The Astrophysical Journal, 919:143 (29pp), 2021 October 1 https://doi.org/10.3847/1538-4357/ac1105

The Mass–Metallicity Relation at $z \sim 1$–2 and Its Dependence on the Star Formation Rate

Alaina Henry[1,2], Marc Rafelski[1,2], Ben Sunnquist[1], Norbert Pirzkal[1], Camilla Pacifici[1], Hakim Atek[3], Micaela Bagley[4], Ivano Baronchelli[5], Guillermo Barro[6], Andrew J Bunker[7,8], James Colbert[9], Y. Sophia Dai[10], Bruce G. Elmegreen[11], Debra Meloy Elmegreen[12], Steven Finkelstein[4], Dale Kocevski[13], Anton Koekemoer[1], Matthew Malkan[14], Crystal L. Martin[15], Vihang Mehta[16], Anthony Pahl[14], Casey Papovich[17], Michael Rutkowski[18], Jorge Sánchez Almeida[19], Claudia Scarlata[16], Gregory Snyder[1], and Harry Teplitz[9]

[1] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[2] Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
[3] Institut d'astrophysique de Paris, CNRS UMR7095, Sorbonne Université, 98bis Boulevard Arago, F-75014 Paris, France
[4] Department of Astronomy, The University of Texas at Austin, Austin, TX 78712, USA
[5] Dipartmento di Fisica e Astronomia, Universitá di Padova, Vicolo dell'Osservatorio, I-3 31522, Padova, Italy
[6] Department of Physics, University of the Pacific, 3601 Pacific Avenue, Stockton, CA 95211, USA
[7] Sub-department of Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
[8] Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
[9] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
[10] Chinese Academy of Sciences South America Center for Astronomy (CASSACA), NAOC, 20A Datun Road, Beijing 100101, People's Republic of China
[11] IBM Research Division, T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA
[12] Department of Physics & Astronomy, Vassar College, Poughkeepsie, NY 12604, USA
[13] Department of Physics and Astronomy, Colby College, Waterville, ME 04901, USA
[14] Department of Physics and Astronomy, UCLA, Los Angeles, CA 90095-1547, USA
[15] Department of Physics, University of California, Santa Barbara, CA 93106, USA
[16] Minnesota Institute for Astrophysics, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA
[17] Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843-4343, USA
[18] Department of Physics and Astronomy, Minnesota State University, Mankato, MN 56001, USA
[19] Instituto de Astrofísica de Canarias, La Laguna, Tenerife E-38200, Spain
Received 2021 February 16; revised 2021 June 2; accepted 2021 July 1; published 2021 October 5

Abstract

We present a new measurement of the gas-phase mass–metallicity relation (MZR) and its dependence on star formation rates (SFRs) at $1.3 < z < 2.3$. Our sample comprises 1056 galaxies with a mean redshift of $z = 1.9$, identified from the Hubble Space Telescope Wide Field Camera 3 (WFC3) grism spectroscopy in the Cosmic Assembly Near-infrared Deep Extragalactic Survey and the WFC3 Infrared Spectroscopic Parallel Survey. This sample is four times larger than previous metallicity surveys at $z \sim 2$ and reaches an order of magnitude lower in stellar mass ($10^8\,M_\odot$). Using stacked spectra, we find that the MZR evolves by 0.3 dex relative to $z \sim 0.1$. Additionally, we identify a subset of 49 galaxies with high signal-to-noise (S/N) spectra and redshifts between $1.3 < z < 1.5$, where Hα emission is observed along with [O III] and [O II]. With accurate measurements of SFR in these objects, we confirm the existence of a mass–metallicity–SFR (M–Z–SFR) relation at high redshifts. These galaxies show systematic differences from the local M–Z–SFR relation, which vary depending on the adopted measurement of the local relation. However, it remains difficult to ascertain whether these differences could be due to redshift evolution, as the local M–Z–SFR relation is poorly constrained at the masses and SFRs of our sample. Lastly, we reproduced our sample selection in the IllustrisTNG hydrodynamical simulation, demonstrating that our line flux limit lowers the normalization of the simulated MZR by 0.2 dex. We show that the M–Z–SFR relation in IllustrisTNG has an SFR dependence that is too steep by a factor of around 3.

Unified Astronomy Thesaurus concepts: Metallicity (1031); Galaxy chemical evolution (580)

Supporting material: machine-readable table

1. Introduction

The role of the baryon cycle in galaxy evolution is one of the most critical questions facing studies of galaxy evolution. The growth of galaxies—and the properties that we observe today—are determined by gas accretion from the intergalactic medium and circumgalactic medium (CGM), as well as star formation feedback that heats and removes gas from the interstellar medium (ISM). Theoretical studies of galaxy formation, both from hydrodynamical simulations and semianalytic models, must balance these processes to produce realistic properties of galaxies over cosmic history (Somerville & Davé 2015).

One of the key constraints on galaxy formation models is the correlation between stellar mass and galaxy metallicities (both stellar and gas-phase[20]), known as the mass–metallicity relation (MZR). Because metals are produced by star formation, their production is also regulated by feedback. A large number of models, from analytical approaches (Dalcanton 2007; Erb 2008; Peeples & Shankar 2011; Davé et al. 2012; Dayal et al. 2013; Lilly et al. 2013; Yabe et al. 2015b) to semianalytical models (Lu et al. 2014; Croton et al. 2016) and hydrodynamical simulations (Finlator & Davé 2008; Ma et al. 2016; Davé et al. 2017, 2019; De Rossi et al. 2017; Torrey et al. 2018, 2019), have all modeled the MZR. What is clear is that a large number of physical processes may

[20] We hereafter define "metallicity" to mean gas-phase oxygen abundance throughout the paper unless otherwise noted.

contribute to the shape, normalization, and evolution of the MZR. Supernova-driven outflows are usually considered a key ingredient, setting the slope of the MZR at low masses (Tremonti et al. 2004; Henry et al. 2013a, 2013b; De Rossi et al. 2017). These outflows may be mixed with the ISM or metal-enriched from the sites of star formation and supernovae (e.g., Chisholm et al. 2018). Other studies have argued that the evolution of the MZR is largely tied to increasing gas fractions in lower-mass and higher-redshift systems (Ma et al. 2016; Torrey et al. 2019). Feedback from active galactic nuclei (AGN) in simulations has also been shown to play an important role in shaping the MZR at high masses (De Rossi et al. 2017). Complicating all of these effects, gas is likely recycled through the CGM; ejected gas may not leave the halo and reaccreted gas may be metal-deficient or metal-enriched (Sánchez Almeida et al. 2014).

Nonetheless, evidence of an interplay between inflows, star formation, metal production, and feedback is observed. The scatter in the MZR is shown to correlate with star formation rates (SFRs) and (in some cases) gas fractions in galaxies out to $z \sim 2$ (Ellison et al. 2008; Lara-López et al. 2010, 2013; Mannucci et al. 2010, 2011; Cresci et al. 2012; Andrews & Martini 2013; Bothwell et al. 2013; Henry et al. 2013a; Salim et al. 2014, 2015; Sanders et al. 2018; Gillman et al. 2021). At fixed stellar mass, galaxies with higher SFRs show lower metallicities, while those with lower SFRs have more enriched ISM. Mannucci et al. (2010) dubbed this relation the "fundamental metallicity relation" (FMR). Using the limited data available at the time, they argued that the FMR did not evolve out to $z \sim 2.5$; galaxies merely evolve within it. From a theoretical perspective, the interpretation that the FMR is produced by variations in accretion, SFR, gas fractions, and feedback is supported by both semianalytic models and numerical simulations (Yates et al. 2012; Dayal et al. 2013; Davé et al. 2017, 2019; De Rossi et al. 2017; Torrey et al. 2018, 2019; De Lucia et al. 2020).

In the low-redshift universe ($z \lesssim 0.3$), the MZR and FMR have been well characterized down to approximately $10^8\,M_\odot$, using the large sample of galaxies covered by the spectroscopic Sloan Digital Sky Survey (SDSS; Tremonti et al. 2004; Andrews & Martini 2013; Brown et al. 2016; Curti et al. 2020). Additionally, extensions of the MZR to low masses have been made using nearby galaxies (Lee et al. 2006; Berg et al. 2012; Izotov et al. 2012; Ly et al. 2016). Moreover, at intermediate redshifts, obtaining statistical samples is still practical. The MZR derived from optical multiobject spectroscopic surveys reaches $M_* \sim 10^8$–$10^9\,M_\odot$ for samples of around 1000 galaxies at $z \sim 0.5$–1.0 (Zahid et al. 2011; Maier et al. 2015; Guo et al. 2016a). However, at $z > 1$, observational constraints on the evolution of the MZR require near-infrared spectroscopy and have therefore been slower to develop (but see early single object spectroscopy studies; Erb et al. 2006; Maiolino et al. 2008; Wuyts et al. 2012; Belli et al. 2013; Masters et al. 2014). Nevertheless, in recent years, infrared grism spectroscopy with the Wide Field Camera 3 on the Hubble Space Telescope (MacKenty et al. 2010), along with ground-based multiobject infrared spectrometers, has opened up a new window on the evolution of galaxy metallicities, extending measurements of statistical samples to $z \sim 2$ (Yabe et al. 2012, 2015a; Henry et al. 2013b; Zahid et al. 2014; Salim et al. 2015; Sanders et al. 2015, 2018; Grasshorn Gebhardt et al. 2016; Kacprzak et al. 2016; Wuyts et al. 2016; Kashino et al. 2017; C. Papovich et al. 2021, in preparation; Topping et al. 2021; Sanders et al. 2021). Notably, numerous studies show that a mass–metallicity–SFR (M–Z–SFR) relation, similar to the local FMR, is also present out to $z \sim 2$.

One of the challenges faced by high-redshift ($z \sim 1$–2) surveys is limited sensitivity, which leads to a survey design that favors brighter, higher-mass targets. As such, the limiting mass reached by current ground-based studies is around 10^9–$10^{10}\,M_\odot$ at $z > 1$ (Salim et al. 2015; Sanders et al. 2015, 2018; Yabe et al. 2015a; Kacprzak et al. 2016; Kashino et al. 2017; Topping et al. 2021; Sanders et al. 2021). However, as we showed in Henry et al. (2013b), without preselection, slitless grism spectroscopy with WFC3 can measure metallicities from galaxies an order of magnitude smaller in stellar mass. Critically, it is at the lowest masses that the effects of star formation feedback are the most prominent, as ejected ISM can more easily escape the weak gravitational potential of dwarf galaxies. Consequently, extending measurements of the MZR and M–Z–SFR relation to lower masses can provide more powerful constraints on models of galaxy formation. Therefore, the goal of this paper is to provide new measurements of galaxy metallicity evolution at masses ten times lower than in recent ground-based studies using statistical samples.

In this paper, we build on our previous work in Henry et al. (2013b), where we presented the first MZR from WFC3 IR grism spectroscopy. Our earlier measurement was based on stacking spectra of 83 galaxies at $1.3 < z < 2.3$, drawn from the WFC3 Infrared Spectroscopic Parallel (WISP) Survey (Atek et al. 2010; Colbert et al. 2013) and the grism coverage of the Hubble Ultra Deep Field (HUDF; van Dokkum et al. 2013). The advantage of the WISP Survey, in particular, is the inclusion of both WFC3 IR grisms, G102 and G141. The wavelength coverage from $0.85\,\mu\mathrm{m} < \lambda < 1.7\,\mu\mathrm{m}$ allows for metallicity measurements using [O II], [O III], and Hβ over a wide range of redshifts ($1.3 < z < 2.3$) and consequently larger samples than G141 alone ($2.0 < z < 2.3$ only).

Since Henry et al. (2013b), available WFC3 grism samples have grown dramatically. In the WISP survey, the area covered by both WFC3 IR bands (G102 and G141) has increased by more than a factor of 3. Likewise, in the well-studied fields of the Cosmic Assembly and Near-Infrared Deep Extragalactic Legacy Survey (CANDELS; Grogin et al. 2011; Koekemoer et al. 2011), fully reduced 3D-HST spectroscopic data (G141) are now available, including extensive redshift catalogs (Momcheva et al. 2016). Furthermore, G102 coverage of GOODS-N (HST-GO 13420), as well as the CANDELS Lyα Emission at Reionization Survey (CLEAR; Estrada-Carpenter et al. 2019; R. Simons et al. 2021, in preparation), expands the sample to include the $1.3 < z < 2.0$ range over a subset of CANDELS. Altogether, these data increase the sample size from Henry et al. (2013b) by more than a factor of 10, allowing us to address evolution over $1.3 < z < 2.3$ and measure the M–Z–SFR relation at these redshifts. Moreover, our new sample provides a meaningful constraint on theoretical models. For the first time, we carry out a realistic comparison between observations and theory by emulating our sample selection in the IllustrisTNG hydrodynamical simulation (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018; Pillepich et al. 2018; Springel et al. 2018).

This paper is organized as follows. In Section 2 we review the data included in this study, describing our sample selection and measurements on individual spectra. We describe our mass

measurements, identification, and removal of AGN, and provide an overview of the sample characteristics. In Section 3 we describe our methods for analysis, including our technique for creating composite spectra and measuring their emission lines. In this section we also present the arguments behind our choice of metallicity calibration (Curti et al. 2017), and our assessment of the current systematic uncertainties in metallicity measurements. In Section 4 we present our resulting MZR and M–Z–SFR relations; then we compare our results to IllustrisTNG in Section 5. Finally, Section 6 contains our conclusions. The appendices include a description of our equivalent width estimates (Appendix A), an assessment of different methods for dust-correcting stacked spectra (Appendix B), and an outline of our Bayesian method for inferring metallicities (Appendix C). We use AB magnitudes and a Planck 2015 cosmology (Planck Collaboration et al. 2016) throughout.

2. Data and Sample

2.1. Survey Overview

We select galaxies from multiple WFC3 grism spectroscopic surveys that also have multifilter HST imaging. We require spectroscopic coverage from [O II] $\lambda\lambda$3726, 3729 to [O III] $\lambda\lambda$4959, 5007 in order to measure metallicities. This requirement results in the selection of galaxies at $1.3 < z < 2.3$ for fields that have observations in the G102 and G141 grisms, and $2.0 < z < 2.3$ for fields with only G141 spectroscopy. Our science goals require the masses of the galaxies, and therefore we select fields with multiband photometry so that we can conduct spectral energy distribution (SED) fitting. To this end, we select spectroscopic surveys in the CANDELS fields, all of which have significant photometric data (e.g., Skelton et al. 2014), and fields in the WISP survey that include sufficient imaging.

In detail, the five CANDELS fields are covered by multiple imaging and spectroscopic surveys. The photometric data are extensive, with many bands and coverage from the U band through Spitzer/IRAC (Galametz et al. 2013; Guo et al. 2013; Skelton et al. 2014; Nayyeri et al. 2017; Stefanon et al. 2017; Barro et al. 2019). For the spectroscopic data, we use G141 spectroscopy from the 3D-HST survey, which observed GOODS-S, EGS, UDS, and COSMOS to two-orbit depth and included GOODS-N G141 observations from the A Grism Hα SpecTroscopic (AGHAST) Survey (PI Weiner, PID 11600; two-orbit depth). Likewise, the CANDELS survey itself included deep, 21 orbit G141 spectroscopy in GOODS-S in one pointing. In contrast to the excellent G141 coverage, no uniform G102 spectroscopy data exist. Still, we include shallow G102 spectroscopy of GOODS-N (Barro et al. 2019; two-orbit depth) and deep G102 spectroscopy from the CLEAR survey (Estrada-Carpenter et al. 2019; 12 orbit depth) covering parts of GOODS-N and GOODS-S (which also includes multiple position angles to disambiguate contamination from overlapping spectra). If there is overlap in the G102 spectroscopy for a target, then we use the CLEAR data. Because the reduced CLEAR data are not public, we conducted our own reduction of these data (see Section 2.4). In total, these five fields cover an area of around 680 arcmin^2. We refer to these objects as the CANDELS/3D-HST sample.

In addition to observations in the CANDELS fields, the WISP survey includes 483 individual WFC3 pointings obtained in HST's pure parallel mode. In these fields, we limit our usage to the 151 fields that have G102 and G141 spectroscopy and include sufficient HST imaging for SED fits. The imaging that supplements the spectroscopy for these 151 fields comprises IR imaging (typically F110W and F160W) and un-binned optical imaging with WFC3, typically in either 475X and F600LP, or F606W and F814W.[21] Of these 151 fields, 100 also include Spitzer/IRAC channel 1 imaging (Fazio et al. 2004), providing photometric coverage from 0.5 to 4 μm. Although the WISP photometry is not as extensive as in CANDELS, these data combined with the known redshifts of the sources are sufficient to obtain reliable mass estimates. These 151 fields[22] cover around 540 arcmin^2. Typically, the spectroscopy is at least two orbits in G102 and one orbit in G141, although a range of parallel visit lengths are included in the survey (Bagley et al. 2020; M. Bagley et al. 2021, in preparation).

2.2. Sample Selection

We select galaxies with signal-to-noise ratio S/N > 5 in [O III] $\lambda\lambda$4959,5007 in the 3D-HST catalog from Momcheva et al. (2016) and from the WISP emission-line catalog (M. Bagley et al. 2021, in preparation). In addition to this S/N cut, we also require the detection of multiple emission lines in the WISP sample. While Baronchelli et al. (2020) show that machine learning can identify the redshifts of single-line emitters in WISP with around 80% accuracy, we conservatively opt for a more rigorous selection to ensure the correct redshift. We adopt the criterion that at least one additional line must be confirmed by visual inspection. These lines include Hα, Hβ, [O II], or an asymmetric [O III] line profile indicative of blended [O III] $\lambda\lambda$4959,5007. For the CANDELS/3D-HST sample, on the other hand, we allow single-line emitters because the high-quality ancillary data enable photometric redshifts that are sufficient to identify the emission line and confirm the redshift (see Section 2.5). In total, we select a parent sample of 1431 galaxies from the CANDELS fields and 384 galaxies from WISP, which are then reviewed (see Section 2.5). We acknowledge that an [O III]-based selection and the requirement for multiple emission lines in some (but not all) of the sample introduce biases in metallicity. These are addressed in Sections 4.5 and 5.

2.3. Spectroscopy

The CANDELS, 3D-HST, and AGHAST spectra are publicly available from the 3D-HST collaboration on the Mikulski Archive for Space Telescopes (MAST)[23,24] and are described in Momcheva et al. (2016). To these data, we add G102 observations in GOODS-N using the fully reduced spectra presented in Barro et al. (2019). Additionally, we include G102 observations from CLEAR, processing the data using the simulation-based extraction pipeline created to reduce

[21] Binned UVIS data cannot be corrected for charge transfer inefficiency, neither can it be properly calibrated. Therefore, we exclude the 11 WISP fields where UVIS optical imaging was taken in binned mode.

[22] The effective area of a pure parallel grism pointing is reduced from the full WFC3/IR area of 4.6 arcmin^2 to around 3.6 arcmin^2. Emission lines on the right-hand side of the detector are excluded from the WISP catalog because imaging outside the field of view would be required to disambiguate emission lines and zero-order images.

[23] https://archive.stsci.edu/prepds/3d-hst/

[24] https://archive.stsci.edu/prepds/candels/

the Faint Infrared Grism Survey (FIGS) data (Pirzkal et al. 2017). To summarize, a custom background subtraction was performed to remove the dispersed background from each individual observation, correcting for any variation in the background during the course of an exposure. The CLEAR imaging data were astrometrically corrected to match that of the CANDELS mosaics, and these corrections were propagated to the G141 observations. Full simulations of each of the G141 observations were then generated using the available broadband photometry and object footprints in these fields. These simulations were used to produce realistic estimates of any contamination caused by overlapping spectra and to later calculate extraction weights to use for optimal spectral extraction (Horne 1986). The spectra from multiple roll angles in the CLEAR observations were combined, excluding regions of spectral contamination. A detailed and complete description of this process is described in Pirzkal et al. (2017).

The WISP data are also publicly available from the WISP collaboration on MAST,[25] although we also made use of proprietary WISP data products (e.g., drizzled data at different pixel scales, as discussed in Section 2.4). The pipeline for spectroscopic reduction is based on the one described in Atek et al. (2010), which uses the aXe slitless grism reduction package (Kümmel et al. 2009) but is updated to include improved calibrations and methodologies (I. Baronchelli et al. 2021, in preparation). In short, the new pipeline uses Astrodrizzle, custom dark calibrations, and charge transfer efficiency corrections for WFC3/UVIS as described in Rafelski et al. (2015), as well as corrections for scattered light in the WFC3/IR data and improved source detection using all available WFC3/IR direct images. We also made use of the WISP team's upgraded line identifications and flux measurements to select galaxies covering the full set of fields as described in Bagley et al. (2020) and M. Bagley et al. (2021, in preparation).

In all of the above surveys, the resolution of the spectra are low. The G102 and G141 grisms have $R \sim 210$ and 130 for point sources. In practice, the galaxies under consideration in this paper have even lower resolution, with a line spread function that is broadened by the morphology of the source. This results in spectra where [N II] $\lambda\lambda$6548,6583 is completely unresolved from Hα, and the [O III] $\lambda\lambda$4959,5007 doublet is at best marginally resolved.

2.4. Photometry

We require a photometric catalog of all galaxies in our sample to enable measurements of masses in a uniform and consistent fashion using the latest SED fitting methodologies (e.g., Pacifici et al. 2016). We incorporate the photometry catalogs from CANDELS (Galametz et al. 2013; Guo et al. 2013; Nayyeri et al. 2017; Stefanon et al. 2017; Barro et al. 2019), 3D-HST (Skelton et al. 2014), and WISP to create a combined photometry catalog for our selected CANDELS/3D-HST and WISP sources.

For the CANDELS fields we utilize the CANDELS photometric catalogs. We give preference to these measurements, rather than those from 3D-HST, as we found the CANDELS catalogs to be more reliable when including ground-based and Spitzer/IRAC data. This difference is likely due to the CANDELS team's use of isophotal apertures, combined with TPHOT (Merlin et al. 2015, 2016; Nayyeri et al. 2017), rather than the fixed aperture photometry used for the 3D-HST measurements (Skelton et al. 2014). However, our parent sample does include 42 sources without matching CANDELS photometry; for these, we instead use the 3D-HST photometry.

The WISP photometry is obtained from the catalog in M. Bagley et al. (2021, in preparation) and is described in more detail therein. Here we provide a brief overview. All HST images are drizzled onto the same pixel scale optimized for the WFC3/UVIS images ($0\farcs04\,\mathrm{pixel}^{-1}$) and cleaned of bad pixels, chip gaps, and noisy edges. The images are then convolved with a kernel to match the point-spread function (PSF) in the F160W filter, using IRAF's `PSFMATCH`. As part of the WISP reduction pipeline, SExtractor (Bertin & Arnouts 1996) is used to generate a segmentation map from the F110W and F160W detections. This segmentation map is resampled onto the same pixel scale and used as the source definitions. The photometry is performed using `photutils` in Astropy to derive isophotal fluxes in all HST bands (Bradley et al. 2021; Astropy Collaboration et al. 2013, 2018), with local sky subtraction performed in $10''$ rectangular apertures. All source flux is masked out of the sky apertures. SExtractor photometry is also performed on WFC3/IR $0\farcs08\,\mathrm{pixel}^{-1}$ images (prior to PSF matching) in order to define the spectroscopic isophotal extraction region and also to obtain the total magnitudes via SExtractor's MAG_AUTO. Aperture corrections are derived from the difference between MAG_AUTO and the isophotal magnitude in the reddest HST band (generally F160W). Finally, the IRAC photometry is obtained with TPHOT matched to the F160W image. The resultant catalog contains PSF-matched photometry in all filters in various apertures.

We make modifications to the WISP photometric catalog to standardize and clean the measurements of our WISP sources before adding them to our combined catalog as follows. First, we omit IRAC entries for around 10% of sources, where the TPHOT flag values indicate a blended source or a source near the image edge. Second, for the HST photometry, we start with the isophotal PSF-matched photometry and apply a magnitude correction to calculate their corresponding MAG_AUTO magnitudes (not included in the catalog for WFC3/UVIS photometry, only for the WFC3/IR $0\farcs08\,\mathrm{pixel}^{-1}$ images). These MAG_AUTO magnitudes are the total magnitudes that are needed for consistency with the CANDELS photometry but are only provided for the near-infrared photometry in M. Bagley et al. (2021, in preparation).

Third, for a dozen sources with negative fluxes in the WISP catalog, where no individual aperture correction is possible, we apply a magnitude correction to the flux value equal to the median magnitude correction of sources with S/N between 0 and 1 in the given filter. This strategy enables us to apply an aperture correction to negative flux entries and derive upper limits instead of discarding them. Additionally, we omit any entries for sources with a high F160W magnitude error (MAGERR_AUTO_F160W > 15) or those without both isophotal and AUTO F160W coverage (near edges) as these result in poor magnitude corrections. We do not apply this same aperture correction strategy to negative IRAC flux entries because the WISP catalog does not contain any flux measurements for the IRAC channels (just magnitudes), and these negative entries are therefore omitted.

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

The result is a photometric catalog for both the CANDELS fields and the WISP fields with consistent aperture and PSF-matched photometry. This catalog is the input for the SED fitting described in Section 2.6.

2.5. Inspection and Measurement of Individual Grism Spectra

We require accurate line and continuum measurements for every object in order to measure metallicities from individual objects and also to facilitate spectral stacking (see Section 3.1). Therefore, we (A.H. and M.R.) carried out interactive line and continuum fitting for all of the galaxies in the sample, using custom software[26] (M. Bagley et al. 2021, in preparation). Our software interactively fits cubic splines to the continuum, with a separate spline function for G141 and G102 when both are present. This approach is ideal for handling data where contamination from overlapping spectra can produce unusual spectral shapes, even when attempts are made to model and subtract it. In these measurements, the contamination model was not cleaned from the continuum spectra but was instead fit and subtracted with the continuum. Likewise, an interactive inspection of each galaxy in the sample allowed us to mask out regions of contamination from zero-order images or extremely bright overlapping spectra that could not be fit well by our automated method. It also allowed us to adjust the dividing wavelength between G102 and G141 spectra (which overlap slightly), as the optimal transition wavelength varied from object to object. Finally, emission-line fluxes were measured at the same time by fitting Gaussian profiles where the FWHM (in pixels) was held constant among the lines.[27] For this measurement, we fit both lines of the [O III] $\lambda\lambda$4959, 5007 doublet, with a fixed doublet ratio of 2.9:1. However, for [O II] $\lambda\lambda$3726, 3729 and Hα + [N II] $\lambda\lambda$6548, 6583 we fit single Gaussians. In the former case, the doublet is at best marginally resolved for the resolution of our grism spectra; in the latter case, the [N II] lines are generally weak, and the S/N of the emission-line blend is not high enough to detect any non-Gaussianity in the line profile. Importantly, this analysis combined the CLEAR data with those from 3D-HST/AGHAST, ultimately providing a uniform set of line measurements and continuum+contamination subtracted spectra that we can stack.

As part of the interactive fitting, visual inspection of emission lines also ensured a clean sample. While the WISP catalog only includes sources that were verified by at least two people in a previous visual inspection, the CANDELS/3D-HST sample was not previously examined by members of our team. In this step, 35 out of 384 sources were rejected from WISP and 383 out of 1431 sources were removed from the CANDELS/3D-HST sample. The reasons were varied but included severe contamination due to spectral overlap with a very bright source, the absence of any visible emission lines, spectra near the edges of detectors, and cases where emission lines from multiple objects were possibly present in the 1D spectrum. This reanalysis provided a clean sample with 349 galaxies from the WISP survey and 1048 from the five CANDELS/3D-HST fields.

In addition to line fluxes, measurements of emission-line equivalent widths are required so that we can correct the Balmer line flux ratios for stellar absorption when we measure metallicities. We estimate equivalent widths of the lines using broadband photometry to estimate the continuum, as we describe in Appendix A. In this appendix, we show that these measurements agree with spectroscopic measurements of equivalent width with a 1σ scatter of 40%. However, the broadband continuum estimates are more complete for faint objects, so we adopt this method for the sake of consistency within our sample.

We verified our measured [O III] fluxes and flux errors by comparing them to the 3D-HST catalog (Momcheva et al. 2016). The line fluxes that we measured from Gaussian fits were on average 15% lower than those from Momcheva et al. (2016), where the spectral line profile is a convolution of the source morphology and the point-source line-spread function.[28] Likewise, the measurement errors from our fits were, on average, 8% larger than those in the 3D-HST catalogs. Combined, these systematics imply that our measurements yield emission-line S/Ns that are, on average, 80% of what is given in the 3D-HST catalogs. We conclude that this level of agreement is reasonable.

Similarly, we compared our redshift identification to those from the 3D-HST catalog. Here, we followed Momcheva et al. (2016), calculating the normalized absolute median deviation ($\sigma_{\rm NMAD}$) of the difference between our measured redshift and that from 3D-HST: $\Delta z = z_{\rm measured} - z_{\rm 3DHST}$. We take

$$\sigma_{\rm NMAD} = 1.48 \times {\rm median}\left(\left|\frac{\Delta z - {\rm median}(\Delta z)}{1 + z_{\rm measured}}\right|\right), \quad (1)$$

as given by Brammer et al. (2008). We find $\sigma_{\rm NMAD} = 0.0055$ for sources with $H_{160} < 24$. If we restrict this redshift comparison to sources where our method finds [O III] S/N > 5, we find $\sigma_{\rm NMAD} = 0.0042$. Alternatively, for a subset of 85 of our sources with [O III] S/N > 5 and robust spectroscopic redshifts from the MOSFIRE Deep Evolution Field (MOSDEF) Survey (category 6 or 7; Kriek et al. 2015), we calculate $\sigma_{\rm NMAD} = 0.0014$. Much of this uncertainty is attributable to the low resolution of WFC3/IR grism spectroscopy, where a single pixel corresponds to $\Delta z = 0.009$ (for [O III] λ5008). For comparison, Momcheva et al. (2016) report a higher $\sigma_{\rm NMAD}$ of 0.0015–0.0045 when the 3D-HST redshift accuracy is assessed against MOSDEF and other ground-based near-infrared spectroscopic follow-up (e.g., Wisnioski et al. 2015). We speculate that the 3D-HST redshift uncertainties may be higher due to inaccuracies that could be removed with more human supervision of the emission-line-fitting process (see also, Rutkowski et al. 2016).

Our comparison with the MOSDEF spectroscopic redshift catalog verifies the accuracy of the redshifts of the single emission-line objects from the CANDELS/3D-HST sample. Of the 85 sources in common between the two samples, we note only 3 objects with $|\Delta z| > 0.15$. These objects are all in

[26] https://github.com/HSTWISP/wisp_analysis

[27] In slitless grism spectroscopy (and slit spectroscopy where the source does not fill the aperture), the line-spread function is set by the morphology of the source. Hence, the FWHM of the lines is expected to be the same in pixels in G102 and G141. Because the dispersion in G102 is two times higher than in G141, the same FWHM in pixels translates to two times higher spectral resolution in G102.

[28] We compared several of our line profile fits with those from 3D-HST and find that the latter are often slightly too broad. Because the 3D-HST spectral line profiles are fixed by the broadband source morphology, this discrepancy could be an indication that the emission-line regions are more compact than the stellar continuum. Nonetheless, on an object-by-object basis, the line fluxes measured by the two methods agree within the uncertainties.

GOODS-N, with IDs 8537, 13286, and 21398 in the MOSDEF and 3D-HST v4.1 catalogs. Of these, two of the MOSDEF spectra are included in the 2021 January public data release.[29] We reviewed these two spectra and found that the emission-line detections and measured spectroscopic redshifts were not particularly convincing. Therefore, we conclude that we cannot determine at this time whether the MOSDEF or WFC3/IR grism spectroscopic redshifts are correct. In either case, 82/85 objects show excellent agreement, suggesting at least ∼96% accuracy on the grism+photometric redshifts of our *vetted* CANDELS/3D-HST sample.

Finally, as we noted above, the [O III] S/N that we measure in our reanalysis is typically lower than the threshold (S/N > 5) that we used to select the sample. Therefore, we removed 4r galaxies from the WISP sample and 265 galaxies from the CANDELS/3D-HST sample to preserve our S/N > 5 threshold. In total, our final sample comprises 345 galaxies from WISP and 783 galaxies from CANDELS/3D-HST, for a total of 1128 objects. Because we have thoroughly vetted both the WISP and CANDELS/3D-HST spectroscopy, our sample is of much higher quality than the unsupervised 3D-HST spectroscopic catalog.

2.6. SED Fitting: Stellar Mass and SFR

Stellar masses are derived by SED fitting to the photometry described in Section 2.4. We opt to derive stellar masses for our sample instead of using published mass catalogs from the CANDELS or 3D-HST teams (Skelton et al. 2014; Mobasher et al. 2015). This approach has a clear advantage, in that we can apply a uniform approach to the WISP and CANDELS/3D-HST data. Likewise, prior estimates of stellar mass do not account for emission-line contamination to broadband photometry, which can cause masses to be overestimated (Atek et al. 2011). Rederiving stellar masses gives us the opportunity to use the most up-to-date SED fitting methodology, making use of nonparametric star formation histories. Indeed, Lower et al. (2020) show that nonparametric star formation histories produce more accurate stellar masses than parametric models.

We adopt the approach described in Pacifici et al. (2012, 2016). In brief, we generate model SEDs assuming star formation and chemical enrichment histories from a semianalytical model, which allows us to span a wide range of star formation histories. Stellar population models are taken from the 2011 version of Bruzual & Charlot (2003), and nebular emission lines are modeled consistently with the stars (see Pacifici et al. 2012, 2016). The attenuation by dust is computed using a two-component dust model to allow for different dust geometries and galaxy orientations (Charlot & Fall 2000). Fixing the redshifts to those derived from the grism spectroscopy, each model SED is compared to the observed photometry. From this analysis, we derive estimates and confidence ranges for the stellar mass and the SFR using a Bayesian approach. A Chabrier (2003) IMF is assumed. The average 68% confidence range for stellar mass is 0.12 dex for the CANDELS/3D-HST subsample and 0.37 dex for WISP. Likewise, the average 68% confidence interval for the SED-derived SFRs are 0.40 and 0.63 dex for the two subsamples, respectively. The differences between the SED fit uncertainties in WISP versus CANDELS/3D-HST reflect the significantly larger number of filters used in the latter.

2.7. AGN

We aim to remove AGN from our sample so that our metallicity measurements are not contaminated by nonstellar ionizing sources. With the low resolution of our spectra, the low S/N, and the wavelength coverage that does not always reach Hα and [N II], discriminating between star-forming galaxies and AGN can be challenging. The traditional "BPT" diagram (Terlevich et al. 1981; [N II]/Hα versus [O III]/Hβ) cannot be used to distinguish between star-forming galaxies and AGN. Fortunately, there are alternatives that are applicable to our data, each of which we consider here. First, all of the spectra in the CANDELS/3D-HST fields have X-ray observations, as well as Spitzer/IRAC photometry, both of which can identify AGN. Second, we include AGN identified by their *z*-band photometric variability in GOODS-N and GOODS-S (Villforth et al. 2010). Third, we can use the "Mass–Excitation" (MEx) diagram, which replaces the [N II]/Hα ratio in the BPT diagram with stellar mass (Juneau et al. 2011, 2014). Fourth, and finally, at $z < 1.5$, we can consider a modified BPT diagram, using [S II]/ ([N II]+ Hα) versus [O III]/Hβ.

Figure 1 shows the MEx diagram for the sources in the CANDELS/3D-HST portion of our sample, divided into two panels for $1.3 < z < 2$ and $2 < z < 2.3$. Here, we exclude the WISP data, as we aim to validate the MEx diagram using the known AGN in the CANDELS fields. The contours in this figure show the density of galaxies in the low-redshift relation, taken from the MPA/JHU SDSS DR7 catalog.[30] We have overplotted X-ray-identified AGN taken from the CANDELS survey (Kocevski et al. 2018, D. Kocevski et al. 2021, private communication), as well as two AGN identified through their photometric variability in Villforth et al. (2010). We also applied the Spitzer/IRAC color selection from Donley et al. (2012) to identify additional galaxies that were not in the X-ray or variable source catalogs. For the IRAC photometry, we took measurements from the 3D-HST photometric catalog (Skelton et al. 2014), requiring 3σ detections in all four IRAC bands and less than 50% contamination (from blending) to the IRAC photometry. We considered relaxing the infrared photometry selection to include lower S/N, more contamination, or objects falling less than 1σ outside the Donley et al. (2012) color cut. However, many of these additional objects fell well within the star-forming locus in Figure 1, suggesting that a relaxed selection returned mostly false positives.

A comparison of AGN selection methods, including X-ray, IRAC colors, BPT, and MEx has previously been carried out for $z \sim 2.3$ galaxies by Coil et al. (2015). Figure 1 is consistent with their conclusion that the $z \sim 0$ version of the MEx diagram cannot be used at high redshift. Because stellar mass serves as a proxy for [N II]/Hα and metallicity evolves with redshift, the MEx AGN selection should also evolve. Coil et al. (2015) propose, based on their X-ray- and IR-identified AGN, that the $z \sim 0$ AGN threshold should be shifted by around 0.75 dex to higher stellar mass. This shift is similar to the 1 dex shift that we proposed in Henry et al. (2013b); however, because Coil et al. (2015) calibrated the shift based on known AGN, we judge this estimate to be more accurate. We show this modified selection in Figure 1. All of the known IR and X-ray AGN are near or above this MEx threshold, or have lower limits on [O III]/Hβ that are consistent with an MEx AGN classification. Of the photometrically variable AGN, one is an X-ray AGN

[29] http://mosdef.astro.berkeley.edu/for-scientists/data-releases/

[30] https://home.strw.leidenuniv.nl/~jarle/SDSS/

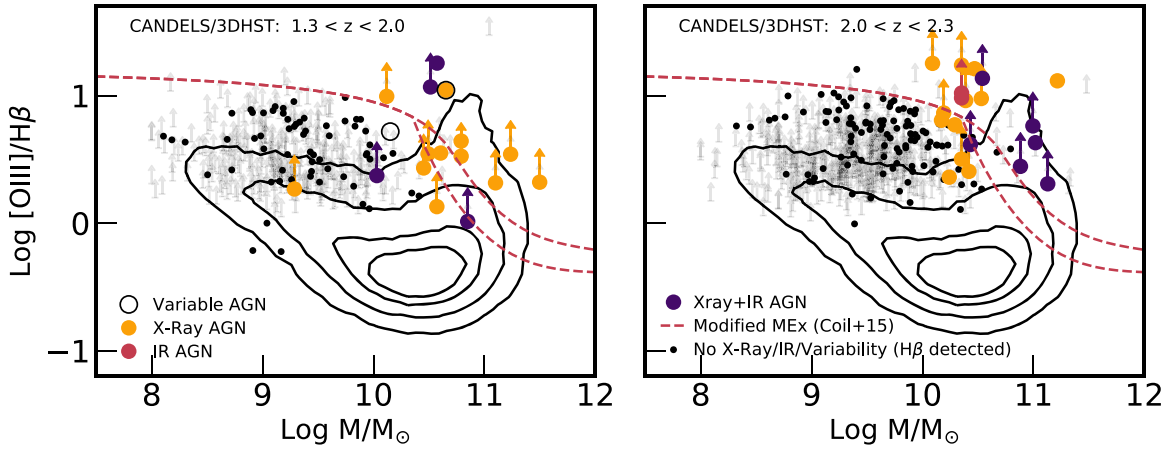

Figure 1. The mass–excitation (MEx) diagram can be used to discriminate between star-forming galaxies and AGN. In this diagnostic, AGN fall above the red dashed line, with the lower and upper curves representing lower and higher probabilities for hosting AGN (see Juneau et al. 2014). Here, we show galaxies from the CANDELS/3D-HST fields only; this approach facilitates a comparison with the known AGN in these fields. These AGN are identified as X-ray-detected objects (Kocevski et al. 2018; D. Kocevski 2019, private communication), objects with nonstellar Spitzer/IRAC colors (Donley et al. 2012), and AGN identified from their variable z-band photometry (GOODS-N and GOODS-S only; Villforth et al. 2010). Several objects meet multiple criteria. The small black points (Hβ detections, $>3\sigma$) and gray arrows ([O III]/Hβ 3σ lower limits) show objects that are not classified as AGN by the X-ray, IR, or variability methods, although they may still be classified as AGN by the MEx method. The contours show the density of low-redshift galaxies and are derived from the JHU/MPA SDSS DR7 catalog. The dashed line to distinguish galaxies from AGN is from Coil et al. (2015), which follows the shape of the $z \sim 0$ relation from Juneau et al. (2014), but is shifted by their recommendation of 0.75 dex to higher mass (to the right) to account for redshift evolution.

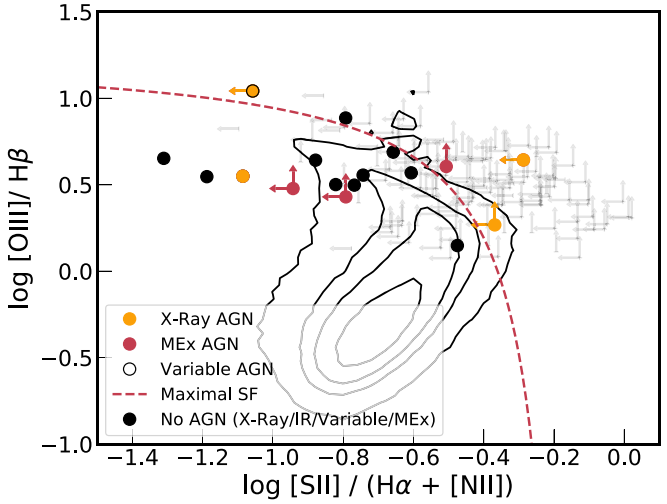

Figure 2. The modified BPT diagram that is measurable from the WFC3/IR grism spectra, for both WISP and CANDELS/3D-HST sources at $1.3 < z < 1.5$. Both lines of the [O III], [S II], and [N II] doublets are included in the line ratios plotted here. X-ray AGN from the CANDELS/3D-HST fields are shown (one of which is variable in its z-band photometry; Villforth et al. 2010), along with MEx AGN from the full survey, selected via the Coil et al. (2015) modified MEx threshold. Black points and gray arrows (3σ upper/lower limits) show galaxies that are not classified as AGN by any method. The dashed curve shows a maximal star-forming threshold, calculated from Cloudy (Ferland et al. 2017) models, as given in Equation (2). The contours are derived from the JHU/MPA SDSS DR7 catalog.

and also falls above the MEx threshold, while the other appears in the star-forming region.

Figure 1 also breaks the sample into two redshift bins—$1.3 < z < 2$ and $2 < z < 2.3$—in case of evolution within our sample. However, we see no need for different MEx thresholds to be applied in the low and high-redshift bins. Because there is only 800 Myr between the mean redshift in the two panels of Figure 1 ($z = 1.68$ versus $z = 2.15$), and we also do not detect significant metallicity evolution in our sample (Section 4.2), we conclude that there is no strong evidence for the evolution of the MEx AGN selection within the redshift range that we consider. In summary, the known AGN confirm the result seen by Coil et al. (2015): the MEx diagnostic with a 0.75 dex shift works reasonably well at $z \sim 2$.

To complement the MEx classification, we also consider a more conventional emission-line-only AGN diagnostic. This test is only possible for the portion of our sample at $1.3 < z < 1.5$, where we have coverage of Hα and [S II]. While the resolution of the grism spectra blends Hα and the [N II] lines, [N II] is likely weak for most galaxies in our sample (Erb et al. 2006), implying that [S II]/(Hα+ [N II]) $\sim$ [S II]/Hα. This AGN diagnostic diagram is plotted in Figure 2. As in Figure 1, we compare to known AGN: X-ray-identified objects and one photometrically variable AGN at this redshift are plotted, along with those that meet the MEx criteria (for both WISP and CANDELS/3D-HST). There are no IR-selected AGN in this redshift range. Lastly, we plot a theoretical maximum for star-forming galaxies, accounting for blended [N II]. We calculated this threshold from the same Cloudy (v17; Ferland et al. 2017) models used in Henry et al. (2018); the maximum is set by the hardest plausible ionizing BPASS v2.0 (Eldridge & Stanway 2016) stellar spectrum, with $Z = 0.001$, an IMF extending to 300 $M_\odot$, and a constant SFR. We estimate that star-forming galaxies fall below a curve following the typical functional form:

$$\log\left(\ [\mathrm{O\ III}]\ 4959,\ 5007/\mathrm{H}\beta\right) = \frac{0.28}{x + 0.14} + 1.27, \quad (2)$$

where

$$x = \log \frac{[\mathrm{S\ II}]\ 6716,\ 6731}{(\mathrm{H}\alpha\ +\ [\mathrm{N\ II}]\ 6548,\ 6583)}. \quad (3)$$

As previously observed by Coil et al. (2015), we find one clear X-ray AGN showing weak [S II] emission and low [O III]/Hβ, placing its BPT measurements squarely in the star-forming region. All of the remaining AGN have limits on their weaker lines, such that they could fall within the star-forming

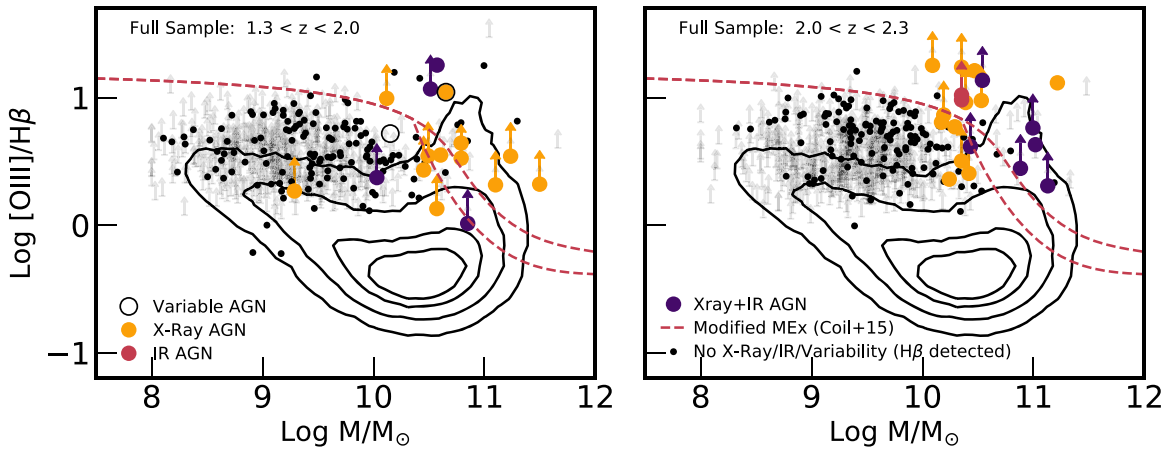

Figure 3. The MEx diagram for the combined WISP and 3D-HST sample. All markers have the same meaning as in Figure 1, but in this case, we also include the galaxies from WISP. For these objects, we do not have additional AGN diagnostics from X-ray or infrared observations. Comparing to Figure 1, we see a clear presence of objects with no prior AGN identification, but lying in the same region as the X-ray- and IR-selected AGN. The MEx diagram identifies these AGN from the WISP survey, where we have no other indicators.

region. Therefore, we concur with the conclusions in Coil et al. (2015): the [S II] BPT diagnostic seems to be a poor method for discriminating between AGN and star-forming galaxies at high redshifts. While the reason for this breakdown is unclear, it should be noted that weak [S II]/Hα in galaxies with otherwise normal [N II]/Hα ratios have been proposed as a means of selecting galaxies with optically thin, Lyman continuum (LyC) leaking ISM (Alexandroff et al. 2015; Wang et al. 2019). When the ISM is density bounded, the outer regions of low-ionization0state gas ([S II], [O II]) can be small or absent. This scenario can also apply to AGN; indeed, Wang et al. (2019) found that two of five LyC emitter candidates at $z \sim 0.3$, when selected to have weak [S II], were actually AGN. Hence, it is plausible that the evolving conditions in the ISM of high-redshift galaxies and AGN may make [S II] an unreliable AGN diagnostic. We therefore use the MEx diagram for both WISP and CANDELS/3D-HST, along with the known X-ray, IR, and photometrically variable AGN in the CANDELS/3D-HST fields. Figure 3 shows this diagnostic diagram for our full sample. We now see many objects in the AGN part of the diagram from the WISP survey, where the MEx diagram is the only available indicator of AGN activity.

Finally, we note that Figures 1–3 reveal that a large number of sources are ambiguous, due to Hβ (and [S II]) nondetections. In particular, the lower limits on [O III] /Hβ and [S II]/Hα are consistent with being both above and below the AGN thresholds. However, at low masses, this problem is not severe. A number of authors have now shown that, when the high-redshift BPT diagram is considered, very few sources at low [N II]/Hα have [O III] /Hβ consistent with AGN (Steidel et al. 2014; Sanders et al. 2016; Strom et al. 2017; Kashino et al. 2019). Because our analysis is based on stacking, a small minority of contaminating AGN will have a negligible impact. However, above $M \sim 10^{10}\,M_{\odot}$, where the MEx curves fall to lower [O III]/Hβ, the nature of the sources with [O III] /Hβ lower limits is less clear. Only 32 galaxies at $M > 10^{10}\,M_{\odot}$ have 3σ Hβ detections and [O III]/Hβ ratios consistent with star formation, while 47 are ambiguous. Because X-ray and IR AGN samples are likely incomplete at all but the highest masses, it is not clear how significant the AGN contamination is for $M > 10^{10}\,M_{\odot}$ in our sample. Therefore, as part of our stacking analysis in Section 3.1, we tested the effect of excluding these 47 galaxies with unknown AGN. In our stack of the 32 galaxies that are clearly star-forming, the ratio of [O III]/Hβ is decreased from 3.60 to 2.61, which corresponds to an increase in metallicity of 0.08 dex. Much of this increase is likely a bias toward galaxies with stronger Hβ emission and consequently lower [O III]/Hβ ratios and higher metallicities. Nonetheless, this test quantifies the possible impact of AGN contamination in our highest-mass stack. At most, an unknown contribution from AGN increases the [O III]/Hβ ratio by 40% and lowers the metallicity by 0.08 dex. Ultimately, while we will assume that galaxies with ambiguous lower limits on [O III]/Hβ are star-forming, we urge caution when interpreting the highest-mass bin in our sample.

Altogether, we remove 72 AGN, of which 63 meet the MEx classification, 12 are IR AGN, 35 X-ray AGN, and 2 have measured variability in their z-band photometry. The remaining sample comprises 1056 galaxies.

3. Analysis

3.1. Composite Spectra

In order to measure metallicity, we require robust measurements of multiple emission lines in our spectra. Therefore, we create composite spectra by stacking to achieve higher S/N. Our procedure is similar to that of Henry et al. (2013b). First, we subtract the continuum that was determined in our interactive fitting (Section 2.5). Then, in order to avoid weighting the stacks toward galaxies with stronger emission lines, we normalized each spectrum by the measured flux of the [O III] $\lambda\lambda$4959, 5007 emission. While this method of normalization does not remove biases owing to a range of dust extinction being present in each stack, we show in Appendix B that the impact on metallicity from this effect is negligible. Next, we deredshift the spectra, using a linear interpolation to shift them onto a common set of rest wavelengths. Finally, we take the median of the normalized fluxes at each wavelength. As a rule, we only consider the Hα + [S II] wavelength region if the stack is restricted to galaxies at $z \leqslant 1.5$, where these lines are covered in the G141 spectra. Figure 4 shows a stack of the entire sample, alongside a stack for the subset at $z \leqslant 1.5$. In addition to providing robust measurements of [O II] and Hβ for metallicity inferences, we detect a handful of weak emission

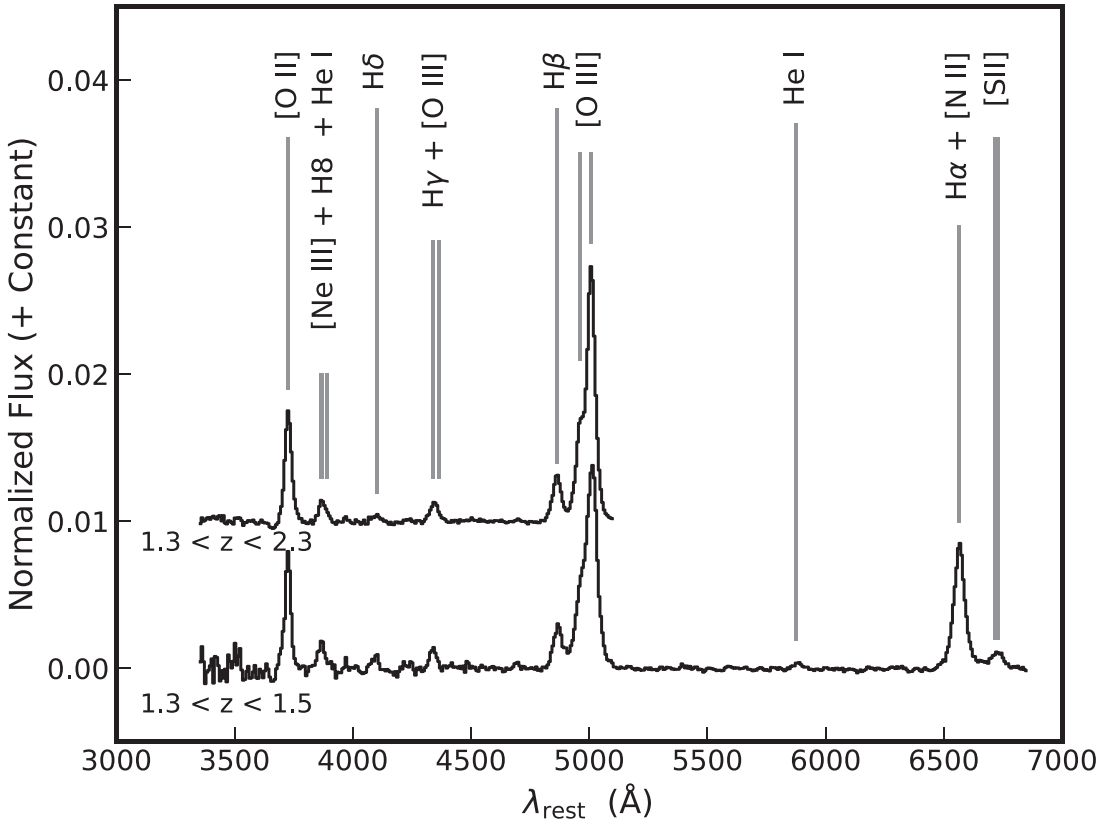

Figure 4. Stacked spectra for the full sample (top; 1056 galaxies), and the subset of the sample with coverage of Hα and [S II] (bottom; 151 galaxies). Numerous weak lines (and blends of lines) are detected.

lines: a blend of Hγ and [O III] λ4363; Hδ; blended [Ne III] λ3869, He I λ3867, and H8 (λ3889); He I λ5875; and blended [S II] $\lambda\lambda$6716, 6731.

The sample was divided into subsets for creating stacked spectra. First, we considered stellar mass, using five bins of 0.5 dex: $\log\ M/M_{\odot} < 8.5$, $8.5 < \log\ M/M_{\odot} < 9.0$, $9.0 < \log\ M/M_{\odot} < 9.5$, $9.5 < \log\ M/M_{\odot} < 10.0$, and $\log\ M/M_{\odot} > 10.0$. Figure 5 shows the composite spectra for these five mass bins. In addition, the large numbers of galaxies in these bins allow us to test subdivision by other properties, keeping the mass bins fixed. Therefore, we made stacks in several subsamples:

1. We tested for evolution, dividing into two bins at $1.3 < z < 1.7$ and $1.7 < z < 2.3$ (discussed in Section 4.2).
2. We tested whether galaxies with higher SED-derived SFRs had lower metallicities than galaxies with lower SFRs. We divided the sample into sources above and below the median SFR in each mass bin (Section 4.3).
3. We created stacks for galaxies in each mass bin with [O III] equivalent widths both above and below the median in that mass bin (Section 4.3).
4. We also explored whether the different sample selections for WISP and CANDELS/3D-HST would result in different metallicities, creating separate sets of stacks for each survey. Because the WISP and CANDELS/3D-HST samples have different mean redshifts, we also created a set of stacks (in the same five mass bins) for each survey, at redshifts above and below 1.7 (Section 4.5).

The number of galaxies in each of the five mass bins for these subsamples is given in Table 1.

In all stacks, the emission-line fluxes were measured by fitting a set of Gaussian profiles to the lines in the stacked spectra. We simultaneously fit [O II], [O III] $\lambda\lambda$ 4959, 5007, Hα (blended with [N II]), Hβ, Hγ (blended with [O III] λ4363), Hδ, [S II], He I 5876, and a blend of lines around Ne III λ3869. Furthermore, we fixed the doublet ratios of [O III] λ5007/[O III] λ4959 to 2.9:1 but did not include separate lines for the closely spaced blends of [O II] $\lambda\lambda$3726, 29, Hα + [N II] $\lambda\lambda$6548, 6583, [S II] $\lambda\lambda$6716, 6731, and Hγ + [O III] λ4363. We also considered whether He I λ6678 might be contributing to the excess flux that is sometimes visible between the Hα + [N II] blend and [S II]. However, this line is intrinsically 3.6 times fainter than He I λ5876 (Porter et al. 2012), which is already very weak in our stacked spectra. Therefore, we conclude that contributions from this line are negligible. Additionally, because the spectral resolution of the stacked spectra is higher at blue wavelengths,[31] we do not require the lines to have the same FWHM, except for closely spaced pairs ([O III] and Hβ, Hα and [S II]). We did, however, require the FWHM of the individual lines to be within a factor of 2 of the [O III] line width. We also allowed a small shift of the emission-line centroids, within ± 10 Å in the rest frame, in

[31] In Section 2.5, we noted that the emission lines in the individual spectra were fit with Gaussians, where the FWHM is the same, in pixels, for all the lines. The G102 grism has a dispersion and spectral resolution two times higher than G141, so this constraint implies that the bluest lines are fit with FWHM, which are two times smaller in angstroms. In the stacked spectra, the wavelengths shortward of Hβ have varying contributions from G102 and G141, resulting in a spectral resolution that increases at blue wavelengths.

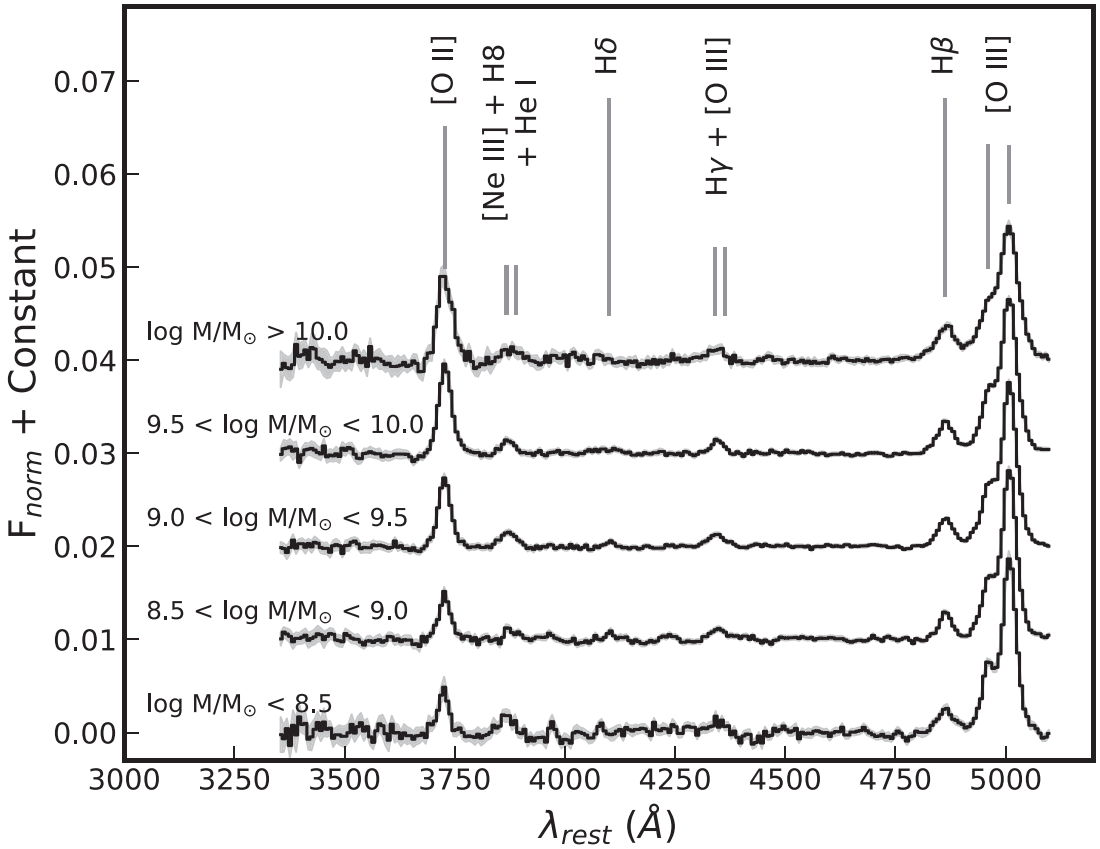

Figure 5. Stacked spectra are shown for five bins of stellar mass. All galaxies between $1.3 < z < 2.3$ with [O III] S/N $> 5\sigma$ are shown. These composite spectra are truncated just beyond [O III], as only the lower-redshift portion of the sample contributes to the longer wavelengths. The gray shaded areas show the 1σ uncertainty (the error on the mean), which is sometimes very small due to the large number of galaxies in the stack.

Table 1
Number of Galaxies in Subsample Stacks

Sample	*N*
All galaxies	68, 223, 379, 307, 79
$1.3 < z < 1.7$	24, 58, 103, 69, 24
$1.7 < z < 2.3$	44, 165, 276, 238, 55
High SFR[a]	34, 111, 189, 153, 39
Low SFR[a]	34, 112, 189, 153, 40
High [O III] equivalent width[b]	34, 111, 189, 153, 39
Low [O III] equivalent width[b]	34, 112, 189, 153, 40
WISP	29, 77, 114, 75, 30
CANDELS/3D-HST	39, 146, 265, 232, 49
WISP, $1.3 < z < 1.7$	13, 30, 49, 29, 17
CANDELS/3D-HST, $1.3 < z < 1.7$	11, 28, 54, 40, 7
WISP, $1.7 < z < 2.3$	16, 47, 65, 46, 13
CANDELS/3D-HST, $1.7 < z < 2.3$	28, 118, 211, 192, 42

Notes. Subsamples that we used to create stacks are given, alongside the number of galaxies in five mass bins for each. The mass bins are the same throughout: $\log M/M_\odot < 8.5$, $8.5 < \log M/M_\odot < 9.0$, $9.0 < \log M/M_\odot < 9.5$, $9.5 < \log M/M_\odot < 10.0$, and $\log M/M_\odot > 10.0$.

[a] The high and low-SFR subsamples are determined by dividing at the median SED-derived SFR in each mass bin, as given in Table 4.

[b] The high and low [O III] equivalent width subsamples are determined by dividing at the median [O III] equivalent width in each mass bin, as given in Table 2.

order to accommodate systematic uncertainties in the grism wavelength solution. Finally, because we subtracted the continuum in the individual spectra, we generally did not need to account for it when measuring the lines in the stacked spectra. However, occasionally we see a small residual continuum that might be present around Hγ + [O III] λ4363. Therefore, we modeled a flat residual continuum, spanning several hundred angstroms, under these particular lines.

Figure 6 shows that, for any given line, a single Gaussian is a poor representation of the line profile in our stacked spectra. The single Gaussian does not reach the peak of the line profile, and it is too narrow in the line wings. This mismatch between the data and the single Gaussian model is characteristic of all of the stacks that we present in this paper. In retrospect, it is not surprising that the individual slitless spectra—whose line profiles are a convolution of the source morphology and the point-source line spread function—do not make a perfect Gaussian profile when they are stacked to reach high S/N. Therefore, we added a second broader Gaussian component to the model fitting in order to get an accurate sum of the line fluxes. The FWHM of the broad component, relative to the narrow component, is required to be the same for all of the lines, while the amplitudes of the broad components are allowed to vary (among positive values). Visual inspection of all of the stacked spectra used in this paper shows excellent fits when this second component is included. Figure 6 shows a representative example of the improvements gained by adding a secondary component.

Measurement uncertainties on the line fluxes in the stacked spectra are obtained by bootstrapping with replacement. In brief, for each sample of *N* galaxies that are stacked, we draw *N* random galaxies from that sample, allowing individual objects to be selected more than once. Then we create a new stack from

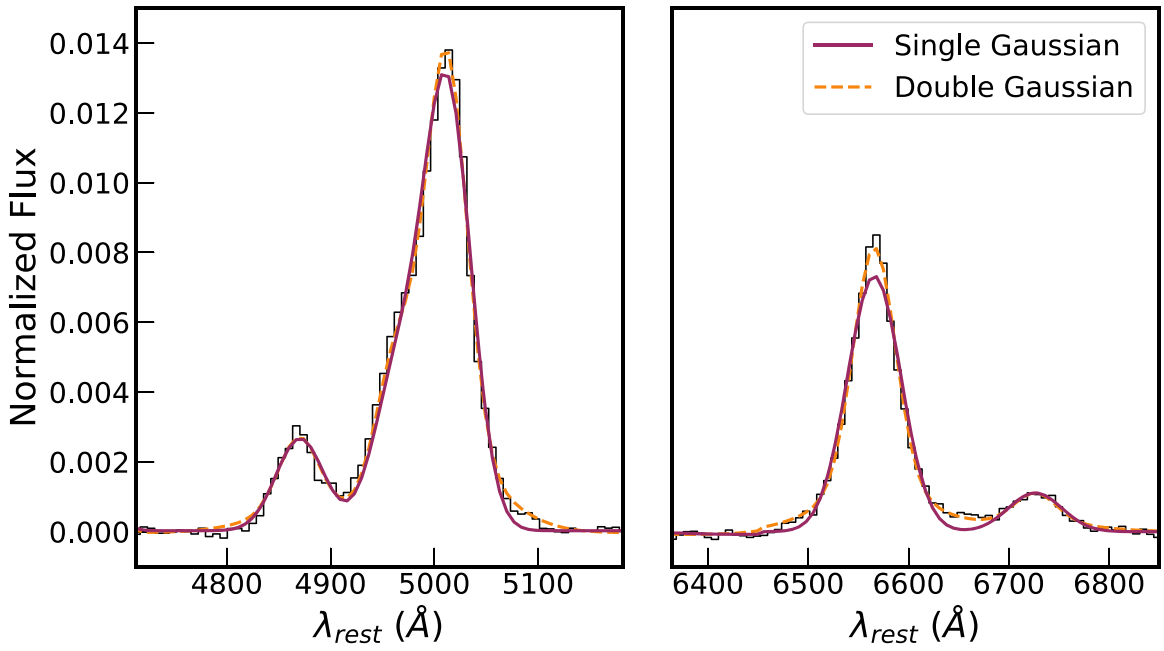

Figure 6. The stacked spectra require two Gaussians for each individual emission line in order to fit the line profiles and provide reliable flux ratios. Here, we show the [O III] and Hβ lines (left) and Hα +[N II] + [S II] lines (right) for all the galaxies in our sample with $z < 1.5$. The single Gaussian fails to reproduce the peaks and wings of the lines, while two Gaussians show a better fit.

Table 2
Measurements from Stacked Spectra

log ($M/M_\odot$) (1)	N (2)	[O III]/Hβ (3)	[O II]/Hβ (4)	[O III]/[O II] (5)	Hγ/Hβ (6)	Hδ/Hβ (7)	F([O III]) (8)	W(Hβ) (9)	W([O III]) (10)
<8.5	68	6.86 ± 1.18	0.88 ± 0.22	7.78 ± 1.43	0.83 ± 0.22	1.16 ± 0.69	7.4	170 ± 54	1210 ± 320
8.5–9.0	223	8.95 ± 0.97	1.39 ± 0.21	6.43 ± 0.65	0.61 ± 0.16	0.23 ± 0.10	8.3	80 ± 14	740 ± 100
9.0-9.5	379	5.98 ± 0.41	1.50 ± 0.12	3.97 ± 0.19	0.48 ± 0.10	0.09 ± 0.03	8.8	57 ± 7	350 ± 40
9.5–10.0	307	5.20 ± 0.26	1.74 ± 0.11	2.99 ± 0.13	0.26 ± 0.05	0.16 ± 0.09	9.9	36 ± 15	200 ± 80
>10.0	79	3.60 ± 0.39	1.31 ± 0.17	2.75 ± 0.22	0.21 ± 0.08	0.03 ± 0.05	14	37 ± 5	140 ± 10

Note. (1) The stellar mass range for each bin. (2) The number of galaxies in each bin. (3)–(7) Flux ratios measured from the stacked spectra shown in Figure 5. Ratios represent observed quantities only and are not corrected for dust or stellar absorption. Ratios involving [O III] and [O II] include both lines of the doublets. The Hγ measurement includes a contribution from [O III] λ4363. (8) The median [O III] flux for the galaxies in the stack, in units of 10^{-17} erg s^{-1} cm^{-2}; both lines of the doublet are included. (9) Inferred rest-frame equivalent width of Hβ emission in angstroms, estimated as described in Appendix A. No correction for stellar absorption is applied. (10) Median rest-frame equivalent width of [O III] for galaxies in the stack, in units of angstroms. The uncertainty represents the error on the mean.

these objects and measure the lines. We repeat this procedure 1000 times and calculate the standard deviation on the line fluxes that are measured from each stack. Measurements from the composite spectra in five mass bins are given in Table 2.

Finally, we note that equivalent widths in Table 2 are estimated for the emission lines in the stacked spectra using broadband photometry and the sample average line fluxes from the stacks. Our method is described in more detail in Appendix A.

3.2. Individual Spectra

While the majority of the sample has low S/Ns, a modest-sized subset has sufficient quality for analysis without stacking. In particular, we find 49 objects with Hα S/N > 10 and $1.3 < z < 1.5$, where we have full spectral coverage from [O II] to Hα. The high S/N of these objects ensures meaningful constraints on Hα/Hβ ratios, dust-corrected Hα luminosities, and SFRs while minimizing Eddington bias from low-S/N sources scattering into the sample (Eddington 1913). While the median S/N on the Hβ emission-line flux is low (~2.5), our analysis method marginalizes over this uncertainty, and the Hα SFRs that we obtain are still more precise than the SED-derived SFRs. Five representative examples of these objects are shown in Figure 7.

The emission-line fluxes for these galaxies are taken from the fitting procedure that we described in Section 2.5, and equivalent widths are measured from broadband photometry (see Appendix A). These measurements are given in Table 3 for a subset of our sample and provided for all of the 49 high-S/N objects in a machine-readable format. To derive nebular gas properties, we use the Bayesian method described in Appendix C. This technique simultaneously constrains dust, metallicity, and contamination of the Hα emission by [N II] while marginalizing over uncertainties due to Balmer line stellar absorption. In particular, metallicity is measured using the Curti et al. (2017) calibration and dust extinction is inferred from the Hα/Hβ ratio, using a Calzetti et al. (2000) extinction curve. Then we use the Kennicutt (1998) calibration to obtain SFRs from Hα luminosities, dividing by 1.8 to convert to a Chabrier (2003) IMF.

Table 3
Measurements and Derived Quantities from Individual Spectra with Hα S/N > 10

ID (1)	R.A. (2)	Decl. (3)	z (4)	[O II] (5)	Hβ (6)	[O III] (7)	Hα+[N II] (8)	W([O III]) (9)	W(Hα) (10)
Par 76–23	201.83432	44.519665	1.3595	48.4 ± 4.8	22.1 ± 6.2	108.0 ± 5.2	74.1 ± 3.4	322	221
GOODS-N 4551	189.14907792	62.16037039	1.3873	8.3 ± 2.3	4.1 ± 2.0	19.6 ± 1.7	16.5 ± 1.1	196	216
GOODS-S 7472	189.32295809	62.1790002	1.3360	8.0 ± 1.6	5.7 ± 1.1	18.0 ± 1.2	13.9 ± 0.6	245	276
Par 377–107	255.382965	64.136452	1.3022	18.4 ± 4.0	9.1 ± 2.8	70.10 ± 2.7	32.4 ± 1.2	1630	755
Par 96–176	32.362873	−4.718067	1.3798	3.1±0.8	3.9 ± 0.9	14.6±1.0	6.3 ± 0.5	841	361

$\log M/M_\odot$ (11)	$\log \mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1}$ (SED) (12)	$\log \mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1}$ (Hα) (13)	$E(B-V)_{\mathrm{gas}}$ (14)	12 + log(O/H) (15)
$10.38^{+0.01}_{-0.03}$	$1.80^{+0.02}_{-0.06}$	$1.66^{+0.15}_{-0.11}$	$0.17^{+0.12}_{-0.12}$	$8.36^{+0.04}_{-0.03}$
$9.88^{+0.02}_{-0.02}$	$1.05^{+0.06}_{-0.02}$	$1.10^{+0.23}_{-0.17}$	$0.26^{+0.19}_{-0.22}$	$8.39^{+0.07}_{-0.07}$
$9.51^{+0.06}_{-0.02}$	$0.65^{+0.14}_{-0.35}$	$0.87^{+0.08}_{-0.07}$	$0.14^{+0.07}_{-0.07}$	$8.39^{+0.04}_{-0.03}$
$9.33^{+0.07}_{-0.04}$	$1.17^{+0.09}_{-0.10}$	$0.23^{+0.17}_{-0.08}$	$0.10^{+0.16}_{-0.10}$	$8.25^{+0.06}_{-0.06}$
$8.74^{+0.05}_{-0.05}$	$0.73^{+0.09}_{-0.02}$	$0.51^{+0.05}_{-0.04}$	$0.00^{+0.04}_{-0.00}$	$8.11^{+0.07}_{-0.09}$

Note. Measurements for the 49 spectra with Hα S/N > 10, as described in Section 3.2. Columns are defined as follows: (1) The field name and object ID. (2)–(3) R.A. and decl., J2000, given in decimal degrees. (4) Redshifts, as measured from the grism spectra. (5)–(8) Line fluxes, in units of 10^{-17} erg s^{-1} cm^{-2}; both lines of the [O III], [O II], and [N II] doublets are included. No corrections for dust extinction or stellar absorption are applied. (9)–(10) Rest-frame equivalent width of the Hα and [O III] emission in angstroms, calculated as described in Appendix A. No correction for stellar absorption is applied. Uncertainties on the emission-line equivalent widths of individual objects are around 40%. (11)–(12) Stellar mass and SFR from SED fits, as described in Section 2.6; (13) SFR derived from Hα, as described in Section 3.2; (14)–(15) Nebular dust extinction and metallicity, derived simultaneously, as described in Appendix C. A measurement uncertainty of zero (in one direction) indicates that the most likely solution was found at the edge of the physically allowed parameter space.

(This table is available in its entirety in machine-readable form.)

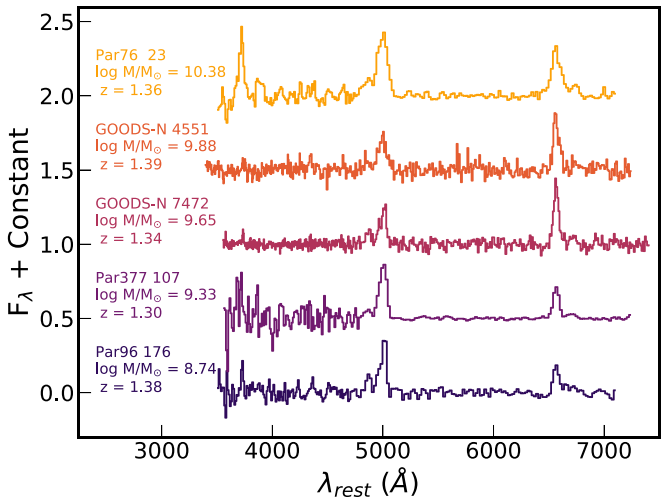

Figure 7. Continuum-subtracted spectra of five objects with Hα S/N > 10. These objects are representative of the 49 individual spectra at $1.3 < z < 1.5$ that we consider in this paper. Three objects—Par 96–176, Par 377–107, and Par 76–23—are from WISP, while GOODS-N 7472 and GOODS-N 4551 have G141 spectra from 3D-HST and G102 spectra from CLEAR. The spectra are in units of 10^{-17} erg s^{-1} cm^{-2} Å^{-1} and are continuum subtracted (with an arbitrary constant added to display them). The spectrum of Par 96–176 is multiplied by a factor of 4 for visualization purposes.

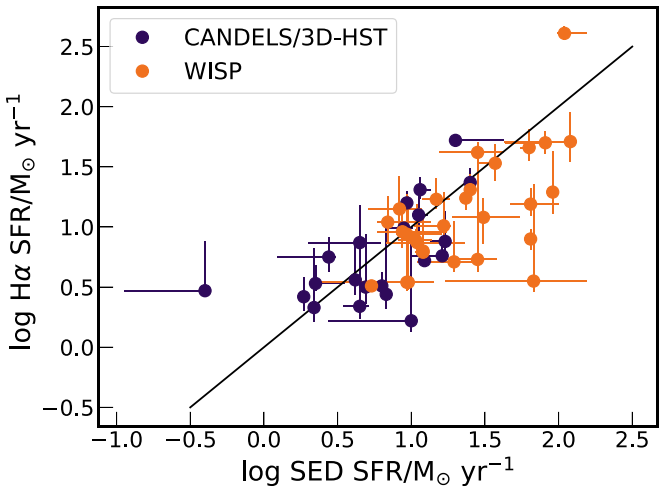

Figure 8. The SED-derived SFRs are compared to Hα derived SFRs for the 49 objects at $1.3 < z < 1.5$ with Hα S/N > 10. Both SFRs are corrected for dust extinction. The black line shows the 1:1 relation. Compared to the CANDELS/3D-HSTHST sample, the WISP measurements are shifted up and to the right: their SED-derived SFRs are systematically 0.25 dex higher than their Hα SFRs, and their Hα SFRs are, on average, 0.35 dex higher than those of the CANDELS/3D-HST objects. Note that the error bars are asymmetric in this plot, and in some cases, the most likely solution is near the 1σ upper or lower bound on the SFR.

Because most of our sample has only SED-derived SFRs, we can use the Hα-derived SFRs to assess the accuracy of the SED-based measurements. In Figure 8, this comparison shows good agreement between the two methods for the CANDELS/3D-HST sample. There is no systematic offset between the two methods; the SED-derived SFRs are larger than the Hα SFRs, on average, by only 0.02 dex. The scatter between the two methods is 0.36 dex, which is somewhat larger than the uncertainties on the SED-derived SFRs (half of the 68% confidence interval on the SED-derived SFRs for CANDELS/3D-HST is 0.2 dex). However, additional scatter can easily be explained by different timescales sampled by Hα emission and continuum emission from young stars (e.g., Lee et al. 2011; Guo et al. 2016b; Mehta et al. 2017; Emami et al. 2019). For the WISP sources, on the other hand, the SED-derived SFRs are systematically higher than the Hα SFRs by 0.25 dex; this difference is not surprising, as the WISP fields have only five

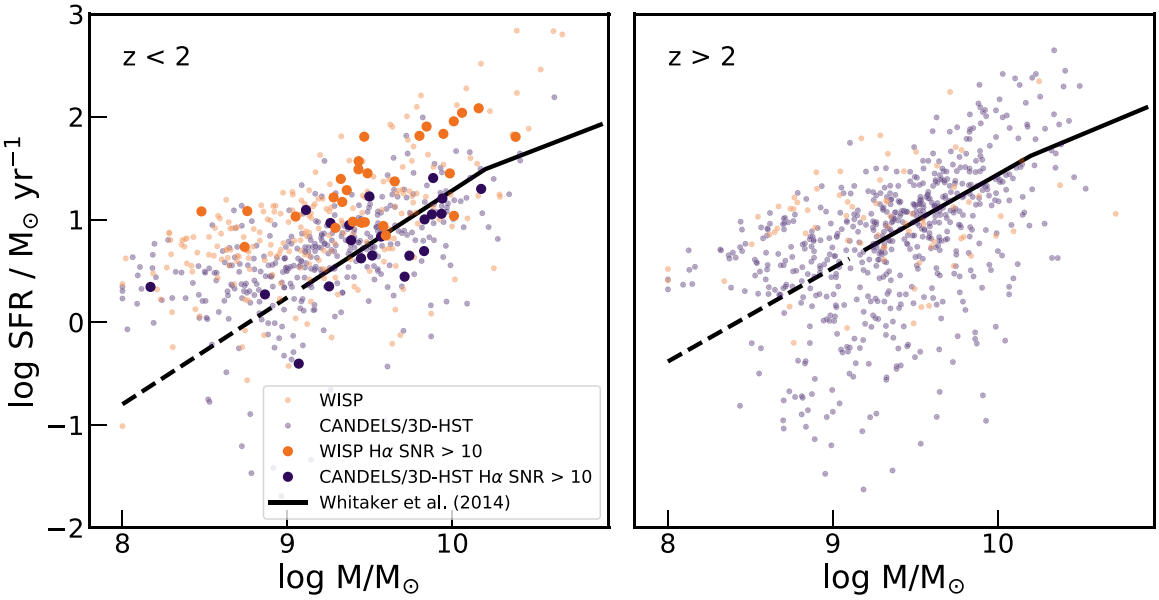

Figure 9. The relationship between SFR and M_* is shown for star-forming objects in our sample at $1.3 < z < 2$ (left), and $2 < z < 2.3$ (right). Both stellar masses and SFRs are derived from SED fits. The dark purple points indicate objects from the CANDELS/3D-HST sample, while the orange points show objects from the WISP survey. The larger points in the left panel highlight the 49 objects with Hα S/N > 10, still showing the SED-derived SFRs. The star-forming main sequence from Whitaker et al. (2014) is shown by the black curves, for $1.5 < z < 2.0$ (left) and $2.0 < z < 2.5$ (right). The dashed line indicates an extrapolation of this measurement.

bands of imaging, compared to the extensively sampled SEDs in the CANDELS/3D-HST fields. For the 49 objects under consideration here, the SED fits tend toward higher extinction in WISP compared to 3D-HST (mean $A_V = 0.40$ versus $A_V = 0.26$). This difference is likely a systematic error in the SED fitting rather than a physical difference in the two samples; propagated to ultraviolet wavelengths, it explains the higher SFRs in the WISP objects. We take the systematic uncertainties on SED-derived SFRs into account when we consider the M–Z–SFR relation from stacked spectra in Section 4.3.

3.3. Sample Characteristics

Figure 9 shows the star-forming main sequence for our sample. Here, the SFRs and masses are derived using SED fitting, as described in Section 2.6. We show $z < 2$ and $z > 2$ in the left and right panels, respectively, and show the WISP and CANDELS/3D-HST samples as gold and dark purple points, respectively. We also highlight the 49 objects at $1.3 < z < 1.5$ with high-S/N spectra discussed in Section 3.2 with larger points (left panel only). These objects do not show any clear difference from the parent sample in the stellar mass–SFR space. We compare our results to the main sequence from Whitaker et al. (2014), which we tentatively extrapolate below $\log M/M_\odot \sim 9.2$. This comparison shows that our full sample is somewhat biased toward higher SFRs, especially at the lowest masses. This bias also appears strongly for WISP galaxies, especially at $z < 1.5$; however, as we showed in Section 3.2, some of this effect may be due to a systematic overestimation of the SED-derived SFRs in the WISP data.

Figures 8 and 9 compare objects from the WISP survey with objects from CANDELS/3D-HST. This distinction shows that the [O III] emission-line selection is different for these two surveys: the objects from the WISP survey have higher SFRs than the 3D-HST galaxies, by an average of 0.35 dex (from the Hα SFRs in Figure 8). This result is due to different selection techniques. For WISP, the objects are required to have clear redshift identification, on the basis of a second line. For CANDELS/3D-HST, on the other hand, single lines are included when their photometric redshift indicates that the line is [O III]. As noted in Section 2.5, this strategy is possible in CANDELS/3D-HST (but not WISP), because the extensive photometry in the CANDELS fields provides robust photometric redshifts. Consequently, relative to CANDELS/3D-HST, the WISP survey is less complete to objects with overall weaker emission lines and lower SFRs.

Figure 10 shows the distributions of [O III] equivalent width, redshift, and stellar mass for the star-forming sample. The [O III] equivalent widths, shown in the left panel, are given for the sum of the λ4959 and λ5007 lines. Histograms are shown for the CANDELS/3D-HST objects, WISP objects, and the 49 objects with Hα S/N > 10. As noted above, the inclusion of single-line emitters with robust photometric redshifts from CANDELS/3D-HST adds objects with lower equivalent widths, whereas the higher equivalent width tail is similar for the two surveys. While the WISP sample appears more biased toward highly star-forming objects, the redshift distribution in the center panel of Figure 10 highlights its importance. Due to the inclusion of the G102 grism in all of the WISP fields, the broader wavelength coverage nearly doubles the available sample at $1.3 < z < 2.0$ (when combined with the GOODS-N and CLEAR G102 data). In comparison, the high-S/N objects are similar to the parent sample in [O III] equivalent width and redshift but are somewhat higher in stellar mass.

Lastly, the distribution of masses in the right panel of Figure 10 gives a sense of the mass completeness of the samples. Previously, Whitaker et al. (2014) reported that the star-forming main sequence from CANDELS imaging was complete to $\log M/M_\odot \sim 9.2$–9.3 at these redshifts. As evident by the turnover in the mass distribution, our sample is roughly consistent with this completeness for both WISP and CANDELS/3D-HST. Hence, at the lowest masses in our sample, we are sensitive to the sources with only the highest SFRs. This quality is also apparent in Figure 9, where the

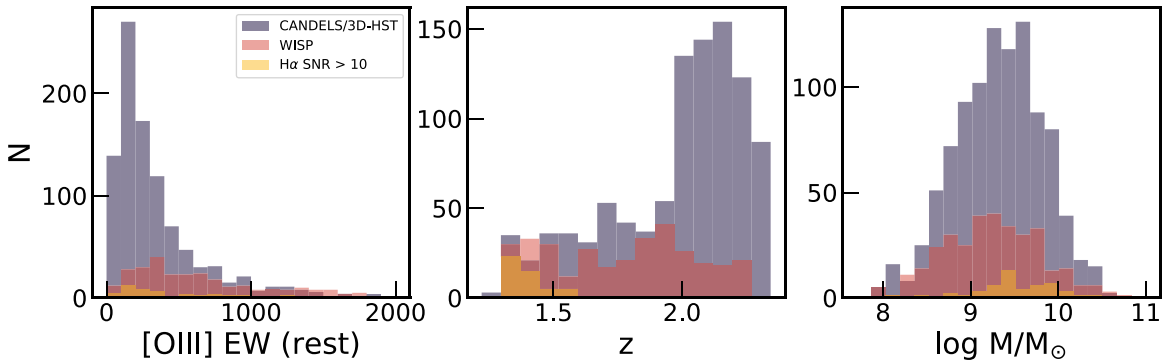

Figure 10. The distribution of rest-frame [O III] equivalent width ($\lambda\lambda$ 4959,5007 summed; left), redshifts (center), and masses (right) are shown. The WISP sample has somewhat different properties than the CANDELS/3D-HST objects, due to observational design and sample selection. The gold histograms show the same properties, but for the 49 objects with Hα S/N > 10, where we obtain constraints on dust and metallicity without stacking. These objects have similar properties to the parent sample but are skewed toward slightly higher masses.

Table 4
Derived Quantities from Stacked Spectra

$\log(M/M_\odot)$ (1)	N (2)	$\langle\log(M/M_\odot)\rangle$ (3)	$\langle\log(\mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1})\rangle$ (4)	$E(B-V)$ (5)	$W(\mathrm{H}\beta)^*$ (6)	[O III] $\lambda4363/\mathrm{H}\gamma$ (7)	$12+\log(\mathrm{O/H})$ (8)
<8.5	68	8.27	0.51	$0.0^{+0.21}_{-0.00}$	$0.0^{+3.75}_{-0.0}$	$0.46^{+0.00}_{-0.24}$	$7.98^{+0.11}_{-0.08}$
8.5–9.0	223	8.77	0.61	$0.0^{+0.13}_{-0.00}$	$5.85^{+0.0}_{-3.45}$	$0.40^{+0.24}_{-0.22}$	$8.07^{+0.04}_{-0.04}$
9.0–9.5	379	9.26	0.80	$0.33^{+0.06}_{-0.06}$	$3.0^{+1.5}_{-1.2}$	$0.28^{+0.18}_{-0.20}$	$8.28^{+0.02}_{-0.02}$
9.5–10.0	307	9.73	1.11	$0.43^{+0.06}_{-0.07}$	$5.85^{+0.0}_{-2.1}$	$0.00^{+0.20}_{-0.00}$	$8.37^{+0.01}_{-0.02}$
>10.0	79	10.20	1.45	$0.66^{+0.06}_{-0.06}$	$3.15^{+1.80}_{-1.35}$	$0.00^{+0.24}_{-0.00}$	$8.45^{+0.01}_{-0.02}$

Note. Derived quantities for stacked spectra in five mass bins. (1) The stellar mass bin for each stack. (2) The number of galaxies in each bin. (3) The mean stellar mass of the galaxies in each bin. (4) The mean SED-derived SFR for the galaxies contributing to the stack. (5)–(8) Properties from our Bayesian inference of nebular dust extinction, stellar Hβ absorption equivalent width, the [O III] $\lambda4363/\mathrm{H}\gamma$ ratio, and metallicity. Errors on each parameter denote the 68% confidence interval and are marginalized over the three parameters that are not under consideration. A measurement uncertainty of zero (in one direction) indicates that the most likely solution was found at the edge of the physically allowed parameter space (see Appendix C for a definition of the allowed parameter space).

sample lies primarily above (albeit, an extrapolation of) the star-forming main sequence at $M \lesssim 10^9\,M_\odot$. We take this incompleteness into account by modeling our sample selection in the IllustrisTNG simulation in Section 5.

3.4. Considerations on Strong-line Metallicity Calibrations

The measurement of gas-phase metallicities from the spectra of galaxies has been a subject of debate. Primarily, two types of calibrations have been used to infer metallicities from strong emission lines: theoretical calibrations, based on photoionization models (Kewley & Dopita 2002; Kobulnicky & Kewley 2004; Strom et al. 2018), and empirical calibrations tied to direct-method metallicities derived from electron temperature (T_e) sensitive auroral lines (Pettini & Pagel 2004; Pilyugin & Thuan 2005; Pilyugin et al. 2012; Pilyugin & Grebel 2016; Curti et al. 2017). Critically, large systematic errors are apparent between the different calibrations, even among the theoretical/empirical classes. Kewley & Ellison (2008) showed that the MZR of SDSS galaxies differs in shape and normalization when different calibrations are used, with a systematic offset as high as 0.7 dex. While it is plausible that the photoionization models represent an oversimplification, some authors have also argued that metallicities based on the direct method are biased, as emission can be dominated by regions with higher T_e (e.g., Stasińska 2002; Bresolin 2007; García-Rojas & Esteban 2007; Peimbert et al. 2007). Other authors have argued that photoionization models can be brought in line with direct-method metallicities if the electron energies follow a κ-distribution, rather than a Maxwell–Boltzmann distribution (Binette et al. 2012; Nicholls et al. 2012; Dopita et al. 2013).

Despite these challenges, recent studies have begun to converge on a range of reasonable calibrations in the local universe. In H II regions and nearby galaxies, comparisons between direct metallicities and supergiant metallicities show similar values (Bresolin et al. 2016; Kudritzki et al. 2016; Davies et al. 2017). These results imply that empirical T_e-based metallicity calibrations (e.g., Pettini & Pagel 2004; Curti et al. 2017) should be preferable to photoionization models.

For high-redshift galaxies, the possibility for evolution of local metallicity calibrations is a cause for concern. It is suspected that the physical conditions in H II regions are different at high redshifts, as high-redshift galaxies are offset from the low-redshift locus in the [N II]/Hα versus [O III]/Hβ line diagnostic diagram (the BPT diagram; Baldwin et al. 1981). This offset implies that, in high-redshift galaxies, metallicities derived from [N II]/Hα will differ from metallicities derived using oxygen-based indicators—even if a self-consistent empirical calibration is used for the two diagnostics (e.g., Maiolino et al. 2008; Curti et al. 2017). Several explanations for this evolution have been suggested, including contamination by AGN (Wright et al. 2009; Trump et al. 2011, 2013), higher electron density (Shirazi et al. 2014), higher ionization potential (Kewley et al. 2013, 2015), harder ionizing spectra coupled with nonsolar O/Fe ratios (Steidel et al. 2014, 2016; Strom et al. 2017, 2018), and elevated N/O ratios (Masters et al. 2014, 2016; Shapley et al. 2015; Strom et al. 2017). These effects could impact metallicity

measurements if low-redshift strong-line calibrations are applied blindly to high-redshift galaxies.

Taking these systematics into account, Strom et al. (2018) derived a new strong-line calibration for high-redshift galaxies. They used photoionization modeling of around 200 galaxies at $z \sim 2$–3 in the Keck Baryonic Structure Survey (Steidel et al. 2014). The key differences between their approach and earlier photoionization modeling (e.g., Kewley & Dopita 2002) were the inclusion of harder ionizing spectra that include binary stars (e.g., BPASS; Eldridge & Stanway 2016; Eldridge et al. 2017), a decoupling of the nebular metallicity and ionizing stellar metallicity (to emulate variations in O/Fe ratios), and allowing the N/O ratio to vary independently of O/H. Strom et al. (2018) derived the metallicity for each galaxy in their sample and then provided relations between their derived metallicity and commonly observed line ratios. While their [N II]/Hα calibration shows some evidence for evolution, their R_{23} calibration[32] is similar to the one derived from local H II regions reported by Pilyugin et al. (2012), after the latter is corrected upwards by 0.24 dex.

Overall, the agreement between the latest photoionization models and nearby H II regions suggests a convergence of oxygen abundance indicators, with systematic uncertainties greatly reduced from the 0.7 dex reported by Kewley & Ellison (2008). Further supporting this conclusion, recent detections of [O III] λ4363 in high-redshift galaxies confirm that low-redshift empirical strong-line calibrations agree with direct-method based measurements, within the uncertainties (Jones et al. 2015a; Gburek et al. 2019; Sanders et al. 2020). These findings suggest that empirically derived strong-line calibrations, tied to direct-method metallicities, are applicable at high redshifts. Given this assessment, we adopt the empirical calibration from Curti et al. (2017), as it is well suited to the Bayesian methodology that we describe in the next section. The R_{23} calibration from Curti et al. (2017) gives metallicities that are around 0.2 dex lower than those from Strom et al. (2018), more in line with the H II regions from Pilyugin et al. (2012). Some of this offset may be attributable to the different handling of dust depletion in photoionization models compared to direct-method measurements.

3.5. Bayesian Inference of Metallicity and Dust Extinction

We use a Bayesian methodology to derive nebular gas properties from our measurements. A detailed description of our calculation is presented in Appendix C. In brief, we model metallicity, dust extinction, Balmer line stellar absorption, and contamination of Hγ by emission from [O III] λ4363. As noted above, metallicities are derived using the Curti et al. (2017) calibration. We do not apply a correction for diffuse ionized gas (DIG), as the contribution is expected to be minimal for highly star-forming, compact galaxies at the redshifts of our sample (Sanders et al. 2017). The dust extinction is calculated from the Balmer decrement, assuming a Calzetti et al. (2000) extinction curve. For sources at $1.3 < z < 1.5$, we use Hα/Hβ, as well as measurements or upper limits on Hγ/Hβ and Hδ/Hβ. For the higher-redshift stacks, we do not have coverage of Hα, so the dust constraints from the Balmer lines alone are poor. In these cases, we adopt a prior based on the lower-redshift measurements of dust extinction in stacked spectra for $1.3 < z < 1.5$ (see Appendix C). We note that the Hβ stellar absorption and the relative strength of [O III] λ4363 are poorly constrained by this method. Therefore, we marginalize over these nuisance parameters to provide realistic uncertainties on metallicity and dust extinction. We do not consider these poorly constrained quantities further.

We applied this methodology to the stacked spectra discussed in Section 3.1 and shown in Figure 5, as well as the 49 individual high-S/N spectra described in Section 3.2. Results for stacks of the full sample, divided into five mass bins, are given in Table 4. Likewise, results for the individual high S/N spectra are highlighted in Table 3 and presented in machine-readable format.

4. Results

In this section, we present the results of our MZR and M–Z–SFR measurements for both stacks and individual galaxies. In Section 4.1, we compare them to previous results at similar redshifts, while in Section 4.2 we discuss the evolution of the MZR. Then, we present the M–Z–SFR relation from stacked spectra in Section 4.3, and the 49 high-S/N individual objects in Section 4.4. Finally, we address biases in our sample selection in Section 4.5.

4.1. The MZR at z ∼ 1–2

Figure 11 shows the MZR that we derive for our stacked spectra at $1.3 < z < 2.3$ and individual galaxies at $1.3 < z < 1.5$. The mean redshift of the full sample is $z = 1.9$. While we aim to compare our measurements with others in the literature at similar redshifts ($z \sim 1$–2), much of this work relies on [N II] (e.g., Erb et al. 2006; Wuyts et al. 2012; Yabe et al. 2015a; Kashino et al. 2017; Gillman et al. 2021). Given the uncertainties surrounding nitrogen abundances in high-redshift galaxies (Masters et al. 2014; Shapley et al. 2015; Strom et al. 2018), we restrict our comparison to oxygen-based indicators. This limits the comparison considerably to data from Sanders et al. (2018), Henry et al. (2013b), and Grasshorn Gebhardt et al. (2016). For these data, we use the published emission-line fluxes (or flux ratios) to recalculate metallicities using the Curti et al. (2017) calibration in order to maintain consistency with our measurements. We use a simple maximum likelihood estimator to derive metallicities and their uncertainties from reported dust- and stellar-absorption-corrected R_{23} and O_{32} ratios[33] and measurement errors. Curiously, the Henry et al. (2013b) and Grasshorn Gebhardt et al. (2016) samples show metallicities that are higher than the present MZR. We believe this to be a sample bias resulting from the requirement for the detection of multiple emission lines, which we discuss further in Section 4.5. Therefore, in Figure 11, we focus on a comparison with the results from Sanders et al. (2018).

Figure 11 shows that our results are in excellent agreement with the stack-based measurements from Sanders et al. (2018) at masses where the samples overlap, even though they have slightly different mean redshifts ($z = 1.9$ for the present data and $z = 2.3$ for Sanders et al. 2018). Critically, we extend the measurement an order of magnitude lower in stellar mass. Additionally, the 49 objects with high-S/N spectra show good

[32] $R_{23} = ([\text{O III}]\ \lambda\lambda 4959, 5007 + [\text{O II}]\ \lambda\lambda 3726, 3729)/\text{H}\beta$

[33] Frequently, O_{32} is defined as $\lambda 5007/[\text{O II}]\ \lambda\lambda 3726, 29$, excluding [O III] λ4959. The choice of definition does not matter, as long as one is self-consistent. In this paper, we use $O_{32} \equiv [\text{O III}]\ \lambda\lambda 4959, 5007/[\text{O II}]\ \lambda\lambda 3726, 39$ since we do not resolve the doublet. The O_{32} metallicity calibration from Curti et al. (2017) is adjusted accordingly.

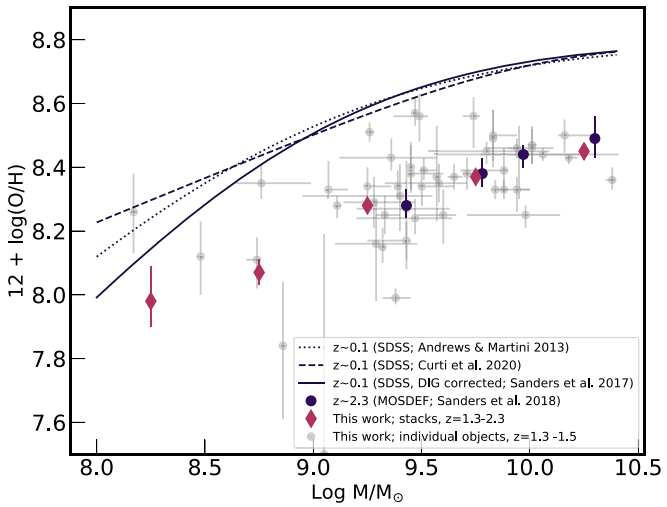

Figure 11. The MZR is shown for our stacked spectra, as well as the objects at $1.3 < z < 1.5$ with Hα S/N > 10. The mean redshift of the sample in the stacks is $z = 1.9$. Our results show excellent agreement with the MOSDEF survey, as reported by Sanders et al. (2018); we have recalculated the metallicities from their stacked measurements, using the Curti et al. (2017) calibration. Other samples at $z \sim 2$ are not shown, as they require the use of [N II]-based metallicity calibrations, which may evolve with redshift (Masters et al. 2014; Shapley et al. 2015; Masters et al. 2016; Strom et al. 2017).

agreement with the stacked spectra, even though they are at a lower redshift ($1.3 < z < 1.5$).

4.2. Redshift Evolution of the MZR

Figure 11 also shows the evolution of our MZR from $z \sim 2$ to $z \sim 0.1$, by comparison with results from the SDSS. We highlight two measurements, both of which are empirically tied to direct-method T_e metallicities: Andrews & Martini (2013), which measured the metallicity in stacked spectra where auroral lines are detected, and Curti et al. (2020), which applies the Curti et al. (2017) strong-line calibration to individual SDSS galaxies. These two SDSS MZR measurements are very similar. We also show an updated MZR reported by Sanders et al. (2017), which corrects the Andrews & Martini (2013) MZR for the DIG (and also to use more recent atomic data; see references in Sanders et al. 2017). This local relation is similar to the Andrews & Martini (2013) and Curti et al. (2020) measurements over most of the mass range in Figure 11, although it has a steeper slope at low masses. Above log $M/M_\odot \sim 8.5$, we find a metallicity evolution of around 0.3 dex. The shape of the $z \sim 2$ MZR appears similar to the low-redshift relation from Andrews & Martini (2013) and Curti et al. (2020), but may be flatter than the DIG-corrected MZR from Sanders et al. (2017). The larger metallicity error in the lowest mass bin makes it difficult to ascertain whether the $z \sim 2$ MZR takes a different shape than the local relation.

Given our large sample size and large redshift range, we also have the ability to measure metallicity evolution within our sample. Therefore, we divided the sample into two redshift bins: $1.3 < z < 1.7$ and $1.7 < z < 2.3$. Each redshift bin is divided into the same mass bins that we used for the whole sample, as indicated in Tables 2 and 4. We see no evidence for metallicity evolution in the redshift range that we probe. The higher- and lower-redshift MZRs are consistent with one another, as well as the MZR for the full sample (Table 4), within the measurement uncertainties. (We do not show these results in tabular form or a figure, as they are indistinguishable from Table 4 and Figure 11). We conclude that any evolution over the redshifts spanned by our sample must be smaller than (or comparable to) our measurement uncertainties. This result is sensible if metallicity evolution is (to first order) linear with time. We measure only 0.3 dex (a factor of 2) of metallicity evolution over the 9 Gyr between $z = 1.9$ and $z = 0.1$, while the time span between the mean redshifts of our high- and low-redshift bins ($z = 1.5$ and $z = 2.0$) is only 1 Gyr. Hence, we might expect a 0.05 dex increase in metallicity between our high- and low-redshift bins, which is indeed comparable to our uncertainties.

4.3. The M–Z–SFR Relation from Stacked Spectra

At lower redshifts, the MZR shows a secondary dependence on SFRs (or gas fractions), such that, at fixed mass, galaxies with higher SFRs (more gas-rich objects) have lower metallicities (Ellison et al. 2008; Lara-López et al. 2010, 2013; Mannucci et al. 2010, 2011; Cresci et al. 2012; Andrews & Martini 2013; Bothwell et al. 2013; Henry et al. 2013a; Salim et al. 2014; Hirschauer et al. 2018; Curti et al. 2020). In this section, we aim to quantify this relation using stacked spectra with our full sample at $1.3 < z < 2.3$.

We begin by asking whether our MZR relation is consistent with a nonevolving M–Z–SFR relation. Figure 12 shows the metallicity residuals between our stack measurements and the local relations from Andrews & Martini (2013) and Curti et al. (2020). Here, we adopt the median of the SED-derived SFRs for the galaxies in each mass bin.[34] For comparison to Andrews & Martini (2013), we identify the corresponding mass and SFR bin in their tabulated relation, whereas for Curti et al. (2020), we evaluated their relation for the masses and SFRs of our stacks. We choose the version of the Curti et al. (2020) relation for total (aperture-corrected) SFRs in order to obtain a measure of the global properties of galaxies. This choice also ensures consistency with Andrews & Martini (2013). We do not apply a correction for contribution from the DIG in local galaxies, as the galaxies in the portion of the local M–Z–SFR relation that match our high-redshift sample are expected to have a minimal DIG fraction and negligible shift in metallicities (Sanders et al. 2017).

In comparison to the local M–Z–SFR relation, Figure 12 shows that our metallicities from stacked spectra (open diamonds) agree with Andrews & Martini (2013) within ± 0.1 dex in the four lower-mass bins, but are 0.26 dex lower than the local relation in the highest-mass bin. As we noted in Section 2.7, the contribution from AGN in this bin is uncertain. Excluding galaxies where the AGN contribution is ambiguous increases our measured metallicity in this stack 0.08 dex. This reduction in the residual from the local relation is shown by the gray diamonds in Figure 12. However, even in this case, this bin shows the largest residual compared to the Andrews & Martini (2013) M–Z–SFR relation. Moreover, the increase in metallicity is at least partly due to a bias from including only the strongest Hβ lines in the stacked spectrum. Hence, we

[34] Formally, the local M–Z–SFR relation is derived from Hα SFRs rather than SED-based SFRs. Nonetheless, evidence of the relation has previously been seen when SED-based SFRs are used (Henry et al. 2013a). As we showed in Figure 8, the SED-derived SFRs for the CANDELS/3D-HST subset of our sample (the majority) show no systematic offset from their Hα-derived SFRs, so the median SED-derived SFR in each mass bin should be adequate for comparing to the M–Z–SFR relation.

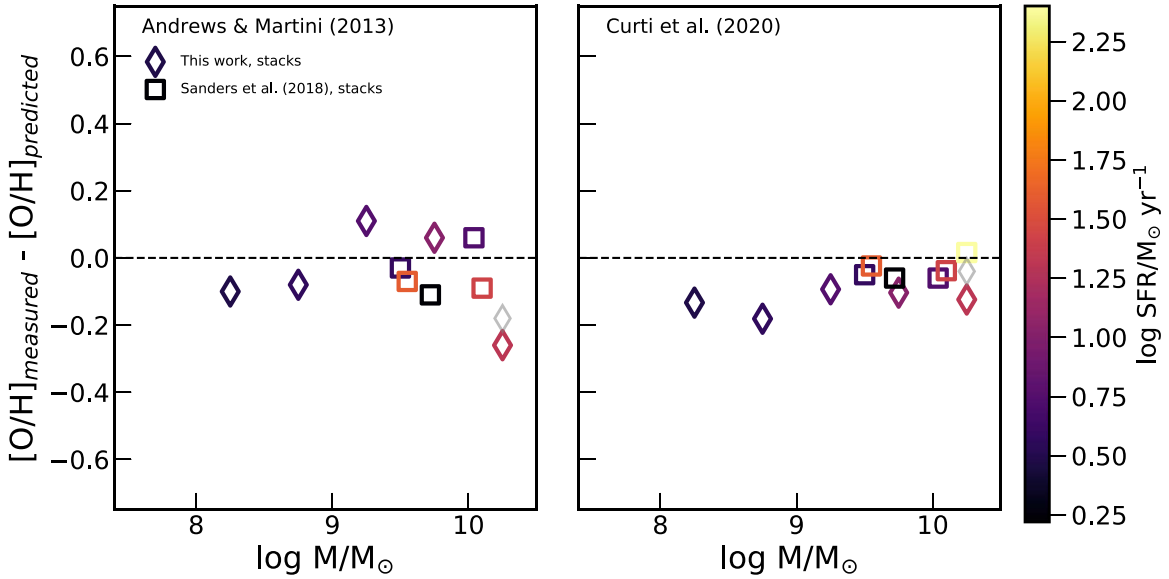

Figure 12. Residuals from the local M–Z–SFR of Andrews & Martini (2013, left) and Curti et al. (2020, right). Points show the results for stacked spectra in this paper (diamonds) and Sanders et al. (2018; squares). The Sanders et al. (2018) results are shown for the stacks divided into three specific SFR (SFR/M_*) bins and two mass bins, with metallicities recalculated on the Curti et al. (2020) calibration. Colors indicate SFR. For the stacks in this paper, the mean SED-derived SFR in each bin is used, whereas for Sanders et al. (2018) the SFRs are derived from Hα. The gray diamond shows the 0.08 dex shift to higher metallicity that we derive when we exclude galaxies with ambiguous AGN contribution.

conclude that the difference between our sample and the local M–Z–SFR relation from Andrews & Martini (2013) at log $M/M_\odot > 10$ is not a result of AGN contamination. Curiously, the large residual at high masses is not mirrored in the right panel of Figure 12, where we compare our measurements to the local parameterization given by Curti et al. (2020). In this case, our metallicities from stacked spectra are systematically lower than the local relation by an amount between 0.10 and 0.17 dex.

Sanders et al. (2018) also compare stacked spectra at $z \sim 2.3$ to the local M–Z–SFR relation given by Andrews & Martini (2013). They report that their observations are offset to metallicities 0.1 dex lower than the local relation. However, Sanders et al. (2018) make this comparison using nitrogen-based metallicity indicators. Therefore, for consistency with this work, we reevaluate the metallicities for their M–Z–SFR stacks using the Curti et al. (2017) calibration for R_{23} and O_{32}. These results are shown as open squares in Figure 12. The bin with the highest mass and SFR is excluded from the left panel, as it does not have a corresponding measurement in Andrews & Martini (2013). Similar to our stacked spectra, the Sanders et al. (2018) metallicities fall within 0.1 dex of the local M–Z–SFR relation from Andrews & Martini (2013) and do not show a systematic difference. On the other hand, their data fall very close to the local M–Z–SFR relation from Curti et al. (2020), showing better agreement than our stacked spectra.

In short, both our stacked spectra for $M/M_\odot < 10$, as well as those from Sanders et al. (2018), agree with the local M–Z–SFR relation from Andrews & Martini (2013) within ± 0.1 dex, but when we use the Curti et al. (2020) measurement of the local M–Z–SFR relation, our stacks show a systematic offset. Hence, it is impossible to determine whether the M–Z–SFR relation evolves because we find different results when comparing to different measurements of the local relation (both from the SDSS). Additionally, the metallicity calibrations used to compare high-redshift samples to the local M–Z–SFR seem to matter, as we do not find the same 0.1 dex metallicity offset reported by Sanders et al. (2018).

We next aim to determine whether we see evidence of an M–Z–SFR relation within our sample. For this test, we divide each of our five mass bins into bins above and below the median SED-derived SFR in that bin. We follow the procedures described in Section 3.1 and Appendix C, creating stacks, measuring the emission lines, and inferring metallicities. Because the SED-derived SFRs are more precise for the CANDELS/3D-HST objects than the WISP objects, we tried this stacking exercise both with and without the WISP objects. Regardless, we find no significant difference between the high-SFR and low-SFR stacks: the high-SFR MZR agrees with the low-SFR MZR at 1σ–2σ. We also tried stacking galaxies with high and low [O III] equivalent widths, dividing each mass bin into galaxies above and below the median [O III] equivalent width in that bin. The result was the same: the metallicities were consistent between the high and low equivalent widths.

The lack of an M–Z–SFR relation in our stacking analysis is somewhat surprising because a relation has been reported at $z \sim 2$ by Salim et al. (2015) and Sanders et al. (2018). One possibility is that our SED-derived SFRs may not be accurate enough to distinguish galaxies with high SFRs from those with low SFRs. We used a simulation to test whether this assessment is correct. In brief, we created a mock sample of galaxies with our stellar mass range, and SFRs defined by the Whitaker et al. (2014) star-forming main sequence at $2.0 < z < 2.5$. Then, we added 0.3 dex of intrinsic scatter to the SFRs and calculated the metallicities using these SFRs with the Curti et al. (2020) M–Z–SFR relation. Then, to model our measurement errors, we added 0.2 dex of additional SFR scatter—half of the typical 68% confidence interval for SED-derived SFRs of the CANDELS/3D-HST objects. We used these SFRs to divide the sample into galaxies, which we select to be above and below the median SFRs. In this way, some galaxies that should be above/below the median SFR are scattered into the opposite SFR bin. We find that the mean metallicities of galaxies that we

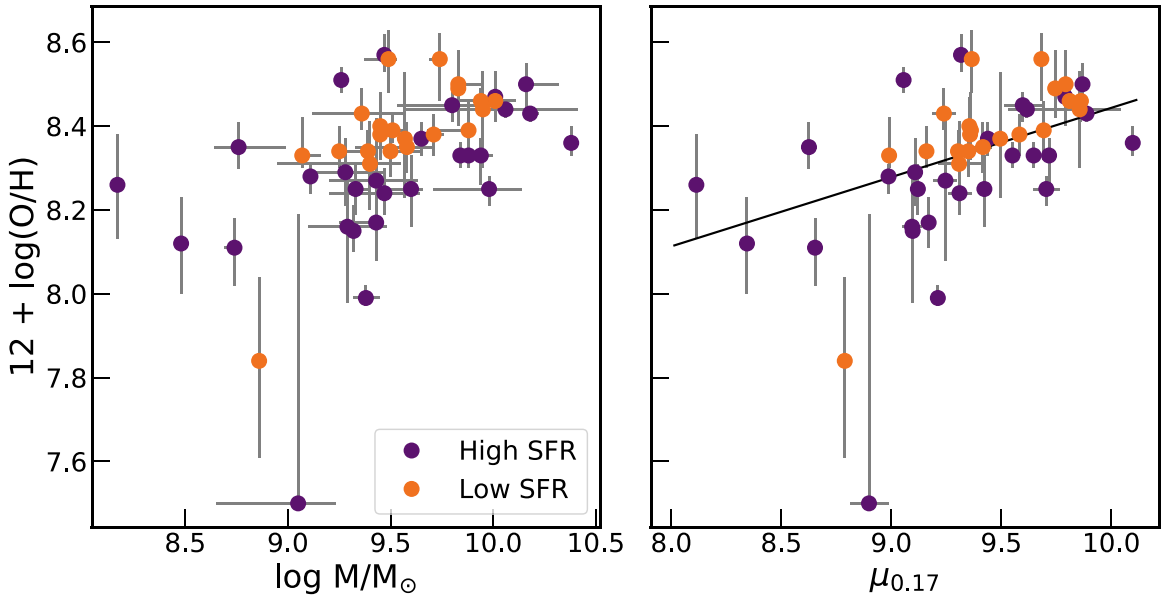

Figure 13. Left: An M–Z–SFR relation is detected in our data for individual galaxies with SFRs derived from Hα. Points show 49 galaxies with Hα S/N > 10. We divide the sample into high- and low-SFR bins, containing objects that are above/below $\log(\mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1}) = 0.68\log(M/\ M_\odot) - 5.56$. This line is determined from a fit to the masses and Hα SFRs in the subset of these galaxies at $\log(M/M_\odot) > 9.2$. Right: The projection of least scatter for the same galaxies. Following Mannucci et al. (2010), we define $\mu_\alpha = \log(M/\ M_\odot) - \alpha\log(\mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1})$; $\alpha = 0.17$ minimizes the scatter. The solid line shows a linear fit to metallicity as a function of $\mu_{0.17}$ and is given in Equation (4).

select to be in the high-SFR bins are only around 0.05–0.06 dex lower than the galaxies in the low-SFR bins. This difference is only somewhat larger than the 1σ uncertainty on the metallicities that we measure in our stacks (Table 4). Hence, we conclude that more accurate measurements of SFR are needed to quantify the M–Z–SFR relation from stacked spectra.

4.4. The M–Z–SFR Relation from Individual Objects at 1.3 < z < 1.5

As noted in Section 3.2, 49 galaxies at redshifts $1.3 < z < 1.5$ have Hα S/N > 10, facilitating dust and metallicity measurements without stacking, as well as providing SFRs from Hα. The median 68% confidence interval on these Hα derived SFRs is 0.3 dex—marginally better than the SED-derived SFRs from the CANDELS/3D-HST data (0.4 dex). Figure 13 (left) shows the M–Z–SFR relation for these galaxies, now using Hα SFRs. We bisect the sample into high- and low-SFR objects, using a linear fit to the data at $\log(M/M_\odot) > 9.2$ (where the sample is larger and the metallicity errors are smaller). This fit yields the relation $\log(\mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1}) = 0.68\ \ \log(M/\ M_\odot) - 5.56$. Galaxies with SFRs above (below) this line are shown as orange (purple) points in Figure 13. In contrast to the stacked spectra, we do see evidence of an M–Z–SFR relation. At a given stellar mass, galaxies with higher SFRs have lower metallicities, particularly at $M \gtrsim 10^9\,M_\odot$, where the numbers of galaxies with high-S/N Hα detections are higher. With only a handful of galaxies at $M < 10^9\,M_\odot$ in Figure 13, detecting an M–Z–SFR relation at these masses and redshifts will require larger samples with higher-S/N spectra. Deeper WFC3/IR grism spectroscopy, as well as new observations with the James Webb Space Telescope (JWST), can improve statistics in this low-mass regime.

We next explore how well these individual objects agree with the M–Z–SFR relations from the literature. As we did with the stacked spectra, we compare them to the local relations from Andrews & Martini (2013) and Curti et al. (2020), showing residuals from these local SDSS relations in Figure 14. Here, we use the published tabular data from Andrews & Martini (2013) in the left panel and the parametric relation for total SFRs from Curti et al. (2020) in the right panel. In the former case, 13 objects do not have matching bins in local M–Z–SFR relation from Andrews & Martini (2013); for these, we extrapolated the parameterized MZR that is given in bins of SFR (see Table 4 of Andrews & Martini 2013). In this comparison, our data show systematic differences from both local measurements. On one hand, our metallicities are 0.1–0.2 dex lower than the local M–Z–SFR relation from Curti et al. (2020), which uses strong-line metallicities. The left-hand panel, on the other hand, reveals a linear trend in the residuals from the M–Z–SFR relation reported by Andrews & Martini (2013), indicating that the M–Z–SFR relation from our data has a different shape than the local one derived with direct-method metallicities. The substantial negative residual for the stack of all galaxies with $\log M/M_\odot > 10$ highlighted in Figure 12 may be a reflection of this trend.

Interpreting a potential disagreement with the local M–Z–SFR relation is difficult. While an offset could imply the evolution of the relation, the masses and SFRs occupied by high-redshift galaxies are not well sampled in either the Andrews & Martini (2013) or Curti et al. (2020) relations. In the latter case, our data fall on an extrapolation of the surface that they fit to their data. Indeed, Curti et al. (2020) point out that the accuracy of their parameterization at low masses and high SFRs—where high-redshift galaxies lie—is only 0.3 dex. It is worth reiterating that the MZR evolves by a similar amount from $z \sim 2$ to $z \sim 0$, so these systematic errors in the local M–Z–SFR are significant. Hence, we arrive at a clear need: While the high-redshift M–Z–SFR relation will be an obvious focus of JWST, measuring its evolution will be challenging unless we can find larger samples of analogous galaxies in the local universe.

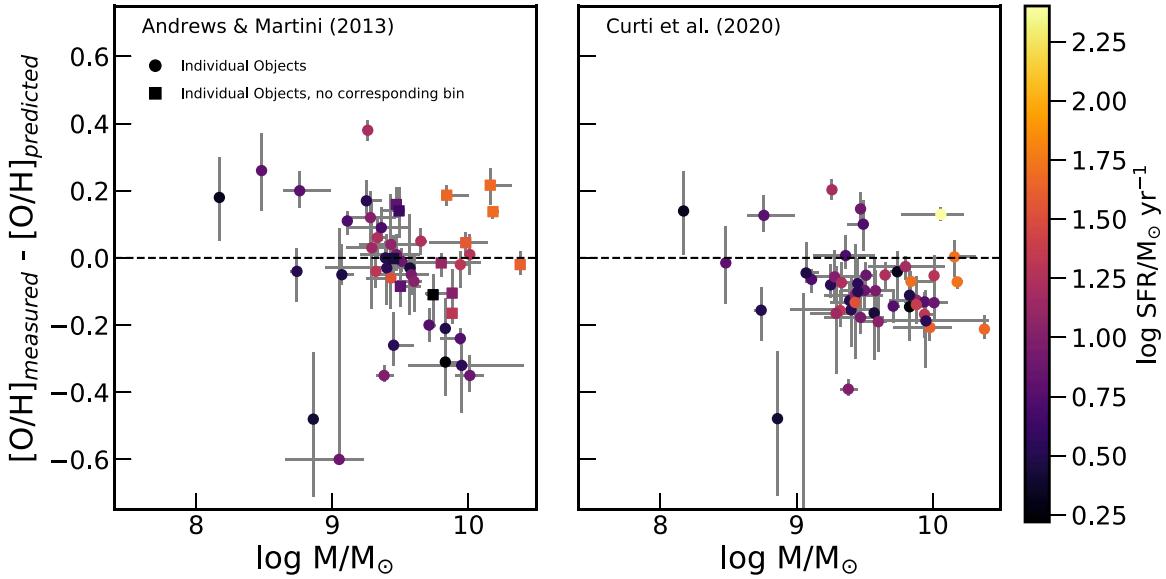

Figure 14. The residuals from the local M–Z–SFR relation are shown for individual objects at $1.3 < z < 1.5$ with Hα S/N > 10 (points). The left panel shows residuals from the Andrews & Martini (2013) relation, where galaxies are matched to the nearest mass and SFR bin in the local sample. The right panels show residuals from the parametric relation using the total (aperture-corrected SFRs) reported by Curti et al. (2020). Squares in the left panel denote the 13 objects that do not have a corresponding bin in Andrews & Martini (2013); the predicted metallicities for these sources are calculated by extrapolating the reported MZR in the appropriate SFR bins (see Table 4 in Andrews & Martini 2013). All SFRs are derived from dust-corrected Hα emission. Lastly, we note that error bars are asymmetric, and in some cases, the statistical uncertainties are smaller than the points.

It is also instructive to consider scatter in metallicities, which is only possible with the individual high-S/N spectra. Both the scatter around the MZR and the scatter relative to the M–Z–SFR relation constrain models for galaxy formation. A key question, which we can address, is how much the MZR scatter can be reduced by considering an M–Z–SFR relation. For this analysis, Mannucci et al. (2010) introduced the projection of least scatter, such that metallicity is a function of $\mu_\alpha = \log(M/M_\odot) - \alpha \log(\mathrm{SFR}/M_\odot\ \mathrm{yr}^{-1})$. Using our data for the 49 galaxies with Hα SFRs, we find that scatter is minimized for $\alpha = 0.17 \pm 0.07$. This value is smaller than the value of $\alpha = 0.32$ that was originally reported for the local M–Z–SFR relation by Mannucci et al. (2010). Other studies that use strong-line metallicity measurements have reported $\alpha = 0.19$ (Yates et al. 2012) and $\alpha = 0.18$ (Hirschauer et al. 2018). Alternatively, Andrews & Martini (2013) find $\alpha = 0.66$ minimizes scatter in stack-based measurements that use direct metallicities.[35] They argue that a larger α—implying that metallicities depend more strongly on SFRs—is a characteristic of direct metallicity measurements of the M–Z–SFR relation. They showed that it is not a consequence of stacking, as the M–Z–SFR relation derived from strong-line calibrations with stacked spectra are characterized by $\alpha = 0.1$–0.3.

The projection of least scatter for our sample is shown in the right panel of Figure 13. Curiously, the value of α that minimizes the scatter in the MZR still leaves the low-SFR galaxies with higher metallicities than the high-SFR galaxies in the projection of least scatter. We verified that this result does not change when α is determined only from the galaxies at $M > 10^9\ M_\odot$ and $12 + \log(\mathrm{O/H}) > 8.0$. This finding may indicate that the projection of least scatter is a poor functional representation of the data. Nonetheless, we provide a linear fit to the data in order to characterize the metallicity as a function of μ_α and facilitate comparisons with other studies:

$$12 + \log(\mathrm{O/H}) = 6.74 + 0.17\mu_{0.17}. \quad (4)$$

The scatter about this relation has an rms of 0.16 dex, reduced from 0.17 dex in the MZR (relative to an MZR with the shape from Andrews & Martini 2013). This small reduction in scatter is characteristic of results that have been derived using strong-line metallicity diagnostics applied to individual galaxies. For example, Curti et al. (2020) find that scatter is reduced from 0.07 dex in the MZR to 0.054 dex relative when SFRs are accounted for and Hirschauer et al. (2018) find a reduction from an rms of 0.182–0.178 for a sample of nearby emission-line-selected galaxies. On the other hand, Andrews & Martini (2013) find that when direct-method metallicities are used, the scatter among stacked measurements is reduced more significantly, from 0.22 dex about the MZR to 0.13 dex in the projection of least scatter. Nonetheless, we acknowledge that scatter among stacks is not the same thing as scatter among galaxies, so these differences are difficult to interpret. Still, a greater reduction in scatter, along with the larger α, suggests a stronger relation between SFR and metallicity (at fixed mass) when direct-method metallicities are used. As Andrews & Martini (2013) point out, this behavior may indicate that systematic errors on strong-line calibrations are correlated with SFR. Our work is not immune to these systematic errors. This points to an increased need for direct metallicities, both at low and high redshifts. The latter will soon be possible with JWST.

4.5. Metallicity Biases from Selection Effects

The existence of an M–Z–SFR relation implies that sample selection effects may impact metallicities, especially for samples that are biased toward high SFRs. As we discussed in Section 3.3, our sample is not mass complete. Below

[35] This measurement of α is not reduced significantly by accounting for the DIG contribution in local galaxies. Sanders et al. (2017) report that a DIG correction decreases α from the Andrews & Martini (2013) measurements only marginally, to $\alpha = 0.63$.

$M/M_{\odot} \lesssim 9.2$–9.3, galaxies fall mostly above an extrapolation of the star-forming main sequence from Whitaker et al. (2014). Likewise, as we showed in Figure 10, the CANDELS/3D-HST sample contains more objects with low [O III] equivalent widths compared to WISP. As we showed in Figure 8, for the 49 objects at $1.3 < z < 2.3$ with Hα S/N > 10, the Hα SFRs from WISP objects are, on average, 0.35 dex higher than those of CANDELS/3D-HST objects.

In order to assess the impact of selection effects in our inhomogeneous sample, we created separate stacked spectra for the WISP and CANDELS/3D-HST objects. We used the same five mass bins that we have used throughout this analysis (see Tables 2 and 4). We find that the metallicities from these stacks all agree at the 1σ–2σ level, and there is no systematic difference between the two surveys.[36]

The agreement between the metallicities in the WISP subsample and the CANDELS/3D-HST subsample can be explained by the fact that the SFR dependence in the M–Z–SFR relation that we measure is not particularly strong. Following Equation (4), if the WISP sources have SFRs that are 0.35 dex higher, we expect their metallicities to be lower than CANDELS/3D-HST by only 0.01 dex. This difference is smaller than the precision of our metallicity measurements (Table 4). Even SFRs that are 1 dex higher than the main sequence—as we might see toward the lowest masses in our sample (see Figure 9)—only result in metallicities that are lower by 0.03 dex.

The dependence of metallicity on SFR may be larger if we were able to use direct metallicities in our high-redshift sample. At $z \sim 0.1$, the M–Z–SFR relation shows a larger spread in metallicity with changing SFR when it is derived by stacking spectra to measure direct metallicities (Andrews & Martini 2013). These authors report that the projection of least scatter has a slope of 0.43, with $\alpha = 0.66$. Therefore, a 1 dex increase in SFR corresponds to a 0.3 dex decrease in metallicity, although the differences between the WISP and CANDELS/3D-HST metallicities would still be small (0.03–0.09 dex). Nonetheless, when we apply a strong-line calibration to our sample of 49 objects with high-S/N spectra, we find a weak relation between the SFR and metallicity (at fixed mass); therefore, we do not expect to observe a measurable bias toward low metallicities in our sample.

In addition to mass completeness, metallicities can also be biased when spectroscopic surveys require the detection of multiple emission lines. In WFC3/IR grism spectra, where the S/N is often not particularly high, secondary emission lines are used to confirm redshifts—typically Hβ or [O II]–often have S/N $\sim 2\sigma$–3σ. In this case, Eddington bias leads to an overestimate of the average fluxes of these lines, due to statistical scatter above the detection threshold (Eddington 1913). Artificially increasing the average Hβ and [O II] fluxes will result in decreased R_{23}, [O III]/Hβ, and O_{32} ratios. These biases tend to increase metallicity inferences (provided that galaxies are on the upper branch of R_{23} and [O III]/Hβ, as they are in most of our sample).

To quantify the bias due to the requirement for multiple emission lines, we recalculated the MZR using the Curti et al. (2020) calibration with the published R_{23} and O_{32} values in two previous WFC3 IR grism studies that were subject to this bias: Henry et al. (2013b) and Grasshorn Gebhardt et al. (2016). The MZRs from these two works are compared to our new results in Figure 15. On average, the metallicities are around 0.1–0.2 dex higher. (Although we note that the current MZR and Henry et al. (2013b) are consistent at around the 2σ–3σ level.) Curiously, this difference contrasts with the WISP-only stacks from our present sample, which we showed to be consistent with CANDELS/3D-HST-only stacks and the MZR of the full sample. A major difference between our prior WISP sample and the current one is that the most recent line lists adopt a higher-S/N threshold for the primary (strongest) emission line, and should include far fewer spurious emission-line galaxies (M. Bagley et al. 2021, in preparation). We conclude that there is a great amount of subjectivity in the sample selection with low S/N spectra when a second line is required to confirm the redshift. The present WISP sample seems to be less biased than Henry et al. (2013b) and Grasshorn Gebhardt et al. (2016), as it shows negligible differences from CANDELS/3D-HST.

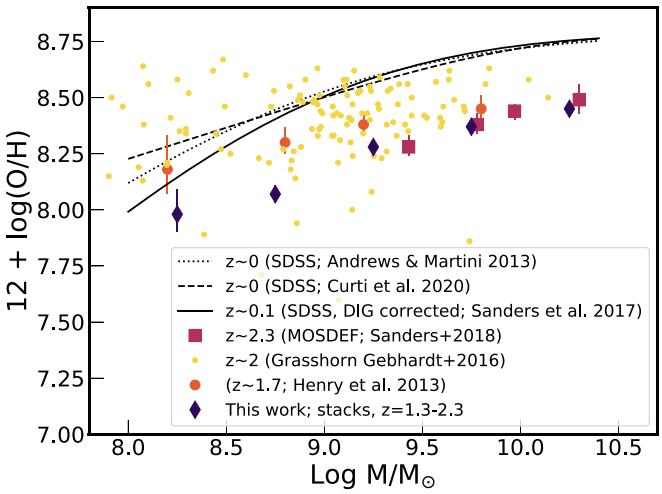

Figure 15. Sample selection effects can impact the metallicity measurements in low S/N spectra when a second line is required to confirm the redshift and measure the metallicity. Here, we compare our new measurements to two WFC3/IR grism-based studies that require confirming emission lines: Henry et al. (2013b) and Grasshorn Gebhardt et al. (2016). This selection effect, coupled with low-S/N spectra, seems to result in an MZR that is biased toward high metallicities. The metallicities from all of the high-redshift studies shown here have been calculated on the Curti et al. (2020) calibration, using published line ratios.

Lastly, we acknowledge that an [O III]-based selection can impart a metallicity bias as the strength of this line is critically sensitive to metallicity. As such, we might expect our sample to be biased toward objects with the highest [O III]/Hβ ratios and metallicities around 12 + log(O/H) ~8.0. Therefore, we next consider the effects of [O III] selection by modeling our survey in the IllustrisTNG hydrodynamical simulation (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018; Pillepich et al. 2018; Springel et al. 2018).

5. Comparison to Theoretical Models

The MZR and M–Z–SFR relation are key observational constraints on theoretical models describing the evolution of galaxies. As we argued in Section 3.4, metallicity calibrations have improved considerably over the past several years, indicating that we are now better poised to compare observations and theoretical models. One thing that remains

[36] Because the WISP and CANDELS/3D-HST sources have different mean redshifts, we also measure the MZR in four sets of stacked spectra: WISP at $z > 2$, $z < 2$, and CANDELS/3D-HST at $z > 2$, $z < 2$. Still, we find no significant differences in the metallicities of these subsamples.

uncertain, however, is how selection effects, in the presence of an M–Z–SFR relation, impact our ability to compare observations and models. While recent simulations report success in matching the slope and evolution of the MZR (Ma et al. 2016; Davé et al. 2017, 2019; De Rossi et al. 2017; Torrey et al. 2019), they do not account for the possibility that observed metallicities may be biased by sample selection effects.

In this section, we address this shortcoming by modeling our sample selection in the IllustrisTNG hydrodynamical simulation (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018; Pillepich et al. 2018; Springel et al. 2018). We use the TNG100-1 simulation, which has a volume of 106.5^3 comoving Mpc^3, and select star-forming galaxies with $M > 10^8\,M_\odot$. We include only central galaxies, as satellite galaxies have metallicities heavily influenced by their environment and are unlikely to be resolved from central galaxies at $z = 2$ (P. Torrey 2020, private communication). From these snapshots, we take masses, SFRs, and metallicities for 24,108 galaxies at $z \sim 0.1$, 30,485 galaxies at $z \sim 1.5$, and 29,578 galaxies at $z \sim 2.0$. We take the SFR-weighted metallicities from the simulation, as these most accurately reflect the luminosity-weighted measurements that are made from emission-line spectra. We refer the reader to Torrey et al. (2019) for details pertaining to the IlustrisTNG MZR.

In order to model emission-line-based selection, we convert these quantities into observed emission-line fluxes. We first use SFRs to determine intrinsic Hα luminosities, assuming the Kennicutt (1998) relation between SFR and Hα luminosity, and multiplying by 1.8 to convert from a Chabrier (2003) IMF used by IllustrisTNG to a Salpeter (1955) IMF used by Kennicutt (1998). Unreddened Hα line fluxes are then calculated from the redshift of the sources in the simulation, assuming a Planck 2015 cosmology (Planck Collaboration et al. 2016).

We next calculate the unreddened Hβ line flux assuming an intrinsic Balmer decrement of $\mathrm{H}\alpha/\mathrm{H}\beta = 2.86$. The unreddened [O III] line flux is inferred from the relation between metallicity and [O III]/Hβ from Curti et al. (2017), adding 0.07 dex of scatter in log (O/H), relative to the calibration, as reported by Curti et al. (2017).

Finally, we dust-redden all three emission lines. In practice, we should be able to estimate dust mass from simulation data, using measures of gas and metal content. However, translating this estimate to dust extinction is highly dependent on uncertain assumptions (e.g., Narayanan et al. 2018). Therefore, we adopt an empirical approach. For $z \sim 0.1$, we use the relation between $A_{\mathrm{H}\alpha}$ and stellar mass in SDSS galaxies (Garn & Best 2010), including 0.28 magnitudes of scatter in the reddening. At higher redshifts, however, measurements of the Balmer decrement at $z \sim 1$–2 do not detect strong correlations between dust extinction and galaxy properties (Domínguez et al. 2013; Theios et al. 2019). Therefore, for $z \sim 1.5$ and 2.0, we assign dust extinction randomly, drawing from an exponential distribution with an e-folding length of $E(B-V)_{\mathrm{gas}} = 0.4$, roughly consistent with the distribution reported by Theios et al. (2019). Reddening is then applied using the Calzetti et al. (2000) extinction relation.

This approach provides us with modeled Hα, Hβ, and [O III] fluxes for each simulated galaxy. We next emulate sample selections by applying cuts to the emission-line fluxes to generate four mock samples of galaxies. For the present WFC3 grism-selected sources that we use in stacks, we adopt $F([\mathrm{O\,III}]) > 5 \times 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$. This flux limit corresponds to the approximate 50% completeness of our catalog. Similarly, for the 49 objects with Hα S/N > 10, we take $F(\mathrm{H}\alpha) > 8 \times 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$. These limits are an approximation, as the WISP data have a range of exposure times, some of which are significantly longer than the CANDELS/3D-HST observations. We also compare these to the flux limits used by Sanders et al. (2018) in the MOSDEF survey, where we estimate[37] that Hα and Hβ fluxes are greater than $1 \times 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$. Lastly, we consider the SDSS for a local comparison sample. We use the 50% completeness in the Hα fluxes reported in the DR7 MPA/JHU catalog, which corresponds to around $F(\mathrm{H}\alpha) > 6 \times 10^{-16}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$. We note that more detailed modeling, accounting for broadband magnitude limits, equivalent width cuts, nonuniform survey sensitivity (a particular concern for WISP), and slit mask/fiber plug plate designs are not addressed here. We leave this more sophisticated analysis to future work.

Figure 16 shows the IllustrisTNG MZR at $z \sim 0.1$ (upper left), $z \sim 1.5$ (upper right), and $z \sim 2.0$ (bottom). The shaded regions show the relation for all star-forming galaxies, with the black solid curve indicating the mean. These are the same relations discussed in detail in Torrey et al. (2019); we have decreased the normalization of the modeled metallicities by 0.2 dex, in order to better match the $z \sim 0.1$ SDSS relation. This renormalization is generally required, as the nucleosynthetic yield of oxygen remains uncertain (Nomoto et al. 2013). The small points show different sample selections that we apply to the simulation, in order to match comparison observations: the SDSS $z \sim 0.1$ measurements from Andrews & Martini (2013) and Curti et al. (2020), upper-left panel, and the $z \sim 2$ measurement from MOSDEF (Sanders et al. 2018, lower left), along with the individual objects at $z \sim 1.3$–1.5 (upper right) and the stacked spectra (lower right; mean $z = 1.9$) in the present work.

At low redshifts, the Hα flux selection does not modify the modeled MZR—i.e., the curves in the upper-left panel of Figure 16 fall on top of each other at all masses. Given the relative brightness of nearby galaxies, the SDSS is sensitive to a representative sample of galaxies at $M > 10^8\,M_\odot$—56% of simulated galaxies have modeled Hα fluxes above the selection threshold. Likewise, IllustrisTNG does a fair job at matching the slope of the SDSS MZR, as the observations overlap closely with the model curves.

At higher redshifts, emission-line flux limits play a more significant role. The observations in this paper, shown in the right-hand panels, show an MZR that is around 0.2 dex lower than the IllustrisTNG MZR derived from the entire simulation. However, when we model the [O III]-based selection in the simulation, we see better agreement between theory (small points) and observations for both the stacked spectra (filled diamonds) and the individual objects with Hα S/N > 10 (open circles). The [O III]-based selection picks out objects with the lowest metallicities, which we expect for galaxies in the mass and metallicity regime that we consider. We also notice in the right-hand panels of Figure 16 that modeling the sample selection removes simulation galaxies at the lowest masses probed by our data. This difference is likely a reflection of our simplified model of sample selection effects, but could also be

[37] Kriek et al. (2015) report typical 5σ sensitivity for MOSDEF of $1.5 \times 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$, and Sanders et al. (2018) select their sample by requiring 3σ detections in Hα and Hβ.

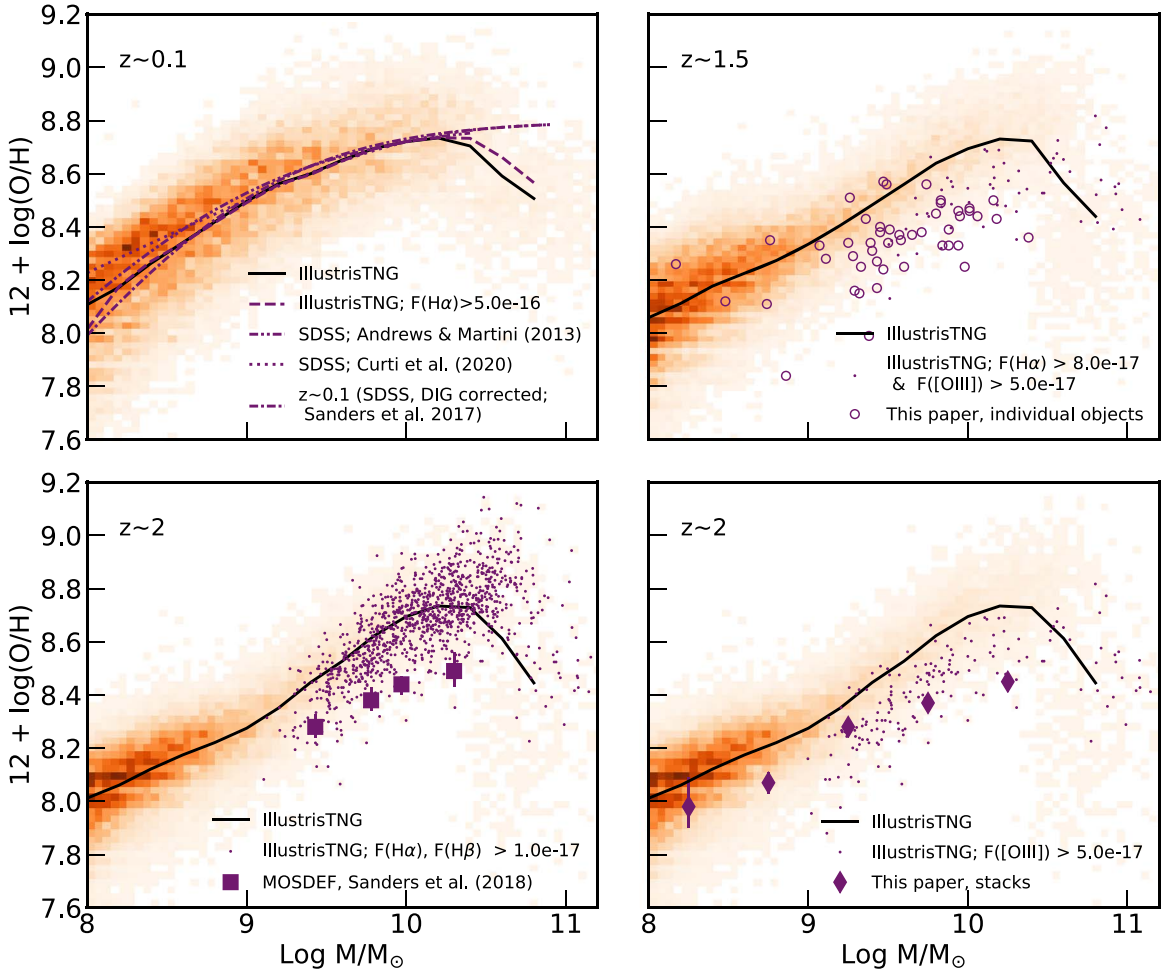

Figure 16. Line-flux-limited surveys can result in a sample selection that is not representative of typical galaxies in a simulation. The MZR from IllustrisTNG (Torrey et al. 2019) is shown as the orange shaded region for $z \sim 0.1$ (upper left), $z \sim 1.5$ (upper right), and $z \sim 2$ (bottom). The black solid line shows the mean of the simulation data. All simulated metallicities have been shifted downwards by 0.2 dex, adjusting for uncertainties in the nucleosynthetic yield of oxygen and matching the $z \sim 0.1$ MZR around $M = 10^{10} M_{\odot}$. Small points illustrate the effect of line-flux-limited selections on the simulation. These represent the SDSS ($F(\mathrm{H}\alpha) > 5 \times 10^{-16}$ erg s^{-1}, cm^{-2}, upper left), MOSDEF ($F(\mathrm{H}\alpha)$ and $F(\mathrm{H}\beta) > 1 \times 10^{-17}$ erg s^{-1} cm^{-2}, lower left), and the present WFC3 grism surveys ($F([\mathrm{O\ III}]) > 5 \times 10^{-17}$ erg s^{-1} cm^{-2} for stacks, lower right, and $F(\mathrm{H}\alpha) > 8 \times 10^{-17}$ erg s^{-1} cm^{-2} for individual objects, upper right). For comparison, observational data are shown. The upper-left panel includes the $z \sim 0.1$ MZR from Andrews & Martini (2013) and Curti et al. (2020), as well as the DIG-corrected SDSS measurement from Sanders et al. (2017). The $z \sim 2$ measurements from MOSDEF are shown in the lower left (Sanders et al. 2018), and the stacks and individual objects from this work are shown in the right panels.

an indication that simulation galaxies do not exist in the right densities for the masses, metallicities, and SFRs probed by this portion of the sample. We leave this question to future work.

In contrast to the results from the present survey, the lower-left panel of Figure 16 shows that the Hα + Hβ selection modeled after MOSDEF selects galaxies that are mostly representative of the full simulation. However, the model data with selection effects applied have metallicities around 0.2 dex higher than the observations. Hence, while accounting for selection effects brings the IllustrisTNG model in line with the MZR in the present work, it does not do so for the MOSDEF MZR. Curiously, even though our MZR shows good agreement with the MOSDEF observations (see Figure 11), when we account for selection effects, IllustrisTNG predicts that our MZR should have a lower normalization than the one from MOSDEF.

The disagreement between the observed and simulated $z \sim 2$ MZRs in Figure 16 suggests a coupling between mass, metallicity, and SFR that is difficult to disentangle from the MZR alone. While modeling the [O III]-based selection lowers the metallicities that are predicted from the simulation, this is not the only effect that is in play. The selection effects, when applied to the simulation, predict that our [O III]-selected sample should have higher SFRs than MOSDEF by about 0.2 dex at $9.75 < \log M/M_{\odot} < 10.0$ (average log SFR$/M_{\odot}$ yr^{-1} = 1.16 versus 0.97). This trend would also lead to lowered metallicities, given the presence of the M–Z–SFR relation in both the observations and the simulation (Torrey et al. 2019). However, our observations show very similar SFRs to those of MOSDEF in the mass range where they overlap (see Table 4 versus Table 1 in Sanders et al. 2018). Therefore, we next compare the simulated M–Z–SFR relation with observations.

Figure 17 shows the simulated M–Z–SFR relation, plotting the deviations from the MZR and the star-forming main sequence, $\Delta\log$ O/H versus $\Delta\log$ SFR$/M_{\odot}$ yr^{-1}, as proposed by Davé et al. (2017). In this case, we use a snapshot from IllustrisTNG at $z = 1.5$, for ease of comparison with our 49 high-S/N objects at $1.3 < z < 1.5$. The slope shows how changes in SFR translate to changes in metallicity. Excluding simulation sources at $\log M/M_{\odot} > 10.2$, where the model for

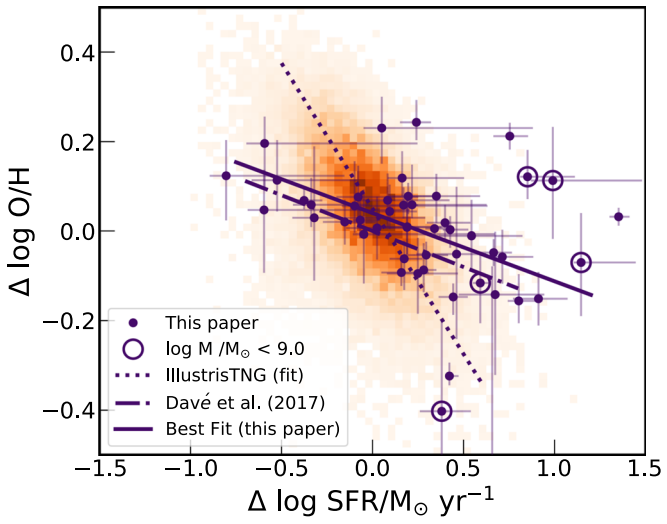

Figure 17. The M–Z–SFR relation at $z = 1.5$ in IllustrisTNG has a dependence on SFR that is too steep when compared to observations. The shaded region shows the simulated deviations from the star-forming main sequence, Δlog SFR vs. the deviations from the MZR, Δlog O/H. Both are calculated with respect to polynomial fits to the simulated scaling relations. The slope of the deviations, shown by the dashed line, is around ~ -0.65. Points show our observations, which are fit with a linear relation having a slope of -0.15 (solid line). The observations are closer to the slope of -0.16 in the simulation reported by Davé et al. (2017, dotted–dashed line).

quenching becomes substantial and the simulations do not match the observations, we find a slope of -1.0 in the simulation data (dotted line). The same quantities for our observations are shown by the points, where we have calculated Δlog O/H relative to a linear fit to the MZR from our stacked spectra, and Δlog SFR/$M_\odot$ yr^{-1} relative to the star-forming main sequence of Whitaker et al. (2014).[38] A linear fit to our data for the 49 high-S/N objects (accounting for errors in both Δlog O/H and Δlog SFR/$M_\odot$ yr^{-1}) yields a slope of -0.15 ± 0.05. This measurement is close to the slope of -0.16 found by Davé et al. (2017).[39]

Figure 17 suggests that our data follow a shallower slope than inferred from the simulated galaxies in IllustrisTNG. A 1 dex increase in SFR (or sSFR) will translate to a 1 dex decrease in metallicity in the simulation, but only a 0.15 dex metallicity decrease in our observations. This weak SFR dependence is similar to what we identified from the projection of least scatter in Section 4.4 and Equation (4). In fact, it follows from Equation (4) (the projection of least scatter) that the slope in Figure 17 should be $0.17 \times \alpha_{0.17} = 0.03$. While the SFR dependence that we find from the projection of least scatter is even weaker than what we see in the Δlog O/H versus Δlog SFR/$M_\odot$ yr^{-1} plot, they are both much weaker than in IllustrisTNG. The slight disagreement between these two observational methods is not surprising, as the calculations have different systematics (e.g., Δlog SFR is calculated with respect to Whitaker et al. 2014; Equation (4) is assumed to be linear).

While the M–Z–SFR relation at low redshift shows a stronger SFR dependence when inferred from stacked spectra with direct metallicity measurements (Andrews & Martini 2013), the amplitude of this effect is not likely to be large enough to explain the discrepancy between our observations and IllustrisTNG. When Andrews & Martini (2013) report the projection of least scatter, they find $\alpha = 0.66$, and the linear relationship between μ_α and 12+log(O/H) has a slope of 0.43; hence their data suggest that a correlation between Δlog O/H and Δlog SFR/$M_\odot$ yr^{-1} in the SDSS should have a slope of approximately -0.3. Similarly, the derivative (with respect to SFR) of the M–Z–SFR relation from Curti et al. (2020) is around -0.17 at $M = 10^{9.0}$–$10^{10.0}\,M_\odot$. Hence, locally, the difference between strong-line calibrations applied to individual galaxies and direct metallicities measured from stacks amounts to a steepening of the slope in Figure 17 of around 0.1. Therefore, it seems unlikely that measuring direct metallicities at $z \sim 2$ would bring the high-redshift M–Z–SFR relation in line with the steeper SFR dependence in IllustrisTNG. The slope of the simulated data in Figure 17 is too steep.

We conclude that modeling selection effects and comparing the M–Z–SFR relation in simulations and observations provide more rigorous tests than the MZR alone. In addition to more careful modeling of selection effects, future comparisons to simulations beyond IllustrisTNG would be informative.

6. Summary and Conclusions

We have measured the MZR and M–Z–SFR relation at $1.3 < z < 2.3$ using a sample of 1056 star-forming galaxies selected based on [O III] line emission in their WFC3/IR grism spectra. This sample, drawn from the WISP survey and CANDELS/3D-HST, has a mean redshift of $z = 1.9$ and is four times larger than previous MZR samples at $z \sim 2$ (Sanders et al. 2018). The WFC3/IR grism selection reaches stellar masses an order of magnitude lower ($10^8\,M_\odot$) than ground-based IR spectroscopic surveys at this redshift. To measure metallicities, we stack the spectra in five mass bins, between $8 < \log M/M_\odot < 10.5$, but also consider 49 objects at $1.3 < z < 1.5$ and Hα S/N > 10, where our observations cover the full suite of optical emission lines from [O II] to Hα + [S II].

The MZR shows good agreement with previous oxygen-based measurements at $z \sim 2$ (Sanders et al. 2018) in the mass regime where they overlap. Stacking our full sample in five mass bins, we find that metallicities evolve by 0.3 dex from $z \sim 2$ to $z \sim 0$. In contrast, we see no evidence for redshift evolution over the 2 Gyr spanned by our sample; stacked spectra at $1.3 < z < 1.7$ and $1.7 < z < 2.3$ show good agreement.

We also measure the M–Z–SFR relation in our data. We use SED-derived SFRs to create stacked spectra having higher and lower SFRs at each mass bin. This exercise shows no evidence for an SFR dependence: the MZR derived from high-SFR galaxies is consistent with the MZR from low-SFR galaxies. We show that uncertainties on SED-derived SFRs make it difficult to fully separate the sample into high- and low-SFR bins. However, if we instead consider the 49 galaxies with more accurate Hα-derived SFRs, we uncover an M–Z–SFR relation at $1.3 < z < 1.5$. In comparison to low redshifts, we find metallicities that are systematically 0.1 dex lower than the local M–Z–SFR relation derived from strong lines in individual galaxy spectra (Curti et al. 2020) but which show a linear trend in their residuals from the M–Z–SFR relation when it is measured from direct metallicities in stacked spectra (Andrews & Martini 2013). It is difficult to determine whether there is

[38] The star-forming main sequence from Whitaker et al. (2014) is reported for $1.0 < z < 1.5$ and $1.5 < z < 2.0$. We take the average of these measurements to represent our sample at $1.3 < z < 1.5$.

[39] We note that Davé et al. (2017) express this diagram in terms of Δlog sSFR/ yr^{-1} rather than Δlog SFR/$M_\odot$ yr^{-1}. These quantities are equivalent, as they are calculated at a fixed mass.

real evolution because small samples of low-redshift galaxies at the masses and SFRs of our sample imply that the local M–Z–SFR relation is poorly constrained in this regime.

The 49 objects with Hα SFRs allow us to measure the projection of least scatter of the M–Z–SFR relation (metallicity as a function of μ_α, where $\mu_\alpha = \log M/M_\odot - \alpha \log \mathrm{SFR}\ /\ M_\odot\ \mathrm{yr}^{-1}$). We find an optimum fit corresponding to $\alpha = 0.17$, although the scatter is reduced only by a small amount: from 0.17 dex about the MZR to 0.16 dex about the M–Z–SFR relation. Both the low value of α and the very small reduction in scatter are consistent with low-redshift studies that use strong emission-line metallicity calibrations (Mannucci et al. 2010; Yates et al. 2012; Andrews & Martini 2013; Hirschauer et al. 2018). The weak SFR dependence in our M–Z–SFR relation implies that our sample, while incomplete to low-SFR objects at low masses, is biased to low metallicities by 0.03 dex at most. However, it is important to note that direct-method abundances could yield a different result. The low-redshift M–Z–SFR appears to have a stronger dependence on SFR when direct-method abundances are used (Andrews & Martini 2013 find $\alpha = 0.66$ using direct abundances in stacked SDSS spectra and $\alpha = 0.1$–0.3 using strong lines in the same stacks). These authors argue that this stronger SFR dependence from direct abundances indicates that strong-line calibrations are subject to systematic uncertainties that correlate with SFRs. We speculate that the M–Z–SFR relation at high redshifts may have a stronger SFR dependence if it were to be measured with direct-method metallicities. As a corollary, our MZR may be more biased by incompleteness than we have inferred from the strong-line M–Z–SFR relation.

Finally, we compare our results with theoretical models of the MZR and M–Z–SFR relation. Because of the presence of an M–Z–SFR relation, directly comparing an observed MZR with models, as is often done, can be misleading. Therefore, we have applied sample selection effects to the IllustrisTNG simulation, thereby more accurately modeling the MZR from the present data, as well as the MZR from Sanders et al. (2018). This exercise suggests that the selection effects in our present sample are uncovering the objects with the highest SFRs and lowest metallicities, whereas the selection used by Sanders et al. (2018) uncovers more typical $z \sim 2$ galaxies. Still, modeling these selection effects brings the high-redshift MZR from IllustrisTNG more in line with the observations that we present in this paper. Curiously, however, IllustrisTNG predicts that the galaxies observed by Sanders et al. (2018) should have higher metallicities than the galaxies in our WFC3/IR-selected sample. Yet, the MZR measurements from these two studies show excellent agreement in the mass range where they overlap.

To further explore this discrepancy between the model and the data, we compared the M–Z–SFR relation in IllustrisTNG with our 49 galaxies with Hα SFRs. Following Davé et al. (2017), we show the relationship between $\Delta\log$ O/H and $\Delta\log \mathrm{SFR}/\ M_\odot\ \mathrm{yr}^{-1}$ for both our observations and the IllustrisTNG simulation data. We find that the slope in the $\Delta\log$ O/H versus $\Delta\log \mathrm{SFR}/\ M_\odot\ \mathrm{yr}^{-1}$ plane is around -1.0 in the simulation data but is -0.15 in the observations. The slope of the simulated data is much too steep; i.e., metallicity depends too strongly on SFR in IllustrisTNG. On the other hand, the shallow slope that we measure shows good agreement with the simulated M–Z–SFR relation from Davé et al. (2017). While direct metallicity measurements at $z \sim 2$ could bring observations closer to IllustrisTNG, we argued in Section 5 that–based on measurements of the local M–Z–SFR relation—the amplitude of this systematic effect is too small to explain our discrepancy with IllustrisTNG.

In conclusion, we reiterate that the metallicities of galaxies remain a key constraint on galaxy formation models. This work has provided new measurements of the evolution of the MZR and the M–Z–SFR relation while also uncovering discrepancies in the simulated M–Z–SFR relation from IllustrisTNG. Nonetheless, we also identify some clear needs. If we are going to understand the evolution of the MZR and M–Z–SFR relations, we need statistical samples of galaxies with direct-method abundances at high redshifts. These data, which will soon be within reach of JWST, will be key to clarifying the strength of the SFR dependence in the M–Z–SFR relation at high redshift.

We are thankful for the efforts of the anonymous referee, whose thoughtful review led to improvements in this manuscript. A.H., M.R., D.E., and B.E. acknowledge support from HST-AR 14580. We thank Paul Torrey for assisting with the interpretation of IllustrisTNG data. We are grateful to Bahram Mobasher and Kit Boyett for thoughtful comments that improved this manuscript. A.J.B. acknowledges funding from the "FirstGalaxies" Advanced Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 789056). Y.-S. D. acknowledges the science research grants from NSFC grant Nos. 10878003 and 11933003, the National Key R&D Program of China via grant No. 2017YFA0402703, and the China Manned Space Project with No. CMS-CSST-2021-A05. This work is based on data obtained from the Hubble Space Telescope, as part of the 3D-HST Treasury Program (GO 12177 and 12328) and the CANDELS Multi-Cycle Treasury Program, as well as GO 11600, 11696, 12283, 12568, 12902, 13352, 13517, 14178, 13420, and 14227. Support for these programs was provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555.

All of the MAST data used in this manuscript can be found at these DOI links:

1. CANDELS: 10.17909/T94S3X,
2. WISP: 10.17909/T9C302,
3. 3D-HST: 10.17909/T9JW9Z,
4. CLEAR: 10.17909/t9-wsnh-yx79, and
5. GOODS-N G102: 10.17909/t9-nypp-mz48.

Software: aXe (Kümmel et al. 2009), IRAF Tody (1993, 1986), SExtractor (Bertin & Arnouts 1996), astropy (Astropy Collaboration 2013, 2018), Photutils (Bradley et al. 2021).

Appendix A
Estimating Emission-line Equivalent Widths

Estimates of emission-line equivalent widths are necessary for both stacked spectra and individual spectra, as the correction for Balmer line stellar absorption is proportional to the emission-line equivalent width. We describe how this correction is applied in Appendix C. Here, we detail the method used to estimate equivalent widths and their uncertainties.

Slitless grism spectra are imperfect for measuring the equivalent width of emission lines, as inaccurate sky

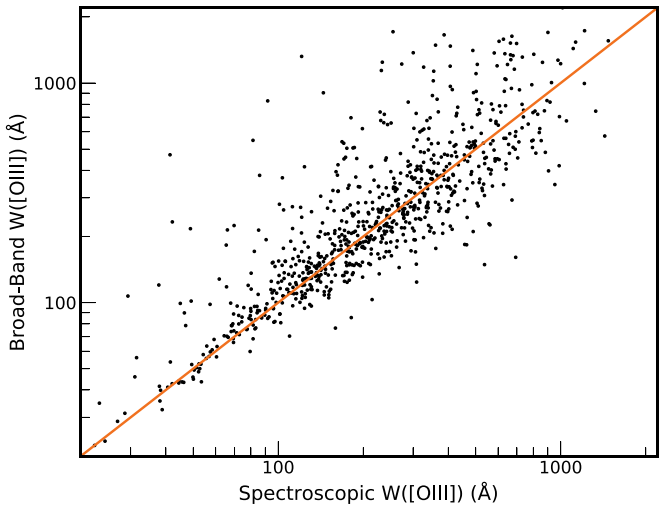

Figure 18. Rest-frame [O III] $\lambda\lambda$4959, 5007 emission equivalent widths are measured using continuum from spectra (abscissa) and continuum from emission-line-corrected broadband photometry (ordinate). Measurements are shown for objects where the continuum is detected in the grism spectroscopy. The solid line shows the one-to-one relation. The broadband continuum measurements give equivalent widths with no systematic difference from the spectra but are more complete to sources with fainter continua.

subtraction or contamination subtraction can result in large systematic errors. Moreover, it is not unusual to detect lines without any continuum, in which case the spectra can only provide a lower limit on the equivalent width. An alternative approach is to measure the continuum from the broadband infrared images. However, this method does not directly measure the continuum flux at the specific wavelength of the emission line, and the extraction aperture is not guaranteed to be the same for the emission lines and the continuum. The broadband fluxes are also likely contaminated by line emission.

Given these caveats, we proceed cautiously and compare two different methods for equivalent width estimation. The first method is to measure equivalent width directly from the spectra. For this measurement, we subtract the contamination model and count spectral continuum detections as cases where the continuum $S/N > 0.5$ per pixel. The second method is to use broadband photometry to estimate the continuum at the wavelength of [O III]. In this case, we use the WFC3/IR photometry: F110W and F160W for WISP, and F125W and F160W for the CANDELS/3D-HST fields. In each galaxy, the broadband photometry is corrected for contamination from emission lines, using our measured fluxes for [O III], [O II], and Hα. We also include a small correction for Hβ and [S II], assuming the average ratios for the sample: [O III] /Hβ = 7.1 and [S II]/(Hα+ [N II]) = 0.16 (with the latter ratio for $z < 1.5$ only). Then, we use a linear interpolation between the corrected continuum fluxes to estimate the continuum at the wavelength of [O III], thereby obtaining the equivalent width in the line. Uncertainties on the continuum at [O III] are estimated by a Monte Carlo simulation, where the WFC3 photometry and the contribution from emission lines were perturbed according to their measured errors.

Figure 18 compares the rest-frame [O III] $\lambda\lambda$4959, 5007 equivalent width measurements from the spectra with the measurements from broadband photometry. Only sources that have continuum detections in the grism spectra with $S/N > 0.5$ per pixel are shown. This analysis shows that there is no systematic difference between the two methods for estimating equivalent widths. Likewise, the fractional error—the rms of $(W_{\rm spec} - W_{\rm broadband})/W_{\rm broadband}$—is 40%. Because we detect continuum from more galaxies when we use the broadband photometry, we adopt this method for all galaxies in the sample.

The approach described above provides an estimate of the lines that are detected in individual spectra. However, when we rely on stacking, [O III] is often the only line that we consistently detect. Therefore, we require a different method to estimate characteristic equivalent widths of the lines that we observe only in the stacked spectra. If the average spectral slope was flat through the rest-frame optical, we could simply scale the median [O III] EW of the stacked galaxies by the measured flux ratios in the stack. However, this is not necessarily a good approximation, except for perhaps Hβ, because it is close in wavelength to [O III]. Therefore, we estimate the continuum at the wavelengths for Hα, Hβ, Hγ, and Hδ, using the same linear interpolation that we described above for estimating the continuum at [O III] from F110W or F125W and F160W. Then, we take (showing Hγ for an example)

$$W_{\rm stack}({\rm H}\gamma) = W_{\rm stack}([{\rm O\ III}])\left(\frac{F({\rm H}\gamma)}{F([{\rm O\ III}])}\right)_{\rm stack} \times \left(\frac{F_\lambda(5007)}{F_\lambda(4341)}\right)_{\rm stack}. \quad \text{(A1)}$$

Here, $W_{\rm stack}([{\rm O\ III}])$ is the median of the [O III] equivalent widths that are included in a given spectral stack, $(F({\rm H}\gamma)/F([{\rm O\ III}]))_{\rm stack}$ is the relative line flux ratio measured from that stack, and $(F_\lambda(5007))/F_\lambda(4341))_{\rm stack}$ is the median of the continuum flux ratios for the galaxies in that stack. In calculating the medians of stacked samples, we exclude the small minority of 25 galaxies where the broadband flux estimates at [O III] are measured at less than 3σ significance.

The uncertainties on the equivalent width estimates for lines other than [O III] are taken by propagating errors in Equation (A1). For $W_{\rm stack}([{\rm O\ III}])$ and $(F_\lambda(5007)/F_\lambda(4341))_{\rm stack}$, we take the error on the mean (the rms of the individual measurements divided by the square root of the sample size in the stack). The uncertainty on the relative flux ratio $(F({\rm H}\gamma)/F([{\rm O\ III}]))_{\rm stack}$ is measured directly from bootstrap simulations, as described in Section 3.1. Overall, this approach provides estimates for the equivalent widths of the Balmer lines in the stacked spectra, as well as their uncertainties, so that we can correct the stack flux ratios for stellar absorption in our measurements of dust and metallicity.

Appendix B
Dust Extinction Correction Before or After Stacking Spectra?

In Section 3.1, we described our methodology for stacking. In short, spectra are continuum subtracted and then normalized by the [O III] line flux to avoid weighting toward the sources with the strongest emission lines. In this approach, the spectra are not corrected for dust extinction before stacking. Rather, an average extinction correction is applied to each stack, based on lines that we measure in the stack (and in some cases a prior based on lower-redshift data; see Appendix C). For most galaxies in the sample, poor knowledge of the individual extinction corrections necessitates this strategy. However, if the galaxies in a given stack exhibit a range in their dust extinction,

then normalizing by [O III] flux can bias the relative [O II] flux to higher values from systems with little dust. A similar effect is present for Hα, for redshifts where it is observed. In this case, the [O III] normalization creates a bias toward strong Hα emission from galaxies with more dust. The consequence of these biases is that dust extinction from Hα/Hβ could be systematically high, while [O III]/[O II] ratios are biased toward galaxies with low dust. Hence, correcting the stack-measured [O II] using the dust extinction also measured from the stack could underestimate the [O III]/[O II] ratio from the combination of these two effects.

We quantified this possible bias by correcting individual spectra for dust extinction prior to stacking, using the subset of galaxies where this is possible. We take the 49 galaxies with Hα S/N > 10 presented in Section 3.2, and divide this subsample into two mass bins with $\log M/M_{\odot} \leqslant 9.49$ and $\log M/M_{\odot} > 9.49$ (25 and 24 objects, respectively). We made stacks of these same galaxies using our normal method, where extinction correction is done after stacking. We find $12 + \log(\mathrm{O/H}) = 8.25 \pm 0.02$ and 8.39 ± 0.01 for the low- and high-mass bins, respectively. On the other hand, when we apply the dust extinction that we derive for these individual galaxies before stacking, we find $12 + \log(\mathrm{O/H}) = 8.27 \pm 0.02$ and 8.39 ± 0.01. For comparison, the means of the individual metallicity measurements for these two mass bins are $12 + \log(\mathrm{O/H}) = 8.25$ and 8.40, in excellent agreement with the stack results from either method. Therefore, we conclude that our stacking method is robust to systematic biases due to variations in dust extinction among the samples.

Appendix C
Bayesian Inference of Metallicity and Dust Extinction

In order to calculate metallicities for both the stacked samples and individual galaxies, the spectra must be corrected for dust extinction and Balmer line stellar absorption. Assessing these corrections and how their uncertainties translate to uncertainties in metallicity is best addressed through a Bayesian analysis. In this way, we can take advantage of Hγ and Hδ detections, or upper limits, while accounting for (or marginalizing over) stellar absorption and the contamination of Hα by [N II] and Hγ by [O III] λ4363. Our method is similar to the approaches used previously for grism spectra by Jones et al. (2015b) and Wang et al. (2017, 2019, 2020).

We use Bayes' theorem to write the posterior probability distribution of our model:

$$p(\theta|R) = \frac{p(R|\theta)p(\theta)}{p(R)}, \tag{C1}$$

where θ is the set of four physical parameters that model our data: metallicity, $E(B-V)_{\mathrm{gas}}$, the equivalent width of Hβ stellar absorption, and the ratio of [O III] λ4363/Hγ fluxes. Likewise, R represents the set of line flux ratios that our model must reproduce: (Hα+ [N II])/Hβ, (Hγ + [O III] λ4363)/Hβ, Hδ/Hβ, as well as $R_2 \equiv \log([\mathrm{O\,II}]/\mathrm{H}\beta)$, $R_3 \equiv \log([\mathrm{O\,III}]/\mathrm{H}\beta)$, and $O_{32} \equiv \log([\mathrm{O\,III}]/[\mathrm{O\,II}])$. Then, $p(\theta)$ is the prior distribution, and $p(R)$ is a constant that normalizes the posterior probability distribution. In most cases, we adopt flat priors, $p(\theta)$, that are bounded by reasonable parameter space. Details regarding the model and priors are as follows:

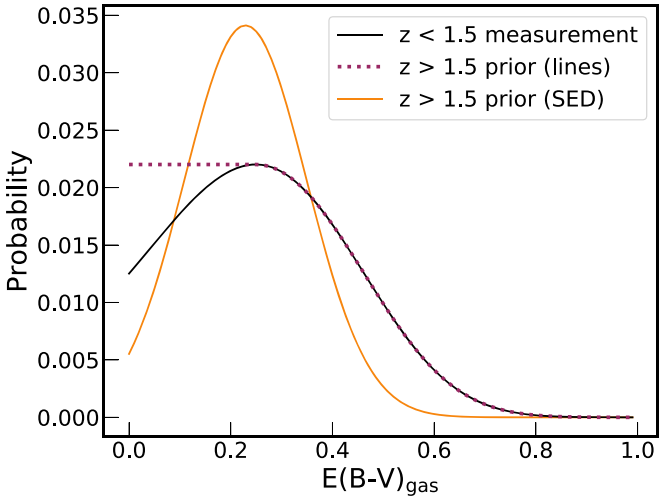

Figure 19. For stacked spectra and individual galaxies at $z > 1.5$, we adopt priors on dust extinction, as the constraints from the weaker Balmer lines are usually poor. Here, we show two possibilities that we considered. First, we derived priors from the posterior probability distributions for stacks at $z < 1.5$. This example highlights the posterior probability distribution for $E(B-V)$ for the stack of galaxies with $9.5 < \log M/M_{\odot} < 10.0$ (black curve). The prior that we use for the higher-redshift galaxies has the same high $E(B-V)$ tail but is flat to lower $E(B-V)$, as illustrated by the magenta dotted curve. Alternatively, we tested a prior based on the extinction derived from SED fits, after a conversion to nebular extinction following Shivaei et al. (2020), shown by the orange curve. Priors are derived for each of the five mass bins that we use in this paper.

Table 5
Dust Extinction in Stacked Spectra

$\log(M/M_{\odot})$ (1)	$N_{z<1.5}$ (2)	$E(B-V)_{z<1.5}$ (3)	$E(B-V)_{\mathrm{all,\ est}}$ (4)
<8.5	18	$0.07^{+0.22}_{-0.07}$	0.12 ± 0.06
8.5–9.0	40	$0.00^{+0.16}_{-0.00}$	0.11 ± 0.05
9.0–9.5	57	$0.13^{+0.10}_{-0.08}$	0.14 ± 0.07
9.5–10.0	27	$0.25^{+0.16}_{-0.18}$	0.23 ± 0.12
>10.0	11	$0.55^{+0.09}_{-0.09}$	0.57 ± 0.29

Note. Two different estimates of extinction were considered for determining priors on dust. (1) The stellar mass bins used for these estimates are the same five that have been presented throughout this paper. (2) The number of galaxies in each mass bin at $z < 1.5$, which we stack to obtain constraints on the dust extinction. (3) Dust extinction derived from the Balmer decrement measured in stacks of galaxies at $z < 1.5$. While the measurements (or limits) for Hγ/Hβ and Hδ/Hβ are included, the most robust constraints come from Hα/Hβ. (4) An estimate of the mean nebular extinction for galaxies in the full redshift range of our sample, determined from their SED fit and multiplied by 2.1 following Shivaei et al. (2020). In this case, we adopt a 50% systematic uncertainty on the stellar-to-nebular extinction correction.

Metallicity—We consider metallicities over the range of $7.6 < 12 + \log(\mathrm{O/H}) < 8.9$, where the Curti et al. (2017) calibration is defined. To produce a model of our data, we calculate line ratios from their calibration, considering R_2, R_3, and O_{32}. We also model [N II]/Hα in order to account for the contamination of our Hα flux by (generally weak) [N II] emission. This factors into our calculation of dust extinction. However, as we discussed in Section 3.4, unlike the oxygen-based line ratios, the relationship between [N II]/Hα and metallicity likely does evolve to high redshifts. Therefore, we adopt the [N II]/Hα calibration from Strom et al. (2018): 12 +

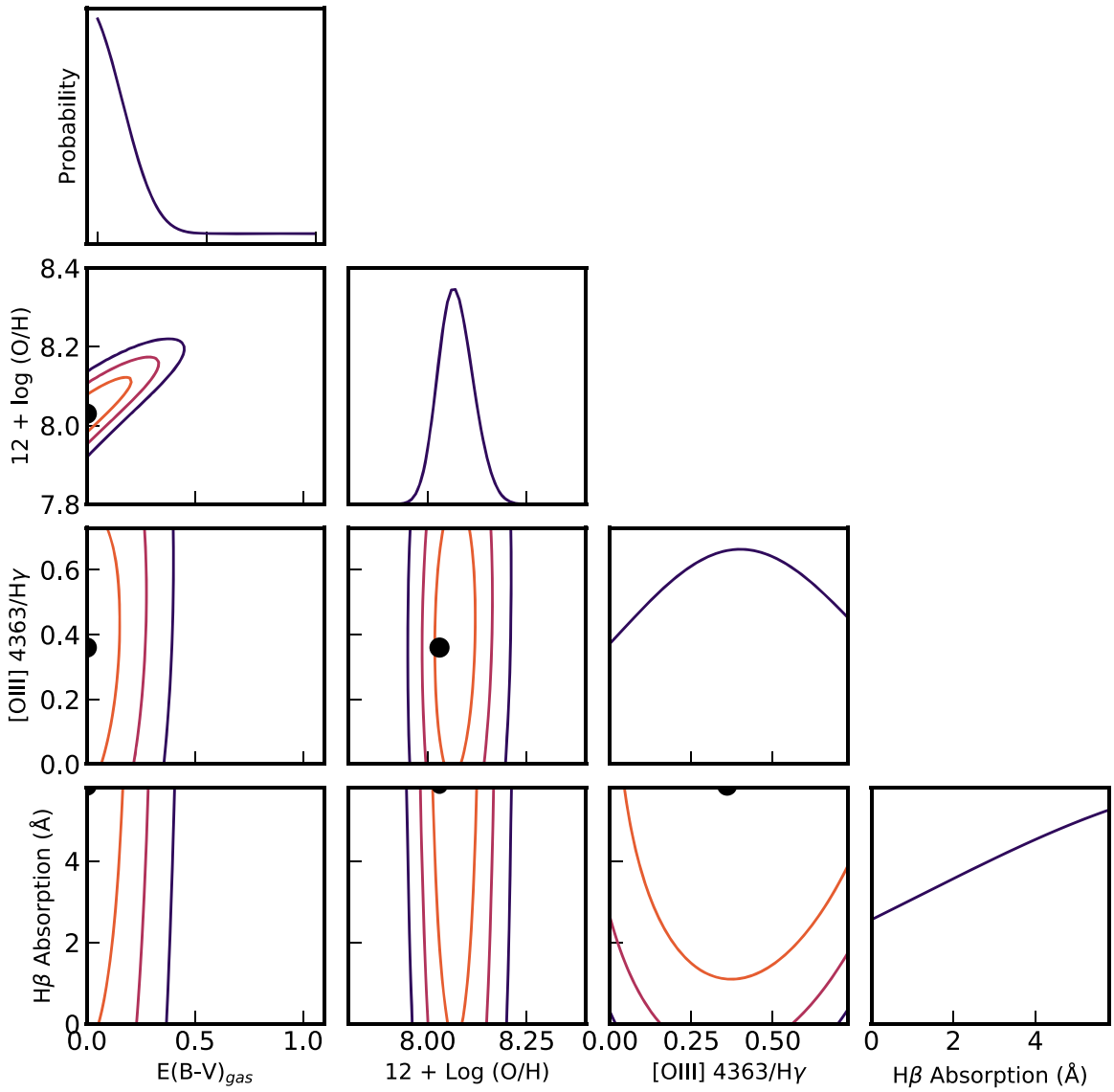

Figure 20. Contours show the 68%, 95%, and 99.7% confidence intervals for the four parameters that describe our model. The filled points show the most likely solution for this stacked spectrum. The one-dimensional posterior probability distribution for each parameter is shown in the panels on the diagonal. This example is for the 223 galaxies with masses between $10^{8.5}$ and $10^{9.0}$ $M_{\odot}$, covering the full redshift range of our sample ($1.3 < z < 2.3$).

$\log(\mathrm{O/H}) = 8.77 + 0.34 \times \log([\mathrm{N\,II}]/\mathrm{H}\alpha)$. Relative to local [N II]/Hα calibrations, this relation implies stronger [N II] at fixed metallicity. Because [N II] is still generally weak, the precise form of this correction does not strongly influence our metallicity inferences; however, the stronger [N II] contamination implied by this choice often results in better agreement between the dust extinction implied from Hα/Hβ and Hγ/Hβ.

Dust Extinction—The line ratios that we evaluate for the metallicity calculations are next reddened, according to Calzetti et al. (2000). We also consider the Balmer line ratios that are not used in the metallicity calibrations described above. We adopt intrinsic Balmer line ratios for $T_e = 10{,}000$ K: Hα/Hβ = 2.86, Hγ/Hβ = 0.468, Hδ/Hβ = 0.259. A hotter T_e does not have a significant impact on our dust extinction and subsequent metallicity estimates. When Hα is included in the spectrum ($z < 1.5$), we use a flat dust prior that extends from $0 < E(B-V)_{\mathrm{gas}} < 1.0$. When Hα is not included, we adopt a prior, which is used in addition to any constraints that come from the Hγ/Hβ and Hδ/Hβ ratios. We tested two methods for deriving the prior, which we illustrate in Figure 19 and quantify in Table 5. First, we derive a prior based on stacks of our $z < 1.5$ sample, using galaxies in the same mass range. Extinction equivalent to or lower than at $1.3 < z < 1.5$ is preferred for higher redshifts, but higher values are still allowed. This prior is illustrated by the dashed magenta curve in Figure 19. Second, we used the mean of the extinction derived from the SED fits of the galaxies, assuming $A_{V,\mathrm{nebular}}/A_{V,\mathrm{stellar}} = 2.1$ (Shivaei et al. 2020), and a 50% systematic uncertainty due to the stellar-to-nebular extinction conversion. An example of this prior is shown as the orange curves in Figure 19. As is evident from Table 5, the nebular extinction inferred from SED fits for galaxies in our entire redshift range ($1.3 < z < 2.3$) is very similar to the dust reddening derived from a stacked sample of galaxies at $z < 1.5$. Ultimately, the method of estimating the prior has a negligible impact on metallicity that we derive (0.01–0.02 dex) We therefore choose to use the prior based on Balmer lines in

the $z < 1.5$ stacks, so that systematic uncertainties on the stellar-to-nebular extinction conversion can be avoided.

Balmer Line Stellar Absorption—All line ratios involving Balmer lines are corrected for stellar absorption, reducing the modeled flux of individual lines to match the observed data. In order to choose a range of stellar absorption for our model, we examine spectra from both continuous and instantaneous burst models in Starburst99 (Leitherer et al. 1999), measuring the absorption equivalent widths in Hα, Hβ, Hγ, and Hδ. For populations with young stars, a reasonable range of Hβ stellar absorption equivalent widths is 0–6 Å, with the other Balmer lines tracking Hβ. Therefore, we adopt stellar absorption equivalent widths in Hγ and Hδ that are equal to that of Hβ, and a Hα stellar absorption equivalent width that is two-thirds that of Hβ. Then, the modeled line fluxes (and ratios) are corrected in proportion to their equivalent widths. For example:

$$f(\mathrm{H}\beta)^{\mathrm{abs}}_{\mathrm{mod}} = f(\mathrm{H}\beta)_{\mathrm{mod}} \frac{|W(\mathrm{H}\beta)^{\mathrm{em}}|}{(|W(\mathrm{H}\beta)^{\mathrm{em}}| + |W(\mathrm{H}\beta)^{*}|)}, \quad \text{(C2)}$$

where $f(\mathrm{H}\beta)^{\mathrm{abs}}_{\mathrm{mod}}$ is the modeled flux that is reduced to account for stellar absorption and $f(\mathrm{H}\beta)_{\mathrm{mod}}$ is the modeled flux, prior to this correction. In this calculation, $W(\mathrm{H}\beta)^{\mathrm{em}}$ and $W(\mathrm{H}\beta)^{*}$ are the equivalent widths of the emission and the stellar absorption, respectively. The emission-line equivalent widths are estimated from broadband photometry; we describe these estimates and assess their accuracy in Appendix A.

Contamination of Hγ *by* [O III] λ4363—If we use Hγ to obtain constraints on dust extinction, we must correct for contamination by emission from [O III] λ4363. This feature is strongest at the low metallicities that we expect for our sample. Therefore, we allow a contribution from this line, spanning from 1/40 to 1/500 of [O III] $\lambda\lambda$4959,5007 for high and low electron temperatures (Osterbrock 1989). Although this emission-line contribution should correlate with metallicity, given the systematic uncertainties discussed in Section 3.4, we instead choose to model it independently.

Using these ingredients to model the line ratios that we observe, R_i, we evaluate the likelihood function:

$$p(R|\theta) = \prod_i e^{-(R_i - R_i^{\mathrm{mod}}(\theta))^2/(2\sigma_i^2)}. \quad \text{(C3)}$$

Here, R_i^{mod} describes the modeled line ratios, which are a function of θ (metallicity, stellar absorption equivalent width, dust extinction, and [O III] λ4363 contamination). Likewise, σ_i represents the statistical error on the observed line flux ratios. For ratios involving Balmer lines, we add in quadrature an additional error to account for the uncertainty on the correction for the stellar absorption. This error arises from the uncertainty on the measured emission-line equivalent widths (see Appendix A).

Figure 20 shows an example of our calculation, highlighting that metallicity is the best-constrained parameter. The contour plots and the one-dimensional posterior probability distributions (on the diagonal panels) show that the Hβ stellar absorption equivalent width and the [O III] λ4363/Hγ flux ratio are not well constrained by our data. This characteristic is shared by all of the stacks that we consider. However, the inclusion of these nuisance parameters in our model is important, as we marginalize over them to arrive at realistic uncertainties on metallicity and dust extinction. Curiously, Figure 20 shows that the metallicity does not depend strongly on dust extinction. This weak dependence is a consequence of the fact that our metallicity inferences require only a relative dust correction between [O III] and [O II], rather than an accurate absolute correction to model the reddened line fluxes.

ORCID iDs

Alaina Henry https://orcid.org/0000-0002-6586-4446
Marc Rafelski https://orcid.org/0000-0002-9946-4731
Camilla Pacifici https://orcid.org/0000-0003-4196-0617
Micaela Bagley https://orcid.org/0000-0002-9921-9218
Ivano Baronchelli https://orcid.org/0000-0003-0556-2929
Guillermo Barro https://orcid.org/0000-0001-6813-875X
James Colbert https://orcid.org/0000-0001-6482-3020
Y. Sophia Dai https://orcid.org/0000-0002-7928-416X
Bruce G. Elmegreen https://orcid.org/0000-0002-1723-6330
Debra Meloy Elmegreen https://orcid.org/0000-0002-1392-3520
Steven Finkelstein https://orcid.org/0000-0001-8519-1130
Anton Koekemoer https://orcid.org/0000-0002-6610-2048
Matthew Malkan https://orcid.org/0000-0001-6919-1237
Crystal L. Martin https://orcid.org/0000-0001-9189-7818
Vihang Mehta https://orcid.org/0000-0001-7166-6035
Casey Papovich https://orcid.org/0000-0001-7503-8482
Michael Rutkowski https://orcid.org/0000-0001-7016-5220
Jorge Sánchez Almeida https://orcid.org/0000-0003-1123-6003
Claudia Scarlata https://orcid.org/0000-0002-9136-8876
Gregory Snyder https://orcid.org/0000-0002-4226-304X
Harry Teplitz https://orcid.org/0000-0002-7064-5424

References

Alexandroff, R. M., Heckman, T. M., Borthakur, S., et al. 2015, ApJ, 810, 104
Andrews, B. H., & Martini, P. 2013, ApJ, 765, 140
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Atek, H., Malkan, M., McCarthy, P., et al. 2010, ApJ, 723, 104
Atek, H., Siana, B., Scarlata, C., et al. 2011, ApJ, 743, 121
Bagley, M. B., Scarlata, C., Mehta, V., et al. 2020, ApJ, 897, 98
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Baronchelli, I., Scarlata, C. M., Rodighiero, G., et al. 2020, ApJS, 249, 12
Barro, G., Pérez-González, P. G., Cava, A., et al. 2019, ApJS, 243, 22
Belli, S., Jones, T., Ellis, R. S., et al. 2013, ApJ, 772, 141
Berg, D. A., Skillman, E. D., Marble, A. R., et al. 2012, ApJ, 754, 98
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Binette, L., Matadamas, R., Hägele, G. F., et al. 2012, A&A, 547, A29
Bothwell, M. S., Maiolino, R., Kennicutt, R., et al. 2013, MNRAS, 433, 1425
Bradley, L., Sipőcz, B., Robitaille, T., et al. 2021, Zenodo
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Bresolin, F. 2007, ApJ, 656, 186
Bresolin, F., Kudritzki, R.-P., Urbaneja, M. A., et al. 2016, ApJ, 830, 64
Brown, J. S., Martini, P., & Andrews, B. H. 2016, MNRAS, 458, 1529
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Chabrier, G. 2003, PASP, 115, 763
Charlot, S., & Fall, S. M. 2000, ApJ, 539, 718
Chisholm, J., Tremonti, C., & Leitherer, C. 2018, MNRAS, 481, 1690
Coil, A. L., Aird, J., Reddy, N., et al. 2015, ApJ, 801, 35
Colbert, J. W., Teplitz, H., Atek, H., et al. 2013, ApJ, 779, 34
Cresci, G., Mannucci, F., Sommariva, V., et al. 2012, MNRAS, 421, 262
Croton, D. J., Stevens, A. R. H., Tonini, C., et al. 2016, ApJS, 222, 22
Curti, M., Cresci, G., Mannucci, F., et al. 2017, MNRAS, 465, 1384
Curti, M., Mannucci, F., Cresci, G., et al. 2020, MNRAS, 491, 944
Dalcanton, J. J. 2007, ApJ, 658, 941
Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827
Davé, R., Finlator, K., & Oppenheimer, B. D. 2012, MNRAS, 421, 98
Davé, R., Rafieferantsoa, M. H., Thompson, R. J., et al. 2017, MNRAS, 467, 115

Davies, B., Kudritzki, R.-P., Lardo, C., et al. 2017, ApJ, 847, 112
Dayal, P., Ferrara, A., & Dunlop, J. S. 2013, MNRAS, 430, 2891
De Lucia, G., Xie, L., Fontanot, F., & Hirschmann, M. 2020, MNRAS, 498, 3215
De Rossi, M. E., Bower, R. G., Font, A. S., et al. 2017, MNRAS, 472, 3354
Domínguez, A., Siana, B., Henry, A. L., et al. 2013, ApJ, 763, 145
Donley, J. L., Koekemoer, A. M., Brusa, M., et al. 2012, ApJ, 748, 142
Dopita, M. A., Sutherland, R. S., Nicholls, D. C., et al. 2013, ApJS, 208, 10
Eddington, A. S. 1913, MNRAS, 73, 359
Eldridge, J. J., & Stanway, E. R. 2016, MNRAS, 462, 3302
Eldridge, J. J., Stanway, E. R., Xiao, L., et al. 2017, PASA, 34, e058
Ellison, S. L., Patton, D. R., Simard, L., et al. 2008, ApJL, 672, L107
Emami, N., Siana, B., Weisz, D. R., et al. 2019, ApJ, 881, 71
Erb, D. K. 2008, ApJ, 674, 151
Erb, D. K., Shapley, A. E., Pettini, M., et al. 2006, ApJ, 644, 813
Estrada-Carpenter, V., Papovich, C., Momcheva, I., et al. 2019, ApJ, 870, 133
Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, ApJS, 154, 10
Ferland, G. J., Chatzikos, M., Guzmán, F., et al. 2017, RMxAA, 53, 385
Finlator, K., & Davé, R. 2008, MNRAS, 385, 2181
Galametz, A., Grazian, A., Fontana, A., et al. 2013, ApJS, 206, 10
García-Rojas, J., & Esteban, C. 2007, ApJ, 670, 457
Garn, T., & Best, P. N. 2010, MNRAS, 409, 421
Gburek, T., Siana, B., Alavi, A., et al. 2019, ApJ, 887, 168
Gillman, S., Tiley, A. L., Swinbank, A. M., et al. 2021, MNRAS, 500, 4229
Grasshorn Gebhardt, H. S., Zeimann, G. R., Ciardullo, R., et al. 2016, ApJ, 817, 10
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24
Guo, Y., Koo, D. C., Lu, Y., et al. 2016a, ApJ, 822, 103
Guo, Y., Rafelski, M., Faber, S. M., et al. 2016b, ApJ, 833, 37
Henry, A., Berg, D. A., Scarlata, C., et al. 2018, ApJ, 855, 96
Henry, A., Martin, C. L., Finlator, K., et al. 2013a, ApJ, 769, 148
Henry, A., Scarlata, C., Domínguez, A., et al. 2013b, ApJL, 776, L27
Hirschauer, A. S., Salzer, J. J., Janowiecki, S., et al. 2018, AJ, 155, 82
Horne, K. 1986, PASP, 98, 609
Izotov, Y. I., Thuan, T. X., & Guseva, N. G. 2012, A&A, 546, A122
Jones, T., Martin, C., & Cooper, M. C. 2015a, ApJ, 813, 126
Jones, T., Wang, X., Schmidt, K. B., et al. 2015b, AJ, 149, 107
Juneau, S., Bournaud, F., Charlot, S., et al. 2014, ApJ, 788, 88
Juneau, S., Dickinson, M., Alexander, D. M., et al. 2011, ApJ, 736, 104
Kacprzak, G. G., van de Voort, F., Glazebrook, K., et al. 2016, ApJL, 826, L11
Kashino, D., Silverman, J. D., Sanders, D., et al. 2017, ApJ, 835, 88
Kashino, D., Silverman, J. D., Sanders, D., et al. 2019, ApJS, 241, 10
Kennicutt, R. C. 1998, ARA&A, 36, 189
Kewley, L. J., & Dopita, M. A. 2002, ApJS, 142, 35
Kewley, L. J., Dopita, M. A., Leitherer, C., et al. 2013, ApJ, 774, 100
Kewley, L. J., & Ellison, S. L. 2008, ApJ, 681, 1183
Kewley, L. J., Zahid, H. J., Geller, M. J., et al. 2015, ApJL, 812, L20
Kobulnicky, H. A., & Kewley, L. J. 2004, ApJ, 617, 240
Kocevski, D. D., Hasinger, G., Brightman, M., et al. 2018, ApJS, 236, 48
Koekemoer, A. M., Faber, S., Ferguson, H., et al. 2011, ApJS, 197, 36
Kriek, M., Shapley, A. E., Reddy, N. A., et al. 2015, ApJS, 218, 15
Kudritzki, R. P., Castro, N., Urbaneja, M. A., et al. 2016, ApJ, 829, 70
Kümmel, M., Walsh, J. R., Pirzkal, N., et al. 2009, PASP, 121, 59
Lara-López, M. A., Cepa, J., Bongiovanni, A., et al. 2010, A&A, 521, L53
Lara-López, M. A., Hopkins, A. M., Lopez-Sanchez, A. R., et al. 2013, MNRAS, 433, L35
Lee, H., Skillman, E. D., Cannon, J. M., et al. 2006, ApJ, 647, 970
Lee, J. C., Gil de Paz, A., Kennicutt, R. C., et al. 2011, ApJS, 192, 6
Leitherer, C., Schaerer, D., Goldader, J., et al. 1999, ApJS, 123, 3
Lilly, S. J., Carollo, C. M., Pipino, A., et al. 2013, ApJ, 772, 119
Lower, S., Narayanan, D., Leja, J., et al. 2020, ApJ, 904, 33
Lu, Y., Wechsler, R. H., Somerville, R. S., et al. 2014, ApJ, 795, 123
Ly, C., Malkan, M. A., Rigby, J. R., et al. 2016, ApJ, 828, 67
Ma, X., Hopkins, P. F., Faucher-Giguère, C.-A., et al. 2016, MNRAS, 456, 2140
MacKenty, J. W., Kimble, R. A., O'Connell, R. W., et al. 2010, Proc. SPIE, 7731, 77310Z
Maier, C., Ziegler, B. L., Lilly, S. J., et al. 2015, A&A, 577, A14
Maiolino, R., Nagao, T., Grazian, A., et al. 2008, A&A, 488, 463
Mannucci, F., Cresci, G., Maiolino, R., et al. 2010, MNRAS, 408, 2115
Mannucci, F., Salvaterra, R., & Campisi, M. A. 2011, MNRAS, 414, 1263
Marinacci, F., Vogelsberger, M., Pakmor, R., et al. 2018, MNRAS, 480, 5113
Masters, D., Faisst, A., & Capak, P. 2016, ApJ, 828, 18
Masters, D., McCarthy, P., Siana, B., et al. 2014, ApJ, 785, 153
Mehta, V., Scarlata, C., Rafelski, M., et al. 2017, ApJ, 838, 29
Merlin, E., Bourne, N., Castellano, M., et al. 2016, A&A, 595, A97
Merlin, E., Fontana, A., Ferguson, H. C., et al. 2015, A&A, 582, A15
Mobasher, B., Dahlen, T., Ferguson, H. C., et al. 2015, ApJ, 808, 101
Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, ApJS, 225, 27
Naiman, J. P., Pillepich, A., Springel, V., et al. 2018, MNRAS, 477, 1206
Narayanan, D., Conroy, C., Davé, R., et al. 2018, ApJ, 869, 70
Nayyeri, H., Hemmati, S., Mobasher, B., et al. 2017, ApJS, 228, 7
Nelson, D., Pillepich, A., Springel, V., et al. 2018, MNRAS, 475, 624
Nicholls, D. C., Dopita, M. A., & Sutherland, R. S. 2012, ApJ, 752, 148
Nomoto, K., Kobayashi, C., & Tominaga, N. 2013, ARA&A, 51, 457
Osterbrock, D. E. 1989, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei (Sausalito, CA: University Science Books)
Pacifici, C., Charlot, S., Blaizot, J., et al. 2012, MNRAS, 421, 2002
Pacifici, C., Kassin, S. A., Weiner, B. J., et al. 2016, ApJ, 832, 79
Peeples, M. S., & Shankar, F. 2011, MNRAS, 417, 2962
Peimbert, M., Peimbert, A., Esteban, C., et al. 2007, RvMxAA, 29, 72
Pettini, M., & Pagel, B. E. J. 2004, MNRAS, 348, L59
Pillepich, A., Nelson, D., Hernquist, L., et al. 2018, MNRAS, 475, 648
Pilyugin, L. S., & Grebel, E. K. 2016, MNRAS, 457, 3678
Pilyugin, L. S., Grebel, E. K., & Mattsson, L. 2012, MNRAS, 424, 2316
Pilyugin, L. S., & Thuan, T. X. 2005, ApJ, 631, 231
Pirzkal, N., Malhotra, S., Ryan, R. E., et al. 2017, ApJ, 846, 84
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13
Porter, R. L., Ferland, G. J., Storey, P. J., et al. 2012, MNRAS, 425, L28
Rafelski, M., Teplitz, H. I., Gardner, J. P., et al. 2015, AJ, 150, 31
Rutkowski, M. J., Scarlata, C., Haardt, F., et al. 2016, ApJ, 819, 81
Salim, S., Lee, J. C., Davé, R., et al. 2015, ApJ, 808, 25
Salim, S., Lee, J. C., Ly, C., et al. 2014, ApJ, 797, 126
Salpeter, E. E. 1955, ApJ, 121, 161
Sánchez Almeida, J., Elmegreen, B. G., Muñoz-Tuñón, C., et al. 2014, A&ARv, 22, 71
Sanders, R. L., Shapley, A. E., Jones, T., et al. 2021, ApJ, 914, 19
Sanders, R. L., Shapley, A. E., Kriek, M., et al. 2015, ApJ, 799, 138
Sanders, R. L., Shapley, A. E., Kriek, M., et al. 2016, ApJ, 816, 23
Sanders, R. L., Shapley, A. E., Kriek, M., et al. 2018, ApJ, 858, 99
Sanders, R. L., Shapley, A. E., Reddy, N. A., et al. 2020, MNRAS, 491, 1427
Sanders, R. L., Shapley, A. E., Zhang, K., et al. 2017, ApJ, 850, 136
Shapley, A. E., Reddy, N. A., Kriek, M., et al. 2015, ApJ, 801, 88
Shirazi, M., Brinchmann, J., & Rahmati, A. 2014, ApJ, 787, 120
Shivaei, I., Reddy, N., Rieke, G., et al. 2020, ApJ, 899, 117
Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, ApJS, 214, 24
Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51
Springel, V., Pakmor, R., Pillepich, A., et al. 2018, MNRAS, 475, 676
Stasińska, G. 2002, RvMxAA, 12, 62
Stefanon, M., Yan, H., Mobasher, B., et al. 2017, ApJS, 229, 32
Steidel, C. C., Rudie, G. C., Strom, A. L., et al. 2014, ApJ, 795, 165
Steidel, C. C., Strom, A. L., Pettini, M., et al. 2016, ApJ, 826, 159
Strom, A. L., Steidel, C. C., Rudie, G. C., et al. 2017, ApJ, 836, 164
Strom, A. L., Steidel, C. C., Rudie, G. C., et al. 2018, ApJ, 868, 117
Theios, R. L., Steidel, C. C., Strom, A. L., et al. 2019, ApJ, 871, 128
Tody, D. 1986, Proc. SPIE, 627, 733
Tody, D. 1993, adass II, 52, 173
Topping, M. W., Shapley, A. E., Sanders, R. L., et al. 2021, MNRAS, 506, 1237
Torrey, P., Vogelsberger, M., Hernquist, L., et al. 2018, MNRAS, 477, L16
Torrey, P., Vogelsberger, M., Marinacci, F., et al. 2019, MNRAS, 484, 5587
Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
Trump, J. R., Konidaris, N. P., Barro, G., et al. 2013, ApJL, 763, L6
Trump, J. R., Weiner, B. J., Scarlata, C., et al. 2011, ApJ, 743, 144
van Dokkum, P., Brammer, G., Momcheva, I., et al. 2013, arXiv:1305.2140
Villforth, C., Koekemoer, A. M., & Grogin, N. A. 2010, ApJ, 723, 737
Wang, B., Heckman, T. M., Leitherer, C., et al. 2019, ApJ, 885, 57
Wang, X., Jones, T. A., Treu, T., et al. 2017, ApJ, 837, 89
Wang, X., Jones, T. A., Treu, T., et al. 2019, ApJ, 882, 94
Wang, X., Jones, T. A., Treu, T., et al. 2020, ApJ, 900, 183
Whitaker, K. E., Franx, M., Leja, J., et al. 2014, ApJ, 795, 104
Wisnioski, E., Förster Schreiber, N. M., Wuyts, S., et al. 2015, ApJ, 799, 209
Wright, S. A., Larkin, J. E., Law, D. R., et al. 2009, ApJ, 699, 421
Wuyts, E., Rigby, J. R., Sharon, K., et al. 2012, ApJ, 755, 73
Wuyts, E., Wisnioski, E., Fossati, M., et al. 2016, ApJ, 827, 74
Yabe, K., Ohta, K., Akiyama, M., et al. 2015a, PASJ, 67, 102
Yabe, K., Ohta, K., Akiyama, M., et al. 2015b, ApJ, 798, 45
Yabe, K., Ohta, K., Iwamuro, F., et al. 2012, PASJ, 64, 60
Yates, R. M., Kauffmann, G., & Guo, Q. 2012, MNRAS, 422, 215
Zahid, H. J., Kashino, D., Silverman, J. D., et al. 2014, ApJ, 792, 75
Zahid, H. J., Kewley, L. J., & Bresolin, F. 2011, ApJ, 730, 137