Publications of the Astronomical Society of the Pacific, 134:044501 (11pp), 2022 April https://doi.org/10.1088/1538-3873/ac59bf

The Sensitivity of GPz Estimates of Photo-z Posterior PDFs to Realistically Complex Training Set Imperfections

Natalia Stylianou[1], Alex I. Malz[2], Peter Hatfield[3], John Franklin Crenshaw[4], and Julia Gschwend[5,6]
[1] School of Physics and Astronomy, University of Leicester, University Road, LE1 7RH, UK
[2] German Centre for Cosmological Lensing, Astronomisches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, Germany
[3] Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
[4] DIRAC Institute and Department of Physics, University of Washington, Seattle, WA 98195, USA
[5] Laboratório Interinstitucional de e-Astronomia (LIneA), Rua General José Cristino, 77, Rio de Janeiro, Brazil
[6] Observatório Nacional, Rua General José Cristino, 77, Rio de Janeiro, RJ, 20921-400, Brazil
Received 2021 December 10; accepted 2022 February 23; published 2022 April 26

Abstract

The accurate estimation of photometric redshifts is crucial to many upcoming galaxy surveys, for example, the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST). Almost all Rubin extragalactic and cosmological science requires accurate and precise calculation of photometric redshifts; many diverse approaches to this problem are currently in the process of being developed, validated, and tested. In this work, we use the photometric redshift code GPz to examine two realistically complex training set imperfections scenarios for machine learning based photometric redshift calculation: (i) where the spectroscopic training set has a very different distribution in color–magnitude space to the test set, and (ii) where the effect of emission line confusion causes a fraction of the training spectroscopic sample to not have the true redshift. By evaluating the sensitivity of GPz to a range of increasingly severe imperfections, with a range of metrics (both of photo-z point estimates as well as posterior probability distribution functions, PDFs), we quantify the degree to which predictions get worse with higher degrees of degradation. In particular, we find that there is a substantial drop-off in photo-z quality when line-confusion goes above $\sim$1%, and sample incompleteness below a redshift of 1.5, for an experimental setup using data from the Buzzard Flock synthetic sky catalogs.

Unified Astronomy Thesaurus concepts: Astrostatistics (1882); Astrostatistics techniques (1886); Photometry (1234); Redshift surveys (1378)

1. Introduction

The estimation of the redshift of distant astronomical sources (mainly galaxies and Active Galactic Nuclei, AGN) is a crucial part of modern cosmology (Hoyle et al. 2018) and extragalactic science (Miyaji et al. 2015). However with increasingly large data sets, in the era of high-precision cosmology, the requirements on the quality of galaxy redshift estimation can be very high (Mitra & Linder 2021). If, for example, in the instance of cosmological inference (say 3×2 pt analysis, Zuntz et al. 2021), the redshifts were systematically a few percent higher or lower than the true redshifts (unknown to us and not included in the modeling), there could be a risk of inferring the incorrect cosmological model. Furthermore, redshifts are critical for extragalactic science, including galaxy formation and evolution, since they provide the third dimension and time evolution (see Etherington et al. 2017 and Fontana et al. 2000). Their use in a variety of science disciplines, therefore, leads to a strong need to understand their accuracy and precision.

There are two main observational approaches to estimating redshifts, both of which have advantages and disadvantages; spectroscopy and photometry (Fernández-Soto et al. 2001). Spectroscopic redshifts ("spec-z") are measured by identifying an emission/absorption feature in a galaxy's spectrum and comparing it to the known rest-frame wavelength. Spec-z estimations typically provide highly accurate redshift values, but can be expensive in terms of telescope time and are thus limited by the sample size, and also typically are more challenging to obtain for high redshift and low luminosity sources.

By contrast, photometric redshifts ("photo-z") make use of photometry. When a spectrum is redshifted, spectral features move in and out of different photometric bands, giving changing measured magnitudes; thus, the photo-z technique relies on the capacity to isolate the wavelength position of redshifted continuum features (e.g., Balmer or Lyman breaks). Hence instead of having a spectrum, we have a certain number

of discrete photometric bands which must be mapped onto a redshift value. The need for multiple photometric bands is due to the redshift degeneracies present when one color corresponds to multiple redshift values (e.g., confusion between different breaks). Wide multiwavelength coverage is necessary for photo-z surveys to limit this effect. The primary benefit of using photo-z is the derivation of redshift measurements for a much larger number of sources detected in imaging surveys, typically to higher magnitudes and redshifts. These low-cost photo-z estimates, however, are typically much less precise than spec-z estimates.

For many existing and forthcoming galaxy surveys like the Dark Energy Survey (The Dark Energy Survey Collaboration 2005), The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST Science Collaboration et al. 2009), and Euclid (Laureijs et al. 2011), the majority of galaxy redshifts will be based on photometry.

Photo-z's themselves have multiple different approaches to their calculation, including template-fitting, machine learning (ML), and hybrid techniques. In the case of template fitting, a series of galaxy templates are selected and a chi-square fitting-like approach is performed where they are shifted in order to see what fits the photometry best. Some examples of template fitting codes include HYPERZ (Bolzonella et al. 2000), LE PHARE (Ilbert et al. 2006), and EAZY (Brammer et al. 2008).

The ML approach instead utilizes a sample of galaxies with *both* photometric and spectroscopic values as the training set of the algorithm. The ML code then "learns" how to map the photometric color-magnitudes of the training data onto the redshift values. Salvato et al. (2019) discuss several of these approaches to accurately estimating photo-z's, while also commenting on the challenge in achieving high redshift precision in large-scale galaxy surveys. Hybrid approaches typically attempt to combine template fitting and machine methods (see Duncan et al. 2018 and Hatfield et al. 2020).

There are now several studies seeking to rigorously assess and compare photo-z performance. Schmidt et al. (2020) investigated twelve photo-z algorithms, where the codes were tested in ideal training and test data scenarios with mock data created for the Rubin Dark Energy Science Collaboration (DESC). Similarly, in Euclid Collaboration et al. (2020), thirteen photo-z methods using either template-fitting or ML techniques were examined on Euclid-like data, providing a detailed comparison of their metrics. The goal of these studies is to develop photo-z techniques appropriate to each survey, and to understand in advance what systematic biases need to be modeled and mitigated.

One important aspect of the ML approach is a dependence on a reliable spectroscopic sample. This is one of the primary sources of systematic bias affecting photo-z estimation with ML. We hence need to consider the *representativity* of the training data, since commonly, the spectroscopic redshifts do not span the full color-space that the target data set might (see Beck et al. 2017). This can lead to reduced performance in parts of color–magnitude space poorly represented in the training set. Beyond representativity, another source of systematic error is incorrectly labeled training data i.e., incorrect spectroscopic redshifts, normally as a result of emission-line confusion. This can risk the ML algorithm incorrectly "claiming" it is giving good predictions, because it is giving photo-z predictions that agree with the spectroscopy—but if the spec-z are themselves inaccurate, then it is very difficult to evaluate the true performance.

The challenge addressed by this paper is to try and understand the impact on photo-z estimation (and specifically on the calculated posterior PDFs) in the non-idealized scenario where (i) the training and test data have dramatically different distributions in color–magnitude space and (ii) some fraction of the spectroscopic data is mislabeled. We use the ML code GPz, which has already been tested for Legacy Survey of Space and Time (LSST) and Euclid-like scenarios (Schmidt et al. 2020 and Euclid Collaboration et al. 2020). A number of photo-z metrics are evaluated for a range of degradations, not just for point estimates, but also photo-z PDF compared with true redshift, and photo-z PDF compared with true redshift PDF. In Section 2 we will discuss the data used in this work (and what degradations we applied to them), in Section 3 we will describe how we calculated our photo-z estimates, in Section 4 we describe the metrics used, in Section 5 we show our results, we discuss them in Section 6, and finally we conclude in Section 7. The code for this work is available on GitHub at https://github.com/nataliastylianou/Photo-z.

2. Data Generation

2.1. The Buzzard Simulation

The data set used in this work is a sample of 100,000 galaxies from the Buzzard Flock synthetic sky catalogs[7] (DeRose et al. 2019), with redshifts in the range $0 < z < 2.3$ and photometry in the LSST ugrizy bands.[8] The Buzzard catalogs are constructed by first adding galaxies onto a dark-matter-only N-body simulation (in such a way as to be consistent with known lower-redshift luminosity functions), and then "observing" them by imposing a realistic set of observational properties and systematics. We chose this data set in order to have a mock catalog of sufficient size with observational properties similar to LSST, and did not need some of the more complex physics that other simulations and mock catalogs might capture (as we are specifically focusing on the impact of degradation of the spectroscopic training set, not other effects e.g., stellar contamination).

[7] Also used in Schmidt et al. (2020).

[8] In particular we use the sample from https://github.com/jfcrenshaw/pzflow.

2.2. Normalizing Flows

In this work, we will at points compare *predicted* redshift PDF to *true* redshift PDF. Galaxies only have one redshift, but given a set of photometric observations (bands and depths), there is a PDF that would represent the perfect extraction of redshift information. If the joint $N+1$ (where N is the number of bands) dimensional redshift magnitude distribution $p(z, \boldsymbol{m})$ (where z is the redshift, and $\boldsymbol{m}$ was the photometry) were known perfectly, then the "true" redshift PDF for a set of observed magnitudes would be $p(z|\boldsymbol{m})$, the best possible estimation of the redshift (even though individual galaxies only have a single value for redshift). In a realistic observational scenario, we will not a priori know $p(z, \boldsymbol{m})$ — but in this paper, we will test how well we can reconstruct $p(z|\boldsymbol{m})$ in a scenario where we do know the full distribution from the simulation.

The joint distribution $p(z, \boldsymbol{m})$ is closely linked to the redshift distribution for the whole population, $N(z)$ (which is required for some science goals). First, if $p(z, \boldsymbol{m})$ is perfectly known, the univariate distribution for z can be found with a simple marginalization: $N(z) = \int p(z, \boldsymbol{m}) \mathrm{d}\boldsymbol{m}$. Second, given a set of galaxies with "true" redshift PDFs $p(z|\boldsymbol{m})$, and the corresponding population distribution of color-magnitudes $p(\boldsymbol{m})$ (itself a function of the underlying luminosity functions and the observational properties of the survey), $N(z)$ can again be recovered by performing the relevant weighted integral ("stacking"): $N(z) = \int p(z|\boldsymbol{m}) p(\boldsymbol{m}) \mathrm{d}\boldsymbol{m}$. Finally, if we knew $p(z, \boldsymbol{m})$ perfectly, then we could treat it as a prior in redshift for an unseen galaxy before any magnitudes were observed (or if only some of the magnitudes were known). For more realistic observational cases where we are merely estimating the true PDF, and do not perfectly know $p(z, \boldsymbol{m})$, a set of more complex approaches have been developed to convert a set of PDF estimates back to $N(z)$ (e.g., see Malz 2021, which discusses more rigorously when a simple stacking approach is appropriate, and when it is not).

In the case of a simulation, we perfectly know both (a) the intrinsic joint distribution of magnitudes and redshifts for the galaxies, and (b) the selection function, which is applied in the simulation to construct the mock sky catalog. However, in the case of real observations, we neither know the intrinsic $p(z, \boldsymbol{m})$ distribution, which is often what we intend to measure, nor do we typically perfectly understand the selection effects, although we usually have some knowledge of certain constraints (e.g., detection thresholds).

The sample of sources from the Buzzard Flock synthetic sky catalogs all have single redshift values. In order to construct true redshift PDFs to compare against, we use a Normalizing Flow (Jimenez Rezende & Mohamed 2015) to model the $p(z, \boldsymbol{m})$ of the Buzzard sample. Normalizing Flows are tools that use sequences of invertible mappings to convert simple distributions into more complex ones. From the resulting Normalizing Flow we can sample galaxies with redshifts, galaxies with redshift PDFs, and we can even apply transformations that correspond to degradations to create new Normalizing Flows, which we can also sample from. To further build a Normalizing Flow we use *pzflow*, which is a package that models normalizing flows (Crenshaw 2021). The approach here uses code from the Redshift Assessment Infrastructure Layers code (RAIL[9]) and borrows heavily in its approach from the example flow in *pzflow*.[10]

We use this Normalizing Flow sample with photometry, redshifts, and true redshift PDFs as our "no-degradation" sample. This data acts as our training, validation, and test data for the "no-degradation" case, and also the test data for when the algorithm is trained on the degraded data.

2.3. More Realistically Complex Training Set Imperfections

Ideally, for an ML-based calculation of photo-z, the training set would consist of perfectly redshift labeled sources with the exact same color–magnitude distribution as the test data. In practice, this is never achieved.

This paper will focus on two sources of training-set imperfections with the aid of two degraders (taken from RAIL[11]): Inverse Redshift Incompleteness, which introduces sample incompleteness (i.e., where the training set is not representative of the test data), and the Emission Line Confusion, which includes spectroscopic systematics (i.e., training spectroscopic data are labeled with incorrect redshifts).

Other training set imperfections not discussed in this paper include AGN variability—consider that if the magnitudes of sources are changing over time then photometric redshift estimates risk becoming more unreliable (as the source redshift does not change) (Simm et al. 2015). Similarly, the blending of sources, and dust reddening depending on galactic coordinates, all result in inaccurate photometric redshift estimates (Calzetti et al. 2000).

2.3.1. Inverse Redshift Incompleteness

The Inverse Redshift Incompleteness Degrader attempts to replicate redshift incompleteness by applying a selection function inversely proportional to redshift. Its selection function probability is described by:

$$p(z) = \min\left(1, \frac{z_p}{z}\right) \tag{1}$$

where the z_p term defines the pivot redshift, specifying the redshift where the incompleteness begins.

[9] https://github.com/LSSTDESC/RAIL

[10] https://github.com/jfcrenshaw/pzflow/blob/main/pzflow/examples/examples.py

[11] https://github.com/LSSTDESC/RAIL/tree/master/rail/creation/degradation

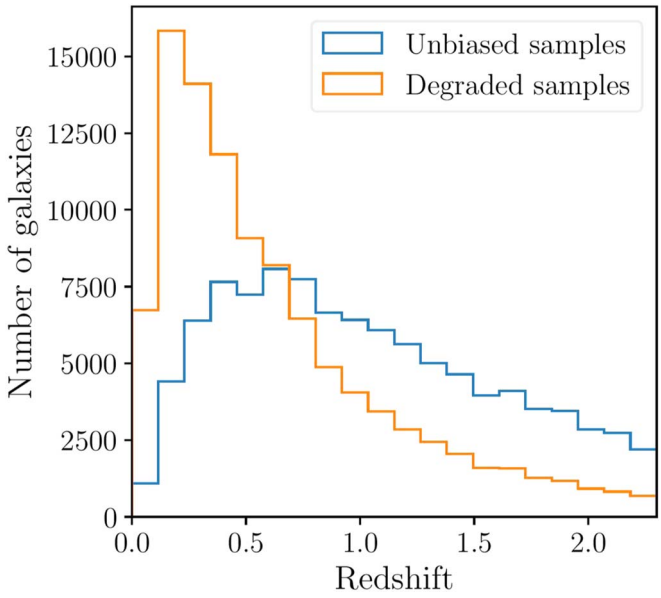


Figure 1. Redshift distributions for degraded and non-degraded (unbiased) data sets.

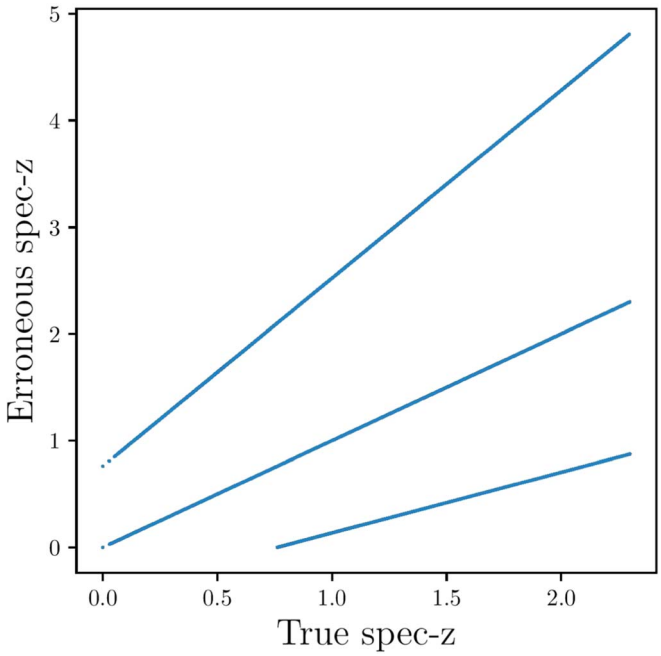


Figure 2. Degraded spectroscopic redshifts against non-degraded true redshifts data sets for the Buzzard photometry used in this study.

Figure 1 shows the two different redshift distributions for the unbiased representative sample set and the degraded sample set through the Inverse Redshift Incompleteness Degrader. We set the pivot redshift for this degradation equal to $z_p = 0.10$. As seen in Figure 1, the degraded data set has most of its galaxies concentrated in the lower redshifts and very few lying in the high-redshift area. Such an effect can be caused by the difficulty in making spectroscopic measurements for high-redshift galaxies with faint photometric magnitudes (although of course, in general, low-luminosity sources at low-redshift may also be affected).

2.3.2. *Emission Line Confusion*

The Emission Line Confusion Degrader mimics the effect of spectroscopic systematic errors by simulating the confusion of different emission lines. Specifically, we used the Emission Line Confusion degrader to misidentify between 0.2% and 10% of O II lines as Hα lines and vice versa (for a discussion of plausible line-confusions and percentages see Euclid Collaboration et al. 2021). Since O II emission lines have a wavelength of 3727 Å and Hα lines a wavelength of 6563 Å, this confusion would consequently result in a larger spectroscopic redshift and the opposite misidentification would result in a smaller spectroscopic redshift. Euclid Collaboration et al. (2021) has considered more complex models of emission line confusion, such as the misclassification of Hα as O II lines for redshifts lower than 0.5 and between 1.4 and 2, O II lines as Hα or Lyα lines for redshifts between 0.5 and 1.4, and for redshifts above 2 then Lyα lines are misidentified as O II lines. For the purposes of this paper, we proceeded to only use the O II and Hα line confusion since further complexity would increase the realism by a small amount, as which line confusions are key will depend heavily on what spectroscopic training set is actually used (what resolution the spectrographs have, what lines were used for redshift measurement, what flag tolerance was used etc.).

Figure 2 shows the degraded spec-z's with the emission line errors against the true non-degraded redshifts. The percentage of degradation used in Figure 2 is 5% (which we will refer to as a "badness" parameter of 0.05). The central line (along the diagonal representing equality) in the plot represents where the true spec-z's and the degraded spec-z's are equal and there was no confusion in the line identification. Conversely, the two lines diverging from the one-to-one line illustrate where the O II and Hα have been misclassified, resulting in a larger or smaller spec-z value. The more degradation the spec-z data endure, the more prominent the two diverging lines would be and the weaker the identity line would appear.

3. Photo-z Estimation

The ML code we use in this work to estimate photo-zs is called GPz. It is a sparse Gaussian process code described in Almosallam et al. (2016a) and Almosallam et al. (2016b). GPz produces a point estimated mean and a variance that incorporates both the uncertainty due to intrinsic output noise, as well as due to low data density. Hence it accounts for

training data insufficiency and galaxy magnitude degeneracies. GPz has been observed to be a fast, high-performing ML code that typically performs well for a range of metrics, normally particularly bias metrics. One weakness is that it outputs only point estimates with uncertainty as opposed to more general PDFs (uni-modal Gaussians instead of multi-modal PDFs), which can prove problematic for certain sources.

Ultimately, as with analogous ML-based codes, the performance of GPz is dependent on the spectroscopic training data used. As mentioned in Section 1, Schmidt et al. (2020) and Euclid Collaboration et al. (2020) tested GPz in the context of a photo-z data challenge. In addition to this, GPz has been used in a number of separate studies, including Gomes et al. (2018) who tested the inclusion of source size information, and Hatfield et al. (2020) who combined Gaussian mixture models (GMMs) with GPz to improve redshift estimation and ultimately accelerate photo-z computation.

4. Evaluation Metrics

To assess the quality of photometric redshift estimations, there are a large number of different metrics that characterize in different ways how successful the estimates have been at predicting the true redshift. Metrics might compare just the photo-z point estimate to the true redshift, the photo-z PDF to the true redshift, or ultimately the photo-z PDF to the true PDF.

4.1. Point Estimate Metrics

We consider the following three metrics (see Section 6.3 of Almosallam et al. 2016a) to evaluate the point estimated outputs compared to single-valued true redshifts; these are (i) the root mean squared error (RMSE), (ii) the fraction retained for 15% (FR15) and iii) the Bias.

Specifically, the RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{1}^{n}\left(\frac{z_{\mathrm{spec}} - z_{\mathrm{photo}}}{1 + z_{\mathrm{spec}}}\right)^2} \tag{2}$$

The FR15 classifies the fraction of catastrophic outliers with a 15% threshold defined as:

$$\mathrm{FR15} = \frac{100}{n}\left\{\left|\frac{z_{\mathrm{spec}} - z_{\mathrm{photo}}}{1 + z_{\mathrm{spec}}}\right| < 0.15\right\} \tag{3}$$

Finally, the Bias demonstrates how the photometric redshift deviates systematically from the true redshift.

$$\mathrm{Bias} = \frac{1}{n}\sum_{1}^{n}\frac{z_{\mathrm{spec}} - z_{\mathrm{photo}}}{1 + z_{\mathrm{spec}}} \tag{4}$$

All of these quantities can be calculated for the population as a whole, or considered as a function of redshift (or some other parameter).

4.2. PDF Metrics Relative to True Redshifts

Beyond assessing the quality of point-estimates of the redshift, we might also wish to assess the quality of our uncertainty estimates, and the realism of the PDFs. We can hence consider (i) the Probability Integral Transform (PIT) and (ii) the Conditional Density Estimate (CDE) loss.

There is a broad range of other metrics in the literature to assess the quality of regression analyzes that could be used to characterize the relationship between photo-z PDFs and the true redshifts. One possibility is the coefficient of determination (R^2, Wright 1921) metric, which characterizes how much of the variation in the data is captured by a predictive model (where $R^2 = 1$ corresponds to perfect predictive power, $R^2 = 0$ would correspond to predicting the population mean each time, and $R^2 < 0$ corresponds to predictions poorer than simply guessing the population mean for each data point). This is in contrast to RMSE measurements for example, which can give comparisons of quality between two regressions, but do not solely by themselves give an indication of whether the model is capturing the full variance present in the data. R^2 has been used in astronomy for a range of applications, for example modeling light curves (Shoji et al. 2020) and testing goodness of fit in studies of magnetohydrodynamical turbulence (González-Casanova et al. 2018). In terms of comparing PDFs rather than point estimates, R^2 was generalized to the Bayesian context in Gelman et al. (2019). The R^2 metric captures several aspects of the quality of fit well, but can give misleading conclusions in others (see for example Lewis-beck & Skalaban 1990 for a discussion). We decided not to use R^2 in this work, as we felt the RMSE and PIT metrics together quantified the same behavior R^2 captures (RMSE describing quality of fit, and PIT characterizing how much of the true variance was captured by the calculated uncertainties). For example, when uncertainties on predictions are too small, the histogram of the PIT values shows a characteristic peak at 0 and 1, a peak at 0.5 when uncertainties are too large, and perfectly calibrated photo-z PDFs give a uniform distribution (see Section 4.2.1). In addition, the PIT distribution has been used to quantify the performance of photo-z PDF methods in many prior instances (Freeman et al. 2017; Tanaka et al. 2018 and Polsterer et al. 2016), making it more straightforwards to compare our work to other studies. However, using (Bayesian) R^2 to compare photometric redshift PDFs would be an interesting study for future work.

4.2.1. Probability Integral Transform (PIT)

The PIT metric seeks to assess the "realism" of PDFs for estimates for a population. For each prediction it takes the integral of individual PDFs from zero up to the true redshift, and then plots the distribution (a histogram) of those values. A histogram of PIT values is commonly used to assess how "realistic" a population of photo-z PDFs is compared with the

true redshifts. Ideally, the histogram would appear as a uniform distribution, which corresponds to the PDFs being perfectly calibrated.

Although the PIT distribution is ideally flat, it is expected to be less flat for the biased training data set than for the representative data set. We can hence use this to see the impact of degradation on the predictions. The test statistics of the PIT distribution tell us about the deviation from that ideal flat distribution and are expected to be more discrepant for more biased training sets.

Outliers at high-redshifts often have underestimated means with large variances, but at low-redshifts they typically have overestimated means with small variances. There are more of the latter, so we expect an overabundance of low PIT values.

Note that the PIT is assessing the realism of the PDFs, not their information content per se; Schmidt et al. (2020) showed that uninformative PDFs could get high PIT scores because they were very well-calibrated PDFs, but did not give any information.

4.2.2. Conditional Density Estimate (CDE) loss

The CDE loss is the discrepancy of the ensemble of PDFs relative to the true redshift-photometry distribution. See Section 4.2 in Schmidt et al. (2020) (also Dalmasso et al. 2020) for a full description and more detailed definition, but it is essentially the root-mean-square-error of the difference between the true and the predicted PDFs. However, in the absence of knowledge of the true PDF, it can still be determined up to a constant, even in the absence of knowledge of the true underlying redshift-photometry distribution.

The CDE loss metric essentially approximates the true posterior PDF from the estimated posterior PDF evaluated at the true redshift. Therefore, the lower the CDE loss is, the better the predictions. The CDE loss might typically be expected to get worse the more biased and degraded the training set data is.

4.2.3. Summary Statistics

The exact shapes of the PIT curves contain information about which way the estimated PDFs are biased. However, these curves can be summarized further if a single number is required.

A Kolmogorov–Smirnov statistic test (KS test) can be used to find the maximum difference between the true and estimated cumulative distributions of the PIT values. The KS test outputs values from zero to one and the closer the KS value is to zero, the more uniform the PIT distribution is. Therefore, the KS values of the representative data sets are expected to be lower than for the biased data sets.

Similarly, a variant of the KS test is the Cramer-von Mises test (CvM test) which represents the mean-square difference between the cumulative distribution functions of estimated and true PDFs, and again would be expected to be lower for the representative training data sets.

Finally, a modification to the KS test is the Anderson-Darling test (AD test). The AD test describes the weighted mean-squared difference and gives more weight to discrepancies in the PIT distribution tails.

See Schmidt et al. (2020) and references within for the PIT, CDE loss, and summary statistics tests.

4.3. PDF Metric Relative to True PDFs

Finally, a comparison of the estimated and true photo-z posterior PDFs is possible, which we will do here with the Kullback Leibler Divergence (KLD) metric. This metric has been used for photo-z PDF evaluation in Malz et al. (2018).

4.3.1. Kullback Leibler Divergence (KLD)

The KLD is the information content difference between the predicted PDF and the true PDF from which the data was generated (in our case via the Normalizing Flow), and it is estimated for each PDF in the sample. Ideally, the KLD value for each galaxy would be very small, signifying a low information loss from using estimated photo-z PDFs instead of the true PDFs.

$$\mathrm{KLD} = \int_{-\infty}^{\infty} \mathrm{PDF}_{\mathrm{true}}(z) \log\left(\frac{\mathrm{PDF}_{\mathrm{true}}(z)}{\mathrm{PDF}_{\mathrm{estimated}}(z)}\right) dz \qquad (5)$$

5. Results

To estimate photo-z's with GPz, samples of 100,000 sources from the NF (either the degraded or non-degraded NF, as appropriate) are created. These data sets are then split into 3 subsequent sets. 20% of the data was used for the training of the algorithm, 20% for the validation process, and the remaining 60% of the data was used for the testing. For the non-degraded case, GPz is trained and validated on non-degraded data, and tested on non-degraded data. However, for the degraded case, GPz is trained and validated on the degraded data, but is still tested on the non-degraded data (i.e., both non-degraded and degraded predictors are applied to the same data). This allows us to probe the impact of training data imperfections on the quality of the predictor.

The GPz performance with the NF data under no degradation is illustrated in Figure 3. We demonstrate the estimated photo-z's in contrast to the spec-z's, colored by number density. It is evident that most galaxies have accurate photo-z's consistent with their spec-z's (lying on the diagonal), especially for $z < 1$.

In the following subsections we describe the deviations from the best prediction performance quantitatively via the metrics previously discussed, as degradations are introduced.

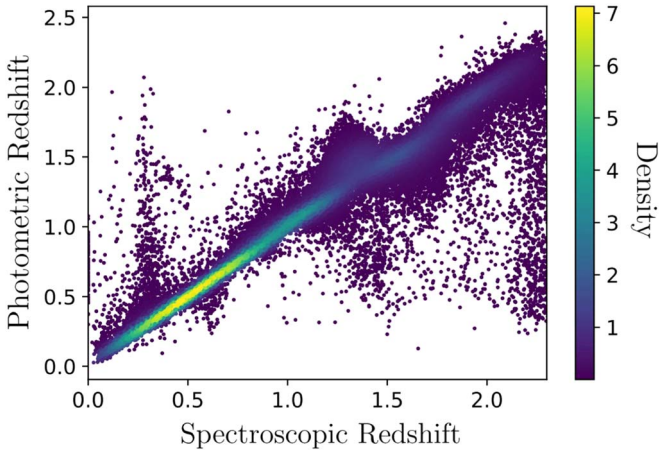


Figure 3. Photo-z predictions with the NF data.

5.1. Point Estimates

Figure 4 exhibits how the RMSE, the FR15, and the Bias metrics vary with representative and non-representative data for the two degraders. The plots correspond to a degradation of 5% for the Emission Line Confusion Degrader and a pivot redshift of 0.92 for the Inverse Redshift Incompleteness Degrader. The RMSE of the biased data is larger than for the representative data in both degraders. Similarly, the biased data shows a lower FR15 in the two degraders than for the representative data.

Finally, the non-degraded data has a very flat Bias close to zero in both cases, whereas the degraded data has biases that deviate from zero in the negative and positive direction respectively for each degrader.

Importantly however, throughout all of these metrics, the Emission Line Confusion degrader plots illustrate a significantly larger discrepancy between the representative and degraded data (at least for the experimental setup considered here).

5.2. PDF Relative to True Redshift

In Figure 5 we show the two plots of the PIT distribution of degraded and non-degraded NF data for degradations of 0.05 badness and a 0.92 pivot redshift, respectively. Both plots indicate a nearly flat distribution for the representative data with the exception of a spike in the very beginning and a considerably smaller spike at the very end of the distribution. The biased data in the Emission Line Confusion Degrader shows a peak at about 0.5, while for the Inverse Redshift Incompleteness Degrader the biased data has similar PIT values to the representative data.

The summary statistics of PIT (KS, CvM and AD) are shown in Figure 6 for different badness and pivot redshifts parameters. The badness parameter was varied from 0.002 to 0.10 (from best to worst) and the pivot redshift from 0.10 to 2.5 (from worst to best) to evaluate performance. For both degraders, the metrics can be seen to be inconsistent with the non-degraded case (apart from the extreme where no degradation is taking place). All of the representative data points are identical in every degradation scenario; therefore the standard deviation from their scatter was calculated and included as a representative error bar for all the summary statistics values.[12] In the Emission Line Confusion Degrader's case, the metrics get very rapidly poorer as the badness degradation parameter increases, and then flattens off for higher values. In the Inverse Redshift Incompleteness Degrader's case, there is a distinct rise in the values of the summary statistics for the lower pivot redshifts (more extreme degradation).

The CDE loss is shown as a function of badness and pivot redshift in Figure 7 with the standard deviation of the representative values as error bars, hence two different trends are observed. The CDE loss seems to in fact be very similar between the degraded and non-degraded predictions for the Inverse Redshift Incompleteness Degrader case. Conversely, for the Emission Line Confusion Degrader case, the CDE loss is low for the representative data and high for the biased data (even for very low fractions of line confusion).

5.3. Estimated PDF Relative to True PDF

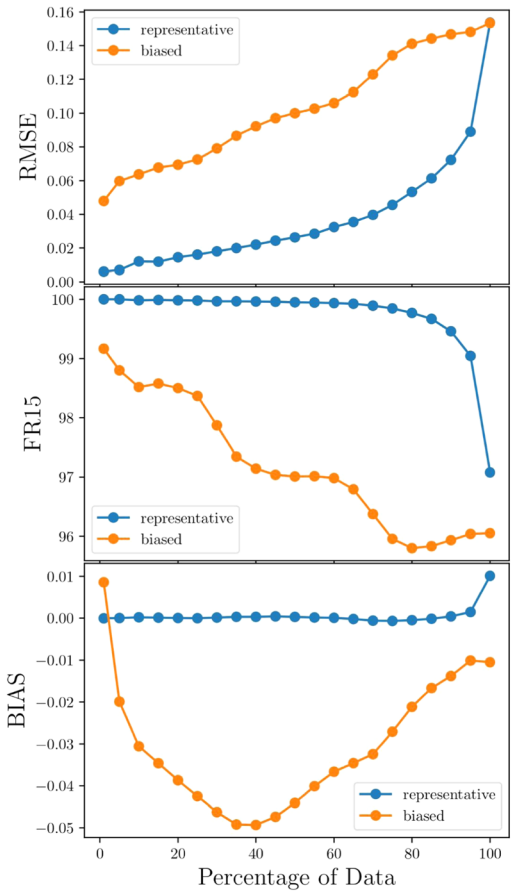
The logarithmic distribution of the KLD values for the two degraders is shown in Figure 8, for badness 0.05 and pivot redshift 0.92 respectively. The representative and bias data in the Inverse Redshift Incompleteness Degrader have comparable KLD distributions. However, in the Emission Line Confusion Degrader case, the representative data have a significant number of their KLD values below zero, unlike their corresponding biased data.

6. Discussion

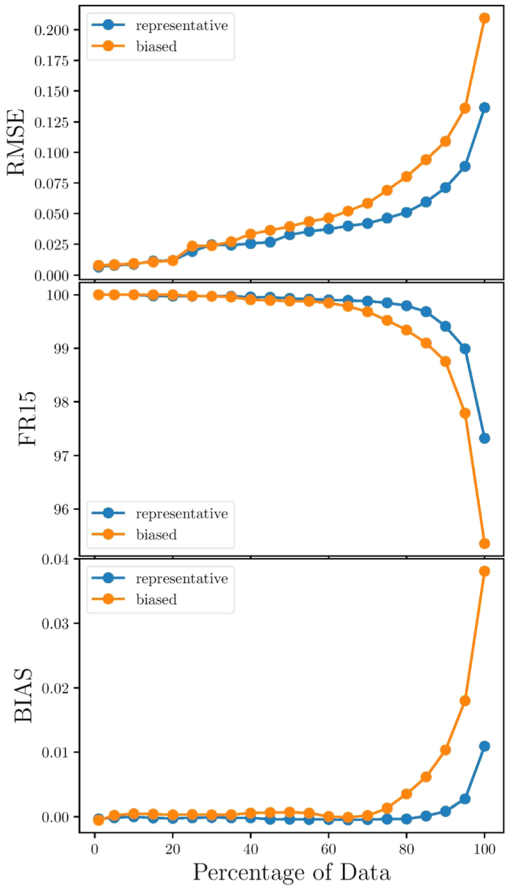
The overall results for the two degraders exhibit similar behavior, in that both of the degraded data show a decline in performance with increasing bias in the training set, while the non-degraded, representative data follow a consistent and generally good performance outcome for all metrics. It is also noted that the degradation of the Emission Line Confusion is generally more extreme than the Inverse Redshift Incompleteness degradation, and hence especially the point estimate plots of the former show greater discrepancies than of the latter.

The overabundance of low PIT values is found to be as expected (see Section 4). The under-representation of high PIT values indicates that GPz is slightly too conservative with the variances (see performance in the DC1 experiment, see Schmidt et al. 2020).

[12] We could have run this whole analysis a large number of times to calculate error bars but this would have been highly computationally expensive.

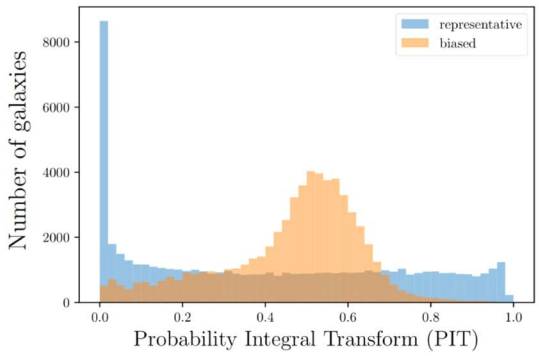


(a) Emission Line Confusion Degrader

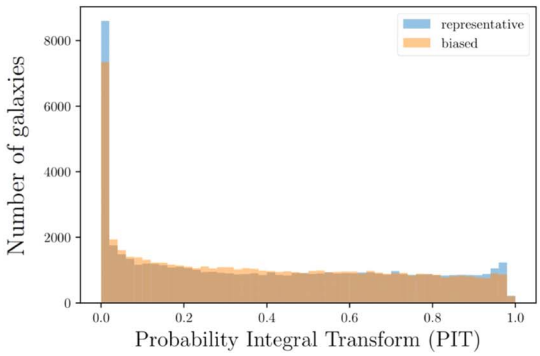


(b) Inverse Redshift Incompleteness Degrader

Figure 4. Plots showing the RMSE, FR15 and BIAS metrics in terms of the Percentage of Data (galaxies ranked by uncertainty on prediction, with 0 being the lowest uncertainty) for the representative and biased samples of the two degraders.



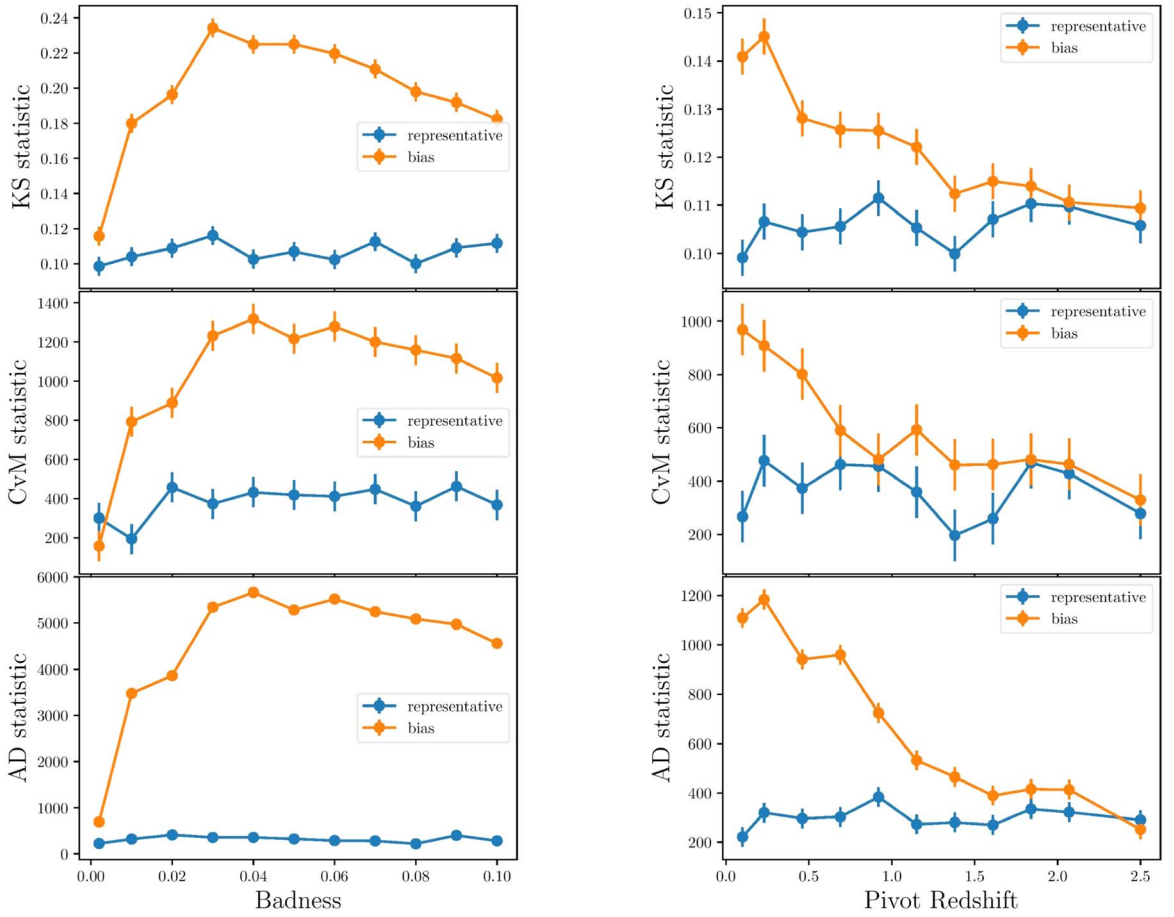
(a) Emission Line Confusion Degrader



(b) Inverse Redshift Incompleteness Degrader

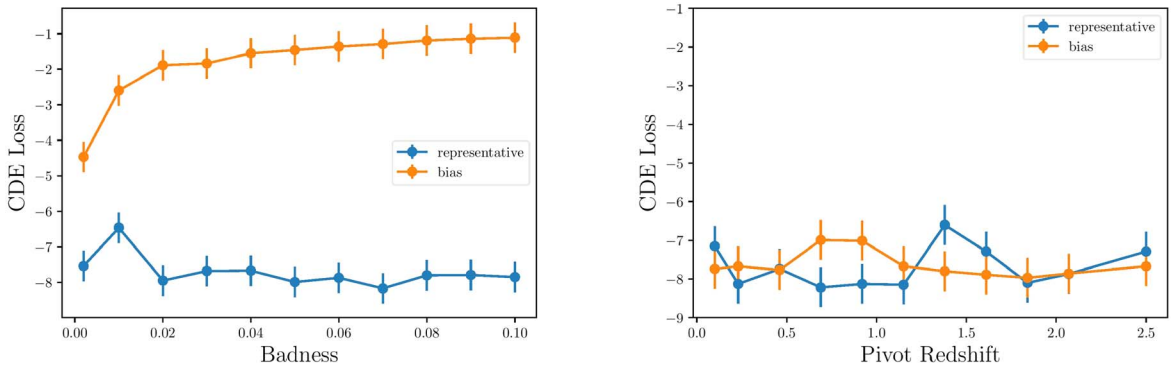
Figure 5. Histograms of PIT values for the representative and biased data sets of the two degraders.

Based on the summary statistics and the CDE loss for the Emission Line Confusion, the metrics get worse quite fast from degradations of 0.2% to about 3% and then seem relatively flat from a degradation of 3% onwards. This shows that there are only modest gains to be found from decreasing spec-z contamination fraction when above $\sim$3%—it appears it must be brought below $\sim$1%–2% for the real improvements to show (at least for this experimental setup).



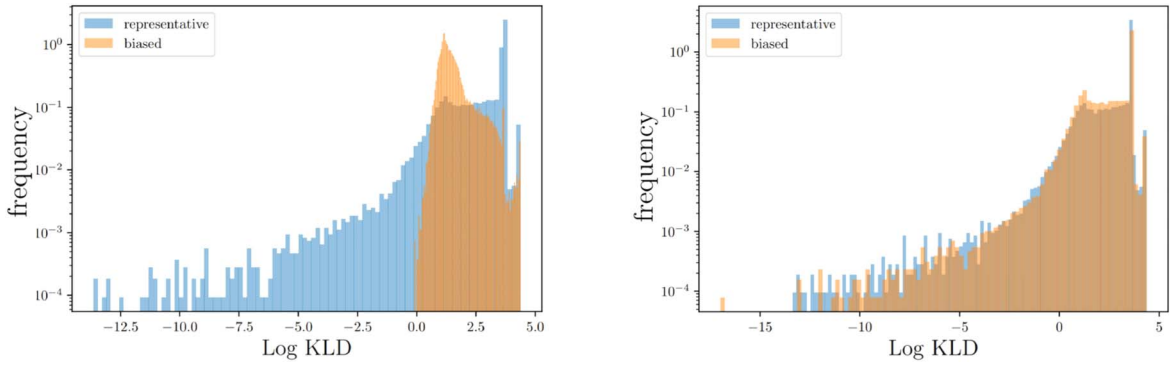
(a) Emission Line Confusion Degrader (b) Inverse Redshift Incompleteness Degrader

Figure 6. Plots showing the PIT Summary Statistics for the two degraders.



(a) Emission Line Confusion Degrader (b) Inverse Redshift Incompleteness Degrader

Figure 7. Plots of the CDE loss against a gradient of degradations corresponding to the two degraders.



(a) Emission Line Confusion Degrader (b) Inverse Redshift Incompleteness Degrader

Figure 8. Plots showing the KLD metric for the two degraders.

We note that for the Inverse Redshift Incompleteness case, we would potentially expect the biased set to have higher CDE loss for lower pivot redshifts, where the bias of the training set is stronger, and lower CDE loss at higher pivot redshift, where the distributions are very similar. This is not observed here—instead, the bias and the representative data show a consistent performance outcome with low CDE loss throughout the training set degradations (although degradation did affect the other metrics). In other words, CDE loss appears not to be affected by redshift incompleteness. Similar behavior is seen for the KLD metric; Emission Line confusion impacts this metric strongly, preventing Log KLD values below 0 from being achieved for any galaxies, whereas for the incompleteness degradation the distribution of KLD values are comparable.

Regarding the performance sensitivity of GPz to the inverse redshift incompleteness, we can say that above a 1.5 pivot redshift the bias data point metrics and summary statistics are generally good for the representative data, as expected, but they get increasingly worse for pivot redshifts below 0.5 for the biased data. This redshift threshold is dependent on the data set used and hence it would likely be different if the original sample had a different redshift and color–magnitude distribution. Nonetheless, for this and any comparable training set distributions, we would caution on sample incompleteness reaching below 0.5 redshift, as the impact on the relevant metrics then starts to become very substantial.

Our results are in agreement with the findings of Cunha et al. (2014), where the impact of incompleteness and incorrect spectroscopic redshifts was investigated using *N*-body-spectro-photometric simulations, although we study a broader range of metrics, including evaluations of PDF quality. They also found redshift incompleteness was potentially not as impactful as the emission line confusion in terms of impact on photo-z estimator performance. Cunha et al. (2014) demonstrated that incorrect redshifts have the most severe impact on the accuracy of photo-z estimators due to their significant degradation on the training set. In particular, they also found that 1% is approximately the tolerable fraction for spectroscopic line confusion (before critically affecting cosmological biases), which is in accordance with our results.

7. Conclusion

To simulate imperfections in spectroscopic redshift training sets for photo-z's estimation we used two degraders to replicate emission line confusion and inverse redshift incompleteness. We compared photo-z based on these biased data sets of increasing degradation with a set of representative data (drawn from the Buzzard Flock synthetic sky catalogs, constructed to be comparable to Rubin-LSST), and calculated a range of metrics that quantified how much poorer degradation of the training data made the resulting photo-z estimates.

It is clear that, broadly, the more biased the data are, and the larger the mismatch between the training and test set is, the worse the metrics and the overall performance of GPz (and likely any other ML-based photo-z estimator) is. Typically, the emission line confusion had a much greater impact on the metrics, with the CDE-loss and KLD metrics, in particular, being only very weakly impacted by redshift incompleteness.

We have shown that for samples comparable to those used in this study, the incompleteness pivot redshift (for a sample spanning approximately $0 < z < 2.3$) should not reach below a redshift of 0.5, as this greatly affects the accuracy. Similarly, we have shown that the emission line confusion fraction may only be worth improving if it can be reduced below 1%–2%, since the decline in the metrics performance is dramatic after that, but relatively flat before.

There are a large number of training set imperfection scenarios, each of which typically will affect photo-z quality to a greater or lesser degree, depending on the exact properties of a specific survey. In this paper we have considered the sensitivity of ML-based photometric redshift estimation under

two training set imperfection scenarios, specifically looking toward the upcoming Rubin-LSST survey. Our results provide an insight into the level of tolerance of training set degradation needed for future large-scale studies, and how badly photo-z predictions can be affected if not mitigated.

NS was supported by a 2021 LSSTC Enabling Science grant, as an "ISSC Ambassador", a scheme to support student researchers working on projects connected to Rubin/LSST science. The program aims to build links between the Informatics and Statistics Science Collaboration (ISSC) and the other Rubin/LSST Science Collaborations (in this case the Dark Energy Collaboration, DESC). This work was built upon a hack-tutorial at "Quarks to Cosmos with AI", a conference supported by the NSF AI Institute: Physics of the Future, NSF PHY-2020295. AIM acknowledges support from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. PH acknowledges generous support from the Hintze Family Charitable Foundation through the Oxford Hintze Centre for Astrophysical Surveys. JFC is also an ISSC Ambassador under a 2021 LSSTC Enabling Science grant, and is supported by the U.S. Department of Energy, Office of Science, under Award DE-SC0011665, as well as the National Science Foundation, Division Of Astronomical Sciences, under Award AST-1715122. This work received software contribution offered from LIneA to DESC via the LSST international in-kind contribution program. The authors thank all developers of the RAIL package, which was key to the development of this project.

Software: pzflow, RAIL, GPz, cdetools, cde-diagnostics (Zhao et al. 2021), qp, astropy (Astropy Collaboration et al. 2013, 2018).

ORCID iDs

Natalia Stylianou https://orcid.org/0000-0002-6029-005X
Alex I. Malz https://orcid.org/0000-0002-8676-1622
Peter Hatfield https://orcid.org/0000-0002-3065-457X
John Franklin Crenshaw https://orcid.org/0000-0002-2495-3514
Julia Gschwend https://orcid.org/0000-0003-3023-8362

References

Almosallam, I. A., Jarvis, M. J., & Roberts, S. J. 2016a, MNRAS, 462, 726
Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., & Roberts, S. J. 2016b, MNRAS, 455, 2387
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Beck, R., Lin, C. A., Ishida, E. E. O., et al. 2017, MNRAS, 468, 4323
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Crenshaw, J. F. 2021, jfcrenshaw/pzflow, v2.0.0, Zenodo, doi:10.5281/zenodo.4679913
Cunha, C. E., Huterer, D., Lin, H., Busha, M. T., & Wechsler, R. H. 2014, MNRAS, 444, 129
Dalmasso, N., Pospisil, T., Lee, A. B., et al. 2020, A&C, 30, 100362
DeRose, J., Wechsler, R. H., Becker, M. R., et al. 2019, arXiv:1901.02401
Duncan, K. J., Jarvis, M. J., Brown, M. J. I., & Röttgering, H. J. A. 2018, MNRAS, 477, 5177
Etherington, J., Thomas, D., Maraston, C., et al. 2017, MNRAS, 466, 228
Euclid Collaboration, Desprez, G., Paltani, S., et al. 2020, A&A, 644, A31
Euclid Collaboration, Ilbert, O., de la Torre, S., et al. 2021, A&A, 647, A117
Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Pascarelle, S. M., & Yahata, N. 2001, ApJS, 135, 41
Fontana, A., D'Odorico, S., Poli, F., et al. 2000, AJ, 120, 2206
Freeman, P. E., Izbicki, R., & Lee, A. B. 2017, MNRAS, 468, 4556
Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. 2019, The American Statistician, 73, 307
Gomes, Z., Jarvis, M. J., Almosallam, I. A., & Roberts, S. J. 2018, MNRAS, 475, 331
González-Casanova, D. F., Lazarian, A., & Cho, J. 2018, MNRAS, 475, 3324
Hatfield, P. W., Almosallam, I. A., Jarvis, M. J., et al. 2020, MNRAS, 498, 5498
Hoyle, B., Gruen, D., Bernstein, G. M., et al. 2018, MNRAS, 478, 592
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Jimenez Rezende, D., & Mohamed, S. 2015, arXiv:1505.05770
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Lewis-beck, M. S., & Skalaban, A. 1990, Political Analysis, 2, 153
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Malz, A. I. 2021, PhRvD, 103, 083502
Malz, A. I., Marshall, P. J., DeRose, J., et al. 2018, AJ, 156, 35
Mitra, A., & Linder, E. V. 2021, PhRvD, 103, 023524
Miyaji, T., Hasinger, G., Salvato, M., et al. 2015, ApJ, 804, 104
Polsterer, K. L., D'Isanto, A., & Gieseke, F. 2016, arXiv:1608.08016
Salvato, M., Ilbert, O., & Hoyle, B. 2019, NatAs, 3, 212
Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, MNRAS, 499, 1587
Shoji, I., Takata, T., & Mizumoto, Y. 2020, MNRAS, 495, 338
Simm, T., Saglia, R., Salvato, M., et al. 2015, A&A, 584, A106
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, PASJ, 70, S9
The Dark Energy Survey Collaboration 2005, arXiv:0510346
Wright, S. 1921, Journal of Agricultural Research, 20, 557
Zhao, D., Dalmasso, N., Izbicki, R., & Lee, A. B. 2021, Uncertainty in Artificial Intelligence, PMLR, arXiv:2102.10473
Zuntz, J., Lanusse, F., Malz, A. I., et al. 2021, OJAp, 4, 13