Identifying Extreme Emission-Line Galaxies in Deep JWST Fields Using Gaussian Processes

Stephen R.D. Wagner[1,2], Alexander de la Vega[3], and Bahram Mobasher[3]

[1]University of Hawai'i at Mānoa, 2500 Campus Rd, Honolulu, HI
[2]Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI
[3]University of California Riverside, 900 University Ave, Riverside, CA

August 16, 2024

Abstract

We fit the continua of 71638 galaxies from JADES GOODS-N and GOODS-S fields using a novel technique in order to detect Extreme Emission Line Galaxies (EELGs). EELGs are important because they provide information on what the conditions of the universe were at high redshifts. This research defines EELGs by their emission lines' high equivalent width (EW). Calculating EW is challenging when a galaxy only has photometric data because photometric data are often wide bands, which means the spectral resolution is low. We use a machine learning technique, Gaussian Processes Regression, to fit the continua of galaxies. After identifying EELGs in our sample, we found that they are less massive and have higher specific star formation rates when compared to the general population of galaxies. We also calculated the number density of EELGs as a function of redshift and found that EELGs are most common at high redshifts; they are ∼3x more common at z ∼ six than at z ∼ two.

1 Introduction

With the help of the James-Webb Space Telescope (JWST), galaxies are being observed at higher redshifts than ever before [Dav+23]. This research focuses on a certain type of galaxies- extreme emission line galaxies (EELGs). EELGs are important because they can elucidate how the universe underwent reionization at high redshifts [Dav+23], [Lle+24]. Additionally, finding out how common EELGs are as a function of redshift can tell us how common the processes that lead to EELGs are. These galaxies' extreme emission lines may be caused by active galactic nuclei or extreme star formation. Additionally, analyzing EELGs' emission-line ratios or spectra can tell us what the conditions of the Interstellar Medium were like at those times.

We define EELGs by their emission lines' high equivalent width (EW).

$$EW = (\frac{Fv - Cv}{Cv})\Delta\lambda \tag{1}$$

Equation 1 shows how EW is calculated. F_ν is the flux in a certain filter, C_ν is the flux of the continuum, and $\Delta\lambda$ is the width of the filter that observed the emission line.

In this research, we require a rest-frame EW of greater than or equal to 1000 for a galaxy to be considered an EELG. Other research uses observed frame EW measurements, such as [Dav+23], who require an observed-frame EW of greater than or equal to 5000 . We use rest-frame EW because it accounts for the broadening of emission lines at higher redshifts. Taking redshift into account makes sure that there isn't a bias in selecting EELGs at high redshifts.

The challenge with identifying EELGs via their emission lines' EW is that most galaxies don't have spectroscopic measurements, only photometric ones. Photometric data primarily uses wide bands, which results in low spectral resolution. Therefore, acquiring a continuum from photometry is difficult. A galaxy's photometry must be fit or assumed in order to get a continuum measurement and calculate EW. To estimate the continuum, galaxies' photometry is usually assumed to be constant or linear (e.g. [Lle+24], [Dav+23], [Wel+11]). However, the EW calculations that result from these continuum fits may not be appropriate. Especially, the technique employed by [Dav+23] cannot fit galaxies with red continua. Assumption of a constant continuum or linear regression don't always work because continua

aren't usually constant or linear- most of the time, they can be described by power laws [Mor+23]. Using a more reliable fitting procedure is crucial for accurate measurements of EW and the subsequent detection of EELGs. This research uses a machine-learning based interpolation technique called Gaussian Processes.

2 Data

We look for EELGs in the JWST Advanced Deep Extragalactic Survey (JADES), specifically the Great Observatories Origins Deep Survey (GOODS) North and South catalogs (GOODS-N and GOODS-S) [Eis+23a]. GOODS-N and GOODS-S were chosen to find EELGs because they have 11 and 14 NIRCam photometric filters respectively [Eis+23a]. These include medium-band filters, which are important for finding EELGs because they are narrower than wide-band filters so emission lines are easier to "pick out" in them. This research uses both photometry and spectra from JADES NIRCam and NIRSpec instruments [Bun+24], [Hai+24], [Eis+23a], [Eis+23b], [Rie+23], [DEu+24]. The area of the combined GOODS surveys is 124 arcminutes^2.

3 Sample Selection

Using a flag available in the JADES catalogs, we made sure that all objects were galaxies, rather than stars. We also necessitated that objects have an apparent magnitude brighter than 29.5, the limiting magnitude of JADES [Eis+23a], in the F356W filter. We also required that objects have photometric fluxes in at least five filters to ensure a reliable fit of the continuum. We also fit these galaxies using EAZY ([BDC08] to obtain stellar masses and star formation rates. We also imposed a redshift cut of one to seven. This is due to these being the ideal redshifts to detect emission lines Hβ & O[III] and N[II] & Hα, which are the most recognizable emission lines in the rest-frame optical. As seen in Figure 1, a redshift of one to seven is ideal for these filters to detect these emission lines in. Two exceptions are around redshifts $\sim$2.2 and $\sim$3.7, where the gap between filters F210M and F277W leaves these emission lines unable to be detected, resulting in an incomplete sample. The number of galaxies in our final sample is 71638 galaxies.

4 Methods

This work uses a method called Gaussian Processes, which is a machine-learning based interpolation technique that can create nonlinear fits and errors. Gaussian processes allows us to use Gaussian probability distributions for functions, letting us get good representations of galaxy continua that can be described by power laws. [Iye+19]. Usually, Gaussian Process regression is used to interpolate between data points with a low signal to noise ratio [Iye+19]. Gaussian Processes has parameters such as kernel and length scale, which can be optimized to create the most accurate fit. To utilize Gaussian Processes, this research uses a Python package called George [Amb+15].

First, we demonstrate that Gaussian Processes can estimate continua accurately. To do this, we first fit the galaxies that have spectral data so we can compare our Gaussian fits to real spectra. When the Gaussian fit and spectra are overlaid in the same plot, it is clear that the Gaussian fit produces a realistic continuum in most cases. In Figure 2, we show example Gaussian Process fits for two galaxies and how they compare with their JWST spectra. These spectra are flux calibrated to ensure a direct comparison. First, fluxes in the same NIRCam filters are estimated from the spectra. Then, a line is fit to the ratio of these fluxes to those from NIRCam photometry, and this fit line serves as a correction factor to the spectra as a function of wavelength. In Figure 2, the continuum photometry is shown as black circles, with points flagged as emission lines shown as red squares. The spectrum is shown as a light grey line, with the emission lines seen in red. The Gaussian Process fit is the dark grey line with the uncertainty shown as the grey shading. We can see from this fit that our technique is validated from real spectra, and that Gaussian Processes fit curved photometry well.

4.1 Continuum Fitting Process

Before fitting with Gaussian Processes, we must ensure that only the continuum of the galaxy is being fit, not any of its potential emission lines. To do this, we iteratively fit a line to the photometry, excluding the brightest point every time. We do this until no improvement is seen the reduced chi squared of the fit of the line, or more than half of the points have been excluded. Then, the fit of the line is subtracted from the fluxes to get a mean value of near zero. Next, outliers are found using sigma clipping. If a value is more than one normalized median absolute deviation away, it is flagged as being contaminated by an emission line. Lastly, we fit the non-emission points with Gaussian Processes and optimize the fit using least-squares-fitting.

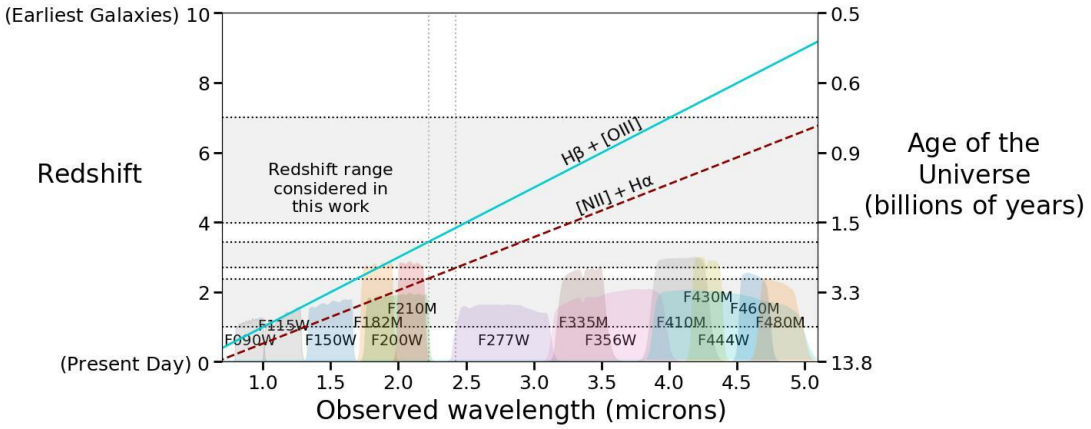

Figure 1: The redshift range (z of one to seven) that this research uses, shown in grey shading. The x axis is the wavelength in microns, and the y axis is the redshift and subsequent age of the universe. The colored shapes are the NIRCam filter throughput curves and their respective coverages. Redshift cuts were chosen based on NIRCam filters detecting emission lines Hβ & O[III] and N[II] & Hα (blue and red lines). We exclude redshifts of $\sim$2.2 and $\sim$3.7 because the gap between filters F210M and F277W leaves these emission lines unable to be detected, resulting in an incomplete sample.

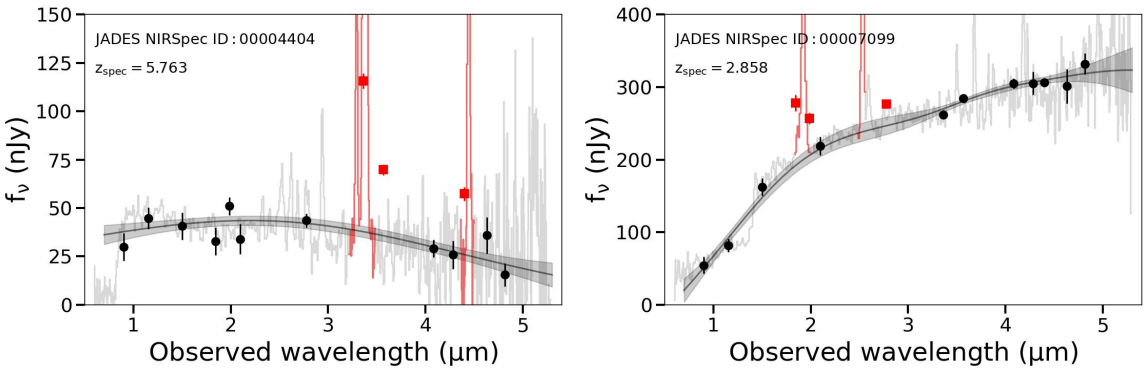

Figure 2: Examples of using Gaussian Processes to fit the continuum of a galaxy. We can see from these fits that our technique is validated from real spectra, and that Gaussian Processes fit curved photometry adeptly.

4.2 Kernel Optimization

As mentioned previously, Gaussian Processes has a parameter known as the kernel. The kernel is also known as the covariance function, and it describes how data points change with respect to each other. To determine which kernel to use, six iterations of Gaussian fitting were conducted on the GOODS-S galaxies. A uniform length scale of four was used for all iterations. Three Gaussian kernels- exponential squared, matern32, and matern52- were tested, along with testing whether or not to exclude certain filters. Excluding bluer filters may be necessary at z $>$ six because bluer wavelength filters can be contaminated by the Lyman Alpha (Lyα) break at high redshifts. This causes the continuum to "break" blueward of Lyα and dip sharply downwards, so including bluer filters at high redshifts can compromise the Gaussian fit. The Lyα break occurs at a rest-frame wavelength of 0.1216 microns. To compensate, we exclude filters with rest-frame wavelengths shorter than Lyα at a given redshift.

The reliability of these six iterations is determined by the log likelihood that George generates for each Gaussian fit. The higher the value of the log likelihood, the better the fit of the Gaussian. In Figure 3, a histogram of the distribution of log-likelihood for each iteration is plotted onto a single figure. From the plot, it is evident that all iterations produce a similar distribution of log likelihoods, meaning that any of the six trials work and the technique is robust. Knowing this, we chose the Matern32 kernel and not to skip certain filters for all of our subsequent Gaussian fits.

We then compared how many EELGs were found in each iteration. For this test, we only imposed an EW cut of ≥ 1000, with no redshift or mass cut. Our results are shown in the table below. We express both the number of EELGs we find and the percentage of the total that EELGs make up, rounded to 2 decimal places. "Skip" refers to skipping certain filters due to the Lyα break.

Kernel Used	**Number of EELGs**	**EELG % of Total**
Exponential Squared	6582	22.37%
Exponential Squared Skip	6576	22.35%
Matern32	6449	21.89%
Matern32 Skip	6443	21.87%
Matern52	6505	22.08%
Matern52 Skip	6499	22.06%

The fact that there is little discrepancy between the number of EELGs found per iteration of fits proves that our technique robustly identifies EELGs.

4.3 Length Scale Optimization

However, a length scale of four was not optimal for all galaxies' continua. A rare few galaxies have extremely high reported max EWs, with some having values as high as $6 * 10^{305}$. We consider an EW of $\geq 10^5$ suspiciously high. These high EWs are due to both a subpar initial length scale and errors in photometry. We find that using a smaller length scale is the solution for these galaxies, with values of 0.01, 0.5, and 1 resulting in better fits. We refit our galaxies with those three length scales, and the length scale that produced the highest log likelihood is recorded as the best initial length scale. The fit that the best initial length scale produces is the one that is recorded. In Figure 4, a histogram of the distribution of log likelihood for each best initial length scale is shown. This plot confirms that all three length scales have similar distributions of log-likelihood, proving that no one length scale is best for all galaxies.

5 Results

5.1 Number of EELGs Found

Our first result discusses the number of EELGs we found. To find EELGs, we imposed different cuts in EW, redshift, and mass. Different cuts yield different numbers of EELGs. The results of different cuts are shown in the table below.

Equivalent Width	**z**	**Mass**	**Number of EELGs**
10^3 – 10^5, Rest-frame	1-7	N/A	9565
10^3 – 10^5, Rest-frame	1-7	10^8 – $10^{11.5}$ $M_\odot$	1279
5000 – 10^5, Observed-frame	4-9	N/A	4549

If we apply a rest-frame EW cut of $10^3 - 10^5$ and a redshift cut of 1-7, we find 9565 EELGs. However, to ensure completeness, we must also apply a cut in mass. To do this, we applied the same rest-frame EW cut and redshift cut but also required galaxies with masses of $10^8 - 10^{11.5}$ $M_\odot$. Doing this, we found 1279 EELGs. Finally, we compared our findings to [Dav+23]'s EELG search in CEERS. To mimic their

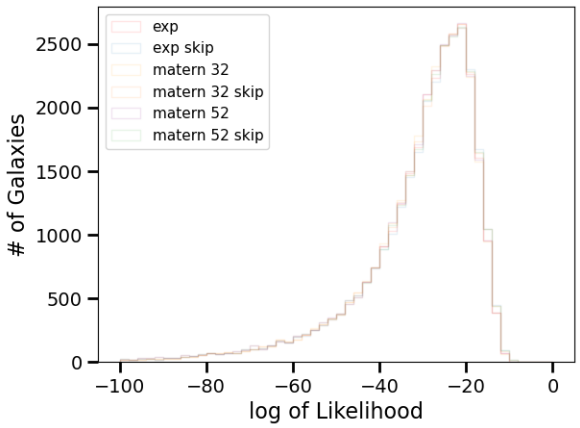

Figure 3: Histogram of the distribution of log likelihoods for GOODS-S galaxies photometry fit with Gaussian Processes. A higher log likelihood indicates a better fit. This graph illustrates that these six iterations produce the same quality of fits.

selection, we required galaxies to have an *observed frame* EW of $5000 - 10^5$ and a redshift of four to nine. We did not impose a mass cut. When we applied these restrictions, we found 4549 EELGs as compared to [Dav+23]'s 1165 EELGs- nearly four times as many EELGs. This research found more EELGs even when the same cuts were applied because JADES's data is deeper than CEERS's data and our continuum fitting technique is flexible.

5.2 Distribution of EELGs' Masses

In Figure 5, the distribution of mass for EELGs and non-EELGs is plotted. We imposed a redshift requirement of one to seven. We can see that non-EELGs' masses peak far higher than EELGs' masses do, which leads us to believe that EELGs are less massive than the general population of galaxies. And indeed, the median of the log of non-EELGs mass is $\sim$7.830$M_\odot$, while the median of the log of EELGs is $\sim$6.878 $M_\odot$. This conclusion of EELGs being less massive than the general population of galaxies agrees with [Lle+24].

5.3 Distribution of EELGs' Specific Star Formation Rates

Using our fits from EAZY, we calculated the specific star formation rate (SSFR) for all of the galaxies in our sample. The distribution of SSFRs for EELGs and non-EELGs is shown in Figure 6. We only plotted the SSFRs of galaxies that have a redshift of one to seven and a mass of $10^8 - 10^{11.5}$ $M_\odot$. We can see that the SSFRs "bunch up" towards the right side of the plot. This is because the EAZY templates we used have limitations, and as as result, we do not have any log SSFRs past -8.3. Future work would involve fitting these galaxy's SEDs using a broader range of models.

As seen in the figure, non-EELGs have a mode in log SSFR around -8.66. However, the distribution of log SSFR for EELGs does not have a clear peak, and rises to the limit in our domain of log SSFR. This suggests that EELGs likely have typical SSFRs that are higher than non-EELGs. Therefore, EELGs must be more active with their star formation than the general population of galaxies. This aligns with the idea that EELG's strong emission lines may be associated with extreme star formation [Wel+11], [Lle+24].

5.4 Number Density of EELGs

Finally, the number density of EELGs is calculated. The redshift bins used are listed in Table 5.4. We exclude certain redshifts because these redshifts put the emission lines Hβ & O[III] and N[II] & Hα out of range of the NIRCam filters.

Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
1-1.5	1.5-2	2.7-3.44	4-5	5-6	6-7

To calculate number density and subsequent error bars, we adopt a technique employed by [Lon+23]. First, each galaxy's redshift is perturbed by generating a random number. This random number has a uniform distribution with an upper bound equal to the redshift plus its uncertainty and a lower bound equal to the redshift minus its uncertainty. This process is done for all galaxies in a certain redshift bin. The number of galaxies that are still in the redshift bin is recorded. This is done for 1000 iterations. The number density is the median of this distribution of

the number of galaxies that remain in the redshift bin. The error bars are the 16th and 84th percentiles.

We can see from Figure 7 that EELGs are about three times more common at redshift $\sim$ six than at redshift $\sim$ two. Our measurements are shown as black circles with error bars. Some of the errors are within the circles. We also plot the number densities of different works. It should be noted that these studies adopt varying definitions of what an EELG is, so interpreting this plot is difficult without making identical cuts as these studies do. [Wel+11]'s number density calculation is shown as the blue triangle, [Car+09]'s is shown as the orange plus, [Ate+11]'s is shown as the green star, and [Shi+11]'s is shown as the yellow square. We note that number density peaks at higher redshifts, then decreases slowly, and finally decreases sharply between redshifts two and zero. This trend is further shown by the results from [Ate+11] and [Car+09]. We can conclude that EELGs are more common at higher redshifts. This suggests that the light from early galaxies is dominated by ionizing radiation. Based on the high number densities at high redshift, the likely cause is intense star formation, with some contribution from active galactic nuclei.

6 Conclusions

This research used Gaussian Processes to fit the photometric continuum of galaxies. The aim of this was to calculate galaxies' emission lines' EWs and introduce a more reliable way to detect EELGs. EELGs are critical to our understanding of the universe in its early stages, such as how it underwent reionization. We used data from JADES' GOODS-N and GOODS-S surveys to search for EELGs. We also used EAZY templates to get mass and star formation rate values for these galaxies. We find that EELGs have lower masses and higher specific star formation rates than the general population of galaxies. Finally, we calculated the number density of EELGs for multiple redshift bins and found that their number density is higher at higher redshifts.

As mentioned previously, future work would involve fitting the GOODS galaxy's SEDs using a broader range of EAZY models. This would result in more accurate values for mass and star formation rate. EELGs could be characterized even better with this information, and therefore more could be learned about the early universe.

References

[BDC08] Gabriel B. Brammer, Pieter G. van Dokkum, and Paolo Coppi. "EAZY: A Fast, Public Photometric Redshift Code". In: *The Astrophysical Journal* 686.2 (Oct. 2008). arXiv:0807.1533 [astro-ph], pp. 1503–1513. ISSN: 0004-637X, 1538-4357. DOI: 10.1086/591786. URL: http://arxiv.org/abs/0807.1533 (visited on 08/26/2024).

[Car+09] Carolin N. Cardamone et al. "Galaxy Zoo Green Peas: Discovery of A Class of Compact Extremely Star-Forming Galaxies". In: *Monthly Notices of the Royal Astronomical Society* 399.3 (Nov. 2009). arXiv:0907.4155 [astro-ph], pp. 1191–1205. ISSN: 00358711, 13652966. DOI: 10.1111/j.1365-2966.2009.15383.x. URL: http://arxiv.org/abs/0907.4155 (visited on 08/17/2024).

[Ate+11] H. Atek et al. "Very Strong Emission-Line Galaxies in the WISP Survey and Implications for High-Redshift Galaxies". In: *The Astrophysical Journal* 743.2 (Dec. 2011). arXiv:1109.0639 [astro-ph], p. 121. ISSN: 0004-637X, 1538-4357. DOI: 10.1088/0004-637X/743/2/121. URL: http://arxiv.org/abs/1109.0639 (visited on 08/17/2024).

[Shi+11] Hyunjin Shim et al. "z~4 Halpha Emitters in GOODS : Tracing the Dominant Mode for Growth of Galaxies". In: *The Astrophysical Journal* 738.1 (Sept. 2011). arXiv:1103.4124 [astro-ph], p. 69. ISSN: 0004-637X, 1538-4357. DOI: 10.1088/0004-637X/738/1/69. URL: http://arxiv.org/abs/1103.4124 (visited on 08/17/2024).

[Wel+11] A. van der Wel et al. "Extreme Emission Line Galaxies in CANDELS: Broad-Band Selected, Star-Bursting Dwarf Galaxies at z>1". In: *The Astrophysical Journal* 742.2 (Dec. 2011). arXiv:1107.5256 [astro-ph], p. 111. ISSN: 0004-637X, 1538-4357. DOI: 10.1088/0004-637X/742/2/111. URL: http://arxiv.org/abs/1107.5256 (visited on 08/17/2024).

[Amb+15] Sivaram Ambikasaran et al. *Fast Direct Methods for Gaussian Processes*. arXiv:1403.6015 [astro-ph, stat]. Apr. 2015. DOI: 10.48550/arXiv.1403.

6015. URL: http://arxiv.org/abs/1403.6015 (visited on 08/26/2024).

[Iye+19] Kartheik G. Iyer et al. "Non-parametric Star Formation History Reconstruction with Gaussian Processes I: Counting Major Episodes of Star Formation". In: *The Astrophysical Journal* 879.2 (July 2019). arXiv:1901.02877 [astro-ph], p. 116. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/ab2052. URL: http://arxiv.org/abs/1901.02877 (visited on 08/29/2024).

[Dav+23] Kelcey Davis et al. *A Census from JWST of Extreme Emission Line Galaxies Spanning the Epoch of Reionization in CEERS*. arXiv:2312.07799 [astro-ph]. Dec. 2023. DOI: 10.48550/arXiv.2312.07799. URL: http://arxiv.org/abs/2312.07799 (visited on 07/09/2024).

[Eis+23a] Daniel J. Eisenstein et al. *Overview of the JWST Advanced Deep Extragalactic Survey (JADES)*. arXiv:2306.02465 [astro-ph]. June 2023. DOI: 10.48550/arXiv.2306.02465. URL: http://arxiv.org/abs/2306.02465 (visited on 07/09/2024).

[Eis+23b] Daniel J. Eisenstein et al. *The JADES Origins Field: A New JWST Deep Field in the JADES Second NIRCam Data Release*. arXiv:2310.12340 [astro-ph]. Oct. 2023. DOI: 10.48550/arXiv.2310.12340. URL: http://arxiv.org/abs/2310.12340 (visited on 08/27/2024).

[Lon+23] Arianna S. Long et al. *Efficient NIRCam Selection of Quiescent Galaxies at $3 < z < 6$ in CEERS*. arXiv:2305.04662 [astro-ph]. June 2023. DOI: 10.48550/arXiv.2305.04662. URL: http://arxiv.org/abs/2305.04662 (visited on 08/17/2024).

[Mor+23] Alexa M. Morales et al. *Rest-Frame UV Colors for Faint Galaxies at $z \sim 9-16$ with the \textit{JWST} NGDEEP Survey*. arXiv:2311.04294 [astro-ph]. Nov. 2023. DOI: 10.48550/arXiv.2311.04294. URL: http://arxiv.org/abs/2311.04294 (visited on 08/26/2024).

[Rie+23] Marcia J. Rieke et al. *JADES Initial Data Release for the Hubble Ultra Deep Field: Revealing the Faint Infrared Sky with Deep JWST NIRCam Imaging*. arXiv:2306.02466 [astro-ph]. Sept. 2023. DOI: 10.48550/arXiv.2306.02466. URL: http://arxiv.org/abs/2306.02466 (visited on 08/27/2024).

[Bun+24] Andrew J. Bunker et al. *JADES NIRSpec Initial Data Release for the Hubble Ultra Deep Field: Redshifts and Line Fluxes of Distant Galaxies from the Deepest JWST Cycle 1 NIRSpec Multi-Object Spectroscopy*. arXiv:2306.02467 [astro-ph]. May 2024. DOI: 10.48550/arXiv.2306.02467. URL: http://arxiv.org/abs/2306.02467 (visited on 08/27/2024).

[DEu+24] Francesco D'Eugenio et al. *JADES Data Release 3 – NIRSpec/MSA spectroscopy for 4,000 galaxies in the GOODS fields*. arXiv:2404.06531 [astro-ph]. Apr. 2024. DOI: 10.48550/arXiv.2404.06531. URL: http://arxiv.org/abs/2404.06531 (visited on 07/26/2024).

[Hai+24] Kevin N. Hainline et al. *The Cosmos in its Infancy: JADES Galaxy Candidates at $z > 8$ in GOODS-S and GOODS-N*. arXiv:2306.02468 [astro-ph]. Jan. 2024. DOI: 10.48550/arXiv.2306.02468. URL: http://arxiv.org/abs/2306.02468 (visited on 08/27/2024).

[Lle+24] M. Llerena et al. *Physical properties of extreme emission-line galaxies at $z\sim 4-9$ from the JWST CEERS survey*. en. arXiv:2403.05362 [astro-ph]. Mar. 2024. URL: http://arxiv.org/abs/2403.05362 (visited on 07/15/2024).

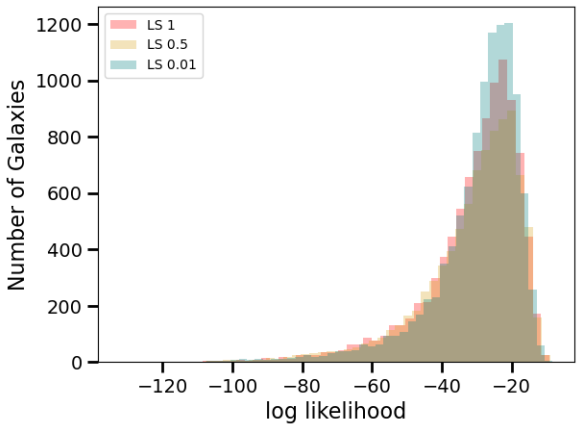

Figure 4: A histogram of the distribution of log likelihoods based on the best initial length scale. All three initial length scales give a similar distribution of log likelihood. This means that all three length scales give good fits and a uniform initial length scale may not be appropriate.

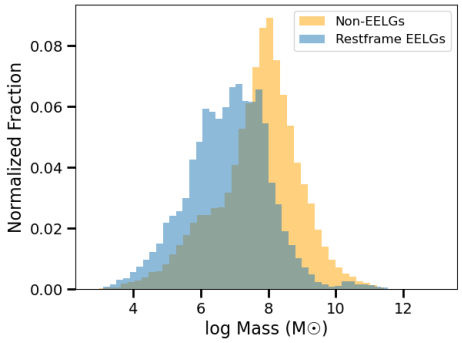

Figure 5: The log of mass in solar masses vs. the normalized fraction of galaxies in our sample that have this mass. EELGs are shown in blue, and non-EELGs are shown in orange. We can see that EELGs have a lower mode in mass, meaning that they are less massive than the general population of galaxies.

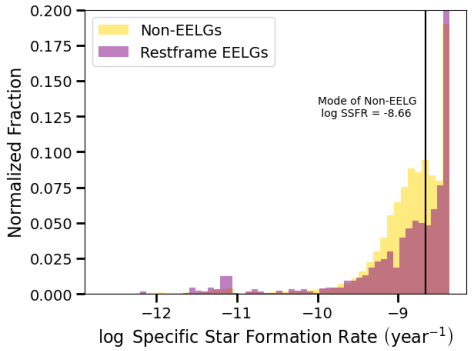

Figure 6: The log of specific star formation rate (SSFR) vs normalized fraction of galaxies in our sample that have this SSFR. EELGs are shown in yellow, and non-EELGs are shown in purple. This data was acquired using EAZY, and the templates we used have limitations, causing a "bunching up" of SSFRs towards the right side of the graph. Nonetheless, it is seen that that non-EELGs have a mode SSFR of about -8.66, while EELGs do not have a peak in the domain of our SSFR values. This tells us that EELGs most likely have higher star formation rates than the general population of galaxies. This is consistent with the idea that EELG's strong emission lines may be caused by extreme star formation [Lle+24].

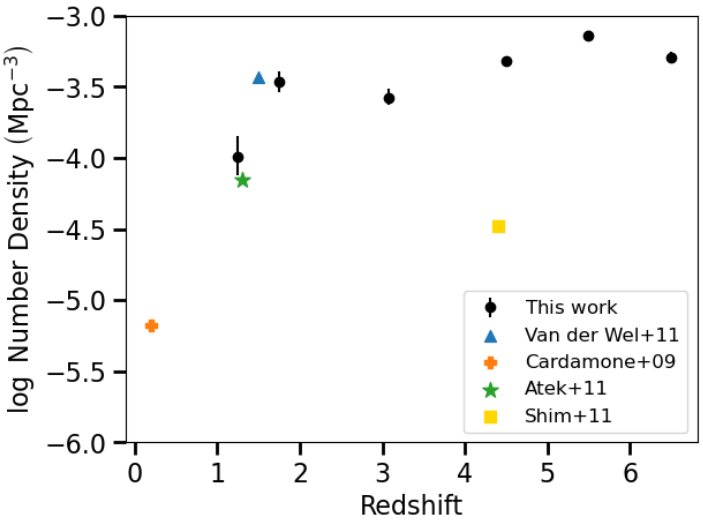

Figure 7: Redshift vs. log number density of EELGs. This work is shown in black circles with error bars (some errors are within the black circles). [Wel+11]'s number density calculation is shown as the blue triangle, [Car+09]'s is shown as the orange plus, [Ate+11]'s is shown as the green star, and [Shi+11]'s is shown as the yellow square. We can see that EELGs are much more abundant at high redshifts than at low redshifts.