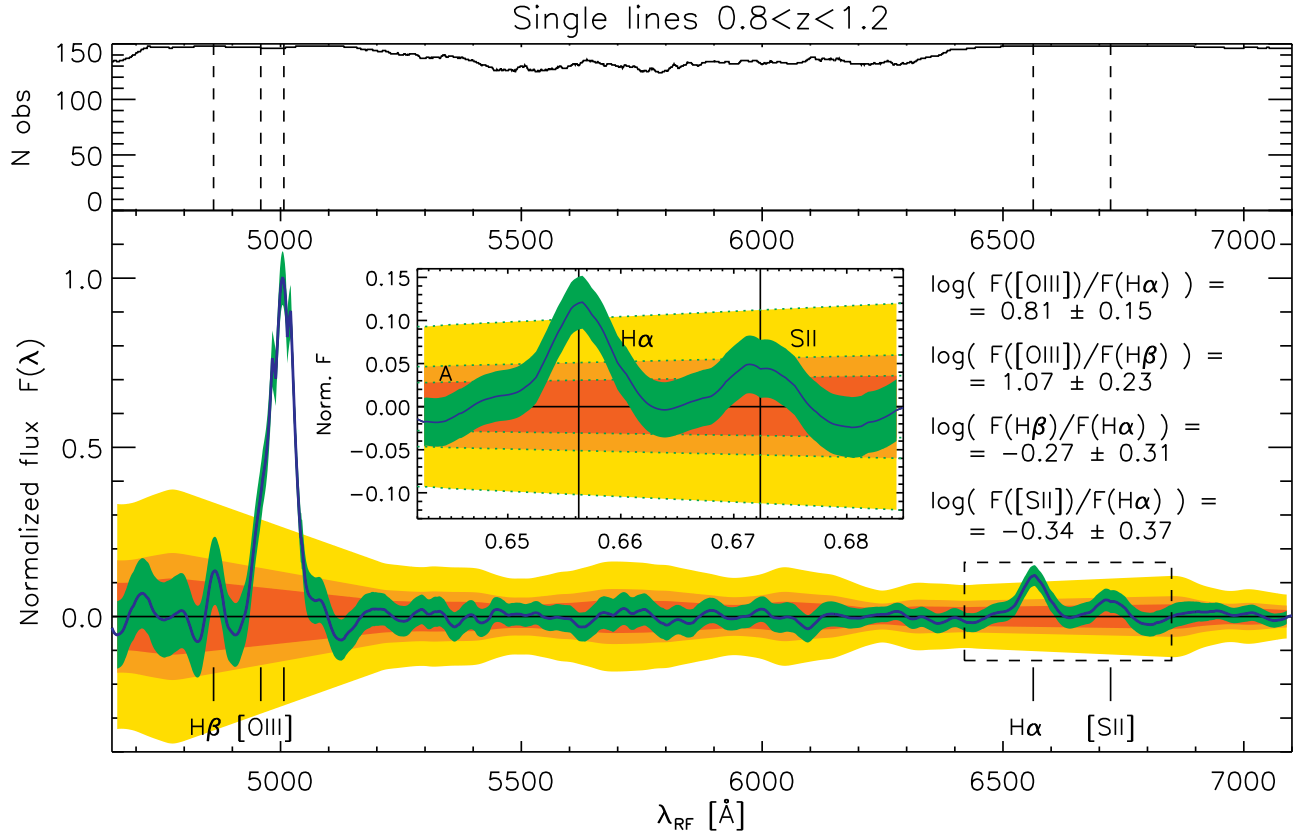


Figure 21. Stacked spectra of “single-line” sources located in the redshift range $0.8 < z < 1.2$. The spectra are stacked consistently with the solution suggested by the algorithm. The position of the main spectral emissions is indicated by black vertical lines (in order of increasing wavelength: H β , [O III], H α , [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 H α 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.

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 H α , as shown in Figure 9. Hence, the sample used for the stacking is mostly made by sources characterized by a particularly suppressed H α emission. In fact, if H α 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.

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 $S/N \sim 7.67$, 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 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

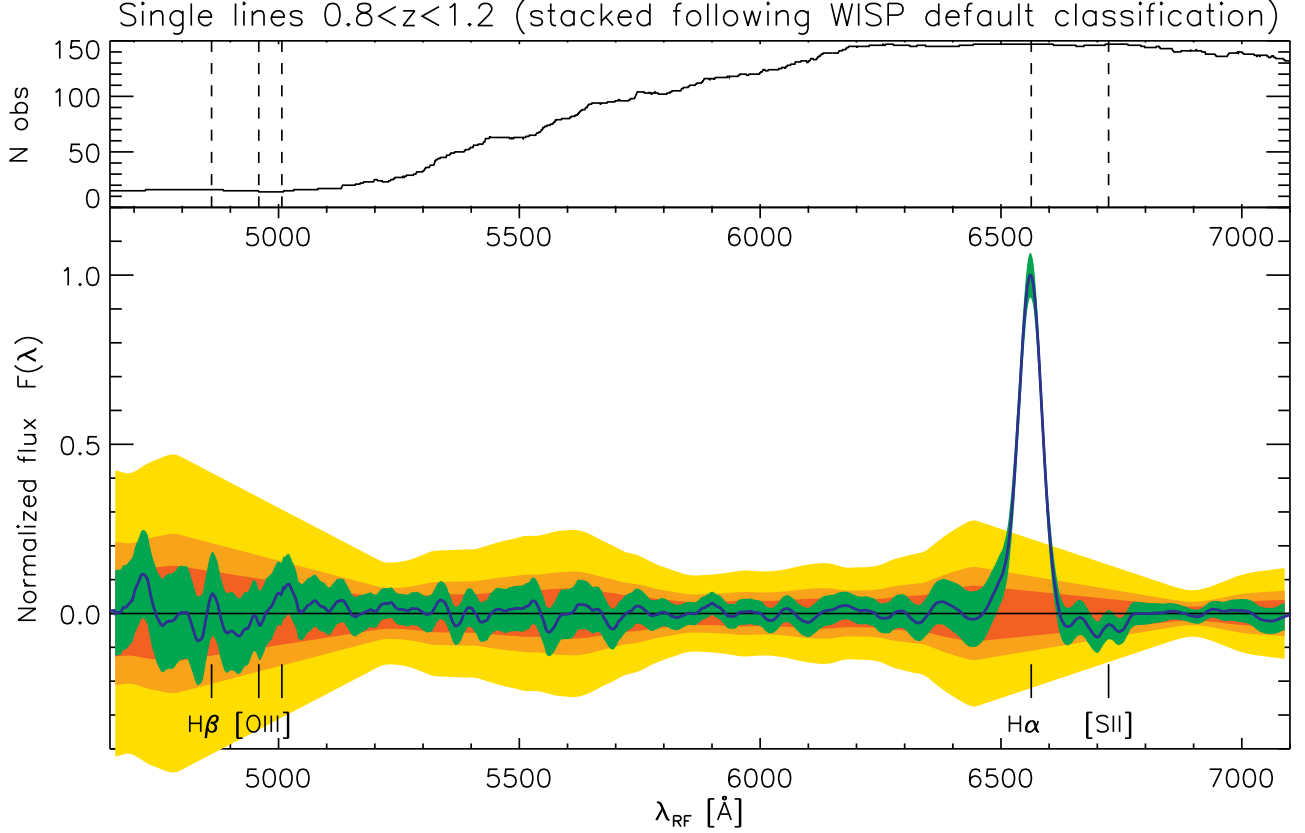


Figure 22. Spectrum obtained by stacking the same “single-line” sources used for Figure 21. In this case, the WISP default classification is assumed. The same fine recentering strategy is used (see Appendix B). In this case, [S II] is not even visible, whereas the same line is detected above 3σ if we follow the classification performed by our algorithm. Instead, following the WISP default classification, H β 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.

$S/N(H\alpha) < 2$, corresponding to the WISP detection threshold¹².

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 $8500\text{\AA} < \lambda < 11100\text{\AA}$ and $11400\text{\AA} < \lambda < 16700\text{\AA}$, 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 H α 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 H α), 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 bias 1). 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 (H α or [O III]), which we set to $S/N=7.67$. The amount of false identifications is set considering the expected purity of the H α and [O III] samples (last column in Table 3), and in accordance with the fraction of actual [O III] and H α 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 identifica-

¹² 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]	
$\log([\text{O III}]/\text{H}\alpha)$	0.30	0.56 ± 0.02	0.59 ± 0.02	0.79 ± 0.03	0.85 ± 0.03	0.81 ± 0.15
$\log([\text{O III}]/\text{H}\beta)$	0.79	0.77 ± 0.04	0.84 ± 0.04	0.98 ± 0.06	1.06 ± 0.18	1.07 ± 0.23
$\log(\text{H}\beta/\text{H}\alpha)$	-0.49	-0.21 ± 0.03	-0.25 ± 0.04	-0.22 ± 0.07	-0.23 ± 0.09	-0.27 ± 0.31
$\log([\text{S II}]/\text{H}\alpha)$	-0.79	-0.49 ± 0.06	-0.52 ± 0.08	-0.57 ± 0.15	-0.75 ± 0.23	-0.34 ± 0.37

^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~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~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 H α 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 can not 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.

tions 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]/H α ratio is more consistent with a 30% contamination, while [O III]/H β agrees better with a 35% contamination. Due to the 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 two 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 an easy addition, replacement, and removal of each single module (and block). Hence, the algorithm can be eas-

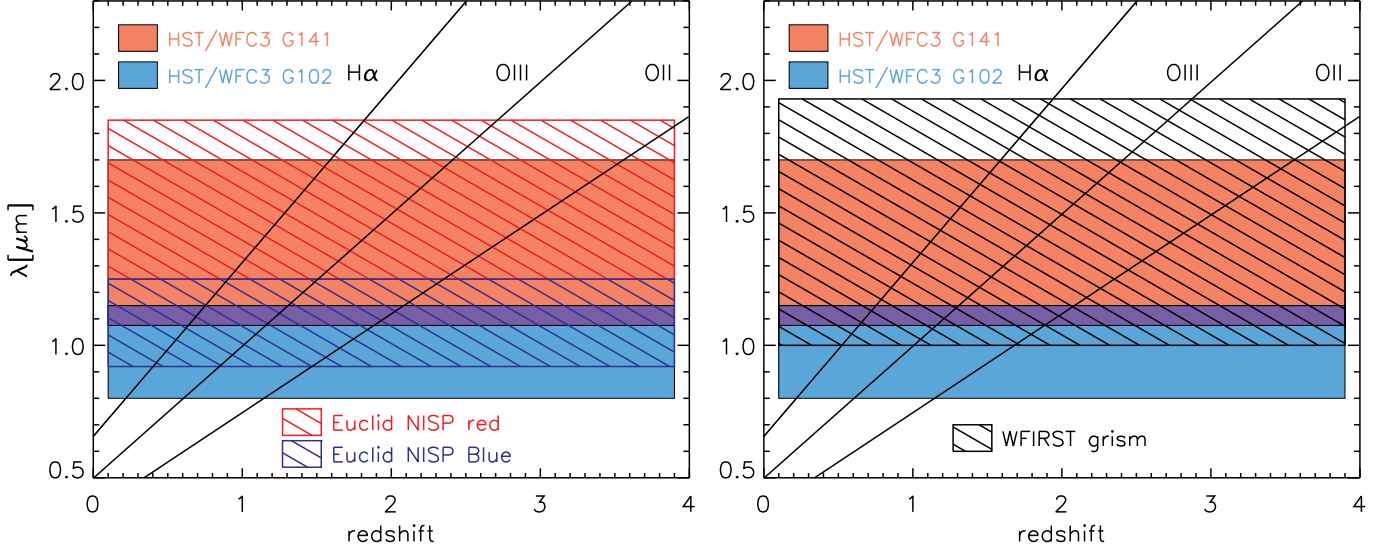
ily adapted to different surveys. Given the similarities between WISP and the Euclid and WFIRST surveys, the experiment that we performed represents an important pilot study in the context of these future missions, in particular for what concerns the maximization of their scientific return. Euclid and 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 H α 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 (~ 5 -10% catastrophic failures, $\sigma_z/(1+z) \leq 0.03$ -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

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

**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 ($H\alpha$, $[O III]$ and $[O II]$) are represented as a function of the redshift, using thick black lines.

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 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 barionic acoustic oscillations (BAO)¹³. To this purpose, the spectroscopic redshift precision required on single sources is $\sigma_{\Delta z} \leq 0.001(1+z)$. 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 $\lambda \sim 5500\text{\AA}$ and $\lambda \sim 9000\text{\AA}$. Additional ground based observations will be ob-

¹³ Euclid is optimized also to study the dark matter distribution in the universe through weak gravitational lensing (WL).

tained 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 than that explored by 3D-HST and AGHAST combined (at the moment, the available WISP data covers a total area of ~ 1520 arcmin²). 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 it has recently been done in Bagley et al. (2019, submitted).

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 H α 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 dataset, 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 ($N_{\text{TGS}}^{[\text{OII}]}=24$) characterized by an [O II] emission stronger than that of H α and [O III]. Besides the inaccurate calibration, given the uneven distribution of sources in the H α , [O II], and [O III] samples, a small fraction of misidentified H α 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 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 ($\sim 82.6\%$), determining the nature of most of the single emission lines is crucial for many reasons. For example, it makes it possible to determine the accurate redshifts of the faintest sources (i.e. the most difficult to explore), or to improve the definition of the faint end of the H α 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, equivalent width, 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 H α emission lines in WISP with good accuracy ($\sim 82.6\%$ 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 datasets 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 test the algorithm, in Tables 6 and 7 we report the accuracy measured by dividing the “*gold*” sample in two independent subsamples. In the first test (Table 6), we use one of the two subsamples to calibrate the algorithm and the other to assess the accuracy. For the second test (Table 7), 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 test, respectively. In Tables 6 and 7 we report only the accuracy obtained for the most relevant transitions (H α , [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.

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 24 show the steps of the processing that precede the stacking phase. Initially, the contamination is removed from the original spectra¹⁴. Then, we estimate the continuum at each λ_i (central red curve in the top panel of Figure 24). for this purpose, we linearly fit the observed flux measured in the wavelength intervals immedi-

ACKNOWLEDGMENTS

I.B. thanks Alvio Renzini, Alberto Franceschini, Paolo Cassata, and Andrea Grazian for the useful discussion about the thematics discussed in the paper. C.S. acknowledges financial support from NASA, through the STScI program number HST-AR-14311.001-A. STScI is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. M.B. acknowledges support from INAF under PRIN SKA/CTA FORE-CaST and from the Ministero degli Affari Esteri della Cooperazione Internazionale - Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02. M.R. acknowledges financial support from NASA, through a grant from the Space Telescope Science Institute (STScI), program number HST-GO-14178.026-A.

ately 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 the two 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 24) 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

¹⁴ The contamination as a function of λ is obtained directly from the original WISP data

Table 6. Alternative accuracy test 1^[a]

species/ transitions	λ_{obs} [Å]	N_{TGS}	N_{I}	N_{CI}	N_{WI}	Completeness ($N_{\text{CI}}/N_{\text{TGS}}$)	Contamination ($N_{\text{WI}}/N_{\text{I}}$)	Accuracy (Purity) ($N_{\text{CI}}/N_{\text{I}}$)
H α	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%

^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 7. Alternative accuracy test 2^[a]

species/ transitions	λ_{obs} [Å]	N_{TGS}	N_{I}	N_{CI}	N_{WI}	Completeness ($N_{\text{CI}}/N_{\text{TGS}}$)	Contamination ($N_{\text{WI}}/N_{\text{I}}$)	Accuracy (Purity) ($N_{\text{CI}}/N_{\text{I}}$)
H α	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%

^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 6)

each spectral line. For example, $CF_{H\alpha}(\lambda_1)$ quantifies the relative probability associated with the H α emission being centered at $\lambda_{\text{obs}}^{H\alpha} = \lambda_1$. This can be immediately translated into the relative probability for the source to be located at $z = \lambda_1/\lambda_{\text{int}}^{H\alpha} - 1$. 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 24 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 25 schematically describes how the contrast factor is computed for the OIII 5007 line that is visible in the spectrum already shown in Figure 24, but the same description can be extended to all the 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

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

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

$$C(\lambda_i) = \int_0^\infty G_{\lambda_i}^+(\lambda) \frac{F(\lambda)}{\sigma_F(\lambda)} d\lambda \quad (\text{B3})$$

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:

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

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 $CF_{\text{TOT}}(z)$. The spectrum is recentered considering the value of z that maximizes the function $CF_{\text{TOT}}(z)$.

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 the line identification process. In fact, the contrast factor is

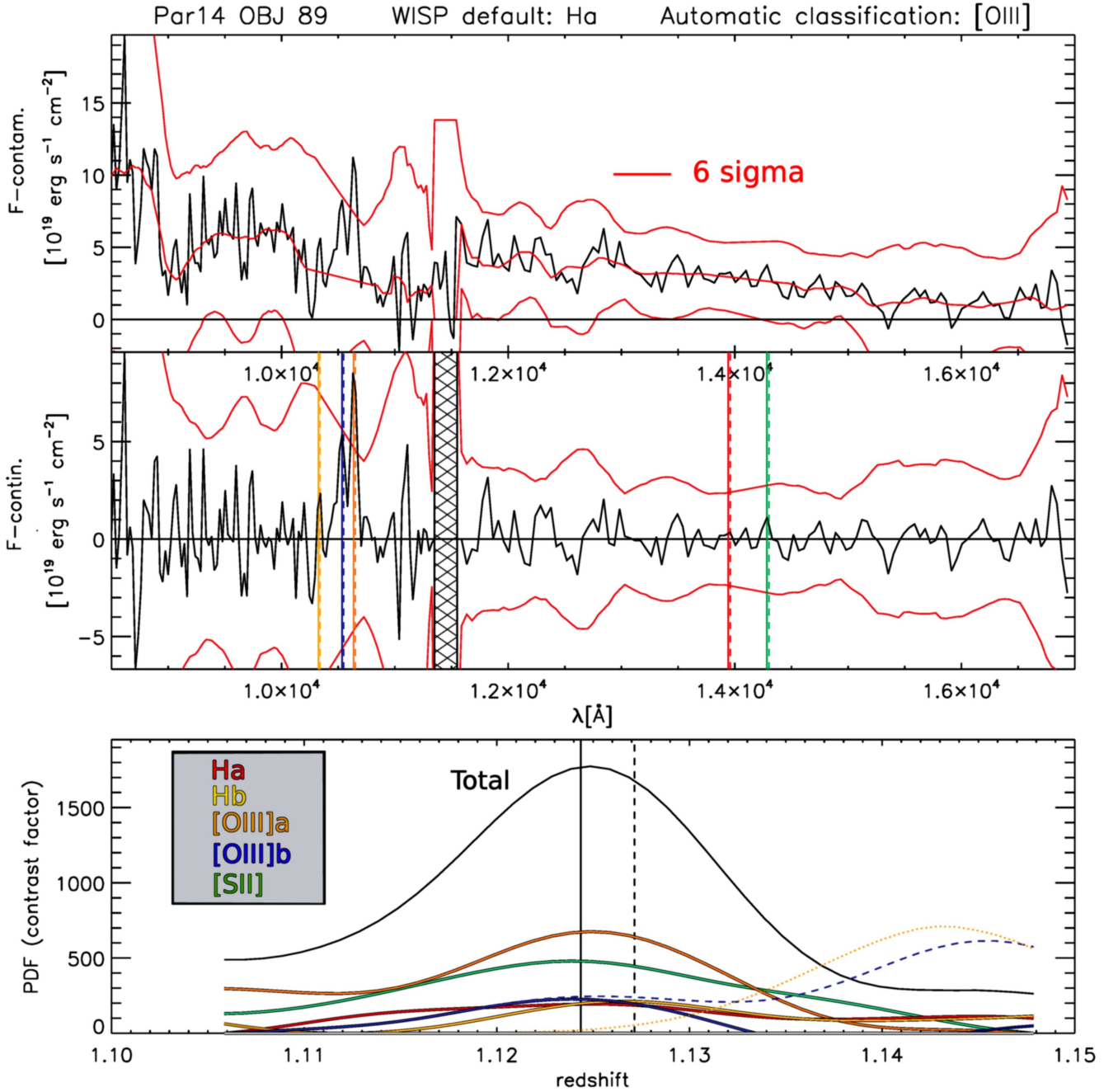


Figure 24. Example of “single-line” spectrum used to compute the stacked spectrum described in Section 5.4 and shown in Figure 21. The contamination-subtracted spectrum is shown in the **top panel**. We considered locally computed $\pm 6\sigma$ deviations (upper and lower red curves) from the median (central red curve) to exclude unrealistic values, unless when the flux excess was located in the 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 25 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 , H β in yellow, O III 4959 in blue, O III 5007 in orange, H α in red, and [S II] in green).

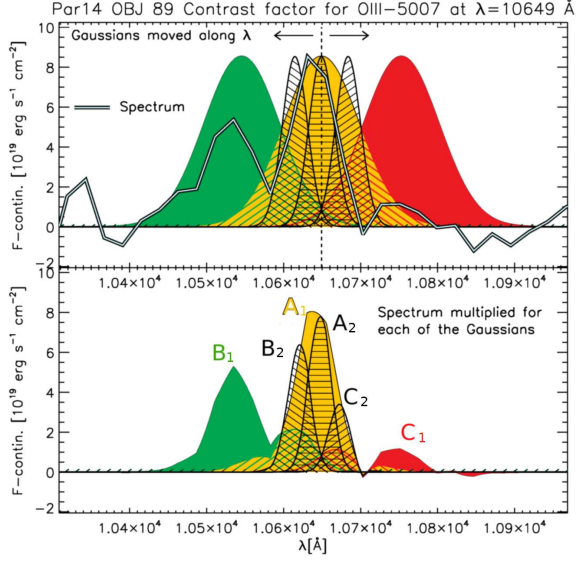


Figure 25. Before the stacking, every spectrum is finely recentered to the common rest frame using a contrast factor $FC(\lambda)$ that we compute for each of the main emission lines. While a simple Gaussian fit is sensitive to the noise at all scales, our contrast factor acts as a combination of two bandpass filters calibrated on the expected width of the spectral line. This approach guarantees the limited influence of the noise at all the other scales. The **top panel** shows the same spectrum of Figure 24, 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 the two 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 ($FC_{TOT}(z)$: black curve in the bottom panel of Figure 24).

to all effects a redshift probability distribution function that could be integrated into our computations¹⁵.

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 the values of $F(\lambda)$ exceeding the 16th and 84th percentiles from the median value ($\pm 1\sigma$).

REFERENCES

- Addison, G. E., Bennett, C. L., Jeong, D., Komatsu, E., & Weiland, J. L. 2019, ApJ, 879, 15, doi: [10.3847/1538-4357/ab22a0](https://doi.org/10.3847/1538-4357/ab22a0)
- Atek, H., Malkan, M., McCarthy, P., et al. 2010, ApJ, 723, 104, doi: [10.1088/0004-637X/723/1/104](https://doi.org/10.1088/0004-637X/723/1/104)

¹⁵ 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

- Bagley, M. B., Scarlata, C., Henry, A., et al. 2017, *ApJ*, 837, 11, doi: [10.3847/1538-4357/837/1/11](https://doi.org/10.3847/1538-4357/837/1/11)
- Baronchelli, I., Rodighiero, G., Teplitz, H. I., et al. 2018, *ApJ*, 857, 64, doi: [10.3847/1538-4357/aab78b](https://doi.org/10.3847/1538-4357/aab78b)
- Bolzonella, M., Miralles, J., & Pelló, R. 2000, *A&A*, 363, 476
- Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, *ApJ*, 686, 1503, doi: [10.1086/591786](https://doi.org/10.1086/591786)
- Brammer, G. B., van Dokkum, P. G., Franx, M., et al. 2012, *ApJS*, 200, 13, doi: [10.1088/0067-0049/200/2/13](https://doi.org/10.1088/0067-0049/200/2/13)
- Bruzual, G., & Charlot, S. 2003, *MNRAS*, 344, 1000, doi: [10.1046/j.1365-8711.2003.06897.x](https://doi.org/10.1046/j.1365-8711.2003.06897.x)
- Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, *ApJ*, 533, 682, doi: [10.1086/308692](https://doi.org/10.1086/308692)
- Clements, D. L., & Couch, W. J. 1996, *MNRAS*, 280, L43, doi: [10.1093/mnras/280.3.L43](https://doi.org/10.1093/mnras/280.3.L43)
- Dickinson, M. 1998, in *The Hubble Deep Field*, ed. M. Livio, S. M. Fall, & P. Madau, 219
- Green, J., Schechter, P., Baltay, C., et al. 2011, arXiv e-prints, arXiv:1108.1374. <https://arxiv.org/abs/1108.1374>
- . 2012, arXiv e-prints, arXiv:1208.4012. <https://arxiv.org/abs/1208.4012>
- Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, *ApJS*, 197, 35, doi: [10.1088/0067-0049/197/2/35](https://doi.org/10.1088/0067-0049/197/2/35)
- Hasinger, G., Capak, P., Salvato, M., et al. 2018, *ApJ*, 858, 77, doi: [10.3847/1538-4357/aabacf](https://doi.org/10.3847/1538-4357/aabacf)
- Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, *ApJS*, 197, 36, doi: [10.1088/0067-0049/197/2/36](https://doi.org/10.1088/0067-0049/197/2/36)
- Lapi, A., Mancuso, C., Bressan, A., & Danese, L. 2017, *ApJ*, 847, 13, doi: [10.3847/1538-4357/aa88c9](https://doi.org/10.3847/1538-4357/aa88c9)
- Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193. <https://arxiv.org/abs/1110.3193>
- Laureijs, R., Gondoin, P., Duvet, L., et al. 2012, in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Vol. 8442, *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, 84420T, doi: [10.1117/12.926496](https://doi.org/10.1117/12.926496)
- Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, *A&A*, 439, 845, doi: [10.1051/0004-6361:20041960](https://doi.org/10.1051/0004-6361:20041960)
- Madau, P., Ferguson, H. C., Dickinson, M. E., et al. 1996, *MNRAS*, 283, 1388
- Mattig, W. 1958, *Astronomische Nachrichten*, 284, 109, doi: [10.1002/asna.19572840303](https://doi.org/10.1002/asna.19572840303)
- Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, *ApJS*, 225, 27, doi: [10.3847/0067-0049/225/2/27](https://doi.org/10.3847/0067-0049/225/2/27)
- Poggianti, B. M., Calvi, R., Bindoni, D., et al. 2013, *ApJ*, 762, 77, doi: [10.1088/0004-637X/762/2/77](https://doi.org/10.1088/0004-637X/762/2/77)
- Rodríguez-Muñoz, L., Rodighiero, G., Mancini, C., et al. 2019, *MNRAS*, 485, 586, doi: [10.1093/mnras/sty3335](https://doi.org/10.1093/mnras/sty3335)
- Salvato, M., Ilbert, O., & Hoyle, B. 2018, *Nature Astronomy*, doi: [10.1038/s41550-018-0478-0](https://doi.org/10.1038/s41550-018-0478-0)
- Schechter, P. 1976, *ApJ*, 203, 297, doi: [10.1086/154079](https://doi.org/10.1086/154079)
- Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, *ApJS*, 214, 24, doi: [10.1088/0067-0049/214/2/24](https://doi.org/10.1088/0067-0049/214/2/24)
- Spiegel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757. <https://arxiv.org/abs/1503.03757>
- Stevens, R., & Lacy, M. 2001, *MNRAS*, 325, 897, doi: [10.1046/j.1365-8711.2001.04475.x](https://doi.org/10.1046/j.1365-8711.2001.04475.x)
- van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014, *ApJ*, 788, 28, doi: [10.1088/0004-637X/788/1/28](https://doi.org/10.1088/0004-637X/788/1/28)
- 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
- Wittman, D., Bhaskar, R., & Tobin, R. 2016, *MNRAS*, 457, 4005, doi: [10.1093/mnras/stw261](https://doi.org/10.1093/mnras/stw261)