University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Optimized Photometric Redshifts for the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS)

Dritan Kodra,[1,2] Brett H. Andrews,[1,2] Jeffrey A. Newman,[1,2] Steven L. Finkelstein,[3] Adriano Fontana,[4] Nimish Hathi,[5] Mara Salvato,[6] Tommy Wiklind,[7] Stijn Wuyts,[8] Adam Broussard,[9] Nima Chartab,[10] Christopher Conselice,[11] M. C. Cooper,[12] Avishai Dekel,[13] Mark Dickinson,[14] Henry C. Ferguson,[5] Eric Gawiser,[9] Norman A. Grogin,[5] Kartheik Iyer,[15] Jeyhan Kartaltepe,[16] Susan Kassin,[5,17] Anton M. Koekemoer,[5] David C. Koo,[18] Ray A. Lucas,[5] Kameswara Bharadwaj Mantha,[19,20] Daniel H. McIntosh,[21] Bahram Mobasher,[22] Camilla Pacifici,[5] Pablo G. Pérez-González,[23] and Paola Santini[4]

[1]*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA*

[2]*Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA 15260, USA*

[3]*Department of Astronomy, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA*

[4]*Osservatorio Astronomico di Roma, via di Frascati 33, I-00078 Monte Porzio Catone (Roma), Italy*

[5]*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*

[6]*Max-Planck-Institut für extraterrestrische Physik, Giessenbachstr. 1, 85748 Garching, Germany*

[7]*Department of Physics, Catholic University of America, Washington D.C. 20064, USA*

[8]*Department of Physics, University of Bath, Claverton Down, Bath BA2 7AY, UK*

[9]*Department of Physics and Astronomy, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA*

[10]*Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA*

[11]*Jodrell Bank Centre for Astrophysics, University of Manchester, Oxford Road, Manchester, UK*

[12]*Center for Cosmology, Department of Physics & Astronomy, University of California, Irvine, 4129 Reines Hall, Irvine, CA 92697, USA*

[13]*Racah Institute of Physics, The Hebrew University, Jerusalem 91904 Israel*

[14]*NSF's NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA*

[15]*Dunlap Institute for Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada*

[16]*School of Physics and Astronomy, Rochester Institute of Technology, Rochester, NY 14623, USA*

[17]*The William H. Miller III Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*

[18]*Department of Astronomy and Astrophysics, University of California, Santa Cruz, Santa Cruz, CA 95064, USA*

[19]*Minnesota Institute for Astrophysics, University of Minnesota, 116 Church St SE, Minneapolis, MN 55455, USA*

[20]*School of Physics and Astronomy, University of Minnesota, 116 Church St SE, Minneapolis, MN 55455, USA*

[21]*Department of Physics & Astronomy, University of Missouri, Kansas City, Kansas City, MO 64110, USA*

[22]*Department of Physics and Astronomy, University of California, Riverside, 900 University Ave, Riverside, CA 92521, USA*

[23]*Centro de Astrobiología, CAB/CSIC-INTA, Ctra. de Torrejón a Ajalvir, km 4, 28850, Torrejón de Ardoz, Madrid, Spain*

ABSTRACT

We present the first comprehensive release of photometric redshifts (photo-z's) from the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) team. We use statistics based upon the Quantile-Quantile (Q–Q) plot to identify biases and signatures of underestimated or overestimated errors in photo-z probability density functions (PDFs) produced by six groups in the collaboration; correcting for these effects makes the resulting PDFs better match the statistical definition of a PDF. After correcting each group's PDF, we explore three methods of combining the different groups' PDFs for a given object into a consensus curve. Two of these methods are based on identifying the minimum f-divergence curve, i.e., the PDF that is closest in aggregate to the other PDFs in a set (analogous to the median of an array of numbers). We demonstrate that these techniques yield improved results using sets of spectroscopic redshifts independent of those used to optimize PDF modifications. The best photo-z PDFs and point estimates are achieved with the minimum f-divergence using the best 4 PDFs for each object (mFDa4) and the Hierarchical Bayesian (HB4) methods, respectively. The HB4 photo-z point estimates produced $\sigma_{\mathrm{NMAD}} = 0.0227/0.0189$ and $|\Delta z/(1+z)| > 0.15$ outlier fraction = $0.067/0.019$

Corresponding author: Brett H. Andrews
andrewsb@pitt.edu

for spectroscopic and 3D-HST redshifts, respectively. Finally, we describe the structure and provide guidance for the use of the CANDELS photo-z catalogs, which are available at https://archive.stsci.edu/hlsp/candels.

1. INTRODUCTION

Cosmic Noon ($z \sim 2$) saw the most intense star formation in the history of the Universe; the rapid buildup of galactic disks and bulges; and the end of star formation in the most massive galaxies. Deep galaxy surveys investigate this critical epoch of galaxy formation by tracking the evolving galaxy demographics via the galaxy luminosity function (e.g., Cole et al. 2001; Reddy & Steidel 2009; Finkelstein et al. 2015), stellar mass function (e.g., Marchesini et al. 2009; Muzzin et al. 2013; Duncan et al. 2014; Tomczak et al. 2014; Davidzon et al. 2017), and the stellar mass–star formation rate relation (e.g., Noeske et al. 2007; Whitaker et al. 2012; Speagle et al. 2014; Salmon et al. 2015). The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS; Grogin et al. 2011; Koekemoer et al. 2011) combines high spatial resolution Hubble Space Telescope (HST) Wide Field Camera 3 (WFC3)/IR and Advanced Camera for Surveys (ACS) imaging with existing intermediate resolution optical and IR imaging down to faint limits (5σ point source limit of $H \sim 27$ for wide and $H \sim 27.7$ for deep) to extend our knowledge of these important galaxy evolution probes. However, the measurements required for these probes rely on accurate redshifts.

Consequently, there have been multiple concerted efforts on 8–10 m class telescopes to obtain spectroscopic redshifts (spec-z's) for CANDELS galaxies (e.g., VVDS Le Fèvre et al. 2005; zCOSMOS Lilly et al. 2007a; VUDS Le Fèvre et al. 2015; and VANDELS McLure et al. 2018). Spec-z's are the gold standard for accurate redshifts, but the galaxies with successfully measured spec-z's tend to be brighter, at lower redshifts, and preferentially biased towards star-forming galaxies with strong emission lines. Given the resource-intensive nature of measuring spec-z's, spectroscopic followup of a large and representative sample, especially of the faintest CANDELS objects, is impossible with current and near-future facilities (Newman et al. 2015). Grism-based redshifts (grism-z's) from HST (e.g., 3D-HST Momcheva et al. 2016a) can help alleviate some of the issues with ground-based spec-z's because they can cover many more galaxies and probe a more representative population of galaxies (Bezanson et al. 2016). Nevertheless, grism spectroscopy is technically challenging, does not cover the full depth of the CANDELS imaging, and is not available for the whole CANDELS footprint. Thus, we must rely on photometric redshifts (photo-z's) for distance information (Baum 1957) across the full range of galaxies,

albeit at lower accuracy and precision than spec-z's or grism-z's.

Photo-z estimation approaches mostly fall into two main groups—machine learning (ML) based methods and template-based methods—with the optimal approach depending on the amount and quality of training data and science goals (see Salvato et al. 2019 and references therein). ML methods estimate redshift from galaxy photometry by training an ML model on galaxies with spec-z's. ML methods outperform template-based methods in the regime of complete and representative coverage of color–z space, such as Sloan Digital Sky Survey photo-z efforts (e.g., Beck et al. 2016; Pasquet et al. 2019; Hayat et al. 2021; Dey et al. 2022a), where data-driven interpolation is very accurate. Template-based methods involve redshifting spectral energy distribution (SED) templates (or combinations thereof) to find the best fit to a galaxy's SED. Template-based methods are superior to ML methods when spectroscopic coverage is incomplete and/or biased (e.g., in deep fields) because they leverage our knowledge of galaxy physics. Hence, only template-based methods have been employed for CANDELS due to the biased and incomplete spec-z training sets. However, shortcomings in our knowledge of galaxy physics can limit the success of template-based methods; issues include template mismatch (i.e., when the SED templates differ from the observed SEDs), incorrect treatment of dust attenuation, and inappropriate priors (Salvato et al. 2019; Newman & Gruen 2022).

Template libraries fall intro two broad classes, empirical and theoretical, each with its own relative advantages and disadvantages. Empirical templates (e.g., Coleman et al. 1980 and Kinney et al. 1996) are principally derived from spectroscopic observations of nearby galaxies (and supplemented with model spectra to increase wavelength coverage), so they can accurately capture spectral features in real galaxies. However, they are based on a relatively small set of observations at very low redshift, and do not span the full range of potential ages, metallicities, or galaxy types in CANDELS. Empirical templates can be optimized for high redshift galaxies through calibration (e.g., Ilbert et al. 2006), but this process requires a large spectroscopic redshift sample that is usually not available and not representative of the photometric galaxy sample.

Theoretical template sets are constructed from stellar population synthesis models (e.g., Fioc & Rocca-Volmerange 1997a, Bruzual & Charlot 2003, and Maraston 2005a), so they cover more of color–redshift space and perform better at

high redshift (Brammer et al. 2008a). All of the groups in this work used theoretical template libraries. However, theoretical template libraries cannot fully cover color–redshift space, and the mismatch between template and observed SEDs remains one of the biggest sources of systematic uncertainty in photo-z estimation using template-based methods (Newman & Gruen 2022). Furthermore, theoretical template libraries must assume star formation histories for their template SEDs, include (or not) emission lines, and adopt a dust attenuation law. These choices represent a simplified version of the properites of real galaxies, especially given the wide range in redshift, luminosity, color, and galaxy type of CANDELS galaxies. Priors can help break degeneracies between multiple redshift solutions by incorporating knowledge about the galaxy population as a whole (e.g., the abundance of galaxies as a function of luminosity and redshift), but they can introduce biases. For instance, in Figure we show that the stacked redshift PDFs from different codes produce different redshift peaks because of differences in template sets and assumed priors even though they all perform well according to the point estimate metrics used.

Photo-z methods can be evaluated using spectroscopic training sets (e.g., Hogg et al. 1998; Hildebrandt et al. 2008, 2010), clustering measurements (Newman 2008), and mock catalogs generated from simulations (Schmidt et al. 2020), but for CANDELS only the first is a viable option. Most relevant to this work is the comparison by Dahlen et al. (2013) that studied 11 photo-z analyses of the same CANDELS photometry and spec-z data set. These analyses utilized a diversity of codes, template sets, and priors—all of which impact the resulting photo-z PDFs. Dahlen et al. (2013) found that none of the photo-z codes significantly outperformed the others. However, the analyses that used spec-z's to calibrate the photo-z PDFs (e.g., by implementing zero point offsets to the photometry or modifying the templates) achieved better results. They found that photo-z PDFs from individual analyses were generally too narrow to be consistent with the statistical definition of a PDF (as did Bezanson et al. 2016), for which they implemented a correction. Finally, they were able to improve the photo-z PDFs by combining the PDFs from individual analyses via several methods: taking the median, summing PDFs, and a Hierarchical Bayesian approach.

This paper builds off of Dahlen et al. (2013) by comparing a more recent effort by 6 groups running 5 different photo-z codes with a larger spec-z data set and grism-z's to better study performance at fainter magnitudes where the low signal-to-noise ratio of the photometry can drive discrepant results between analyses. We calibrate the PDFs from individual codes and introduce a better method for combining PDFs, the minimum f-divergence(Rényi 1961). We also test the performance of each group's PDFs using point estimates

(specifically, the weighted mean redshift z_{weight}). Our results are presented as the final CANDELS photo-z catalog.

This paper is organized as follows. Section 2 describes the data sets employed in this paper. Section 3 explains the methods used to improve PDFs from individual photo-z codes and assesses the results with a variety of summary statistics. In Section 4 we test several methods of combining PDF results from multiple photo-z codes. Section 5 presents final photo-z catalog for the CANDELS fields. Finally, Section 6 summarizes the major findings of this study and provides guidance on the use of these catalogs.

Throughout the paper we assume AB magnitudes unless differently stated. In order to allow direct comparison with existing works from the literature of X-ray surveys, we adopt a flat ΛCDM cosmology with $h = H_0/[100 \,\text{km s}^{-1}\text{Mpc}^{-1}] = 0.7$; Ω_M=0.3; Ω_Λ=0.7.

2. DATA

2.1. *Photometry*

Throughout this study we use data from the CANDELS collaboration. CANDELS obtained observations in HST/WFC3 F105W, F125W, and F160W as the primary exposures and HST/ACS F606W and F814W as parallel exposures. These observations covered five fields spanning a total area of $\sim 800 \,\text{arcmin}^2$ on the sky with $\sim 125 \,\text{arcmin}^2$ in the deep portion and $\sim 675 \,\text{arcmin}^2$ in the wide portion. The deep portion has a depth of ~ 13 orbits per pointing (spread across three filters) and includes data in two of the CANDELS fields, GOODS-North and GOODS-South. The wide portion consists of ~ 2 orbits per pointing and includes buffer regions around the deep fields as well as three additional fields, COS-MOS, the Extended Groth Strip (EGS), and the UKIDSS Ultra-Deep Survey field (UDS). Detailed descriptions of the sky coverage and observing strategy of CANDELS can be found in Grogin et al. (2011) and Koekemoer et al. (2011).

One of the most powerful aspects of the CANDELS data set is the wealth of complementary photometry for these heavily studied extragalactic fields. We utilized matched-model TFIT (Laidler et al. 2007) photometry (from the v1 photometric catalogs) from ground- and space-based UV to IR imaging as the input for the photometric redshift codes. The sources and bands vary somewhat between fields but are typically u-band through Spitzer/IRAC 8 μm with F160W serving as the selection band for the photometric catalogs. Full descriptions of the catalogs for each CANDELS field are given in Nayyeri et al. (2017, COSMOS), Stefanon et al. (2017, EGS), Guo et al. (2013, GOODS-S), Galametz et al. (2013, UDS), and Barro et al. (2019, GOODS-N). In the interest of uniformity across fields, we only used broad-band photometry. For instance, we did not utilize the medium-band photometry from the SHARDS survey (Pérez-González et al. 2013) for the GOODS-N field, so for certain use cases (e.g., environ-

ment) the higher precision photo-z's from Barro et al. (2019) that use those bands will be preferable.

2.2. *Individual Photometric Redshifts*

Six groups within the CANDELS collaboration have used the TFIT photometric catalogs to estimate photo-z PDFs for each galaxy observed by CANDELS. For each galaxy, the codes fit a set of SED templates in redshift space and minimized χ^2 between the observed and template SEDs. To calculate a redshift posterior PDF, most codes use a magnitude dependent redshift prior for each object combined with a likelihood proportional to $\exp(-\chi^2/2)$. To measure photo-z's, the groups used different codes for calculation, different template SEDs, and/or different criteria when using a code (e.g., different priors). In the following, we briefly explain each of the codes used to estimate photo-z PDFs in this paper.

EAZY (Brammer et al. 2008b)[1] fits a linear combination of stellar population templates to the observed U-to-8 μm SEDs (in flux space) by minimizing the χ^2 statistics. Corrections for absorption by intervening HI clouds are applied following Madau (1995). EAZY allows down-weighting data at rest-frame wavelengths $\geq 2\,\mu$m via a template error function. It incorporates a "luminosity-function $\times$ volume" prior that reduces the probability of low redshifts (due to the small volume) and of high redshifts for objects with bright apparent magnitudes. In addition, an iterative application of photometric zero point offsets improves the match to the available spectroscopic redshifts.

EAZY was run by two separate groups, led by S. Finkelstein and S. Wuyts, respectively. The two groups differed in their choices of parameters and priors. The Wuyts group used the "EAZY_v1.0_lines" templates, which are identical to the original "EAZY_v1.0" templates (based on the PÉGASE stellar population synthesis models; Fioc & Rocca-Volmerange 1997b) but with additional emission lines added following the prescription by Ilbert et al. (2008); the 6 templates contain a subset of 5 "principal component" templates constructed as described by Brammer et al. (2008b), complemented by an additional template to account for dusty galaxies that do not appear in the semi-analytical models that the principal component set was based on. The Finkelstein group used the "EAZY_v1.1_lines" templates, whose first six templates are the same as the templates from the "EAZY_v1.0_lines" set except that all of the PÉGASE emission lines were replaced using the Ilbert et al. (2008) prescription. It also has an additional template to account for massive old red galaxies.

zphot (Giallongo et al. 1998; Fontana et al. 2000) is a code that minimizes χ^2 statistics when fitting template SEDs. It

returns the best fitting values and estimated photo-z uncertainties as well as other parameters of the galaxy template (e.g., age, metallicity, and dust extinction). zphot accepts both fluxes and magnitudes as input with a rigorous treatment of non-detections in the latter case. The code allows for a minimum photometric error to be set in each photometric band and optionally applies a cut to avoid unrealistically large negative fluxes. For the templates, we used the PÉGASE 2.0 models (Fioc & Rocca-Volmerange 1997b), which assumed the stellar initial mass function from Rana & Basu (1992) because they yielded the most accurate photo-z's with the set of parameters explored. Emission lines were not included in the templates. Our adopted template library was constructed with a combination of star formation histories (SFHs). Specifically, we used constant and truncated SFHs as well as a SFH with the star formation rate proportional to the amount of gas. Metallicity evolution is measured self-consistently during the galaxy evolution. We implemented dust extinction with a Calzetti et al. (2000) extinction law with E(B-V) ranging from 0 to 1. Finally, to avoid contamination from non-stellar emission it excludes IRAC bands that probe rest frame $> 5.5\,\mu$m (ch3 at $z \leq 0.15$ and ch4 at $z \leq 0.6$). This code was run by the group led by A. Fontana.

HyperZ (Bolzonella et al. 2000)[2] performs fits by minimizing the χ^2 statistics in flux space while excluding negative fluxes. The user can specify shifts to magnitudes. To avoid unrealistic solutions, a prior for the F160W band absolute magnitude in the range $-30 < M < -9$ (Vega mag) was used. Reddening was included in the form of the Calzetti et al. (2000) reddening law varying A_V from 0 to 3 in steps of 0.2. SED templates are based on the Maraston (2005b) models with exponentially declining SFHs (with e-folding times $\tau = 0.1, 0.3, 1$, and 2 Gyr) at four metallicity values (1/5, 1/2, 1, and 2 $\times$ solar metallicity). The allowed age range of the templates was restricted between 0.1 Gyr and the age of the Universe at the given redshift of the galaxy in question. Finally, the option for a minimum photometric error was set to 0.05 magnitudes. This recipe was introduced in Pforr et al. (2013). This code was run by the group led by J. Pforr.

LePhare (Arnouts & Ilbert 2011)[3] minimizes χ^2 statistics when fitting template SEDs, both in magnitude and in flux space. It has the capability to include luminosity priors; to add extra photometry errors; to use a training sample to optimize the template SEDs and to derive zero-point offsets; and to account for contributions from emission lines (Ilbert et al. 2006, 2009). The median values from the photo-z PDFs are also provided. Here, a prior on the optical absolute magnitude

---

[1] http://www.astro.yale.edu/eazy/

[2] http://webast.ast.obs-mip.fr/hyperz/

[3] http://www.cfht.hawaii.edu/~arnouts/LEPHARE/lephare.html

in the range $-24 < M < -8$ is used, while IRAC ch3 and ch4 are excluded. The SED template set was a combination of the BC03 and Polletta et al. (2007) templates. This code was run by the group led by M. Salvato.

WikZ (Wiklind et al. 2008) fits SED templates to observed photometry in flux space by minimizing χ^2 statistics. The code can be run with two different parameterized SFHs, an exponentially-declining star formation rate or a delayed-τ SFH. In the current application, the code was run using a delayed-τ SFH with the BC03 SED templates. Negative fluxes are not completely excluded, but they add to the χ^2 when the template flux is brighter than the 1σ upper limit, while they are excluded when the flux is lower than the 1σ upper limit. Additionally, it excludes IRAC ch3 and ch4 for $z < 0.5$ and $z < 0.7$, respectively. Finally, it has the capability to add extra smoothing errors to the photometry. This code was run by the group led by T. Wiklind.

2.3. *Spectroscopic and 3D-HST Grism Redshifts*

Spectroscopic redshift are used to train photometric redshift codes (e.g., by identifying zero point offsets in observed-frame photometry that, if removed, will improve fits) and to test the accuracy of the photo-z estimates. The six photo-z participant groups used a training set of 5807 high quality spectroscopic redshifts spanning all five CANDELS fields. The same set of redshifts are used to recalibrate the PDFs in Section 3.

Our primary testing set consists of 4089 high quality spectroscopic redshifts drawn from a variety of sources completely independent of those included in the training set with any overlapping objects removed. We use this set to assess the performance of point statistics (e.g., the redshift of maximum probability) and to test the quality of photo-z PDFs from each code, both before and after the optimization procedure described below.

We also use 3D−HST grism redshifts (grism-z's; Momcheva et al. 2016a) to test the photo-z point estimate and PDF quality. This set consists of 3367 of the highest-quality redshifts spanning all of the CANDELS fields. The grism-z's were determined spectrophotometrically with EAZY, but run by other investigators with different priors and using different photometry. To mitigate the impact of incorporating photo-z information in the grism redshift fits, we restricted to objects where the spectrum, not the photo-z, drove the fit. We also include only objects with redshifts larger than $z_{grism} > 0.6$.

Figure 1 and Figure 2 show the H-band magnitude as a function redshift for the spectroscopic sample, as well as the marginal histograms for each of the five CANDELS fields separately. Table 6 gives details regarding the construction of these three sets, including references to the original catalogs and the specific cuts applied to select only the highest-quality, most secure redshifts.

Due to the order in which spec-z's were obtained, the objects in the training sets used to tune the photometric redshift methods differ in both galaxy properties (such as brightness) and redshift coverage from the testing and 3D-HST sets, as illustrated in Figure 1 and Figure 2. This difference is greatest in the COSMOS and EGS fields. As a result, the performance metrics derived using these independent data sets would be expected to be worse than would be measured using a random training/testing assignment from a common population. In fact, the objects for which we estimate photo-z's extend to fainter magnitudes and at higher redshifts than the training sets, so this more pessimistic assessment is likely more realistic for assessing the performance for the broader population, though we cannot exclude the possibility that photo-z errors are worse in regimes for which we have no spectroscopy.

3. OPTIMIZATION METHODS

The photo-z PDFs produced by the different groups are not analogous to the true probability distributions that meet the statistical definition of a PDF (see e.g., Dey et al. 2021, Dey et al. 2022b). Given the high likelihood that the spectroscopic redshifts used here are correct (>95%), we can treat them to first order as representing the true redshifts of the objects observed. If the statistical definition of a PDF is being obeyed, then we would expect 68% and 95% of them to fall in the 68% and 95% credible intervals of the photo-z PDFs, respectively.

In an earlier test with CANDELS data, Dahlen et al. (2013) found that while some codes had better results than others, none of them performed well when the coverage of credible intervals was assessed in this way. This indicates that the photo-z PDFs must have substantial, qualitative problems. Simple examples of issues would be the presence of bias due to template mismatch, such that the PDFs are shifted to higher or lower redshifts than expected, or inaccuracies in photometric error models that caused the PDFs to be too wide or too narrow.

In this section, we will describe how we optimize the photo-z PDFs produced by each group for each CANDELS field. We correct for (1) systematic shifts in the PDFs and (2) systematically inaccurate PDF widths.

3.1. *Q–Q Plot and Metrics*

In this work, we make use of a set of statistics related to the Quantile-Quantile (Q–Q) plot to optimize the recalibration of the photo-z PDFs produced by each group. While these statistics will prioritize adjustments that improve agreement with the statistical definition of a PDF, they should also deliver more accurate redshift point values and error estimates by reducing biases. Figure 3 depicts the Q–Q plot for a set of objects from the spec-z training set with Finkelstein photo-z's as an illustrative example. This plot shows Q_{Data}—the fraction of the time that the spectroscopic redshift is below
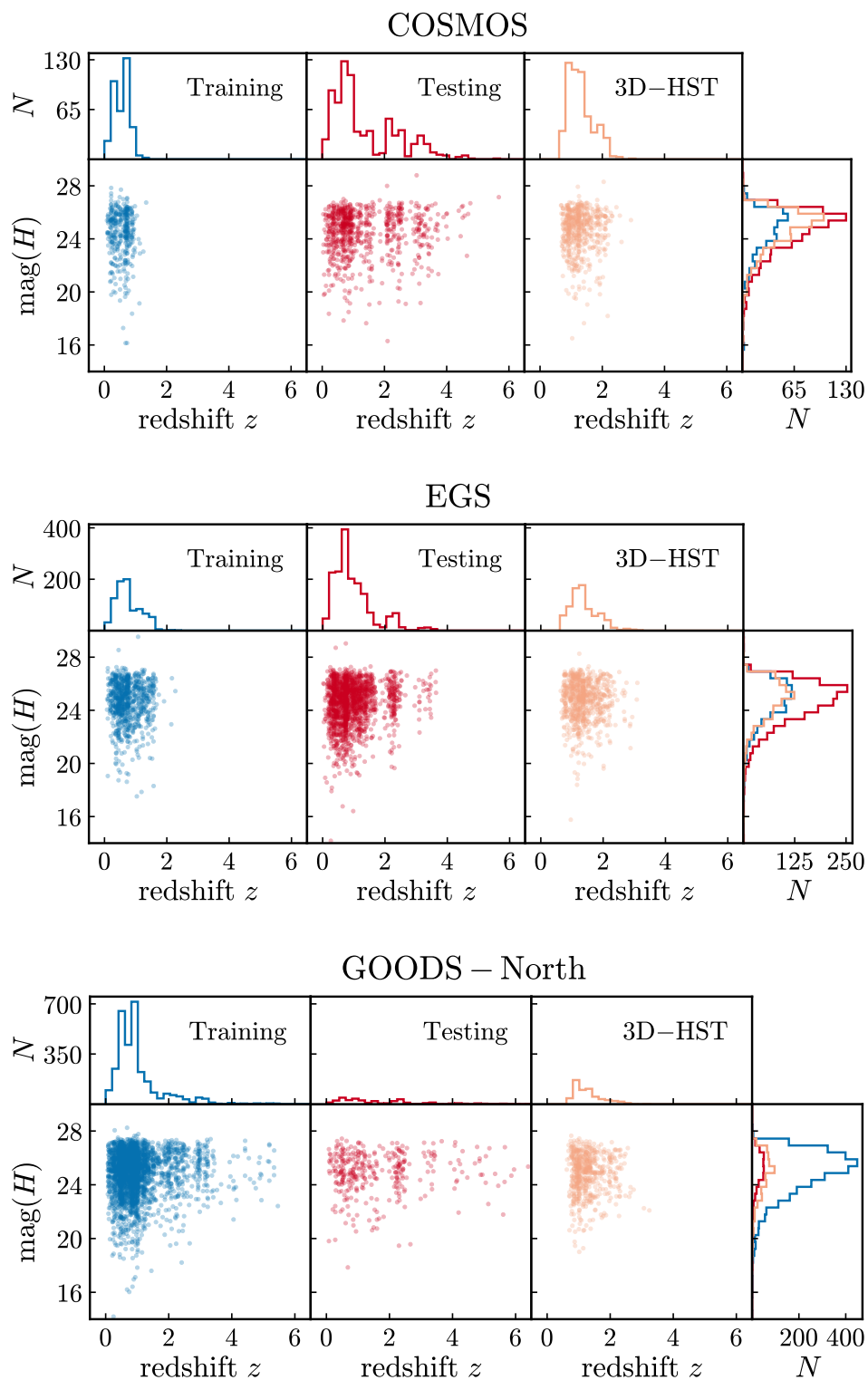
Figure 1. *H*-band magnitude as a function of redshift for objects with spectroscopic redshifts (divided into training and testing sets) and 3D-HST grism redshifts for the COSMOS, EGS, and GOODS-N fields. The redshift and magnitude ranges of the testing and 3D-HST data sets strongly differ from the training set; as a result, they provide highly independent assessments of the quality of photo-*z* PDFs which are optimized based upon the training set.
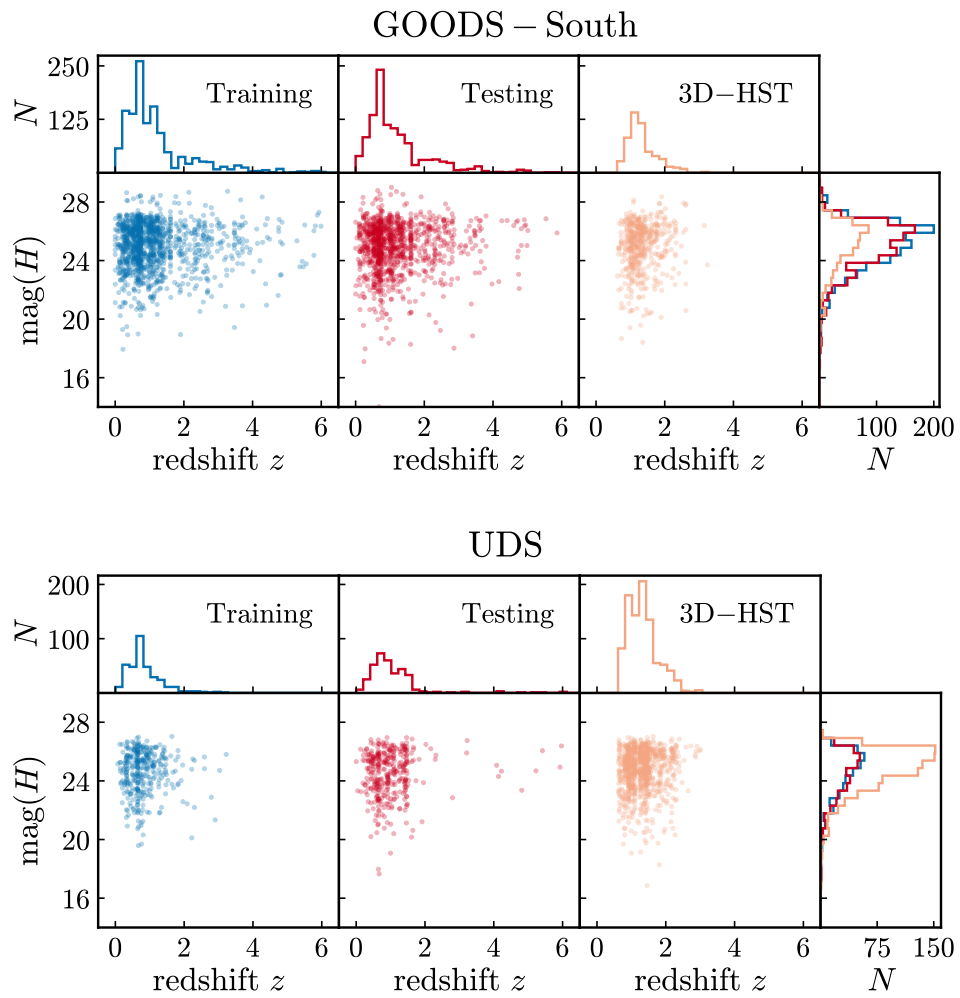
Figure 2. Same as Figure 1 except showing the GOODS-S and UDS fields.

the redshift corresponding to a given quantile in the photo-z cumulative distribution function (CDF)—as a function of the chosen value for the quantile, Q_{Theory}. By construction, the value of Q_{Theory} must range from 0 to 1, just as the CDF does (e.g., $Q_{\text{Theory}} = 0.345$ corresponds directly to the CDF value of 0.345). In our example, $Q_{\text{Data}}(0.345)$ will be the fraction of the time that the spectroscopic redshift of an object is below the redshift where the CDF value for that object is $Q_{\text{Theory}} = 0.345$. If the PDFs satisfy the statistical definition, Q_{Theory} will be the same as Q_{Data} (apart from small variations due to sampling error), so the Q–Q plot will hew closely to the unit line between (0, 0) and (1, 1). To help in interpreting the Q–Q plots shown in this paper, Figure 4 demonstrates how systematic shifts or incorrect widths of PDFs will manifest in them.

We use the normalized ℓ^2-norm to quantify the deviation of the Q–Q plot from the identity line to assess how close a set of PDFs is to the ideal case. The quantity $\Delta Q = Q_{\text{Data}} - Q_{\text{Theory}}$, evaluated at a particular value of Q_{Data}, will correspond to the gap between the Q–Q curve for a given sample and the unity line (doing this as a function of Q_{Data} has computational advantages). The ℓ^2-norm will then correspond to the square root of the sum of this distance squared, $\left(\int (\Delta Q)^2 dQ_{\text{Data}}\right)^{1/2}$. This distance will be zero in the ideal case.

For this analysis, we only include galaxies whose spectroscopic redshift lies between the 0.0001 and 0.9999 quantiles of their CDF in the construction of the Q–Q plot and in the calculation of the ℓ^2-norm. Objects outside that range correspond to catastrophic outliers (as commonly occur in photometric redshift analyses) whose true redshift lies beyond the range that would be expected based on their estimated PDF. We do not want their distribution to affect the recalibration of PDFs for the much more common non-outlier objects. Excluding them also makes our analysis considerably more robust to the possibility of erroneous spectroscopic redshifts. Hence, in the end, we calculate a normalized version of the ℓ^2-norm by summing up the value of $(\Delta Q)^2$ at the Q_{Data} values corresponding to each object in the sample excluding those which were outside these CDF bounds, dividing by the total number of $(\Delta Q)^2$ values which were summed, and then taking the square root.

The second basic statistic we use to characterize the quality of PDFs from a given algorithm is the fraction of objects for which their spectroscopic redshift has a CDF value less than 0.0001 or greater than 0.9999. We label this quantity as f_{op} (for "fraction outside of PDF"). This statistic roughly corresponds in its nature to the catastrophic outlier rate used to characterize the performance of point estimates from photometric redshift algorithms but is based on PDF outputs rather than scalar quantities.

To understand possible redshift and magnitude-dependent biases among the PDFs generated by different participant
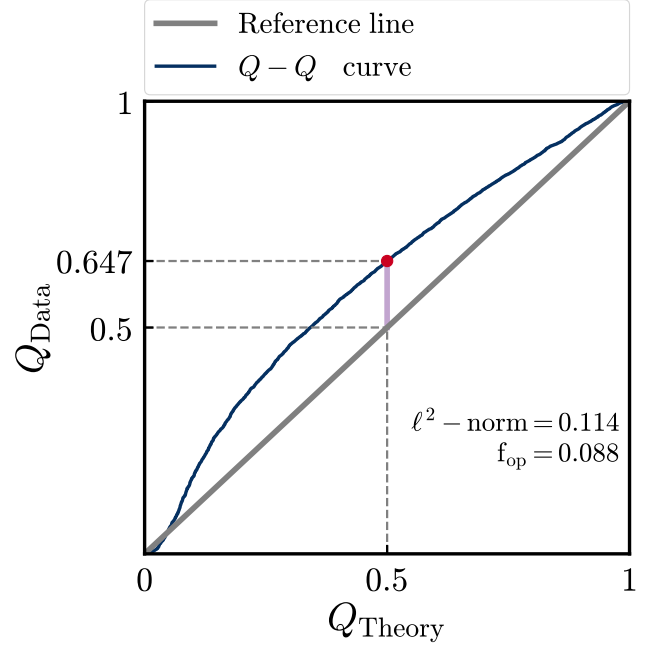

Figure 3. An example Quantile-Quantile (Q–Q) plot constructed using a training set of spec-z's and the corresponding set of photo-z PDFs from Finkelstein. For a given photo-z CDF quantile (Q_{Theory}), Q_{Data} is the fraction of objects for which the spectroscopic redshift is below the redshift corresponding to that photo-z quantile in the CDF for that particular object. For instance, if the spectroscopic redshift is below the 50th percentile point in the photo-z CDF for 64.7% of objects, then (0.5, 0.647) is a point (red circle) along the Q–Q curve (blue line). If photo-z PDFs obey the statistical definition of a probability density function, Q–Q curves will lie along the solid gray reference line from (0, 0) to (1, 1). Their deviation is quantified by the square root of the sum of the vertical distance (pink line) at each point squared (ℓ^2-norm). The ℓ^2-norm is normalized by dividing by the square root of the number of objects. Additionally, the f_{op} gives the fraction of objects with spec-z's that fall outside of the main region of the corresponding PDFs.

codes, we analyze the normalized P(z) in a fixed magnitude bin. In Figure 5, we show the set of the summed PDF curves for EGS objects in the magnitude bin centered at $H = 25$ predicted by each group, both before and after the optimization procedures (described below) have been applied. Even after optimization, the predicted number of objects at low redshifts varies greatly from group to group. Results in this regime may have limited reliability. In Figure 6, we show a set of images constructed from curves such as those shown in Figure 5. These images show redshift as a function of magnitude (using the center of the magnitude bin used to construct the summed PDF curves). The intensity of the color scale indicates the value of the summed PDF at a given magnitude and redshift (corresponding to the y values in Figure 5). Although all of the codes used to produce the photo-z PDFs yield good results when evaluated with standard point statistics for objects with
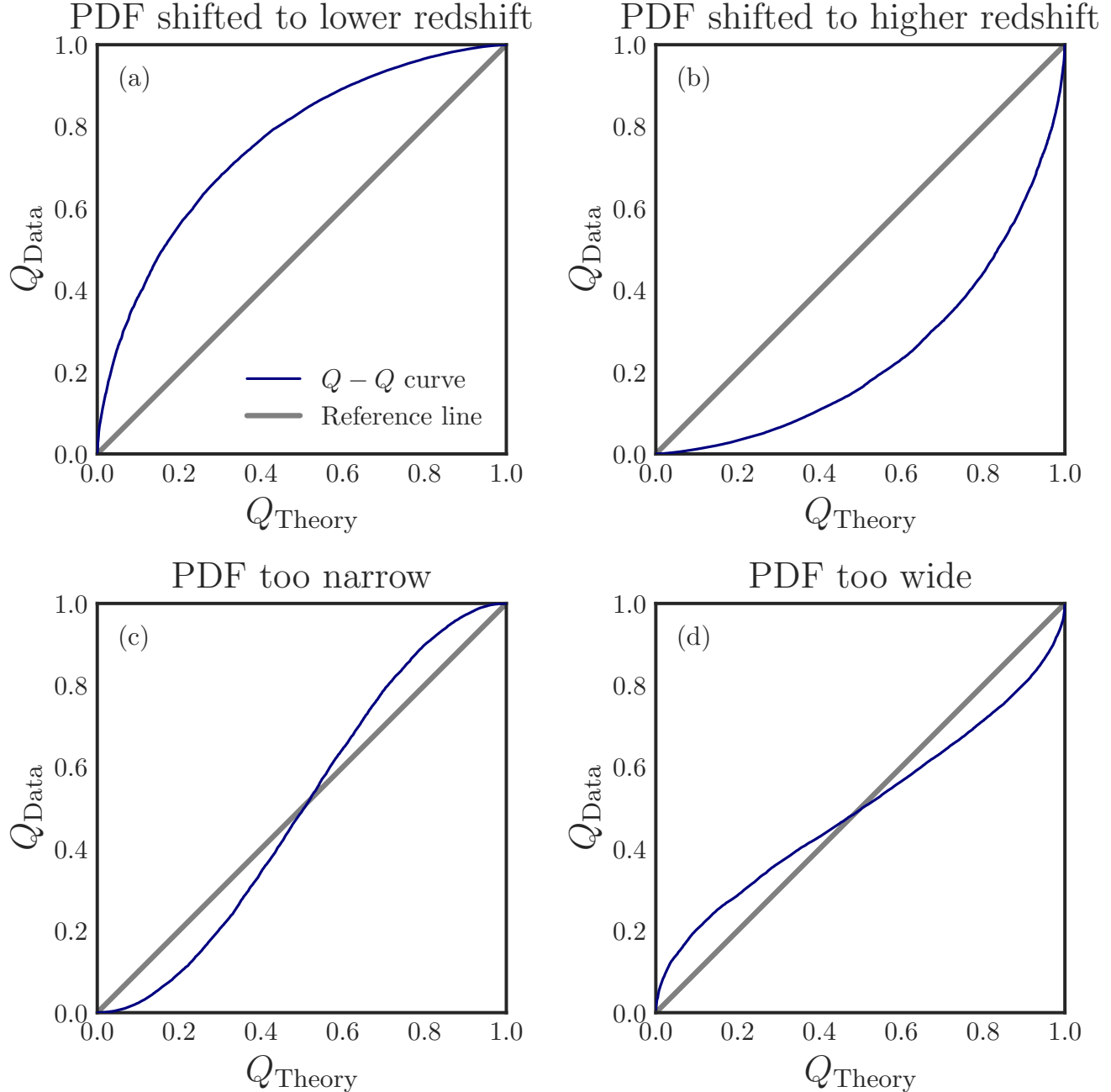
PDF shifted to lower redshift

PDF shifted to higher redshift

PDF too narrow

PDF too wide

Figure 4. Example Q–Q plots showing the effects of photo-z PDFs that are systematically (a) shifted to lower redshift, (b) shifted to higher redshift, (c) too narrow, and (d) too wide. Ideally, PDFs would lie along the gray diagonal reference line.

spectroscopic redshifts (as described below), and all are using the same photometry as inputs, the redshift distributions as a function of magnitude from each group differ in many details, even after optimizing the individual photo-z PDFs such that they better obey the statistical definition. We find that photo-z PDFs by different groups show divergent behavior at $z < 0.3$. Therefore, we exclude these objects for our further analysis.

3.2. *Correcting for Global Shifts*

Using the Q–Q statistics defined above, we can identify any aggregate bias in the photo-z PDFs of a given group by applying negative or positive shifts in the redshift direction (dz) to the PDFs. Our sign convention is such that negative dz values correspond to shifts to the left on a plot of PDF(z), such that the PDFs will peak at lower values of redshifts,

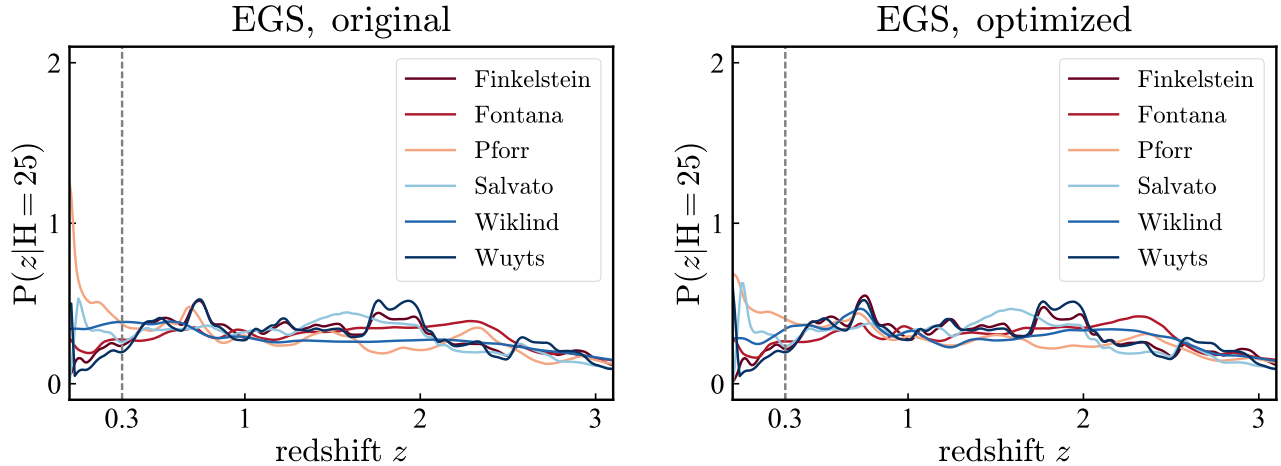## EGS, original

EGS, optimized

Figure 5. Sum of the photo-z PDFs for all objects in a magnitude bin of width 0.5 centered at $H = 25$ mag in the EGS field. The curves correspond to the expectation value for the redshift distribution in this bin. We exclude objects with missing PDFs for any of the groups from the sums to ensure that all curves are equivalent. The disagreement between groups is greatest at low redshifts (below $z = 0.3$, indicated by the gray dashed line). Because of this divergence, we exclude objects with $z \leq 0.3$ from the calculation of Q–Q statistics. Although all codes deliver good performance when PDF peak redshifts are compared to spec-z's, the aggregate predictions for broad galaxy samples differ significantly from code to code.

and vice versa. We apply shifts corresponding to an array of 151 equally-spaced values over the interval $[-0.5, 1.]$, and construct the Q–Q curve in each case after removing objects with spec-z's outside their PDFs. We then identify the shift value that minimizes the normalized ℓ^2-norm; that is, the value which yields a Q–Q curve as close as possible to the ideal unit line. To do this, we interpolate the normalized ℓ^2 – norm values using a quadratic univariate spline (using the `scipy.interpolate.UnivariateSpline` routine with parameters k = 2 and s = 1.) The goal of this interpolation is for greater accuracy and smoother results.

In Figure 7, we present the dependence of the normalized ℓ^2-norm on dz for each group and CANDELS field. Photo-z's from some groups exhibit larger biases than others, with the bias varying from field to field; this may be due to the differences in the imaging bands included in each field and the target selection function of spectroscopic surveys carried out in those fields. While the required shift for the COSMOS field is the smallest for four of the groups, we did not include all of the available medium and narrow band photometry for this field (that would lead to improved photo-z estimates) in the interest of uniformity across fields, so we expect that this feature may be due to the spectroscopy available in the COSMOS field. With only one exception (Pforr results for the GOODS-South field) the shifts are all positive. That indicates that for almost all fields and groups, the photo-z PDFs peak at lower values than they should, biasing point estimates (such as the redshift of the PDF peak) low. Table 1 gives the optimal shift values that should be applied to the photo-z PDFs from each group and each field to yield better performance in the Q–Q plot. We will assess the level of improvement achieved

in both Q–Q statistics and point estimates in sections to follow, using the independent testing set of spectroscopic redshifts as well as 3D-HST.

3.3. *Correcting PDF Width*

To correct for either overestimation or underestimation of photo-z errors, we stretch/sharpen the PDF widths. We do this by raising a PDF to a power of $\alpha = 1/\gamma$ and then renormalizing the PDF to integrate to one. For a Gaussian PDF, this is equivalent to multiplying the σ parameter by $\gamma^{1/2}$, and so provides a way to correct for either overestimates or underestimates of errors; however, the procedure we use can be applied to any PDF whether it is Gaussian or not. Optimal values of the parameter γ that are close to unity imply that little change is needed in the width of the PDFs. Conversely, when the optimal values differ significantly from unity, then large differences in the widths of the PDFs are implied (corresponding to significantly overconfident or underconfident error models).

We determine optimal values of γ after we apply the optimized shifts dz that were determined by the preceding analysis. We consider values in the range $0 < \gamma < 1$ as well as in the range $\gamma \geq 1$, corresponding to $\alpha > 1$ and $0 < \alpha \leq 1$, respectively. The former range corresponds to making photo-z PDFs sharper, since high peaks of the PDF become higher while low PDF values in the tails or valleys between peaks become smaller. Conversely, $\gamma \geq 1$ stretches the photo-z PDFs into broader shapes, since the contrast between the highest peaks and low values of the PDF is reduced.

For each group's PDFs in a particular field, we search a coarse array of γ values spanning $[0.05, 7]$; construct the Q–

Table 1. The optimal values of shifts dz for the PDFs from each group for each field, determined by identifying the minima in the curves shown in Figure 7. Most of the shifts are positive, indicating that PDF(z) curves need to be shifted to the right. Results from some groups exhibit larger biases than others, with Salvato and Wuyts needing the smallest shifts and Wiklind the largest, especially in the EGS field.

Field	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts
COSMOS	0.005	0.030	0.012	0.001	0.020	0.000
EGS	0.015	0.016	0.036	0.002	0.071	0.008
GOODS-N	0.020	0.003	0.001	0.007	0.036	0.015
GOODS-S	0.021	0.011	−0.015	0.013	0.036	0.011
UDS	0.022	0.014	0.031	0.013	0.046	0.008
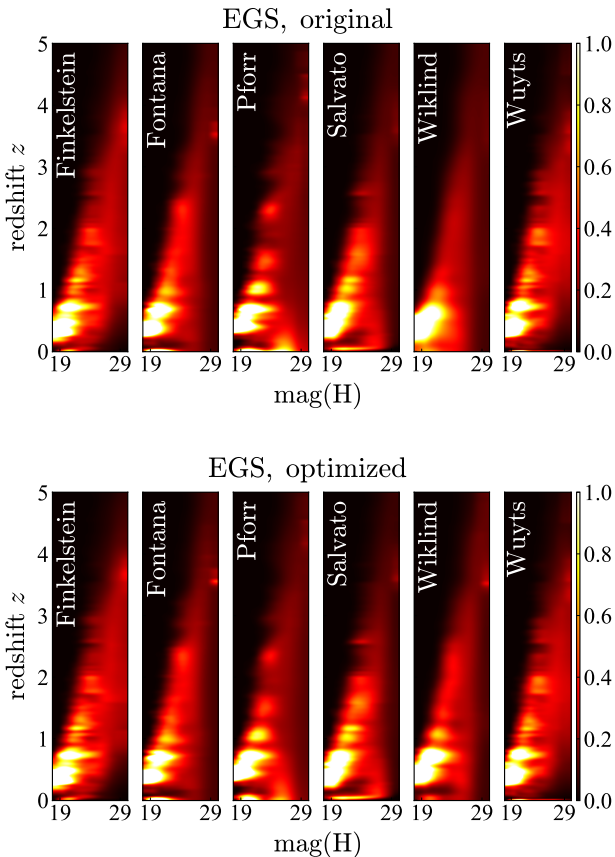
Figure 6. Summed PDFs as a function of magnitude and redshift, before and after the photo-z optimization procedures, for the EGS field. The color scale of the image indicates the summed and renormalized probability that an object in a particular magnitude bin is at the selected redshift, equivalent to the y axis in Figure 5. Objects with missing PDFs for any of the groups are excluded from the sums. The shape and locations of the bright regions (which should correspond to redshifts where there is an excess of objects due to sample/cosmic variance), as well as other detailed features, differ significantly from group to group, even though all groups used identical photometry for the same set of objects to calculate the PDFs used here.

Q plot in each case; and evaluate the normalized ℓ^2-norm for each value of γ. We then use a finer grid of γ values focused around the coarse-grid value that minimizes the ℓ^2-norm to obtain improved precision in the optimal value.

Figure 8 shows the dependence of the ℓ^2-norm on γ for each group and CANDELS field. The Wuyts PDFs are the best calibrated with all fields having best fit γ values near unity. The Salvato PDFs show moderate scatter around $\gamma = 1$ with some fields having PDFs that are too narrow and some too wide. The Finkelstein and Pforr PDFs are consistently too wide and too narrow, respectively. The Fontana and Wiklind PDFs demonstrate larger field-to-field scatter with some fields having PDFs that are too narrow and some too wide. In Table 2, we list the γ values required to optimize the raw photo-z PDFs such that they most closely match the statistical definition.

3.4. *Evaluation Using Q–Q Statistics*

After identifying the optimal values of the dz and γ parameters using the training set of spectroscopic redshifts and their corresponding photo-z PDFs, we can apply the corresponding transformations to all the PDFs from a particular group and field. We can then construct Q–Q plots and evaluate the overall normalized ℓ^2-norm and f_{op} values for each group evaluated with the training data (which should have results that are biased low) as well as with the independent testing set of spectroscopic redshifts and the 3D-HST grism redshifts. Since some fields have relatively few testing redshifts available, we combine the objects from all five CANDELS fields to make the Q–Q plots. We present the Q–Q plots for each group in Figure 9. In each panel, dashed curves correspond to the Q–Q plots for each spectroscopic data set before applying shift and power transformations, and solid curves show the Q–Q plots after transformations. It is clear that all curves more closely follow the diagonal reference line after optimization.

In Figure 10, we show the normalized ℓ^2-norm and f_{op} values before and after optimization of the PDFs. Although in a few cases the f_{op} value becomes slightly larger after optimization, gains in the normalized ℓ^2-norm are substan-
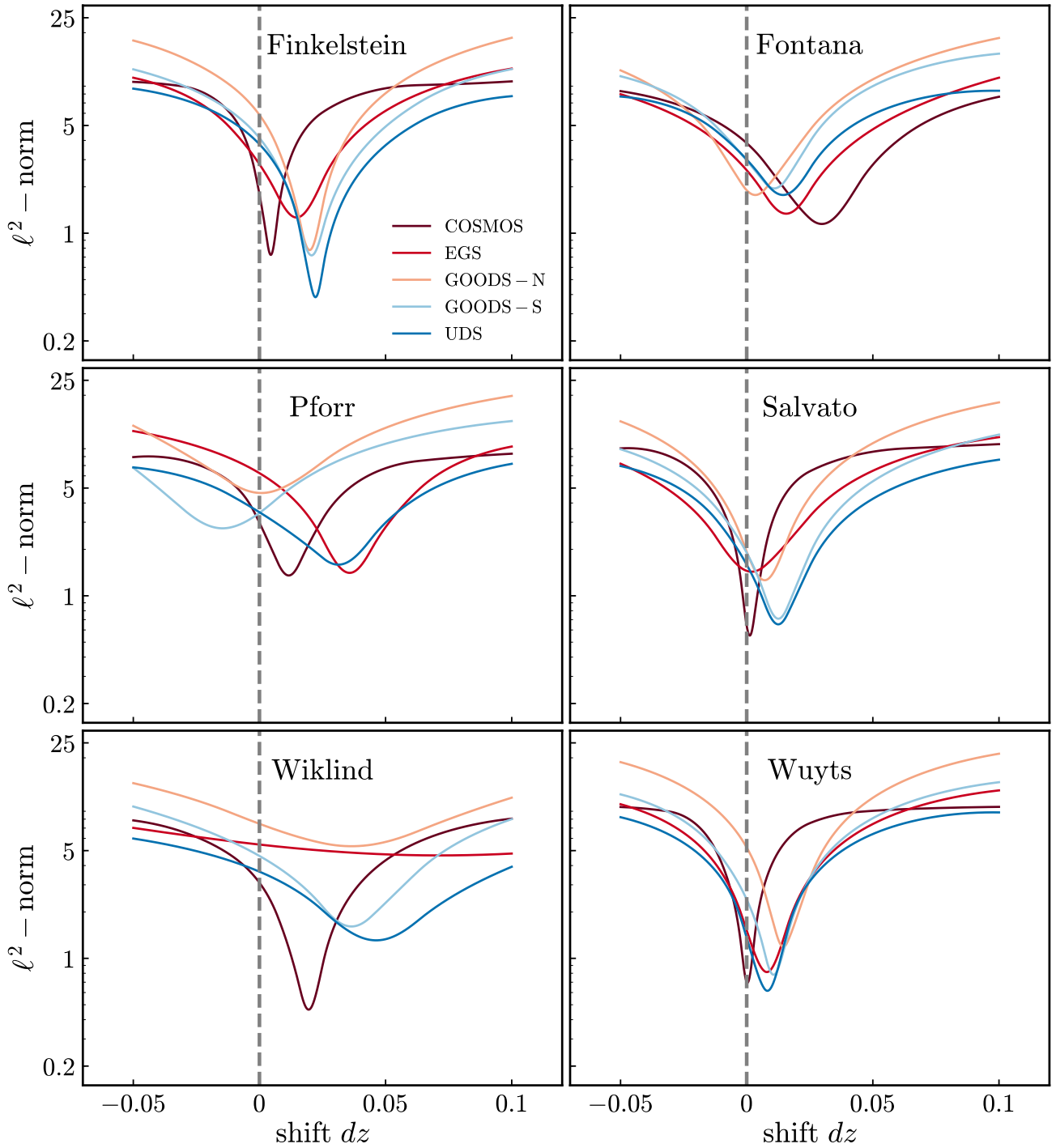
Figure 7. Normalized ℓ^2-norm as a function of the shift dz applied to the photo-z PDFs for each of the five CANDELS fields, calculated using the spectroscopic redshift training set. The normalized ℓ^2-norm will have a smaller value when the PDFs more closely obey the statistical definition of a probability distribution; as a result, we wish to minimize this quantity. The vertical gray dashed line indicates a zero shift, meaning that the calibration of the PDFs from a given group were accurate. A different shift seems to be needed for each group and each field, but nearly all of the shifts are positive. Consequently, the uncorrected photo-z point estimates are biased low.
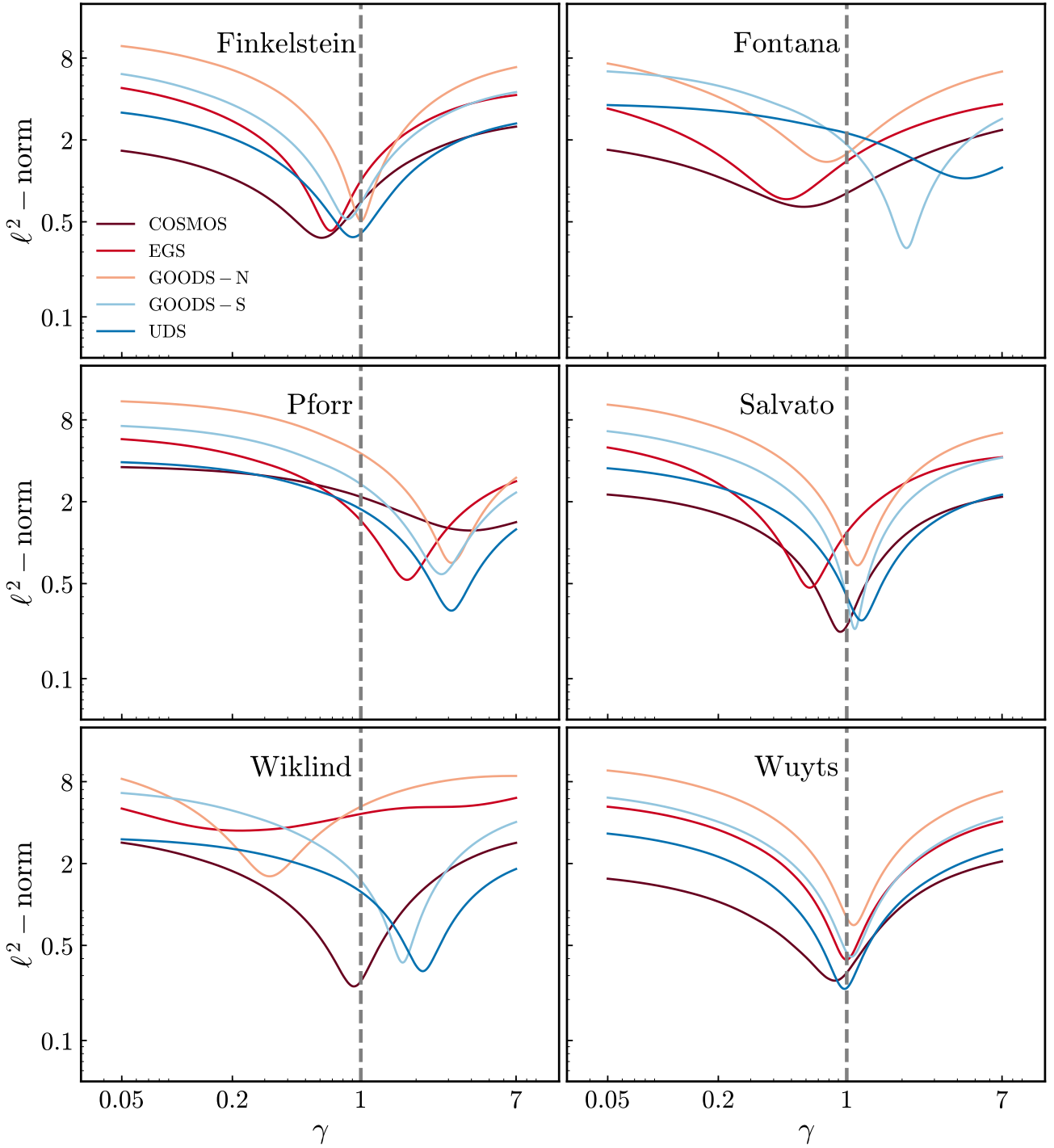
Figure 8. Normalized ℓ^2-norm as a function of γ, a scaling parameter used to improve the widths of the photo-z PDFs. PDFs are raised to the power $1/\gamma$ in this analysis, corresponding to altering the standard deviation parameter of a Gaussian distribution by a factor of $\sqrt{\gamma}$. These plots were constructed using only the training set of spectroscopic redshifts in each of the five CANDELS fields, for each group's PDFs independently, after applying the shifts in redshift listed in Table 1. The vertical gray dashed line indicates where $\gamma = 1$, corresponding to the case where no changes are applied to PDFs. Different values of the parameter γ are needed for each group's PDFs in each field, without any clear pattern.

Table 2. The optimal values of the PDF rescaling parameter γ (used to improve the widths of the PDFs) for each participant group and each CANDELS field. While some groups' results yield γ values that are close to unity, others have significantly lower or significantly larger values, implying substantial changes to the widths of their photo-z PDFs.

Field	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts
COSMOS	0.611	0.582	3.959	0.922	0.916	0.859
EGS	0.686	0.470	1.784	0.628	0.220	0.996
GOODS-N	1.003	0.777	3.118	1.145	0.315	1.090
GOODS-S	0.848	2.115	2.750	1.106	1.683	1.060
UDS	0.905	4.418	3.114	1.203	2.174	0.973

tial and ubiquitous. This is to be expected as the optimized ℓ^2-norm values are lowest for the training set, as by construction our optimization chooses the transformation parameter values which make that quantity as small as feasible. However, it is very encouraging that substantial improvements are also found for the testing and 3D-HST data sets, which have very different coverage in magnitude–redshift space from the training set. We note that the Wiklind analysis produced significantly more objects with missing PDFs, which were not treated as out-of-PDF here. Thus, the low values of $f_{\rm op}$ for that analysis should be considered lower limits on what the performance would be for the full set of objects.

3.5. *Evaluation Using Point Statistics*

Although our primary focus is on producing accurate photometric redshift probability distributions, we also can evaluate the performance of point (or summary) statistics for the photometric redshifts of CANDELS objects before and after optimization. We focus on two different point statistics which are constructed from the photo-z PDFs: the most probable redshift (which we label $z_{\rm peak}$, as it corresponds to the redshift where the highest peak of the photo-z PDF occurs), and the probability-weighted expectation value of the redshift ($z_{\rm weight}$). Following Dahlen et al. (2013), we compute this quantity using only the region surrounding the highest peak of the PDF (specifically, the redshift range within that peak where the PDF value remains above $0.05 \times \mathrm{PDF}(z_{\rm peak})$). To determine this range, we first linearly interpolate the PDF onto a grid of 64001 redshifts in place of the original 1001, then use cubic spline interpolation on this finer grid to find the redshifts around $z_{\rm peak}$ where the PDF is equal to $0.05 \times \mathrm{PDF}(z_{\rm peak})$. If the PDF values are never lower than $0.05 \times \mathrm{PDF}(z_{\rm peak})$, then the bounds used to calculate $z_{\rm weight}$ are the same as the bounds of the grid on which the PDFs are defined. In general, the Dahlen et al. (2013) definition of $z_{\rm weight}$ yields better results than the overall expectation value of the redshift evaluated over the full PDF, as in cases where PDFs have multiple peaks the overall expectation value will often lie between peaks in regions of negligible probability.

We use two quantities to evaluate the accuracy of these point statistics before and after optimization. The first is the normalized median absolute deviation ($\sigma_{\rm NMAD}$) of the differences between photo-z point estimates and spectroscopic redshifts for the corresponding objects, defined by:

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

where $\Delta z = z_{\rm phot} - z_{\rm spec}$ is the difference between a point estimate of the photometric redshift for an object and its spectroscopic redshift. The normalizing factor of 1.48 in the definition of $\sigma_{\rm NMAD}$ causes the expectation value of the NMAD to equal the standard deviation for data that is drawn from a normal distribution.

In addition to the NMAD, we track the fraction of catastrophic outliers in a particular group's results, $f_{\rm co}$, as another useful statistic for characterizing photo-z quality from each group. We define catastrophic outliers as those objects that fulfill the condition:

$$\frac{|\Delta z|}{1 + z_{\rm spec}} > 0.15. \quad (2)$$

Here the limit 0.15 is somewhat arbitrarily chosen; we will show below that typical photometric redshift errors for CANDELS objects with spectroscopic redshifts are $\sim 0.03(1 + z)$, so the threshold of 0.15 approximately corresponds to 5σ outliers.

We find that the NMAD is smaller when computed using $z_{\rm weight}$ rather than $z_{\rm peak}$; that is, $z_{\rm weight}$ provides a superior estimate of the redshift of an object. Therefore, we only show results for statistics computed using $z_{\rm weight}$ in the remainder of this paper, although we still include $z_{\rm peak}$ in our final photo-z catalogs.

Figure 11 shows the scatter plots of $z_{\rm weight}$ vs. $z_{\rm spec}$ for all six groups, using only objects belonging to the training set of spec-z's. The $z_{\rm weight}$ values are calculated using the optimized version of the photo-z PDFs. In the same plots we present the $\sigma_{\rm NMAD}$, the percentages (and number of objects in parentheses) of catastrophic outliers, as well as the percentage (and number of objects in parentheses) of the objects
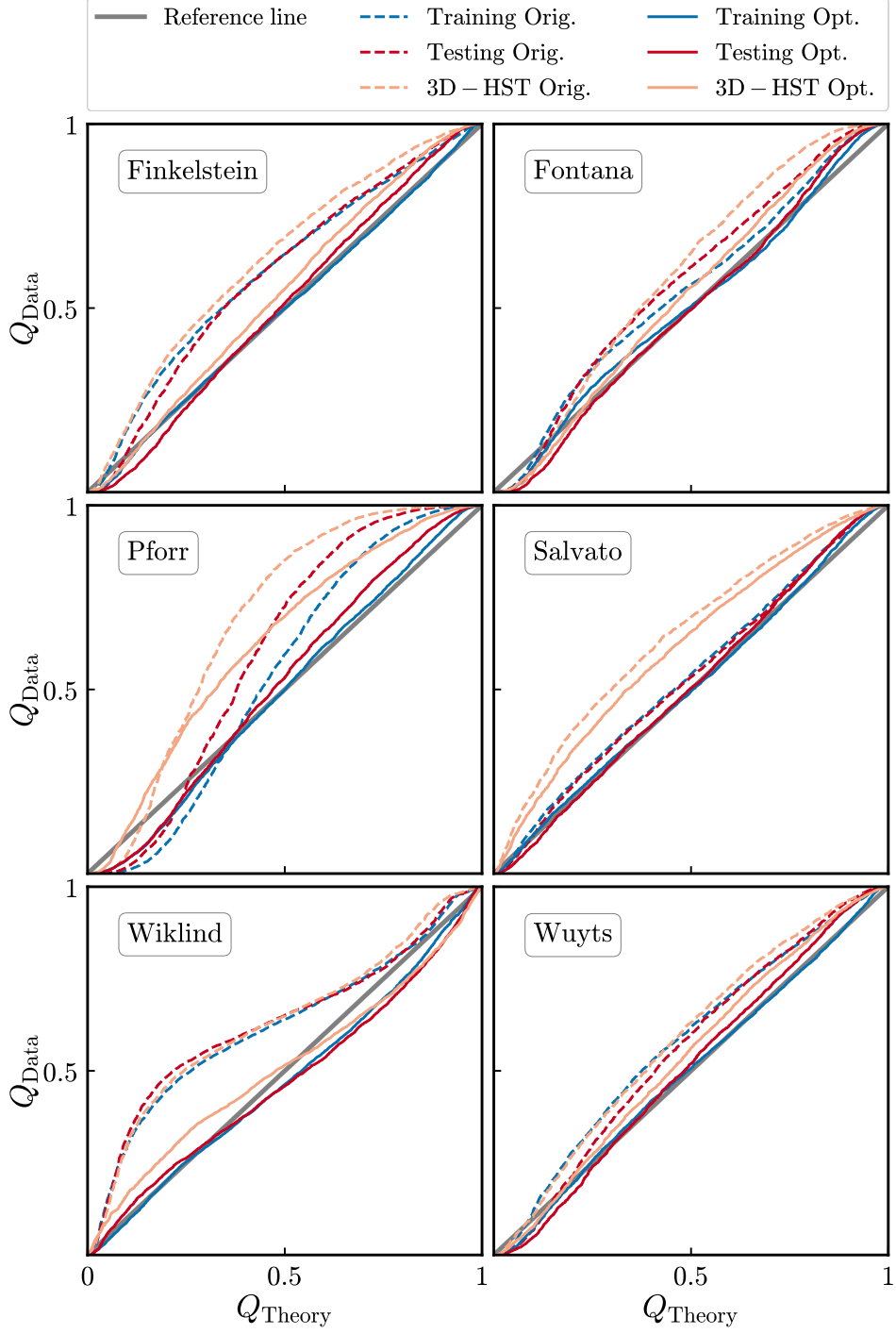
Figure 9. Quantile–Quantile (Q–Q) plots for photo-z PDFs both before ("original", dashed curves) and after ("optimized", solid curves) optimized transformations have been applied. The Q–Q plot shows the CDF value evaluated at the spectroscopic redshift of an object, with $Q_{\rm data}$ corresponding to a given quantile of the set of CDF values, $Q_{\rm theory}$. For instance, if the 30th percentile in the set of CDF($z_{\rm spec}$) values is 0.21, (0.3, 0.21) would be a point along the Q–Q curve. If photo-z PDFs obey the statistical definition of a probability density function, Q–Q curves will lie along the unit line from (0,0) to (1,1). We evaluate the results from each group separately using the independent training and testing sets of spectroscopic redshifts as well as the 3D-HST grism redshifts; results from all five CANDELS fields are combined together here. It is clear that our optimization methods improve the results for each group and for every set of redshifts. The normalized ℓ^2-norm used to evaluate PDF accuracy corresponds to the RMS deviation between a Q–Q curve and the diagonal in the y direction in these plots.

that have missing photo-z PDFs and therefore missing z_{weight} values. Below each scatter plot, we show the dependence of $\Delta z/(1 + z_{spec})$ on z_{spec}, where $\Delta z = z_{weight} - z_{spec}$. In the same plots we also present the redshift dependence of $\sigma_{NMAD} = 1.48 \times \Delta z/(1 + z_{spec})$ (dashed gray line), and the Hodges-Lehmann mean of $\Delta z/(1 + z_{spec})$ (solid gray line), where bins of 0.5 have been used along z_{spec} for their construction. They generally do not seem to vary significantly with redshift, apart from the spikes observed at $z_{spec} \sim 5.5$, where the number of objects becomes very low.

Figure 12 plots the σ_{NMAD} values and catastrophic outlier fractions for the photometric redshifts from each group, calculated using the PDFs both before and after optimization has been applied; results for each spectroscopic sample are shown in separate panels. In general, recalibrating the PDFs significantly improves the scatter between photo-z's and spectroscopic redshifts, while improving or only negligibly degrading the outlier fraction. The Wiklind analysis produced significantly more objects without any point estimates, which were treated as catastrophic outliers in this case. Consequently, the high values of f_{co} for that analysis should be considered upper limits on what the performance would be for the full set.

For the objects belonging to the training or testing sets, the outlier fraction is similar from all the different groups (roughly 8%), while it ranges from 2 to 5% for objects with 3D-HST grism redshifts. There are two possible explanations for this discrepancy. One is that a significant ($\sim 5\%$) fraction of the objects in the training and testing samples have incorrect redshifts assigned to them (potentially due to incorrect matches between the spectroscopic target and CANDELS galaxy). The other possibility is that the smaller outlier fraction for 3D-HST is artificially low due to the grism redshift fits incorporating photo-z information, though we limited to objects whose spectrum drove the grism-z fit to mitigate this possibility.

Finally, we compute the fraction of objects that are both out-of-PDF and catastrophic outliers ($f_{op\,\&\,co}$). These objects are the most problematic because they have both a poorly estimated photo-z and poorly characterized photo-z uncertainty. Generally, around 3% of objects fit into this category across codes (excluding the Wiklind results that should be viewed as lower limits due to their high missing PDF rate driving a low out-of-pdf fraction).

4. METHODS FOR COMBINING PROBABILITY DENSITY FUNCTIONS FROM MULTIPLE CODES

Previous works have found that combining results from multiple photometric redshift codes can yield superior performance over individual codes when testing with spec-z's. Dahlen et al. (2013), Castellano et al. (2016), and Merlin et al. (2021) found that the median of the photo-z's from all
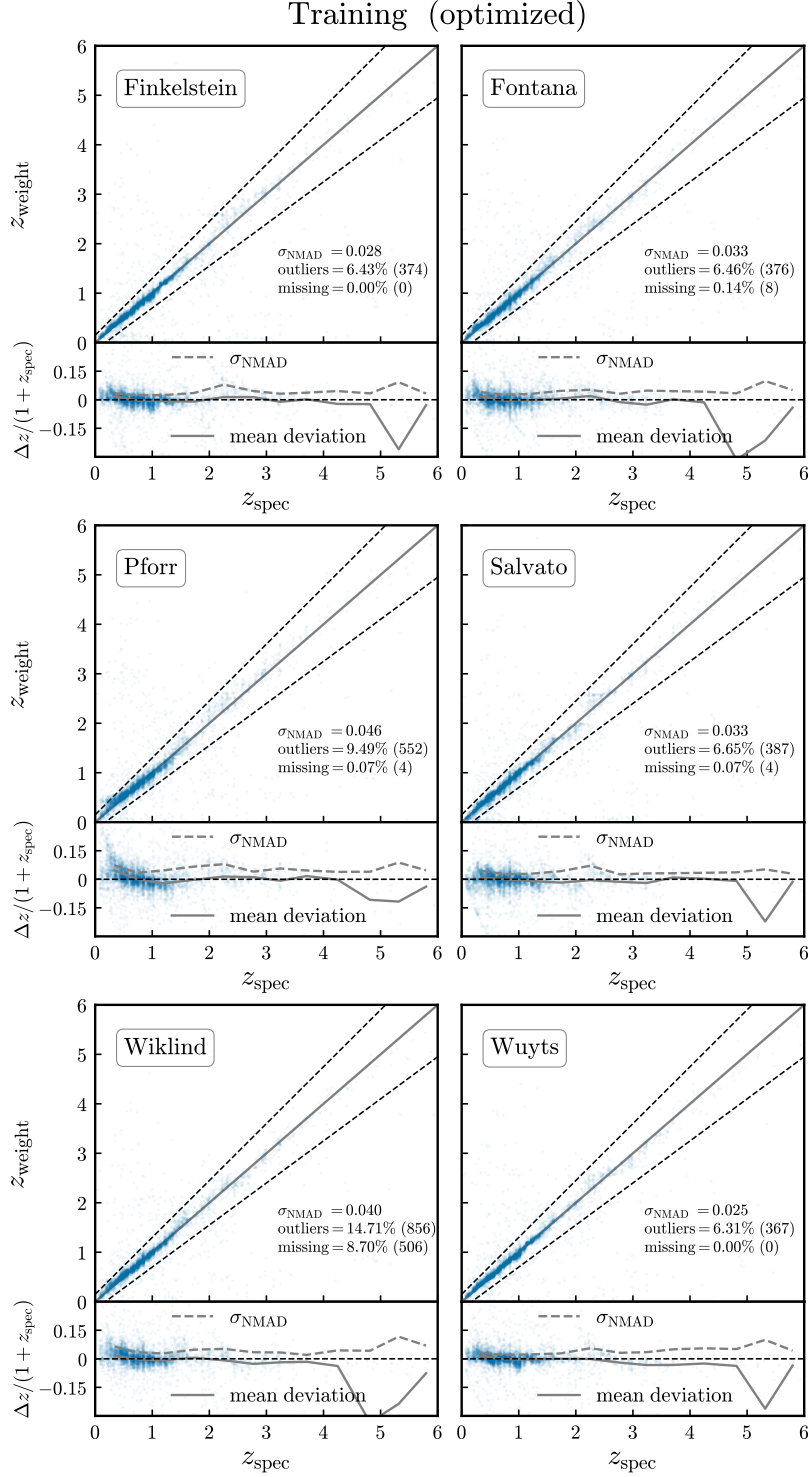
Figure 10. The normalized ℓ^2-norm as a function of the out-of-PDF fraction f_{op} for the PDFs from each group both before (open circles at arrow bases) and after (filled circles at arrow heads) optimization. The three panels show these statistics evaluated using the training set of spectroscopic redshifts (top); the testing set (middle); and 3D-HST grism redshifts (bottom). The ℓ^2-norms are reduced in all cases; the f_{op} values are lower in most cases and only increase slightly otherwise. No single group produced superior results in all cases.

Training (optimized)

Figure 11. Scatter plot of z_{weight} vs. z_{spec} (upper panels), using the optimized version of photo-z PDFs for the calculation of z_{weight} from the six participating groups. The σ_{NMAD}, the outlier percentage, and the percentage of objects with missing photo-z PDFs (and therefore missing z_{weight}) are listed along with the actual number of objects (given in parentheses) for the latter two statistics. Additionally, $\Delta z/(1 + z_{\text{spec}})$ vs. z_{spec} (lower panels) is presented below each scatter plot. The gray dashed and solid lines correspond to σ_{NMAD} and Hodges-Lehmann mean of $\Delta z/(1 + z_{\text{spec}})$, as a function of z_{spec}. Bins of 0.5 in z_{spec} have been used for the calculation of these two curves.
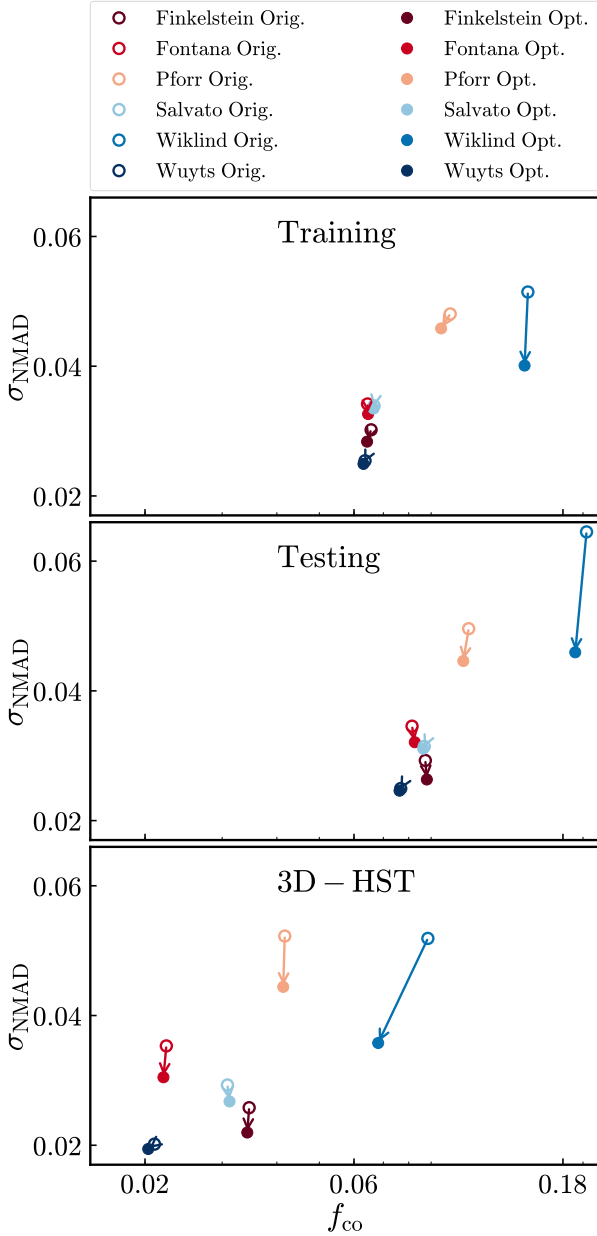
Figure 12. The normalized median absolute deviation between photometric and spectroscopic redshifts, $\sigma_{\rm NMAD}$, as a function of the catastrophic outlier fraction, $f_{\rm co}$ from each group both before (open circles at arrow bases) and after (filled circles at arrow heads) optimization. The three panels show these statistics evaluated using the training set of spectroscopic redshift (top); the testing set (middle); and 3D-HST grism redshifts (bottom). Results shown were calculated using the weighted mean photo-z peak redshifts ($z_{\rm weight}$) from each group. Results from Wuyts outperform the other groups in all cases, especially when evaluated using the 3D-HST data set. We caution that the 3D-HST redshift determination incorporates results from the EAZY code used by Finkelstein and Wuyts as part of the grism redshift determination process, so the level of agreement for them may be artificial.

the groups contributing measurements yielded a lower scatter than any single code. Since we are not only interested in point values (such as $z_{\rm weight}$ or $z_{\rm peak}$) but also desire accurate photo-z PDFs (or, equivalently, accurate error estimates), we require methods that can combine multiple probability density functions. In this paper, we utilize two different methods for combining PDFs: the Hierarchical Bayesian approach (HB) first presented in Dahlen et al. (2013), which produces an entirely new PDF constructed based on the input PDFs (see also Duncan et al. 2018b, Duncan et al. 2018a, and Hatfield et al. 2022), and the minimum f-divergence method (mFD), which selects one of the input PDFs as being the most representative for a given object (much as the median of a set of numbers is the element of that set which approximates its central value).

4.1. *Hierarchical Bayesian Combination of Photometric Redshift PDFs*

The Hierarchical Bayesian approach introduced by Dahlen et al. (2013) combines PDF results from multiple codes based upon the assumptions that estimated PDFs for a given object are not entirely statistically independent (since they are based upon the same photometry) and that some may be inaccurate. This basic framework has a number of antecedents in the literature (Press 1997; Newman et al. 1999; Lang & Hogg 2012).

If we assume that the PDF from a given group is either informative about the true redshift of an object ("good") or uninformative ("bad"), the posterior probability for the redshift of an object from the i'th code, $P_i(z)$, can be described as:

$$P_i(z) = P_i(z|\text{bad})P(\text{bad}) + P_i(z|\text{good})[1 - P(\text{bad})], \quad (3)$$

where $P_i(z|\text{bad})$ is the PDF resulting from an uninformative measurement (which we take here to be a uniform probability distribution from $z = 0$ to $z = 10$); $P(\text{bad})$ is the probability a randomly selected object has an uninformative measurement; and $P_i(z|\text{good})$ is the redshift PDF for a given object predicted by the i'th photo-z code. We assume that $P(\text{bad})$ is the same for all objects, and hence it will be equal to the fraction of all PDF measurements that are uninformative, which we label as $f_{\rm bad}$. Therefore, for a given value of $f_{\rm bad}$, the posterior PDF is given by:

$$P_i(z, f_{\rm bad}) = P_i(z|\text{bad})f_{\rm bad} + P_i(z|\text{good})(1 - f_{\rm bad}). \quad (4)$$

For a given object, the Bayesian posterior combining the information of each (participant) PDF is given by:

$$P(z, f_{\rm bad}) = \prod_i P_i(z, f_{\rm bad})^{1/\alpha}, \quad (5)$$

where we introduce a parameter α that provides a correction for the covariance between the different PDFs. If the PDFs

are all statistically independent, the probabilities from each one would multiply, so $\alpha = 1$. In contrast, if they were completely covariant, we would need α to match the number of PDFs being combined, n_p, so that after multiplication and exponentiation $P(z, f_{\text{bad}})$ would match the result from a single PDF.

Based on tests with the CANDELS data, we find that the optimal value of α is best described by the equation:

$$\alpha = 1 + (n_p - 1) \times 1.1/4, \tag{6}$$

which yields $\alpha = 1$ for $n_p = 1$ and $\alpha = 2.1$ for $n_p = 5$ (matching the results from Dahlen et al. 2013 when $n_p = 5$). This implies that the covariance between results from different participating groups is non-negligible (since $\alpha > 1$ is optimal in general) but also far from complete (since $\alpha < n_p$).

We then marginalize over the fraction of bad measurements to obtain a PDF for redshift alone:

$$P(z) = \int_0^1 P(z, f_{\text{bad}}) \, df_{\text{bad}}, \tag{7}$$

assuming a flat prior distribution for f_{bad} over $[0, 1]$.

The result of this procedure is a PDF that matches the PDFs from each code when they agree but will include extra probability at the redshifts predicted by each individual PDF when they disagree. Hence, HB photo-z constraints are appropriately degraded when the PDFs disagree, in a manner that is agnostic about which PDFs are most accurate.

4.2. *Minimum f-divergence Approach*

In addition to the Hierarchical Bayesian method of combining PDFs, we also test a new approach of choosing the single PDF that minimizes a measure of the total distance to the other PDFs; for the distance calculations, we consider two measures that fall in the general class of f-divergence measures. In this paper, we consider f-divergences calculated either by summing the absolute values of the differences between two PDFs (an ℓ^1 distance) or by taking the square root of the sum of the squares of the differences (an ℓ^2 distance) as measured at many equally-spaced z values:

$$FD_{\text{abs}} = \sum_i |P_1(z_i) - P_2(z_i)| \quad \text{and} \tag{8a}$$

$$FD_{\text{sqr}} = \left\{ \sum_i [P_1(z_i) - P_2(z_i)]^2 \right\}^{1/2}, \tag{8b}$$

where z_i indicates the i'th redshift point at which the distances are evaluated and we use FD_{abs} and FD_{sqr} to label the ℓ^1 and ℓ^2 versions of the f-divergence, respectively. For two given curves $P_1(z)$ and $P_2(z)$ that are defined at the same set of discrete points, the f-divergence is found by combining the difference between the values of the curves at each ordinate
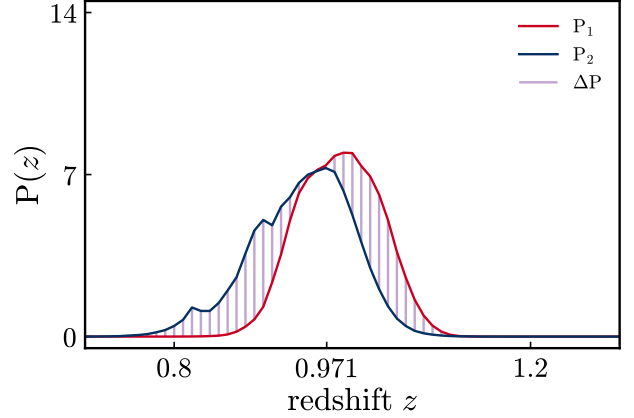

Figure 13. An example illustration of the calculation of the f-divergence distance measures between two photo-z PDFs, $P_1(z)$ and $P_2(z)$, represented by the red and blue curves. The vertical purple hatching indicates the vertical difference between them at each point where they are defined, $\Delta P(z_i) = P_1(z_i) - P_2(z_i)$. We compute the f-divergence between the two curves in two ways: with an ℓ^1 distance (i.e., the sum of the absolute values of the differences) and an ℓ^2 distance (i.e., the square root of the sum of the squared differences).

point, as illustrated in Figure 13. In these calculations, we perform a discrete sum over redshift points, which should closely approximate the integral calculated using a continuous PDF.

In our case, six groups have provided photo-z PDFs for each object in the CANDELS photometric catalogs. We then calculate the sum of the f-divergence between each PDF and each of the other five results, yielding the total distance of that curve from the rest (in the case where the ℓ^2 distance is used we sum in quadrature). We repeat this procedure for each PDF for a given object, and identify the one for which the total distance to the other PDFs is lowest; that is, the one with the minimum f-divergence (mFD) from the other PDFs. This curve will be the most similar to all the rest, and therefore provides a reasonable way to summarize the ensemble of PDFs. This construction is analogous to the use of the median value of an array as a summary statistic; the median minimizes the sum of the absolute values of the deviations from all points in an array.

The ℓ^1 distance (FD_{abs}) will be less affected by the largest excursions between two curves than the ℓ^2 distance (FD_{sqr}) because the latter metric squares the differences before summation, providing more weight to larger deviations. As a result, the minimum f-divergence curve selected using an ℓ^1 distance metric, which we label as "mFDa" as it is based upon the sum of absolute values, is less sensitive to outlier curves than when we use an ℓ^2 metric, which we label as "mFDs" as it is based upon the sum of squared differences. The situation is analogous to the reason why the median statistic is more robust to outliers in a data array than the mean.
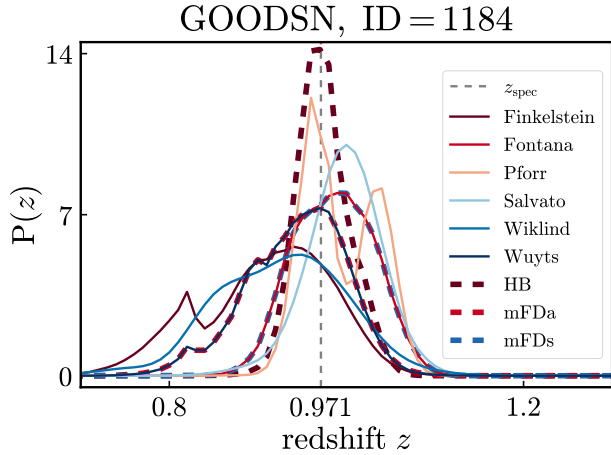
Figure 14. Optimized PDFs from all six participant photo-z groups, as well as the results from three PDF combination methods, for a CANDELS galaxy in the GOODS−North field (ID = 1184). The Hierarchical Bayesian (HB, dark brown dashed curve) method produces a new PDF by combining information from each input probability distribution and accounting for the possibility that some are inaccurate. The minimum f-divergence method identifies the PDF that has the smallest total f-divergence from the other curves. When the sum of absolute values of differences is used as a distance metric, the minimum f-divergence curve corresponds to the Salvato PDF for this object (mFDa, red dashed curve), whereas when the square root of the sum of the squares of separations is used to compute distances the Fontana curve is selected (mFDs, blue dashed curve). The vertical gray dashed line indicates the galaxy's spectroscopic redshift, $z_{spec} = 0.971$, which is consistent with all PDFs shown.

4.3. *Comparing Combination Methods*

Figure 14 illustrates the results of the Hierarchical Bayesian and minimum f-divergence approaches for a galaxy in the GOODS−North field. In this figure, we present the PDFs from the six different groups, as well as the results of each combination method. Whereas the Hierarchical Bayesian method produces a completely new photo-z PDF that is distinct from all the input curves, the minimum f-divergence method selects one of the original PDFs as being the best representative for a given object. For this object, the mFDa and mFDs algorithms select different PDFs, but it is more common that they agree with each other than not.

For each of the three methods of combining photometric redshift PDFs applied here, we also explore how the quality of results changes when only a subset of the groups' PDFs are used as inputs. We note that some groups' PDFs performed significantly better than others when evaluated using the ℓ^2-norm, f_{op}, σ_{NMAD}, or f_{co} for both the testing set of spectroscopic redshifts and the 3D-HST grism redshifts. Given the possibility that some groups' PDFs are more useful for computing combined distributions than others, we have computed PDFs using the HB and mFD methods using only subsets of groups that yielded the lowest σ_{NMAD} and f_{co} values (using

the same set of groups' photo-z's for all objects). We use the label 6 for when all six groups are used, the label 5 for when the five best groups are used, the label 4 for when the four best groups are used, and finally the label 3 for when only the three best groups are used; hence, results labeled HB4 correspond to a Hierarchical Bayesian combination of PDFs from the four best-performing codes.

In Figure 15, we plot results for the ℓ^2-norm and f_{op} for the HB, mFDa, and mFDs methods with the 3, 4, 5, and 6 best-performing codes both before and after PDF optimization, averaging the results for these statistics from the testing set of spectroscopic redshifts and the 3D-HST set of grism redshifts. For the HB combination, HB3 produces the lowest ℓ^2-norm. Note that f_{op} is zero for the HB combined PDFs by construction, as the uniform probability distribution used to model uninformative measurements yields a small amount of probability at all redshifts in the final HB combination. For the minimum f-divergence combinations, mFDa4 and mFDs4 have the lowest ℓ^2-norm values and are our preferred combinations even though mFDa6 and mFDs6 have lower f_{op} values. For all three combination methods, combinations that did not use all six codes produced the lowest ℓ^2-norm values.

Figure 16 takes the best-performing subset for each combination method (i.e., HB3, mFDa4, and mFDs4) and compares them for the training set, spectroscopic testing set, and 3D-HST testing set. It also shows the results from the individual groups for comparison. The mFDa4 and mFDs4 combinations perform similarly well and are clearly better than HB3 (and the individual groups) on the spectroscopic testing set. We prefer the mFDa4 combination because it has the lowest ℓ^2-norm averaged across all three data sets. The mFDa4 combination has slightly higher f_{op} values than the mFDs4 combination for all three data sets. We conclude that the minimum f-divergence (with the absolute value as the distance metric) PDF selected from the four highest-quality results comes closest to meeting the statistical definition of a probability density function for the actual redshift of a galaxy, and therefore is the preferred combination technique for CANDELS when a consensus PDF is desired.

Although mFDa4 is our preferred combination method for PDFs, there are cases when the accuracy of point measures of the redshift is more desirable than the accuracy of the PDF. We therefore evaluate the accuracy of z_{weight} estimates from the same set of combination methods and subsets of the input PDFs considered above, using the σ_{NMAD} and f_{co} statistics to evaluate them. We present the averaged results from the testing set of spectroscopic redshifts and the 3D-HST grism redshifts in Figure 17, combining objects from all CANDELS fields. Among the Hierarchical Bayesian combinations considered, HB4 has the lowest σ_{NMAD} value; mFDa4 and mFDs3 yield the lowest scatters among the minimum f-divergence combinations. We compare the results for these
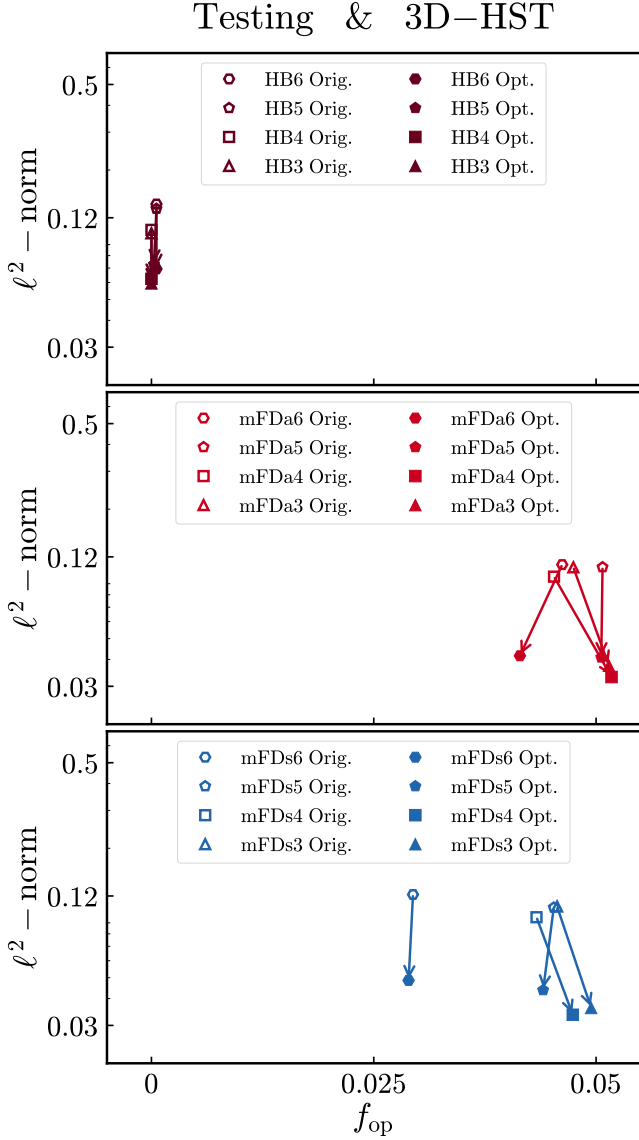
Figure 15. Average values of the normalized ℓ^2-norm and out-of-PDF fraction $f_{\rm op}$ for each PDF combination method considered here, combining results from the testing and 3D-HST redshifts across all CANDELS fields. We show values both without (open symbols at arrow bases) and with (filled symbols at arrow heads) optimization of the PDFs from each group in advance of combination. The panels differ in the method of combination considered; here "HB" indicates Hierarchical Bayesian combination (top), "mFDa" indicates the minimum f-divergence curve selected using an ℓ^1 (sum of absolute values) metric (middle); and "mFDs" indicates the minimum f-divergence curve selected using an ℓ^2 (sum of squares) metric (bottom). The number following the combination type indicates the number of PDFs combined; e.g., in the HB3 case we combine the PDFs from the best 3 groups (ranked according to their performance at point value metrics). Within each type of combination, the lowest values for the ℓ^2-norm are obtained using the optimized PDFs for HB3, mFDa4, and mFDs4, respectively.
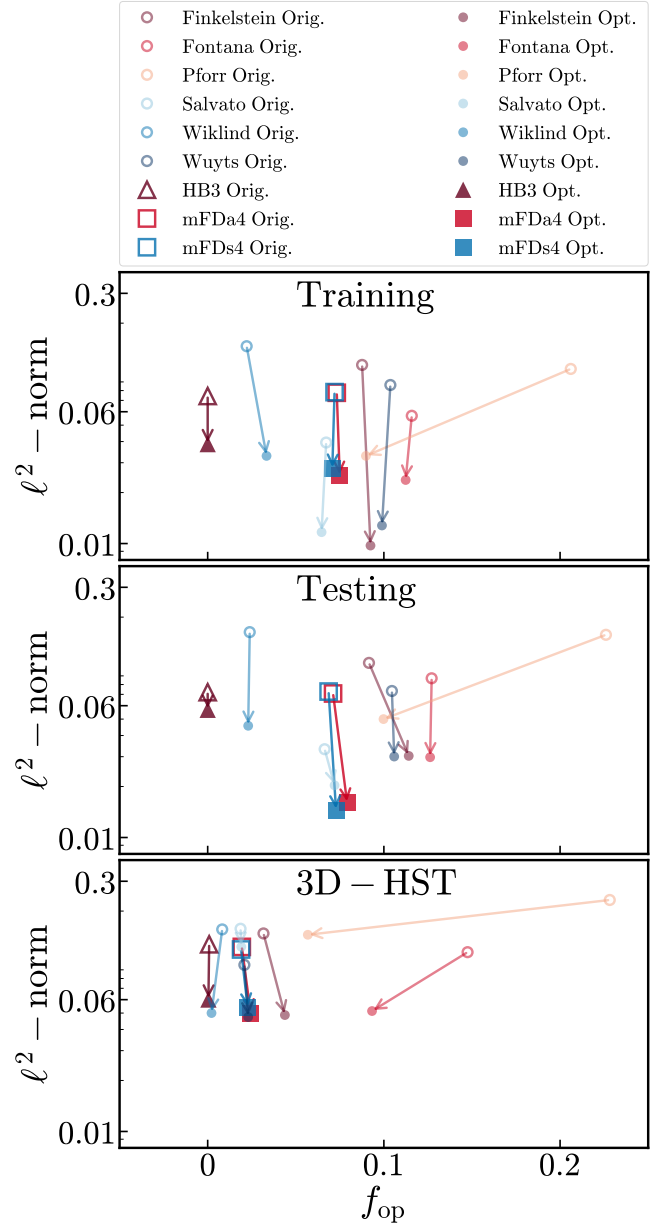
Figure 16. Normalized ℓ^2-norm and $f_{\rm op}$ for the lowest-norm PDF combinations of each type—HB3, mFDa4, and mFDs4—with results from individual groups shown for comparison (same as Figure 10). We show results for the training and testing sets of spectroscopic redshifts and the 3D-HST grism redshifts separately in each panel but combine objects from all CANDELS fields. mFDa4 and mFDs4 produce similarly excellent results for all three data sets, give lower ℓ^2-norm values than HB3 in every case, and outperform the individual groups on the testing set. We prefer the mFDa4 combination because it gives the lowest ℓ^2-norm value averaged across all three data sets, indicating that the minimum f-divergence curve constructed from the four highest-quality PDFs comes closest to meeting the statistical definition of a probability density function for the actual redshift of a galaxy.

three best cases against each other and the individual groups using each set of redshifts separately in Figure 18. In every case, the Hierarchical Bayesian combination of the best four PDFs, HB4, yields the lowest σ_{NMAD}; it is therefore the combination method which provides the most accurate point values of photo-z's, with $\sigma_z \sim 0.02(1+z)$ or better for all three spectroscopic samples. In every case it outperforms the best individual group results, though the improvement over the best individual group (Wuyts) is small.

Table 3 summarizes the performance of the optimized photo-z PDFs from each group and the PDFs produced by the best combination methods of each type (Hierarchical Bayesian [HB], minimum f-divergence computed with an ℓ^1 metric [mFDa], and minimum f-divergence computed with an ℓ^2 metric [mFDs]) for each of the three spectroscopic data sets (training, testing, or 3D-HST) across the five CAN-DELS fields. The method yielding the best performance for a given statistic is highlighted in blue, and the combination method with the best performance on average is indicated in red. The quantities used to evaluate the quality of photo-z PDFs are the normalized ℓ^2-norm between the Q–Q curve for a given set of PDFs and the unity line (ℓ^2-norm) and the fraction of spectroscopic redshifts that lie outside the photo-z PDF for their corresponding object (f_{op}; which are zero by construction for the Hierarchical Bayesian combination method). To test the performance of each group and combination method for the weighted-mean peak redshift, z_{weight}, we use the normalized Median Absolute Deviation (σ_{NMAD}) and the $\Delta z/(1+z) > 0.15$ outlier fraction f_{co}. Finally, we list the fraction of objects for which a PDF file was not provided by a given group as $f_{missing}$ (fraction of missing files). Combinations of different numbers of PDFs (either the best three or the best four) yielded the best results for PDF statistics (where HB3, mFDa4, and mFDs4 proved superior), which differed from the best ones for point statistics (where HB4, mFDa4, and mFDs3 were preferred).

The Finkelstein group's PDFs yielded the smallest average normalized ℓ^2-norm, while the Wiklind PDFs had the smallest out-of-PDF fraction f_{op} (caused by objects with missing PDFs not being considered as out-of-PDF). If objects with missing PDFs are considered out-of-PDF, then Salvato's PDFs would have had the smallest f_{op} of the individual codes. However, the mFDa4 combination method yielded only marginally larger ℓ^2-norm than Finkelstein while having a smaller f_{op}; we therefore recommend the use of this combination if the most accurate PDFs are desired.

Whereas individual groups' results yielded the best performance for some PDF quality statistics, the best point statistics were obtained using combinations of multiple photo-z PDFs. The HB4 (Hierarchical Bayesian combination of the four best
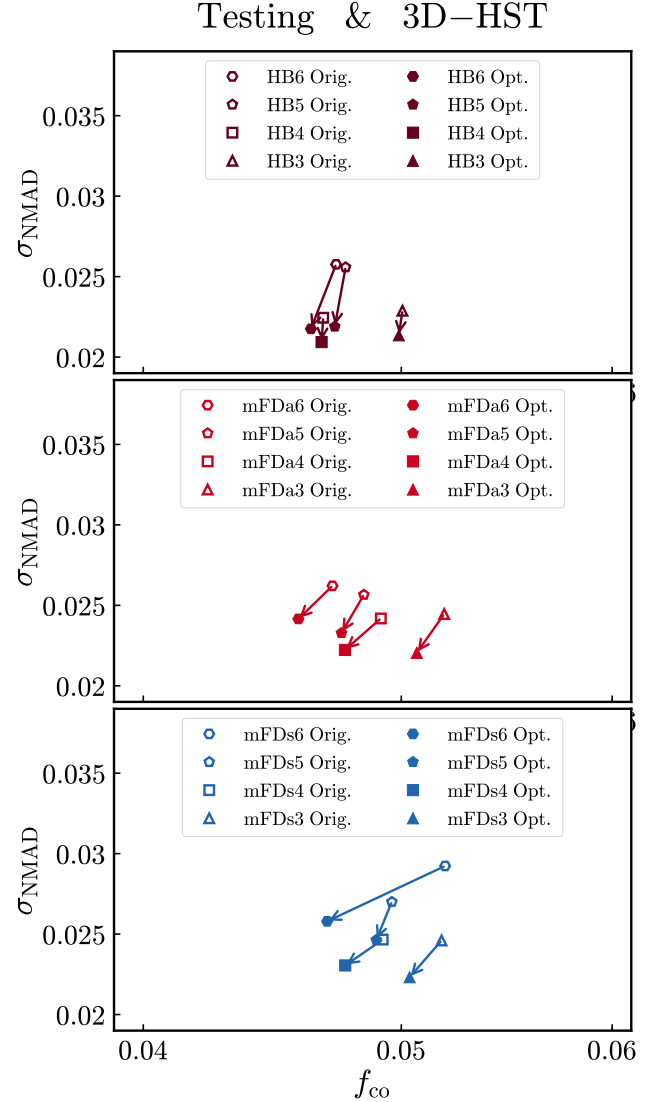

Figure 17. Comparison of photo-z point statistics (z_{weight}) from different combination methods using the NMAD (σ_{NMAD}) and outlier fraction (f_{co}) statistics both before (open symbols at arrow bases) and after (filled symbols at arrow heads) optimization. The results shown are for the testing and 3D-HST redshifts across all five CAN-DELS fields. The HB4, mFDa4, and mFDs3 combination methods yielded the lowest NMAD values, and hence the most accurate point estimates when evaluated with these samples.

PDFs) combination yielded both the smallest average NMAD (0.022) and the lowest average outlier rate (5.1%). However, it is worth noting that the results from the Wuyts group taken on their own were almost as accurate.

Table 3. Table of quantities used to assess the quality of the optimized photometric redshift PDFs and their point statistics. All values multiplied by 100. The lowest average result among the combination methods is shown in red, and the lowest value in a given row is shown in blue (unless that entry is also the lowest average result for the combination methods). We note that the Wiklind analysis has significantly higher fractions of objects with missing photo-z PDFs and point estimates than the other codes, so its summary statistics are sensitive to the treatment of these objects. If objects with missing PDFs are counted as out-of-pdf, then Wiklind would no longer produce the lowest $f_{\rm op}$ and $f_{\rm op\,\&\,co}$ for all three data sets. In fact, it would have the highest $f_{\rm op}$ for the training and testing sets (and an unremarkable $f_{\rm op}$ for the 3D-HST set). On the other hand, if objects with a missing point estimate are not counted as catastrophic outliers, then Wiklind would produce nearly the best $f_{\rm co}$ for the training and testing sets (and an unremarkable $f_{\rm co}$ for the 3D-HST set).

Quantity	Set	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts	HB3	mFDa4	mFDs4
	Training	**0.97**	2.38	3.27	1.16	3.29	1.28	3.95	2.52	2.80
	Testing	3.00	2.97	4.96	2.06	4.52	3.00	5.81	1.61	**1.46**
ℓ^2 − norm	3D−HST	4.85	5.17	14.4	12.2	5.01	**4.75**	6.08	5.01	5.27
	Average	**2.94**	3.51	7.55	5.13	4.28	3.01	5.28	**3.05**	3.18
	Training	9.25	11.3	8.97	6.54	**3.30**	9.95	0	7.43	7.06
	Testing	11.4	12.7	10.1	7.25	**2.33**	10.7	0	7.95	7.22
$f_{\rm op}$	3D−HST	4.41	9.31	5.75	1.94	**0.18**	2.29	0	2.40	2.26
	Average	8.36	11.1	8.26	5.24	**1.94**	7.64	0	5.92	**5.51**
	Training	2.84	3.26	4.58	3.35	4.01	2.41	**2.32**	2.53	2.58
	Testing	2.63	3.21	4.46	3.11	4.59	2.46	**2.33**	2.38	2.41
$\sigma_{\rm NMAD}$	3D−HST	2.20	3.05	4.44	2.68	3.58	1.94	**1.85**	2.07	2.05
	Average	2.56	3.17	4.49	3.04	4.06	2.27	**2.17**	2.33	2.35
	Training	6.43	6.46	9.49	6.65	14.7	6.31	6.08	**6.05**	6.17
	Testing	8.80	8.26	10.4	8.64	19.2	7.62	7.44	**7.39**	7.53
$f_{\rm co}$	3D−HST	3.42	2.20	4.14	3.12	6.81	2.03	**1.90**	2.14	2.54
	Average	6.22	5.64	7.99	6.13	13.6	5.32	**5.14**	5.19	5.41
	Training	5.14	4.81	2.59	3.57	**2.37**	5.28	–	4.83	–
	Testing	6.12	4.28	5.13	3.74	**1.76**	5.81	–	5.15	–
$f_{\rm op\,\&\,co}$	3D−HST	1.79	1.63	2.22	0.89	**0.27**	1.01	–	0.93	–
	Average	3.95	2.96	3.67	2.32	**1.02**	3.41	–	3.04	–
	Training	0	0.11	0.057	0.057	7.51	0	–	–	–
$f_{\rm missing}$	Testing	0	0.18	0.33	0.076	9.59	0	–	–	–
	3D−HST	0	0.030	0.12	0.030	4.51	0	–	–	–

The most dramatic failures of the photo-z codes are objects that are both out-of-PDF and catastrophic outlier point estimates. Generally, there is some overlap between the out-of-PDF objects and catastrophic outlier objects; but there are also a considerable number of objects that are either one or the other but not both. We note that calculation of $f_{\rm op}$, $f_{\rm co}$, and $f_{\rm op\,\&\,co}$ for the Wiklind analysis in particular is sensitive to the treatment of missing PDFs/photo-z's due to the significantly higher $f_{\rm missing}$ for Wiklind compared to the other codes.

5. CANDELS PHOTOMETRIC REDSHIFT CATALOGS

Figure 19 summarizes the procedure from the initial photo-z PDFs obtained from the six different groups ("Original") to the "Shifted" and "Optimized" versions of the individual photo-z PDFs and then to the combined photo-z PDFs used to generate the final photometric redshift catalogs. We provide tabulated "Original," "Optimized," and combined photo-z PDFs and point estimates for all objects in the five CANDELS fields.

First, we provide photo-z PDFs in a separate file for every CANDELS object (with the object identifier specified in the filename, e.g.,
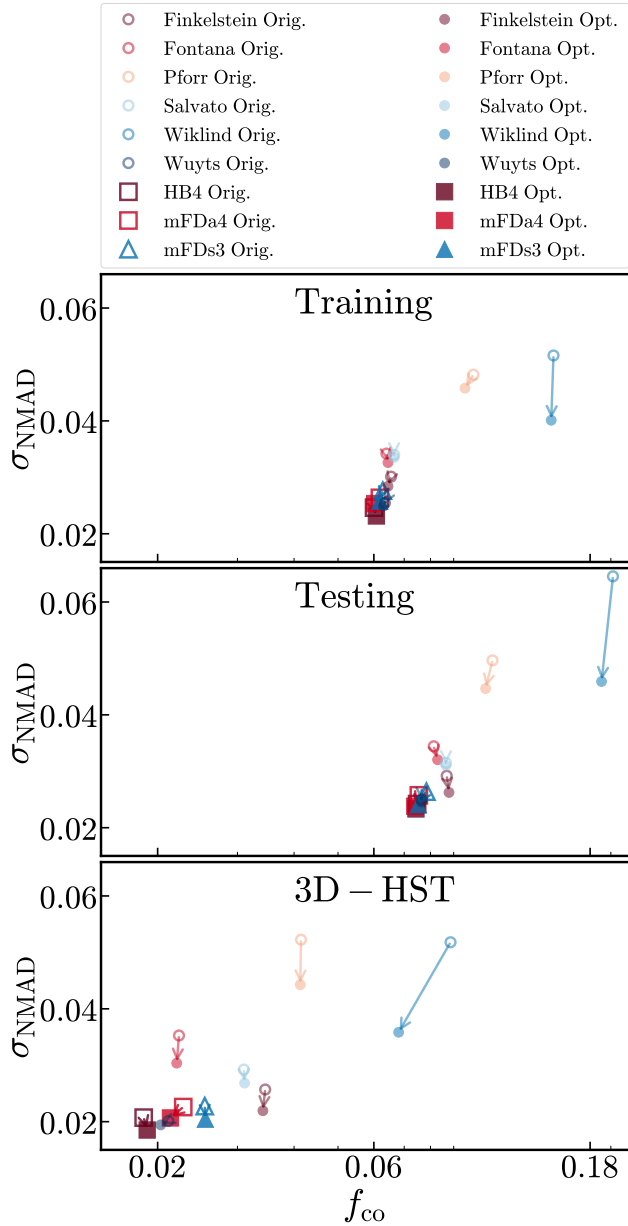
Figure 18. Comparison of the best photo-z point statistic results from each PDF combination method—HB4, mFDa4, and mFDa3—using σ_{NMAD} and f_{co} both before (open symbols at arrow bases) and after (filled symbols at arrow heads) optimization. We also show results from Figure 12 for the individual groups. The three panels show these statistics evaluated using the training set of spectroscopic redshift (top); the testing set (middle); and 3D-HST grism redshifts (bottom). In every case, the Hierarchical Bayesian combination of the best four PDFs, HB4, gives the lowest σ_{NMAD} values for each set and outlier fractions comparable to or lower than other methods, making this the combination that provides the most accurate point estimates. That being said, the improvement over the Wuyts results is slight.

ALL_OPTIMIZED_PDFS_GOODSN_ID00001.pzd). Each file has columns specifying the redshift; the PDF provided by each group after the optimization procedure from Section 3 has been applied; and the PDFs resulting from the two best combination methods, HB4 and mFDa4 (hence the term ALL in the filename). The PDFs cover the redshift interval [0, 10] with a step size of $\Delta z = 0.01$. The details of the columns included in these files are presented in Table 4. We also provide files with the original photo-z PDFs as provided by the six different groups before applying the optimization method. The format of these files is exactly the same as the one with the optimized PDFs, while the file name in this case is ALL_ORIGINAL_PDFS_GOODSN_ID00001.pzd.

Additionally, we provide summary catalogs of objects in each CANDELS field. These catalogs contain photo-z point statistics constructed from the optimized PDFs and the best combinations, as well as spectroscopic and/or grism redshifts where available. These catalogs include estimates of the 1σ and 2σ credible intervals for the photometric redshift constructed from the PDFs. The columns of these files are described in detail in Table 5. The catalogs are available at https://archive.stsci.edu/hlsp/candels.

6. SUMMARY AND DISCUSSION

In this paper, we developed a technique to measure the photo-z PDFs and point estimates for galaxies in the CANDELS field using the final photometric catalogs. Using this technique, we present photometric redshifts for over 150,000 galaxies. We began with probability density functions measured by six groups within the collaboration by applying a variety of template-based methods to the same photometric catalogs. We determined the optimal shift and stretching/sharpening parameters for the PDFs from each group using statistics based upon the Q–Q plot, which we measured using the same training set of spectroscopic redshifts provided to each group to tune their photo-z algorithms.

In tests with a training set of spectroscopic redshifts and with independent sets of spectroscopic and grism redshifts, the optimized PDFs much more closely match the statistical definition of a probability density function than those originally provided by each group, with the normalized ℓ^2-norm statistic (a measure of the accuracy of the photo-z cumulative distribution function) improving by more than a factor of two in some cases. Point estimates of the redshift (e.g., z_{weight}) derived from the optimized photo-z PDFs also exhibit significantly smaller scatter (as measured by the normalized median absolute deviation) and smaller or negligibly worse catastrophic outlier rates, in the best cases yielding photo-z errors of $\sim 0.02(1 + z)$.

After optimizing the results from individual groups, we have explored the gains from three different methods of combining the six PDFs available for each object: the Hierarchical
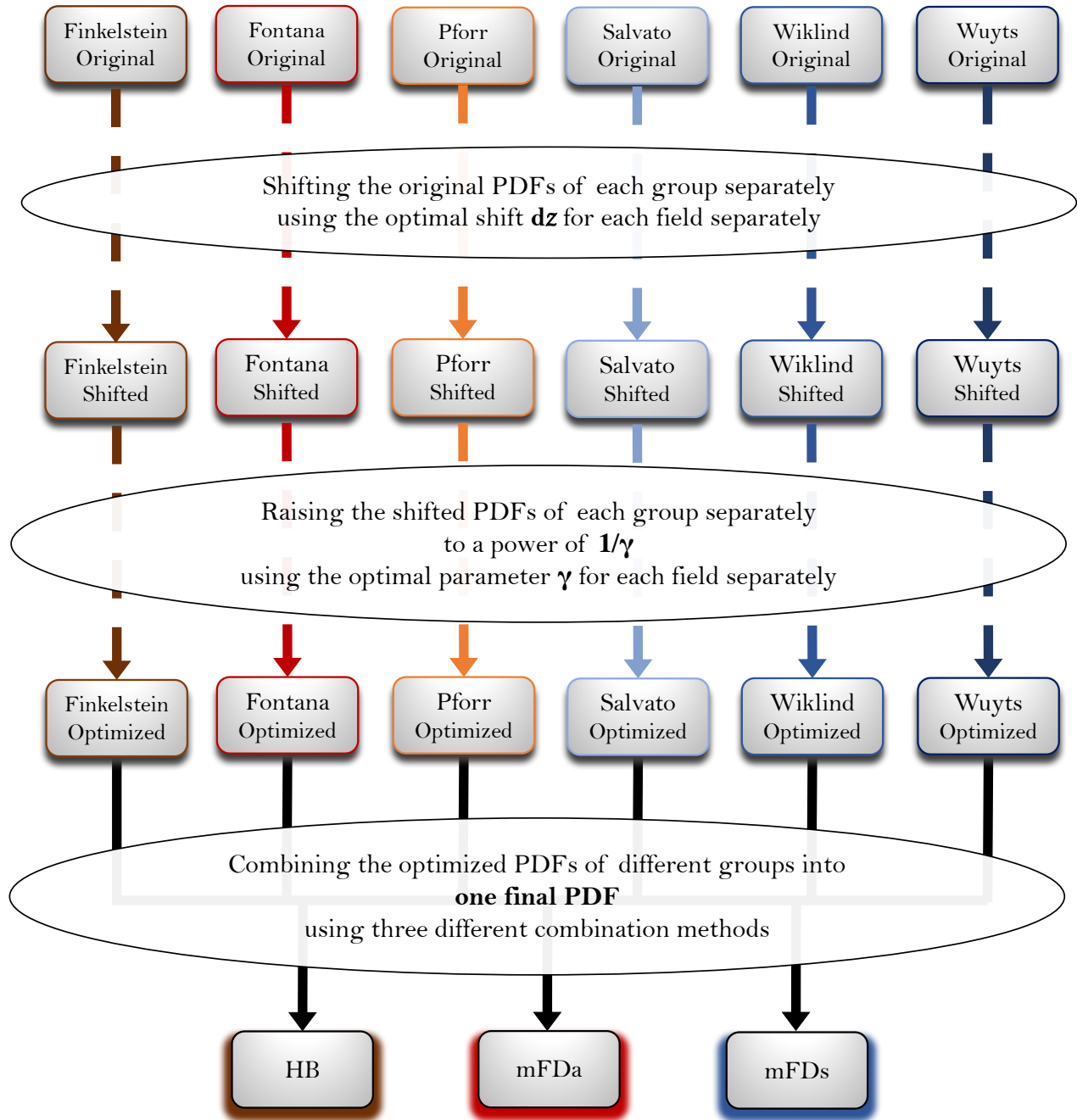
Figure 19. Diagram of the optimization procedure for obtaining the final products of photo-z PDFs, starting from the initially provided PDFs. First the PDFs are shifted, then raised to a power, resulting in optimized PDFs. Then the PDFs from different groups are combined using three different combination methods into final PDFs that can be used for science.

Column	Description
#1 z	Redshift values for the grid on which PDFs are tabulated
#2 Finkelstein	Probability Density Function (PDF) from Finkelstein
#3 Fontana	PDF from Fontana
#4 Pforr	PDF from Pforr
#5 Salvato	PDF from Salvato
#6 Wiklind	PDF from Wiklind
#7 Wuyts	PDF from Wuyts
#8 HB4	PDF from Hierarchical Bayesian combination, constructed using the PDFs from the best-performing four groups
#9 mFDa4	PDF from the minimum Fréchet Distance combination (computed with ℓ^1 distance metric), constructed using the PDFs from the best-performing four groups

Table 4. Detailed description of the files (e.g., `ALL_OPTIMIZED_PDFS_GOODSN_ID00001.pzd`) containing the PDFs from each participant as well as the two best combination methods. The number of models used in the combination methods is reported in the header, as well as the value of the parameter α from the Hierarchical Bayesian method. Note that while the four best participants are included in the evaluation of the combination methods, one or more participants might have missing PDFs for a given object, therefore the total number of PDFs used is not always 4. Both the original and optimized versions of PDFs are provided in separate files.

Column	Description
# 1 file	Name of the PDF file used to estimate photometric point values.
# 2 ID	CANDELS ID of the object as used in the photometric catalogs.
# 3 RA	Right Ascension of object (from photometric catalog).
# 4 DEC	Declination of object (from photometric catalog).
# 5 z_best	Best redshift value which can be spectroscopic, grism, or photometric.
# 6 z_best_type	Type of photometric redshift: s = spec-z, g = grism-z, p = photo-z.
# 7 z_spec	Spectroscopic redshift if available.
# 8 z_spec_ref	Reference of catalog from which the spectroscopic redshift is obtained.
# 9 z_grism	3D-HST grism redshift of object if available
# 10 mFDa4_z_peak	Peak value of mFDa4 PDF
# 11 mFDa4_z_weight	Weighted average value of mFDa4 PDF
# 12 mFDa4_z683_low	Lower boundary of 68.3% (1σ) credible interval of mFDa4 PDF
# 13 mFDa4_z683_high	Higher boundary of 68.3% (1σ) credible interval of mFDa4 PDF
# 14 mFDa4_z954_low	Lower boundary of 95.4% (2σ) credible interval of mFDa4 PDF
# 15 mFDa4_z954_high	Higher boundary of 95.4% (2σ) credible interval of mFDa4 PDF
# 16 HB4_z_peak	Peak value of HB4 PDF
$\vdots$	
# 22 Finkelstein_z_peak	Peak value of Finkelstein PDF
$\vdots$	
# 28 Fontana_z_peak	Peak value of Fontana PDF
$\vdots$	
# 34 Pforr_z_peak	Peak value of Pforr PDF
$\vdots$	
# 40 Salvato_z_peak	Peak value of Salvato PDF
$\vdots$	
# 46 Wiklind_z_peak	Peak value of Wiklind PDF
$\vdots$	
# 52 Wuyts_z_peak	Peak value of Wuyts PDF
$\vdots$	

Table 5. Detailed description of the columns of the CANDELS photometric redshift catalogs (e.g., `zcat_EGS_v2.0.cat`), which provide point statistics constructed from the optimized photometric PDFs, as well as spectroscopic and/or grism redshifts where available, for all objects in each CANDELS field. Each CANDELS object corresponds to one row in the catalog. Statistics based upon the optimized PDFs from all six groups, as well as the two best combination methods, mFDa4 and HB4, are provided within the catalog. The full set of statistics tabulated for the minimum f-divergence (mFDa4) PDF are detailed. Corresponding statistics are provided for each groups' results are included in the catalog, with the column identifier only differing in its prefix (i.e., HB4, Finkelstein, Fontana, Pforr, Salvato, Wiklind, or Wuyts) from the column identifier for mFDa4.

Bayesian (HB) method described in Dahlen et al. (2013) as well as two techniques introduced here that identify the PDF with the minimum f-divergence based on the sum of absolute values (mFDa) and squared differences (mFDs). We construct new PDFs by applying each method to subsets of the six results for each object. Comparing them to each other with the same statistics used to assess individual groups' results shows that combining the PDFs from the four best-performing groups produced the best results. The Hierarchical Bayesian

method yielded the lowest scatter in point statistics, while the minimum f-divergence curve computed with an ℓ^1 metric (mFDa) had the lowest ℓ^2-norm values, indicating that it provides the most accurate PDFs, and hence the most accurate credible intervals as well.

The optimized methods were then used to estimate photo-z PDFs for galaxies in CANDELS. We constructed publicly-available catalogs of optimized PDFs and photometric redshift summary statistics for all objects from the CANDELS

photometric catalogs used to calculate the photo-z's. Instructions on how best to use these catalogs are as follows:

- In general, results from different photometric redshift codes are sufficiently different from each other (as can be seen most clearly in Figure 6) that we recommend performing an analysis multiple times using photo-z point estimates or PDFs from different groups each time to ensure that conclusions are robust to these variations.

- Different columns of the summary catalog are better for different purposes. For instance, if one wants the best estimate for the redshift of an individual object (where uniformity does not matter) the z_best value from the catalog (which is determined from the combined data set of spectroscopic redshifts, 3D-HST grism redshifts, and mFDa4 photometric redshifts) would be most appropriate. If instead the smallest-scatter estimator of redshift for a uniform sample is needed, HB4_z_weight (the z_{weight} value computed from the Hierarchical Bayesian combination of the four best PDFs for each object) would be most appropriate. This photo-z point estimate yielded σ_{NMAD} = 0.0227/0.0189 and $|\Delta z/(1 + z)| > 0.15$ outlier fraction = 0.067/0.019 for the testing spec-z's and 3D-HST grism-z's respectively. We note that redshift uncertainties will never be uniform across galaxy samples—even when restricting uniformly to using only photo-z's—because different objects have different signal-to-noise ratios and different SED shapes with more or less pronounced breaks.

- The mFDa4 (minimum f-divergence curve constructed from the best four PDFs from individual groups) yielded probability density functions that best meet the statistical definition of a PDF. As a consequence, this is the preferred set of PDFs to use when the accuracy of

credible intervals on redshifts is desired. Correspondingly, the mFDa4_z_weight column of the summary table will have the best-characterized error estimates associated with it.

The photometric redshift catalogs presented here represent the culmination of a considerable amount of effort by the CANDELS collaboration to obtain a broad range of imaging data, measure uniform photometry with TFIT, and calculate photometric redshifts. They represent a public legacy of the survey which should contribute to a wide variety of science in the future, such as the estimation of stellar masses of galaxies.

Facilities: HST(WFC3 and ACS)

APPENDIX

A. SPECTROSCOPIC AND GRISM REDSHIFTS

For training and testing the photo-z codes, we compiled spectroscopic and grism redshifts for numerous sources. Table 6 lists the original sources, the relevant CANDELS field, and any quality cuts applied.

Table 6. Details of the spectroscopic redshifts and grism redshifts used in this study. For each data set used we provide the name of the survey or instrument used, where applicable; a reference for the source catalog; the number of redshifts provided in each CANDELS field; and any cuts applied in order to restrict to the most robust redshifts.

Survey / Instrument (Reference)	Number Of Redshifts in each field					Cuts applied
	COSMOS	EGS	GOODS-N	GOODS-S	UDS	
Training, COSMOS (Private Communication: B. Mobasher)	370	–	–	–	–	$z > 0$, flag = 1
Training, EGS (DEEP3, Cooper et al. 2011, 2012a; Zhou et al. 2019)	–	840	–	–	–	$z > 0$, flag = none
Training, GOODS−N (Private Communication: B. Mobasher)	–	–	2994	–	–	$z > 0$, flag $\geq$ 3
Training, GOODS−S (Private Communication: B. Mobasher)	–	–	–	1249	–	$z > 0$, GRISM_FLAG = 0[5]
Training, UDS (Private Communication: B. Mobasher)	–	–	–	–	354	$z > 0$, flag = 1
3D−HST (Momcheva et al. 2016b)	566	771	579	523	928	z_max_grism > 0.6 use_zgrism = 1, use_phot = 1 flag1 = 0, flag2 = 0 z_best_s $\neq$ 0, z_spec $\leq$ 0 z_phot_u68 − z_phot_l68 > 0 z_grism_u68 − z_grism_l68 < 0.01 $\frac{z\_grism\_u68 - z\_grism\_l68}{z\_phot\_u68 - z\_phot\_l68} < 0.1$ z_max_grism > z_phot_l95 z_max_grism < z_phot_u95
MULTIPLE FIELDS						
DEIMOS (Faber et al. 2003, Private Communication: B. Mobasher)	172	–	70	39	179	$z > 0$, flag = 1
MOSDEF (Kriek et al. 2015)	189	268	73	10	26	$z > 0$, flag > 0

Table 6 continued

[5] Flag to avoid grism−z's and only keep spec−z's.

Table 6 (*continued*)

Survey / Instrument (Reference)	Number Of Redshifts in each field					Cuts applied
	COSMOS	EGS	GOODS-N	GOODS-S	UDS	
MOSFIRE (Wirth et al. 2015; Trump et al. 2013)	–	–	22	75	–	$z > 0$, flag ≥ 3
<u>COSMOS FIELD ONLY</u>						
FMOS (Silverman et al. 2015)	9	–	–	–	–	$z > 0$, flag > 3
FORS2 (Private Communication: J. Comparat)	33	–	–	–	–	$z > 0$, flag > 3
Gemini–S (Private Communication: M. Balogh)	16	–	–	–	–	$z > 0$, flag > 3
PRIMUS (Coil et al. 2011)	232	–	–	–	–	$z > 0$, flag > 3
VUDS (Le Fèvre et al. 2015)	101	–	–	–	–	$z > 0$, flag $= 3$ and 4
WFC3 (Krogager et al. 2014, grism-z's)	12	–	–	–	–	$z > 0$, flag > 3
zBRIGHT (Lilly et al. 2009)	2	–	–	–	–	$z > 0$, flag $= 1.5, 2.5, 9.3, 9.5$ 3.x, 4.x, 13.x, 14.x secondary targets
zCOSMOS (Lilly et al. 2007b)	7	–	–	–	–	$z > 0$, flag $= 1.5, 2.5, 9.3, 9.5$ 3.x, 4.x, 13.x, 14.x secondary targets
<u>EGS FIELD ONLY</u>						
DEEP2 (Davis et al. 2003; Davis et al. 2007; Newman et al. 2013)	–	1432	–	–	–	$z > 0$, flag ≥ 3
<u>GOODS – N FIELD ONLY</u>						
Barger (24 μm, ACS, K, UV, Barger et al. 2008, and references therein)	–	–	80	–	–	$z > 0$, flag $= 4$ or A
DEEP3 (Cooper et al. 2011; Zhou et al. 2019)	–	–	3	–	–	$z > 0$, flag ≥ 3
Pirzkal (Pirzkal et al. 2013, grism-z's)	–	–	81	–	–	$z > 0$, flag ≥ 3
Reddy_2006 (Shivaei et al. 2015)	–	–	2	–	–	$z > 0$, flag ≥ 3
Wirth_2004 Wirth et al. 2004	–	–	3	–	–	$z > 0$, flag ≥ 3

Table 6 *continued*

Table 6 *(continued)*

Survey / Instrument (Reference)	COSMOS	EGS	GOODS-N	GOODS-S	UDS	Cuts applied
GOODS – S FIELD ONLY						
ACES (Cooper et al. 2012b)	–	–	–	103	–	z > 0, flag = 0
COMBO−17 (Wolf et al. 2004)	–	–	–	6	–	z > 0, flag = 0
CXO−CDFS (Szokoly et al. 2004)	–	–	–	77	–	z > 0, flag = 0
FW_7 (van der Wel et al. 2004)	–	–	–	7	–	z > 0, flag = 0
FW_10 (Daddi et al. 2004)	–	–	–	3	–	z > 0, flag = 0
FW_13 (Wuyts et al. 2009)	–	–	–	2	–	z > 0, flag = 0
FW_14 (Kriek et al. 2007)	–	–	–	1	–	z > 0, flag = 0
FW_16 (Private Communication: J.-S. Huang)	–	–	–	2	–	z > 0, flag = 0
K20 (Mignoli et al. 2005)	–	–	–	99	–	z > 0, flag = 0
Kurk et al. 2012	–	–	–	49	–	z > 0, flag = 0
Morris et al. 2015	–	–	–	54	–	z > 0, flag = 0
VIMOS_10_MRLR (Balestra et al. 2010)	–	–	–	196	–	z > 0, flag = 0
VLT_2008 (Vanzella et al. 2008)	–	–	–	144	–	z > 0, flag = 0
VLT_IMAG (Ravikumar et al. 2007)	–	–	–	5	–	z > 0, flag = 0
VLT_LBGs (Vanzella et al. 2009)	–	–	–	1	–	z > 0, flag = 0
VUDS_2015 (Le Fèvre et al. 2015)	–	–	–	77	–	z > 0, flag = 0
VVDS (Final, Le Fèvre et al. 2013)	–	–	–	116	–	z > 0, flag = 0
Wuyts_2009 (Wuyts et al. 2009)	–	–	–	3	–	z > 0, flag = 0
Xray_1 (Szokoly et al. 2004)	–	–	–	7	–	z > 0, flag = 0
Xray_2 (Vanzella et al. 2008)	–	–	–	1	–	z > 0, flag = 0
Xray_6 (Popesso et al. 2009)	–	–	–	1	–	z > 0, flag = 0
Xray_12 (Balestra et al. 2010)	–	–	–	1	–	z > 0, flag = 0
Xue_4 (Mignoli et al. 2005)	–	–	–	2	–	z > 0, flag = 0
Xue_5 (Ravikumar et al. 2007)	–	–	–	2	–	z > 0, flag = 0
Xue_6 (Vanzella et al. 2008)	–	–	–	2	–	z > 0, flag = 0
Xue_7 (Popesso et al. 2009)	–	–	–	2	–	z > 0, flag = 0
Xue_8 (Treister et al. 2009)	–	–	–	1	–	z > 0, flag = 0

Table 6 continued

Table 6 (*continued*)

Survey / Instrument (Reference)	Number Of Redshifts in each field					Cuts applied
	COSMOS	EGS	GOODS-N	GOODS-S	UDS	
Xue_9 (Balestra et al. 2010)	–	–	–	1	–	$z > 0$, flag = 0
Xue_10 (Silverman et al. 2010)	–	–	–	2	–	$z > 0$, flag = 0
UDS FIELD ONLY						
Santini et al. 2015	–	–	–	–	61	$z > 0$, flag > 3
Akiyama (Akiyama et al. 2015)	–	–	–	–	23	$z > 0$, flag = A
UDSz (Bradshaw et al. 2013; McLure et al. 2013)	–	–	–	–	63	$z > 0$, flag = 4 or A

Software: EAZY (Brammer et al. 2008b), HyperZ (Bolzonella et al. 2000), LePhare (Arnouts & Ilbert 2011), WikZ (Wiklind et al. 2008), zphot (Giallongo et al. 1998; Fontana et al. 2000)

REFERENCES

Akiyama, M., Ueda, Y., Watson, M. G., et al. 2015, PASJ, 67, 82, doi: 10.1093/pasj/psv050

Arnouts, S., & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate, Astrophysics Source Code Library. http://ascl.net/1108.009

Balestra, I., Mainieri, V., Popesso, P., et al. 2010, A&A, 512, A12, doi: 10.1051/0004-6361/200913626

Barger, A. J., Cowie, L. L., & Wang, W.-H. 2008, ApJ, 689, 687, doi: 10.1086/592735

Barro, G., Pérez-González, P. G., Cava, A., et al. 2019, ApJS, 243, 22, doi: 10.3847/1538-4365/ab23f2

Baum, W. A. 1957, AJ, 62, 6, doi: 10.1086/107433

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371, doi: 10.1093/mnras/stw1009

Bezanson, R., Wake, D. A., Brammer, G. B., et al. 2016, ApJ, 822, 30, doi: 10.3847/0004-637X/822/1/30

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476

Bradshaw, E. J., Almaini, O., Hartley, W. G., et al. 2013, MNRAS, 433, 194, doi: 10.1093/mnras/stt715

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008a, ApJ, 686, 1503, doi: 10.1086/591786

—. 2008b, ApJ, 686, 1503, doi: 10.1086/591786

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000, doi: 10.1046/j.1365-8711.2003.06897.x

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682, doi: 10.1086/308692

Castellano, M., Amorín, R., Merlin, E., et al. 2016, A&A, 590, A31, doi: 10.1051/0004-6361/201527514

Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, ApJ, 741, 8, doi: 10.1088/0004-637X/741/1/8

Cole, S., Norberg, P., Baugh, C. M., et al. 2001, MNRAS, 326, 255, doi: 10.1046/j.1365-8711.2001.04591.x

Coleman, G. D., Wu, C. C., & Weedman, D. W. 1980, ApJS, 43, 393, doi: 10.1086/190674

Cooper, M. C., Aird, J. A., Coil, A. L., et al. 2011, ApJS, 193, 14, doi: 10.1088/0067-0049/193/1/14

Cooper, M. C., Griffith, R. L., Newman, J. A., et al. 2012a, MNRAS, 419, 3018, doi: 10.1111/j.1365-2966.2011.19938.x

Cooper, M. C., Yan, R., Dickinson, M., et al. 2012b, MNRAS, 425, 2116, doi: 10.1111/j.1365-2966.2012.21524.x

Daddi, E., Cimatti, A., Renzini, A., et al. 2004, ApJ, 617, 746, doi: 10.1086/425569

Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93, doi: 10.1088/0004-637X/775/2/93

Davidzon, I., Ilbert, O., Laigle, C., et al. 2017, A&A, 605, A70, doi: 10.1051/0004-6361/201730419

Davis, M., Faber, S. M., Newman, J., et al. 2003, in Proc. SPIE, Vol. 4834, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II, ed. P. Guhathakurta, 161–172, doi: 10.1117/12.457897

Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJL, 660, L1, doi: 10.1086/517931

Dey, B., Andrews, B. H., Newman, J. A., et al. 2022a, MNRAS, doi: 10.1093/mnras/stac2105

Dey, B., Newman, J. A., Andrews, B. H., et al. 2021, arXiv e-prints, arXiv:2110.15209. https://arxiv.org/abs/2110.15209

Dey, B., Zhao, D., Newman, J. A., et al. 2022b, arXiv e-prints, arXiv:2205.14568. https://arxiv.org/abs/2205.14568

Duncan, K., Conselice, C. J., Mortlock, A., et al. 2014, MNRAS, 444, 2960, doi: 10.1093/mnras/stu1622

Duncan, K. J., Jarvis, M. J., Brown, M. J. I., & Röttgering, H. J. A. 2018a, MNRAS, 477, 5177, doi: 10.1093/mnras/sty940

Duncan, K. J., Brown, M. J. I., Williams, W. L., et al. 2018b, MNRAS, 473, 2655, doi: 10.1093/mnras/stx2536

Faber, S. M., Phillips, A. C., Kibrick, R. I., et al. 2003, in Proc. SPIE, Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes, ed. M. Iye & A. F. M. Moorwood, 1657–1669, doi: 10.1117/12.460346

Finkelstein, S. L., Ryan, Russell E., J., Papovich, C., et al. 2015, ApJ, 810, 71, doi: 10.1088/0004-637X/810/1/71

Fioc, M., & Rocca-Volmerange, B. 1997a, A&A, 326, 950

—. 1997b, A&A, 326, 950

Fontana, A., D'Odorico, S., Poli, F., et al. 2000, AJ, 120, 2206, doi: 10.1086/316803

Galametz, A., Grazian, A., Fontana, A., et al. 2013, The Astrophysical Journal Supplement Series, 206, 10. http://stacks.iop.org/0067-0049/206/i=2/a=10

Giallongo, E., D'Odorico, S., Fontana, A., et al. 1998, AJ, 115, 2169, doi: 10.1086/300361

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35, doi: 10.1088/0067-0049/197/2/35

Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24, doi: 10.1088/0067-0049/207/2/24

Hatfield, P. W., Jarvis, M. J., Adams, N., et al. 2022, MNRAS, 513, 3719, doi: 10.1093/mnras/stac1042

Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, ApJL, 911, L33, doi: 10.3847/2041-8213/abf2c7

Hildebrandt, H., Wolf, C., & Benítez, N. 2008, A&A, 480, 703, doi: 10.1051/0004-6361:20077107

Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31, doi: 10.1051/0004-6361/201014885

Hogg, D. W., Cohen, J. G., Blandford, R., et al. 1998, AJ, 115, 1418, doi: 10.1086/300277

Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841, doi: 10.1051/0004-6361:20065138

Ilbert, O., Salvato, M., Capak, P., et al. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamada, & K. Aoki, 169

Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236, doi: 10.1088/0004-637X/690/2/1236

Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, ApJ, 467, 38, doi: 10.1086/177583

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36, doi: 10.1088/0067-0049/197/2/36

Kriek, M., van Dokkum, P. G., Franx, M., et al. 2007, ApJ, 669, 776, doi: 10.1086/520789

Kriek, M., Shapley, A. E., Reddy, N. A., et al. 2015, ApJS, 218, 15, doi: 10.1088/0067-0049/218/2/15

Krogager, J.-K., Zirm, A. W., Toft, S., Man, A., & Brammer, G. 2014, ApJ, 797, 17, doi: 10.1088/0004-637X/797/1/17

Kurk, J., Cimatti, A., Daddi, E., et al. 2012, VizieR Online Data Catalog, 354

Laidler, V. G., Papovich, C., Grogin, N. A., et al. 2007, PASP, 119, 1325, doi: 10.1086/523898

Lang, D., & Hogg, D. W. 2012, AJ, 144, 46, doi: 10.1088/0004-6256/144/2/46

Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845, doi: 10.1051/0004-6361:20041960

Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, A&A, 559, A14, doi: 10.1051/0004-6361/201322179

Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. 2015, A&A, 576, A79, doi: 10.1051/0004-6361/201423829

Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007a, ApJS, 172, 70, doi: 10.1086/516589

—. 2007b, ApJS, 172, 70, doi: 10.1086/516589

Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218, doi: 10.1088/0067-0049/184/2/218

Madau, P. 1995, ApJ, 441, 18, doi: 10.1086/175332

Maraston, C. 2005a, MNRAS, 362, 799, doi: 10.1111/j.1365-2966.2005.09270.x

—. 2005b, MNRAS, 362, 799, doi: 10.1111/j.1365-2966.2005.09270.x

Marchesini, D., van Dokkum, P. G., Förster Schreiber, N. M., et al. 2009, ApJ, 701, 1765, doi: 10.1088/0004-637X/701/2/1765

McLure, R. J., Pearce, H. J., Dunlop, J. S., et al. 2013, MNRAS, 428, 1088, doi: 10.1093/mnras/sts092

McLure, R. J., Pentericci, L., Cimatti, A., et al. 2018, MNRAS, 479, 25, doi: 10.1093/mnras/sty1213

Merlin, E., Castellano, M., Santini, P., et al. 2021, A&A, 649, A22, doi: 10.1051/0004-6361/202140310

Mignoli, M., Cimatti, A., Zamorani, G., et al. 2005, A&A, 437, 883, doi: 10.1051/0004-6361:20042434

Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016a, ApJS, 225, 27, doi: 10.3847/0067-0049/225/2/27

—. 2016b, ApJS, 225, 27, doi: 10.3847/0067-0049/225/2/27

Morris, A. M., Kocevski, D. D., Trump, J. R., et al. 2015, AJ, 149, 178, doi: 10.1088/0004-6256/149/6/178

Muzzin, A., Marchesini, D., Stefanon, M., et al. 2013, ApJ, 777, 18, doi: 10.1088/0004-637X/777/1/18

Nayyeri, H., Hemmati, S., Mobasher, B., et al. 2017, The Astrophysical Journal Supplement Series, 228, 7. http://stacks.iop.org/0067-0049/228/i=1/a=7

Newman, J. A. 2008, ApJ, 684, 88, doi: 10.1086/589982

Newman, J. A., & Gruen, D. 2022, arXiv e-prints, arXiv:2206.13633. https://arxiv.org/abs/2206.13633

Newman, J. A., Zepf, S. E., Davis, M., et al. 1999, The Astrophysical Journal, 523, 506. http://stacks.iop.org/0004-637X/523/i=2/a=506

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5, doi: 10.1088/0067-0049/208/1/5

Newman, J. A., Abate, A., Abdalla, F. B., et al. 2015, Astroparticle Physics, 63, 81, doi: 10.1016/j.astropartphys.2014.06.007

Noeske, K. G., Weiner, B. J., Faber, S. M., et al. 2007, ApJL, 660, L43, doi: 10.1086/517926

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26, doi: 10.1051/0004-6361/201833617

Pérez-González, P. G., Cava, A., Barro, G., et al. 2013, ApJ, 762, 46, doi: 10.1088/0004-637X/762/1/46

Pforr, J., Maraston, C., & Tonini, C. 2013, MNRAS, 435, 1389, doi: 10.1093/mnras/stt1382

Pirzkal, N., Rothberg, B., Ly, C., et al. 2013, ApJ, 772, 48, doi: 10.1088/0004-637X/772/1/48

Polletta, M., Tajer, M., Maraschi, L., et al. 2007, ApJ, 663, 81, doi: 10.1086/518113

Popesso, P., Dickinson, M., Nonino, M., et al. 2009, A&A, 494, 443, doi: 10.1051/0004-6361:200809617

Press, W. H. 1997, in Unsolved Problems in Astrophysics, ed. J. N. Bahcall & J. P. Ostriker, 49–60

Rana, N. C., & Basu, S. 1992, A&A, 265, 499

Ravikumar, C. D., Puech, M., Flores, H., et al. 2007, A&A, 465, 1099, doi: 10.1051/0004-6361:20065358

Reddy, N. A., & Steidel, C. C. 2009, ApJ, 692, 778, doi: 10.1088/0004-637X/692/1/778

Rényi, A. 1961, On measures of entropy and information, Proc. 4th Berkeley Symp. Math. Stat. Probab. 1, 547-561 (1961).

Salmon, B., Papovich, C., Finkelstein, S. L., et al. 2015, ApJ, 799, 183, doi: 10.1088/0004-637X/799/2/183

Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nature Astronomy, 3, 212, doi: 10.1038/s41550-018-0478-0

Santini, P., Ferguson, H. C., Fontana, A., et al. 2015, ApJ, 801, 97, doi: 10.1088/0004-637X/801/2/97

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, MNRAS, 499, 1587, doi: 10.1093/mnras/staa2799

Shivaei, I., Reddy, N. A., Shapley, A. E., et al. 2015, ApJ, 815, 98, doi: 10.1088/0004-637X/815/2/98

Silverman, J. D., Mainieri, V., Salvato, M., et al. 2010, ApJS, 191, 124, doi: 10.1088/0067-0049/191/1/124

Silverman, J. D., Kashino, D., Sanders, D., et al. 2015, ApJS, 220, 12, doi: 10.1088/0067-0049/220/1/12

Speagle, J. S., Steinhardt, C. L., Capak, P. L., & Silverman, J. D. 2014, ApJS, 214, 15, doi: 10.1088/0067-0049/214/2/15

Stefanon, M., Yan, H., Mobasher, B., et al. 2017, ApJS, 229, 32, doi: 10.3847/1538-4365/aa66cb

Szokoly, G. P., Bergeron, J., Hasinger, G., et al. 2004, ApJS, 155, 271, doi: 10.1086/424707

Tomczak, A. R., Quadri, R. F., Tran, K.-V. H., et al. 2014, ApJ, 783, 85, doi: 10.1088/0004-637X/783/2/85

Treister, E., Virani, S., Gawiser, E., et al. 2009, ApJ, 693, 1713, doi: 10.1088/0004-637X/693/2/1713

Trump, J. R., Konidaris, N. P., Barro, G., et al. 2013, ApJL, 763, L6, doi: 10.1088/2041-8205/763/1/L6

van der Wel, A., Franx, M., van Dokkum, P. G., & Rix, H.-W. 2004, ApJL, 601, L5, doi: 10.1086/381887

Vanzella, E., Cristiani, S., Dickinson, M., et al. 2008, A&A, 478, 83, doi: 10.1051/0004-6361:20078332

Vanzella, E., Giavalisco, M., Dickinson, M., et al. 2009, ApJ, 695, 1163, doi: 10.1088/0004-637X/695/2/1163

Whitaker, K. E., van Dokkum, P. G., Brammer, G., & Franx, M. 2012, ApJL, 754, L29, doi: 10.1088/2041-8205/754/2/L29

Wiklind, T., Dickinson, M., Ferguson, H. C., et al. 2008, ApJ, 676, 781, doi: 10.1086/524919

Wirth, G. D., Willmer, C. N. A., Amico, P., et al. 2004, AJ, 127, 3121, doi: 10.1086/420999

Wirth, G. D., Trump, J. R., Barro, G., et al. 2015, AJ, 150, 153, doi: 10.1088/0004-6256/150/5/153

Wolf, C., Meisenheimer, K., Kleinheinrich, M., et al. 2004, A&A, 421, 913, doi: 10.1051/0004-6361:20040525

Wuyts, S., van Dokkum, P. G., Franx, M., et al. 2009, ApJ, 706, 885, doi: 10.1088/0004-637X/706/1/885

Zhou, R., Cooper, M. C., Newman, J. A., et al. 2019, MNRAS, 488, 4565, doi: 10.1093/mnras/stz1866