

Evaluation of probabilistic photometric redshift estimation approaches for LSST

S.J. Schmidt¹, A.I. Malz^{2,3,4}, J.Y.H. Soo^{5,6}, I.A. Almosallam^{7,8}, M. Brescia⁹, S. Cavaudi^{9,10}, J. Cohen-Tanugi¹¹, A.J. Connolly¹², J. DeRose^{13,14,15,16,17}, P.E. Freeman¹⁸, M.L. Graham¹², K.G. Iyer^{19,20}, M.J. Jarvis^{21,22}, J.B. Kalmbach¹², E. Kovacs²³, A.B. Lee¹⁸, G. Longo¹⁰, C.B. Morrison¹², J.A. Newman²⁴, E. Nourbakhsh¹, E. Nuss¹¹, T. Pospisil¹⁸, H. Tranin¹¹, R. Zhou^{15,24}, R. Izbicki^{25,26}

(LSST Dark Energy Science Collaboration)

¹ Department of Physics, University of California, One Shields Ave., Davis, CA, 95616, USA

² German Centre of Cosmological Lensing, Ruhr-Universitaet Bochum, Universitaetsstrasse 150, 44801 Bochum, Germany

³ Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, 10003, USA

⁴ Department of Physics, New York University, 726 Broadway, New York, 10003, USA

⁵ School of Physics, University of Science Malaysia, 11800 USM, Pulau Pinang, Malaysia

⁶ Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

⁷ King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

⁸ Information Engineering, Parks Road, Oxford, OX1 3PJ, UK

⁹ INAF-Astronomical Observatory of Capodimonte, Salita Moiariello 16, 80131 Napoli, Italy

¹⁰ Department of Physics E. Pancini, University Federico II, via Cinthia 6, I-80126, Napoli, Italy

¹¹ Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France

¹² DIRAC Institute and Department of Astronomy, University of Washington, Box 351580, U.W., Seattle WA 98195, USA

¹³ Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA

¹⁴ Berkeley Center for Cosmological Physics, Department of Physics, University of California, Berkeley CA 94720

¹⁵ Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

¹⁶ Kavli Institute for Particle Astrophysics and Cosmology and Department of Physics, Stanford University, Stanford, CA 94305, USA

¹⁷ Department of Particle Physics and Astrophysics, SLAC National Accelerator Laboratory, Stanford, CA 94305, USA

¹⁸ Department of Statistics & Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

¹⁹ Dunlap Institute for Astronomy & Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada

²⁰ Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA

²¹ Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK

²² Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa

²³ Argonne National Laboratory, Lemont, IL 60439, USA

²⁴ Department of Physics and Astronomy and the Pittsburgh Particle Physics, Astrophysics and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA 15260, USA

²⁵ Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil

²⁶ External collaborator

12 December 2019

ABSTRACT

Many scientific investigations of photometric galaxy surveys require redshift estimates, whose uncertainty properties are best encapsulated by photometric redshift ($photo-z$) posterior probability distribution functions (PDFs). A plethora of $photo-z$ PDF estimation methodologies abound, producing discrepant results with no consensus on a preferred approach. We present the results of a comprehensive experiment comparing twelve $photo-z$ algorithms applied to mock data produced for the Large Synoptic Survey Telescope (LSST) Dark Energy Science Collaboration (DESC). By supplying perfect prior information, in the form of the complete template library and a representative training set as inputs to each code, we demonstrate the impact of the assumptions underlying each technique on the output $photo-z$ PDFs. In the absence of a notion of true, unbiased $photo-z$ PDFs, we evaluate and interpret multiple metrics of the ensemble properties of the derived $photo-z$ PDFs as well as traditional reductions to $photo-z$ point estimates. We report systematic biases and overall over/under-breadth of the $photo-z$ PDFs of many popular codes, which may indicate avenues for improvement in the algorithms or implementations. Furthermore, we raise attention to the limitations of established metrics for assessing $photo-z$ PDF accuracy; though we identify the conditional density estimate (CDE) loss as a promising metric of $photo-z$ PDF performance in the case where true redshifts are available but true $photo-z$ PDFs are not, we emphasize the need for science-specific performance metrics.

2 LSST Dark Energy Science Collaboration

1 INTRODUCTION

The current and next generations of large-scale galaxy surveys, including the Dark Energy Survey (DES, Abbott et al. 2005), the Kilo-Degree Survey (KiDS, de Jong et al. 2013), Hyper Suprime-Cam Survey (HSC, Aihara et al. 2018a,b), Large Synoptic Survey Telescope (LSST, Abell et al. 2009), Euclid (Laureijs et al. 2011), and Wide-Field Infrared Survey Telescope (WFIRST, Green et al. 2012), represent a paradigm shift to reliance on photometric, rather than solely spectroscopic, galaxy catalogues of substantially larger size at a cost of lacking complete spectroscopically confirmed redshifts (z). Effective astrophysical inference using the catalogues resulting from these ongoing and upcoming missions, however, necessitates accurate and precise photometric redshift (photo- z) estimation methodologies.

As an example, in order for photo- z systematics to not dominate the statistical noise floor of LSST’s main cosmological sample of several 10^9 galaxies, the LSST Science Requirements Document (SRD)¹ specifies that individual galaxy photo- zs must have root-mean-square error $\sigma_z < 0.02(1+z)$, 3σ catastrophic outlier rate below 10 per cent, and bias below 0.003. Specific science cases may have their own requirements on photo- z performance that exceed those of the survey as a whole. In that vein, the LSST Dark Energy Science Collaboration (LSST-DESC) developed a separate SRD (The LSST Dark Energy Science Collaboration et al. 2018) that conservatively forecasts the constraining power of five cosmological probes, leading to even more stringent requirements on photo- z performance, including those defined in terms of tomographically binned subsample populations² rather than individual galaxies.

Though the standard has long been for each galaxy in a photometric catalogue to have a photo- z point estimate and Gaussian error bar, even early applications of photo- zs in precision cosmology indicate the inadequacy of point estimates (Mandelbaum et al. 2008) to encapsulate the degeneracies resulting from the nontrivial mapping between broad band fluxes and redshift. Far from a hypothetical situation, such degeneracies are real consequences of the same deep imaging that enables larger galaxy catalogue sizes. The lower luminosity and higher redshift populations captured by deeper imaging introduce major physical systematics to photo- zs , among them the Lyman break/Balmer break degeneracy, that did not affect shallower large area surveys like the Sloan Digital Sky Survey (SDSS, York et al. 2000) and Two Micron All Sky Survey (2MASS, Skrutskie et al. 2006).

To fully characterize such physical degeneracies, subsequent photometric galaxy catalogue data releases, (e. g. Sheldon et al. 2012; Erben et al. 2013; de Jong et al. 2017), provide a more informative photo- z data product, the photo- z probability density function (PDF), that describes the redshift probability, commonly denoted as $p(z)$, as a function of a galaxy’s redshift, conditioned on the ob-

served photometry. Early template-based methods such as Fernández-Soto et al. (1999) approximated the likelihood of photometry conditioned on redshift with the relative χ^2 values of template spectra. Not long after, Bayesian adaptations of template-based approaches such as Benítez (2000) combined the estimated likelihoods with a prior to yield a posterior PDF of redshift conditioned on photometry. While the first data-driven photo- z algorithms yielded a point estimate, Firth et al. (2003) estimated a photo- z PDF using a neural net with realizations scattered within the photometric errors.

There are numerous techniques for deriving photo- z PDFs, yet no one method has been established as clearly superior. Consistent experimental conditions enable the quantification if not isolation of their differences, which can be interpreted as a sort of *implicit prior* imparted by the method itself. Comprehensive comparisons of photo- z methods have been made before; the Photo- z Accuracy And Testing (PHAT, Hildebrandt et al. 2010) effort focused on photo- z point estimates derived from many photometric bands. Rau et al. (2015) introduced a new method for improving photo- z PDFs using an ordinal classification algorithm. DES compared several codes for photo- z point estimates and a subset with photo- z PDF information (Sánchez et al. 2014) and examined summary statistics of photo- z PDFs for tomographically binned galaxy subsamples (Bonnett et al. 2016).

This paper is distinguished from other comparisons of photo- z methods by its focus on the evaluation criteria for photo- z PDFs and interpretation thereof. In the absence of simulated data drawn from known redshift distributions, the very concept of a “true PDF” for an individual galaxy is unavailable, and we must instead rely on measures of ensemble behaviour to characterize PDF quality (see § 4 for further discussion). We aim to perform a comprehensive sensitivity analysis of the dependence of photo- z PDFs on the code used to produce them in order to ultimately select those that will become part of the LSST-DESC pipelines, described in the Science Roadmap (SRM)³. In this initial study, we focus on evaluating the performance of photo- z PDF codes using PDF-specific performance metrics in a formally controlled experiment with complete and representative prior information (template libraries and training sets) to set a baseline for subsequent investigations. This approach probes how each code considered exploits the information content of the data versus prior information from template libraries and training sets.

The outline of the paper is as follows: in § 2 we present the simulated data set; in § 3 we describe the current generation codes employed in the paper; in § 4 we discuss the interpretation of photo- z PDFs in terms of metrics of accuracy; in § 5 we show our results and compare the performance of the codes; in § 6 we offer our conclusions and discuss future extensions of this work.

¹ available at <https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17>

² While tomographic samples will play a prominent role in some science cases, estimation of tomographic distributions is a distinct problem with distinct solutions. In this paper we focus solely on individual galaxy redshift estimates.

2 DATA

In order to test the current generation of photo- z PDF codes, we employ an existing simulated galaxy catalogue, described

³ Available at: https://lsstdesc.org/assets/pdf/docs/DESC_SRM_latest.pdf

in detail in Section 2.1. The experimental conditions shared among all codes are motivated by the LSST SRD requirements and implemented for machine learning and template-based photo- z PDF codes according to the procedures of Sections 2.3.1 and 2.3.2 respectively.

2.1 The Buzzard-v1.0 simulation

Our mock catalogue is derived from the BUZZARD-highres-v1.0 catalogue (DeRose et al., in prep). BUZZARD is built on the Chinchilla-400 (Mao et al. 2015) dark matter-only N-body simulation consisting of 2048^3 particles in a 400 Mpc h^{-1} box. The lightcone was constructed from smoothing and interpolation between a set of time snapshots. Dark matter halos were identified using the Rockstar software package (Behroozi et al. 2013) and then populated with galaxies with stellar masses and absolute r -band magnitudes in the SDSS system determined using a sub-halo abundance matching model constrained to match both projected two-point galaxy clustering statistics and an observed conditional stellar mass function (Reddick et al. 2013).

To assign a spectrum to each galaxy, the Adding Density Dependent Spectral Energy Distributions (SEDs) procedure (ADDSEDS, DeRose et al. 2019, Wechsler et al., in prep,) was used. ADDSEDS uses a sample of $\sim 5 \times 10^5$ galaxies from the magnitude-limited SDSS Data Release 6 Value Added Galaxy Catalogue (NYU-VAGC, Blanton et al. 2005) to train an empirical relation between absolute r -band magnitude, local galaxy density, and SED. Each SDSS spectrum is parameterized by five weights corresponding to a weighted sum of five basis SED components using the k-correct software package⁴ (Blanton & Roweis 2007).

Correlations between SED and galaxy environment were included so as to preserve the colour-density relation of galaxy environments. The distance to the spatially projected fifth-nearest neighbour was used as a proxy for local density in the SDSS training sample. For each simulated galaxy, a galaxy with similar absolute r -band magnitude and local galaxy density was chosen from the training set, and that training galaxy's SED was assigned to the simulated galaxy.

2.1.1 Caveats

By necessity, BUZZARD does not contain all of the complicating factors present in real data, and here we discuss the most pertinent ways that these limitations affect our experiment. BUZZARD includes only galaxies, not stars nor AGN. The catalogue-based construction excludes image-level effects, such as deblending errors, photometric measurement issues, contamination from sky background (Zodiacal light, scattered light, etc.), lensing magnification, and Galactic reddening.

The BUZZARD SEDs are drawn from a set of $\sim 5 \times 10^5$ SEDs, which themselves are derived from a five-component linear combination fit to $\sim 5 \times 10^5$ SDSS galaxies; thus the sample contains only galaxies that resemble linear combinations of those for which SDSS obtained spectra, and there are necessarily duplicates. The linear combination SEDs also restrict the properties of the galaxy population to linear

combinations of the properties corresponding to five basis templates, precluding the modeling of non-linear features such as the full range of emission line fluxes relative to the continuum. The only form of intrinsic dust reddening comes from what is already present in the five basis SEDs via the training set used to create the basis templates, and linear combinations thereof do not span the full range of realistic dust extinction observed in galaxy populations.

While these idealized conditions limit the realism of our mock data, they are irrelevant to the controlled experimental conditions of this study, if anything assuring that differentiation in the performance of the photo- z PDF codes is due to the inferential techniques rather than nuances in the data.

2.2 LSST-like mock observations

Given the SED, absolute r -band magnitude, and true redshift of each simulated galaxy, we computed apparent magnitudes in the six LSST filter passbands, $ugrizy$. We assigned magnitude errors in the six bands using the simple model of Ivezić et al. (2008), assuming achievement of the full 10-year depth, with a modification of fiducial LSST total numbers of 30-second visits for photometric error generation: we assume 60 visits in u -band, 80 visits in g -band, 180 visits in r -band, 180 visits in i -band, 160 visits in z -band, and 160 visits in y -band.

As a consequence of adding Gaussian-distributed photometric errors, 2.0 per cent of our galaxies exhibit a negative flux in one or more bands, the vast majority of which are in the u -band. We deem such negative fluxes *non-detections* and assign a placeholder magnitude of 99.0 in the catalogue to indicate to the photo- z PDF codes that such galaxies would be “looked at but not seen” in multi-band forced photometry.

The full dataset thus covers 400 square degrees and contains 238 million galaxies of redshift $0 < z \leq 8.7$ down to $r = 29$. Systematic inconsistencies with galaxy colors at $z > 2$ were observed, so the catalogue was limited to $0 < z \leq 2.0$. To obtain a catalogue matching the LSST Gold Sample, we imposed a cut of $i < 25.3$, which gives a signal-to-noise ratio $\gtrsim 30$ for most galaxies. In order for statistical errors to be subdominant to the systematic errors we aim to probe, we further reduced the sample size to $< 10^7$ galaxies by isolating ~ 16.8 square degrees selected from five separate spatial regions of the simulation. We refer to this final set of galaxies as DC1, for the first LSST-DESC Data Challenge.

2.3 Shared prior information

For the purpose of performing a controlled experiment that compares photo- z PDF codes on equal footing as a baseline for a future sensitivity analysis, we take care to provide each with optimal prior information. Redshift estimation approaches built upon physical modeling and machine learning alike have a notion of prior information considered beyond the photometry of the data for which redshift is to be constrained: that information is derived from a template library for a model-based code and a training set for a data-driven code. In this initial study, we seek to set a baseline for a later comparison of the performance of photo- z PDF codes

⁴ <http://kcorrect.org>

4 LSST Dark Energy Science Collaboration

under incomplete and non-representative prior information that will propagate differently in the space of data-driven and model-based algorithms. However, for the baseline case of perfect prior information, physical modeling and machine learning codes can indeed be put on truly equal footing. We outline the equivalent ways of providing all codes perfect prior information below.

2.3.1 Training and test set division

Following the findings of Bernstein & Huterer (2010), Newman et al. (2015), and Masters et al. (2015) that only $\sim 10^4$ spectra are necessary to calibrate photo-zs to Stage IV requirements, we aimed to set aside a randomly selected training set of $3 - 5 \times 10^4$ galaxies, ~ 10 per cent of the full sample. After all cuts described above, we designated the *DC1 training set* of 44 404 galaxies for which observed photometry, true SEDs, and true redshifts would be provided to all codes and the blinded *DC1 test set* of 399 356 galaxies for which photometry alone would be provided to all codes and photo-z PDFs would be requested. The exact form of LSST photometric filter transmission curves were also considered public information that could be used by any code.

2.3.2 Template library construction

We aimed to provide template-fitting codes with complete yet manageable library of templates spanning the space of SEDs of the DC1 galaxies. We constructed $K = 100$ representative templates from the $\sim 5 \times 10^5$ SEDs of the SDSS DR6 NYU-VAGC by using the five-dimensional vectors of SED weight coefficients described above. After regularizing the SED weight coefficients $\in [0, 1]$, we ran a simple K-means clustering algorithm on the five-dimensional space of regularized SED weight coefficients of the SDSS galaxy sample. The resulting clusters were used to define Voronoi cells in the space of weight coefficients, with centre positions corresponding to weights for the k-correct SED components, yielding the 100 SEDs that comprise the *DC1 template set* provided to all template-based codes. We did not, however, exclude from consideration template-based codes that made modifications in their use of these templates due to architecture limitations (as opposed to knowledge of the experimental conditions that could artificially boost the code's apparent performance), with deviations noted in Section 3.

3 METHODS

Here we summarize the twelve photo-z PDF codes compared in this study, listed in Table 1, which include both established and emerging approaches in template fitting and machine learning. Though not exhaustive, this sample represents codes for which there was sufficient expertise within the LSST-DESC Photometric Redshifts Working Group. Some aspects of data treatment were left to the individual code runners, for example, whether/how to split the available data with known redshifts into separate training and validation sets.

Another key difference is the treatment of non-detections in one or more bands: some codes ignore incomplete bands, while others replace the value with either an

estimate for the detection limit, the mean of other values in the training set, or another default value. There are varying conventions among machine learning-based codes for treatment of non-detections, and no one prescription dominates in the photo-z literature. However, we remind the reader that only 2.0 per cent of our sample has non-detections, almost exclusively in the u -band, and thus should not dominate the code performance differences.

We describe the algorithms and implementations of the model-based and data-driven codes in Sections 3.1 and 3.2 respectively, with a straw-person approach included in Section 3.3.

3.1 Template-based Approaches

We test three publicly available and commonly used template-based codes that share the standard physically motivated approach of calculating model fluxes for a set of template SEDs on a grid of redshift values and evaluating a χ^2 merit function using the observed and model fluxes:

$$\chi^2(z, T, A) = \sum_i^{N_{\text{filt}}} \left(\frac{F_{\text{obs}}^i - A F_{\text{pred}}^i(T, z)}{\sigma_{\text{obs}}^i} \right)^2 \quad (1)$$

where A is a normalization factor, $F_{\text{pred}}^i(T, z)$ is the flux predicted for a template T at redshift z , F_{obs}^i is the observed flux in a given band i , σ_{obs}^i is the observed flux error, and N_{filt} is the total number of filters, in our case the six *ugrizy* LSST filters. The likelihood is a sum of observed flux error σ_b^{obs} -weighted squared differences between the observed flux F_b^{obs} and the normalized predicted flux $F_b^{\text{mod}}(T, z)$ in N_{filt} photometric filters b , which are the LSST *ugrizy* filters in this case. Specific implementation details of each code, e. g. prior form and implementation, are described below.

3.1.1 BPZ

Bayesian Photometric Redshift (BPZ, Benítez 2000) determines the likelihood $p(C|z, T)$ of a galaxy's observed colours C for a set of SED templates T at redshifts z . The BPZ likelihood is related to the χ^2 likelihood by $p(C|z, T) \propto \exp[-\chi^2/2]$. Given a Bayesian prior $p(z, T|m_0)$ over apparent magnitude m_0 and type T , and assuming that the SED templates are spanning and exclusive, BPZ constructs the redshift posterior $p(z|C, m_0)$ by marginalizing over all SED templates with the form $p(z|C, m_0) \propto \sum_T p(C|z, T) p(z, T|m_0)$ (Eq. 3 from Benítez 2000), corresponding to setting the parameter PROBS_LITE=TRUE in the BPZ parameter file. The BPZ prior is the product of an SED template proportion that varies with apparent magnitude $p(T|m_0)$ and a prior $p(z|T, m_0)$ over the expected redshift as a function of apparent magnitude and SED template. We anticipate BPZ to outperform other template-based approaches due to the prior that both comprehensively accounts for SED type and is calibrated to the training set.

Here we test BPZ-v 1.99.3 (Benítez 2000) with the DC1 template set of Section 2.3.2. To keep the number of free parameters manageable, the DC1 template set is pre-sorted by the rest-frame $u - g$ colour and split into three broad classes of SED template, equivalent to the E, Sp and Im/SB types. The Bayesian prior term $p(T|m_0)$ was derived directly from the DC1 training set, and the other term $p(z|T, m_0)$ was

Table 1. List of photo- z PDF codes featured in this study

Published code	Type	Public source code
BPZ (Benítez 2000)	template fitting	http://www.stsci.edu/~dcoe/BPZ/
EAZY (Brammer et al. 2008)	template fitting	https://github.com/gbrammer/eazy-photoz
LePhare (Arnouts et al. 1999)	template fitting	http://www.cfht.hawaii.edu/~arnouts/lephare.html
ANNz2 (Sadeh et al. 2016)	machine learning	https://github.com/IftachSadeh/ANNz2
CMNN (Graham et al. 2018)	machine learning	https://github.com/dirac-institute/CMNN_Photoz_Estimator
Delight (Leistedt & Hogg 2017)	hybrid	https://github.com/ixkael/Delight
FlexZBoost (Izbicki & Lee 2017)	machine learning	https://github.com/tpospisil/flexcode ; https://github.com/rizbicki/FlexCoDE
GP z (Almosallam et al. 2016b)	machine learning	https://github.com/OxfordML/GPz
METAPhoR (Cavuoti et al. 2017)	machine learning	http://dame.dsfs.unina.it
SkyNet (Graff et al. 2014)	machine learning	http://ccforge.cse.rl.ac.uk/gf/project/skynet/
TPZ (Carrasco Kind & Brunner 2013)	machine learning	https://github.com/mgckind/MLZ
trainZ	–	See Section 3.3

chosen to be the best fit for the eleven free parameters from the functional form of Benítez (2000). We use template interpolation, creating two linearly interpolated templates between each basis SED (sorted by rest-frame $u - g$ colour) by setting the parameter `INTERP=2`. Prior to running the code, the non-detection placeholder magnitude was replaced with an estimate of the $1-\sigma$ detection limit for the undetected band as a proxy for a value close to the estimated sky noise threshold.

to match observed colours with those predicted from a template set. The template set can be semi-empirical or entirely synthetic. The reported photo- z PDF is an arbitrary normalization of the likelihood evaluated on the output redshift grid.

Here we use LePhare-v 2.2 with the DC1 template set of Section 2.3.2. Unlike both BPZ and EAZY, LePhare uses generic, SED-independent priors that are not tuned to the DC1 data set.

3.1.2 EAZY

Easy and Accurate Photometric Redshifts from Yale (EAZY, Brammer et al. 2008) extends the basic χ^2 fit procedure that defines template-fitting approaches. The algorithm models the observed photometry with a linear combination of template SEDs at each redshift. The best-fit SED at each redshift is found by simultaneously fitting one, two, or all of the templates via χ^2 minimization, which is distinct from marginalizing across all templates. The minimized χ^2 likelihood at each redshift is then combined with an apparent magnitude prior to obtain the redshift posterior PDF. We note that the utilization of the best-fit SED conditioned on redshift rather than a proper marginalization does not lead to the correct posterior distribution, an implementation issue that has now been identified and will be addressed by the developers in the future.

In contrast with BPZ, EAZY’s apparent magnitude prior is independent of SED, though it was derived empirically from the DC1 training set. The EAZY architecture cannot accept a template set other than the same five basis templates employed by `k-correct` when constructing the DC1 catalogue, but, for consistency with the experimental scope of perfect prior information, EAZY’s flexible `all-templates` mode was used to fit the photometric data with a linear combination of the five basis templates. Though EAZY can account for uncertainty in the template set by adding in quadrature to the flux errors an empirically derived template error as a function of redshift, we set the template error to zero since the same templates were in fact used to produce the DC1 photometry.

3.1.3 LePhare

Photometric Analysis for Redshift Estimate (LePhare, Arnouts et al. 1999; Ilbert et al. 2006) uses the χ^2 of Equa-

3.2 Machine Learning-based Approaches

We compared nine data-driven photo- z estimation approaches, eight of which are described in this section and one of which is discussed in Section 3.3. Because the algorithms differ more from one another and the techniques are relative newcomers to the astronomical literature, we provide somewhat more detail about the implementations below.

3.2.1 ANNz2

ANNz2(Sadeh et al. 2016) supports several machine learning algorithms, including artificial neural networks (ANN), boosted decision tree, and k-nearest neighbour (KNN) regression. In addition to accounting for errors on the input photometry, ANNz2 uses the KNN-uncertainty estimate of Oyaizu et al. (2008) to quantify uncertainty in the choice of method over multiple runs. Using the Toolkit for Multivariate Data Analysis with ROOT⁵, ANNz2 can return the results of running a single machine learning algorithm, a “best” choice of the results from simultaneously running multiple algorithms (based on evaluation the cumulative distribution functions of validation set objects), or a combination of the results of multiple algorithms weighted by their method uncertainties averaged over multiple runs.

In this study, we used ANNz2-v.2.0.4 to output only the result of the ANN algorithm. Photo- z PDFs were produced by running an ensemble of 5 ANNs with a 6 : 12 : 12 : 1 architecture corresponding to the 6 $ugrizy$ inputs, 2 hidden layers with 12 nodes each, and 1 output of redshift. Each of the five ANNs was trained with different random seeds for the initialization of input parameters, reserving half of the training set for validation to prevent overfitting. Undetected

⁵ <http://tmva.sourceforge.net/>

6 LSST Dark Energy Science Collaboration

414 galaxies were excluded from the training set, and per-band
 415 non-detections in the test set were replaced with the mean
 416 magnitude in that band within the entire test set.

417 3.2.2 Colour-Matched Nearest-Neighbours

418 The colour-matched nearest-neighbours photometric red-
 419 shift estimator (CMNN, [Graham et al. 2018](#)) uses a training
 420 set of galaxies with known redshifts that has equivalent or
 421 better photometry than the test set in terms of quality and
 422 filter coverage. For each galaxy in the test set, CMNN identifies
 423 a colour-matched subset of training galaxies using a thresh-
 424 old in the Mahalanobis distance $D_M = \sum_j^{N_{\text{colours}}} (c_j^{\text{train}} -$
 425 $c_j^{\text{test}})^2 / \delta c_{\text{test}}^2$ in the space of available colours c , with colour
 426 measurement errors δc_{test} and $N_{\text{colours}} = 5$ colours j defined
 427 by the *ugrizy* filters, which defines the set of colour-matched
 428 neighbours based on a value of the percent point function
 429 (PPF). As an example, for $N_{\text{filt}} = 5$ with PPF = 0.95, 95 per
 430 cent of all training galaxies consistent with the test galaxy
 431 will have $D_M < 11.07$. Undetected bands are dropped,
 432 thereby reducing the effective N_{filt} for that galaxy. The
 433 photo- z PDF of a given test set galaxy is the normalized dis-
 434 tribution of redshifts of its colour-matched subset of training
 435 set galaxies.

436 Here, we make two modifications to the implementation
 437 of [Graham et al. \(2018\)](#) to comply with the controlled exper-
 438 imental conditions. First, we do not impose non-detections
 439 on galaxies fainter than the expected LSST 10-year limit-
 440 ing magnitude nor galaxies bright enough to saturate with
 441 LSST’s CCDs, instead using all of the photometry for the
 442 DC1 test and training sets. Second, we apply the initial
 443 colour cut to the training set before calculating the Ma-
 444 halanobis distance in order to accelerate processing and use a
 445 magnitude pseudo-prior as in [Graham et al. \(2018\)](#), but for
 446 both we use cut-off values corresponding to the DC1 training
 447 set galaxies’ colours and magnitudes.

448 We make an additional adaptation to enable the CMNN
 449 algorithm to yield accurate photo- z PDFs for all galaxies,
 450 as the original [Graham et al. \(2018\)](#) algorithm is optimized
 451 for photo- z point estimates and is susceptible to less acc-
 452 curate photo- z PDFs for bright galaxies or those with few
 453 matches in colour-space. We use PPF = 0.95 rather than
 454 PPF = 0.68 to generate the subset of colour-matched train-
 455 ing galaxies, whose redshifts are weighted by their inverse
 456 Mahalanobis distances when composing the photo- z PDF
 457 rather than weighting all colour-matched training galaxies
 458 equally. Additionally, when the number of colour-matched
 459 training set galaxies is less than 20, the nearest 20 neigh-
 460 bours in colour-space are used instead, and the output
 461 photo- z PDF is convolved with a Gaussian kernel of vari-
 462 ance $\sigma_{\text{train}}^2 (\text{PPF}_{20}/0.95)^2 - 1$ to account for the correspond-
 463 ing growth of the effective PPF to include 20 neighbours.

464 3.2.3 Delight

465 Delight ([Leistedt & Hogg 2017](#)) is a hybrid technique that
 466 infers photo- z s with a data-driven model of latent SEDs and
 467 a physical model of photometric fluxes as a function of red-
 468 shift. Generally, machine learning methods rely on represen-
 469 tative training data with shared photometric filters, while

470 template-based methods rely on a complete library of tem-
 471 plates based on physical models constructed. Delight aims
 472 to take the best aspects of both approaches by construct-
 473 ing a large collection of latent SED templates (or physical
 474 flux-redshift models) from training data, with a template
 475 SED library as a guide to the learning of the model, thereby
 476 circumventing the machine learning prerequisite of represen-
 477 tative training data in the same photometric bands and the
 478 template fitting requirement of detailed galaxy SED models.
 479 It models noisy observed flux $\hat{\mathbf{F}} = \mathbf{F} + F_b$ as a sum of a noise-
 480 less flux plus a Gaussian processes $F_b \sim \mathcal{GP}(\mu^F, k^F)$ with
 481 zero mean function μ^F and a physically motivated kernel k^F
 482 that induces realistic correlations in flux-redshift space.

483 From a template-fitting perspective, each test set galaxy
 484 has a posterior $p(z|\hat{\mathbf{F}}) \approx \sum_i p(\hat{\mathbf{F}}|z, T_i)p(z|T_i)p(T_i)$ of red-
 485 shift z conditioned on noisy flux $\hat{\mathbf{F}}$, where $p(z|T_i)p(T_i)$ cap-
 486 tures prior information about the redshift distributions and
 487 abundances of the galaxy templates T_i . As in traditional
 488 template fitting, each likelihood $p(\hat{\mathbf{F}}|\mathbf{F})$ relates the noisy flux
 489 $\hat{\mathbf{F}}$ with the noiseless flux \mathbf{F} predicted by the model of a linear
 490 combination of templates, carefully constructed to account
 491 for model uncertainties and different normalization of the
 492 same SED, plus the Gaussian process term.

493 The machine learning approach appears in the inclu-
 494 sion of a pairwise comparison term $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$ for the
 495 prediction of model flux \mathbf{F} at a model redshift z with re-
 496 spect to training set galaxy j with redshift z_j and ob-
 497 served flux $\hat{\mathbf{F}}_j$. Thus the photo- z posterior $p(\hat{\mathbf{F}}|z, T_i) =$
 498 $\int p(\hat{\mathbf{F}}|\mathbf{F})p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)d\mathbf{F}$ may be interpreted as the proba-
 499 bility that the training and the target galaxies have the same
 500 SED at different redshifts. The flux prediction $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$
 501 of the training galaxy at redshift z is modeled via the Gau-
 502 ssian process described above; more detail is provided in
 503 [Leistedt & Hogg \(2017\)](#).

504 In this study, the default settings of Delight were used,
 505 with the exception that the PDF bins were set to be linearly
 506 spaced rather than logarithmically. The Gaussian process
 507 was trained using the full DC1 training set. We used the
 508 full DC1 template set with a flat prior in magnitude and
 509 SED type. Photometric uncertainties from the inputs are
 510 propagated into the code, while non-detections for each band
 511 are set to the mean of the respective bands.

512 3.2.4 FlexZBoost

513 FlexZBoost ([Izbicki & Lee 2017](#); [Dalmasso et al. 2019](#)) is
 514 built on FlexCode, a general-purpose methodology for con-
 515 verting any conditional mean point estimator of z to a condi-
 516 tional density estimator $p(z|\mathbf{x}) \equiv f(z|\mathbf{x})$, where \mathbf{x} here rep-
 517 presents our photometric covariates and errors. FlexZBoost
 518 expands the unknown function $f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x})\phi_i(z)$ us-
 519 ing an orthonormal basis $\{\phi_i(z)\}_i$. By the orthogonality
 520 property, the expansion coefficients $\beta_i(\mathbf{x}) = \mathbb{E}[\phi_i(z)|\mathbf{x}] \equiv$
 521 $\int f(z|\mathbf{x})\phi_i(z)dz$ are thus conditional means. The expecta-
 522 tion value $\mathbb{E}[\phi_i(z)|\mathbf{x}]$ of the expansion coefficients condi-
 523 tioned on the data is equivalent to the regression of the
 524 space of possible redshifts on the space of possible photom-
 525 etry. Thus the expansion coefficients $\beta_i(\mathbf{x})$ can be estimated
 526 from the data via regression to yield the conditional density
 527 estimate $\hat{f}(z|\mathbf{x})$.

528 In this paper, we used `xgboost` ([Chen & Guestrin](#)

529) for the regression; it should, however, be noted that
 530) **FlexCode-RF**⁶, based on Random Forests, generally per-
 531) forms better on smaller datasets. As our basis $\phi_i(z)$, we
 532) choose a standard Fourier basis. The two tuning parame-
 533) ters in our photo- z PDF estimate are the number I of terms
 534) in the series expansion and an exponent α that we use to
 535) sharpen the computed density estimates $\tilde{f}(z|\mathbf{x}) \propto \hat{f}(z|\mathbf{x})^\alpha$.
 536) Both I and α were chosen in an automated way by mini-
 537) mizing the weighted L_2 -loss function (Eq. 5 in Izbicki & Lee
 538) 2017) on a validation set comprised of a randomly selected
 539) 15 per cent of the DC1 training set. While **FlexCode**'s loss-
 540) less native encoding stores each photo- z PDF using the basis
 541) coefficients $\beta_i(\mathbf{x})$, we discretized the final estimates into 200
 542) linearly spaced redshift bins $0 < z < 2$ to match the consis-
 543) tent output format of the experimental conditions.

544) 3.2.5 GPz

545) **GPz**(Almosallam et al. 2016a,b) is a sparse Gaussian process-
 546) based code, a scalable approximation of full Gaussian
 547) Processes (Rasmussen & Williams 2006), that produces
 548) input-dependent variance estimates corresponding to het-
 549) eroscedastic noise. The model assumes a Gaussian poste-
 550) rior probability $p(z|\mathbf{x}) = \mathcal{N}(z|\mu(\mathbf{x}), \sigma(\mathbf{x})^2)$ of the output
 551) redshift z given the input photometry \mathbf{x} . The mean $\mu(\mathbf{x})$
 552) and the variance $\sigma(\mathbf{x})^2$ are modeled as functions $f(\mathbf{x}) =$
 553) $\sum_{i=1}^m w_i \phi_i(\mathbf{x})$ that are linear combinations of m basis func-
 554) tions $\{\phi_i(\mathbf{x})\}_{i=1}^m$ with associated weights $\{w_i\}_{i=1}^m$. The de-
 555) tails on how to learn the parameters of the model and the
 556) hyper-parameters of the basis functions are described in Al-
 557) mosallam et al. (2016b). **GPz**'s variance estimate is composed
 558) of a model uncertainty term corresponding to sparsity of the
 559) training set photometry and a noise uncertainty term en-
 560) compassing noisy photometric observations, enabling quan-
 561) tification of any need for more representative or more pre-
 562) cise training samples. **GPz** may also weight training set sam-
 563) ples by importance according to $|z_{\text{spec}} - z_{\text{phot}}| / (1 + z_{\text{spec}})$ to
 564) minimize the normalized photo- z point estimate error. How-
 565) ever, this function may be adapted to photo- z PDFs, adding
 566) weight to test set galaxies that are not well-represented in
 567) the training set.

568) To smooth the long tail in the distribution of magnitude
 569) errors, we use the logarithm of the magnitude errors, im-
 570) proving numerical stability and eliminating the need for
 571) constraints on the optimization process. Unobserved mag-
 572) nitudes $x_u = \mu_u + \Sigma_{uo} \Sigma_{oo}^{-1} (x_o - \mu_o)$ were imputed from
 573) observed magnitudes x_o and the training set mean μ and
 574) covariance Σ using a linear model. This is the optimal ex-
 575) pected value of the unobserved variables given the observed
 576) ones under the assumption that the distribution is jointly
 577) Gaussian; note that this reduces to a simple average if the
 578) covariates are independent with $\Sigma_{uo} = 0$. We reserved for
 579) validation 20 per cent of the training set and used the Vari-
 580) able Covariance option in **GPz** with 200 basis functions (see
 581) Almosallam et al. (2016b) for details), and did not apply
 582) cost-sensitive learning options.

⁶ <https://github.com/tppospisi/flexcode>;
<https://github.com/rizbicki/FlexCoDE>

583) 3.2.6 METAPhoR

584) Machine-learning Estimation Tool for Accurate Photomet-
 585) ric Redshifts (**METAPhoR**, Cavuoti et al. 2017) is based on
 586) the Multi Layer Perceptron with Quasi Newton Algorithm
 587) (MLPQNA) with the least square error model and Tikhonov
 588) L_2 -norm regularization (Hofmann & Mathé 2018). Photo- z
 589) PDFs are generated by running N trainings on the same
 590) training set, or M trainings on M different random sam-
 591) plings of the training set. Upon regression of the test set,
 592) the photometry m_{ij} of each test set galaxy j in filter i is
 593) perturbed according to $m'_{ij} = m_{ij} + \alpha_i F_{ij} \epsilon$ in terms of
 594) the standard normal random variable $\epsilon \sim \mathcal{N}(0, 1)$, a mul-
 595) tiplicative constant α_i permitting accommodation of multi-
 596) survey photometry, and a bimodal function F_{ij} composed of
 597) a polynomial fit of the mean magnitude errors on the binned
 598) bands plus a constant term representing the threshold be-
 599) low which the polynomial's noise contribution is negligible
 600) (Brescia et al. 2018).

601) In this work, we used a hierarchical KNN to replace
 602) non-detections with values based on their neighbours. The
 603) usual cross-validation of redshift estimates and PDFs was
 604) also omitted for this study.

605) 3.2.7 SkyNet

606) **SkyNet**(Graff et al. 2014) employs a neural network based on
 607) a second-order conjugate gradient optimization scheme (see
 608) Graff et al. 2014, for further details). The neural network
 609) is configured as a standard multilayer perceptron with three
 610) hidden layers and one input layer with 12 nodes correspond-
 611) ing to the 6 photometric magnitudes and their measurement
 612) errors.

613) **SkyNet**'s classifier mode uses a cross-entropy error func-
 614) tion with a 20:40:40 node (all rectified linear units) architec-
 615) ture for each hidden layer and an output layer of 200 nodes
 616) corresponding to 200 bins for the PDF, with a softmax activa-
 617) tion function to enforce the normalization condition that
 618) the probabilities sum to unity. While previous implemen-
 619) tations of the code (see Appendix C.3 of Sánchez et al. 2014;
 620) Bonnett 2015) implement a sliding bin smoothing, no such
 621) procedure was used in this study.

622) We pre-whitened the data by pegging the magnitudes
 623) to (45,45,40,35,42,42) and errors to (20,20,10,5,15,15) for
 624) *ugrizy* filters, respectively. To avoid over-fitting, 30 per cent
 625) of the training set was reserved for validation, and training
 626) was halted as soon as the error rate began to increase on the
 627) validation set. The weights were randomly initialized based
 628) on normal sampling.

629) 3.2.8 TPZ

630) Trees for Photo- z (**TPZ**, Carrasco Kind & Brunner 2013; Car-
 631) rrasco Kind & Brunner 2014) uses prediction trees and ran-
 632) dom forest techniques to estimate photo- z PDFs. **TPZ** re-
 633) cursively splits the training set into branch pairs based on
 634) maximizing information gain among a random subsample of
 635) features, to minimize correlation between the trees, termi-
 636) nating only when a newly created leaf meets a criterion, such
 637) as a leaf size minimum or a variance threshold. The regions
 638) in each terminal leaf node correspond to a subsample of the
 639) training set with similar properties. Bootstrap samples from

8 LSST Dark Energy Science Collaboration

the training set photometry and errors are used to build a set of prediction trees.

To run TPZ, we replaced non-detections with an approximation of the 1σ detection threshold based on the signal-to-noise-based error forecast of the 10-year LSST data, i. e. $dm = 2.5 \log(1 + N/S)$ where $dm \sim 0.7526$ magnitudes for $N/S = 1$ (where N and S are the noise and signal). We calibrated TPZ with the Out-of-Bag cross-validation technique (Breiman et al. 1984; Carrasco Kind & Brunner 2013) to evaluate its predictive validity and determine the relative importance of the different input attributes. We grew 100 trees to a minimum leaf size of 5 using the *ugri* magnitudes, all $u - g, g - r, r - i, i - z, z - y$ colours, and the associated errors, as the z and y magnitudes did not show significant correlation with the redshift in our cross-validation. We partitioned our redshift space into 200 bins and smoothed each individual PDF with a smoothing scale of twice the bin size.

3.3 trainZ: a pathological photo-z PDF estimator

We also consider a pathological photo-z PDF estimation method, dubbed **trainZ**, which assigns each test set galaxy a photo-z PDF equal to the normalized redshift distribution $N(z)$ of the training set, according to

$$p(z|\{z_j\}) \equiv \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \begin{cases} 1 & \text{if } z_k \leq z_i < z_{k+1} \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

Unlike the other methods, the **trainZ** estimator is *independent of the photometric data*, effectively performing a KNN procedure with $k = N_{\text{train}}$, a limiting case of a photo-z PDF estimator dominated by the shared prior information of the training set. In this way, **trainZ** serves as an experimental control that is not a competitive photo-z PDF method that would be used by any real survey.

Though **trainZ** is strongly vulnerable to a nonrepresentative training set, it should optimize performance metrics probing the ensemble properties of the galaxy sample, modulo Poisson error due to small sample size, as the training set and test set are drawn from the same underlying population. We will demonstrate its performance under the metrics of Section 4 and discuss it as an illustrative experimental control case in Section 6.1 to highlight the limitations of our evaluation criteria for photo-z PDFs.

4 ANALYSIS

The goal of this study is to evaluate the degree to which photo-z PDFs of each method can be trusted for a generic analysis. The overloaded “ $p(z)$ ” is a widespread abuse of notation that obfuscates this goal, so we dedicate attention to dismantling it here.

Galaxies have redshifts z and photometric data d drawn from a joint probability space $p(z, d)$ in nature, and each observed galaxy i has a *true posterior photo-z PDF* $p(z|d_i)$. There are a number of metrics that can be used to test the accuracy of a photo-z posterior as an estimator of a true photo-z posterior if the true photo-z PDF is known. However, the true photo-z PDF of observed data is not accessible, and existing mock catalogues produce redshift-photometry pairs (z, d) by a deterministic algorithm that

does not correspond to a joint probability density from which one can take samples. In these cases there is no “true PDF” for an individual object, and most measures of PDF fidelity will necessarily be restricted to probing the quality of the ensemble of photo-z PDFs. (See §6.2 for a discussion of how one might circumvent this limitation.)

Before describing the metrics appropriate to the DC1 data set, we outline the philosophy behind our choices. A photo-z PDF estimator derived by method H must be understood as a posterior probability distribution

$$\hat{p}_i^H(z) \equiv p(z|d_i, I_D, I_H), \quad (3)$$

conditioned not only on the photometric data d_i for that galaxy but also on parameters encompassing prior information I_D shared, in our experiment, among all photo-z PDF codes and I_H that will differ depending on the method H used to produce it. To be concrete, I_D takes the form of a training set for the machine learning codes and a template library for the model fitting codes.

The interpretation of the information I_H is more subtle. This investigation is built upon the knowledge that two codes taking the same approach, among choices of model fitting or machine learning, are nonetheless expected to yield different results even if they take the same external prior information I_D . I_H represents the projection of the code’s architecture onto the estimated posteriors over redshift, specific to each code, and even the tunable parameters or random seeds of a specific run of a code with a random component. We refer to I_H as the *implicit prior*, in contrast with the training set or template library provided to a given code explicitly by the researcher. In simple terms, the implicit prior is the collection of the many different assumptions, coding choices, algorithm selections, and other implementation details that are specific to each code, the ensemble of which results in differing estimates of redshift when combined with the data and prior information in common to all codes.

The presence of the implicit prior in some sense makes a direct comparison of photo-z PDFs produced by different methods impossible; even if they share the same external prior information I_D , by definition they cannot be conditioned on the same assumptions I_H , otherwise they would not be distinct methods at all. In this study, we isolate the effect of differences in prior information I_H specific to each method by using a single training set I_D^{ML} for all machine learning-based codes and a single template library I_D^T for all template-based codes. These sets of prior information are carefully constructed to be representative and complete, so we have $I_D \equiv I_D^{\text{ML}} \equiv I_D^T$ for every method H . Under this assumption, a ratio of posteriors of codes is in effect a ratio of the implicit posteriors $p(z|d_i, I_H')$ since the external prior information I_D is present in the numerator and denominator. Thus comparisons of $\hat{p}_i^H(z)$ isolate the effect of the method used to obtain the estimator, which should enable interpretation of the differences between estimated PDFs in terms of the specifics of the method implementations.

The exact implementation of the metrics theoretically depends on the parametrization of the photo-z PDFs, which may differ across codes and can affect the precision of the estimator (Malz et al. 2018). Even considering a single method under the same parametrization, such as the 200-bin $0 < z < 2$ piecewise constant function used here, the exact

bin definitions must affect the result. The piecewise constant format is chosen because of its established presence in the literature, and the choice of 200 bins was motivated by the approximate number of columns expected to be available for storage of photo-z PDFs for the final LSST Project tables.⁷ We will discuss the choice of photo-z PDF parameterization further in Section 6.

This analysis is conducted using the `qp`⁸ software package (Malz & Marshall 2018) for manipulating and calculating metrics of univariate PDFs. We present the metrics of photo-z PDFs that address our goals in the sections below. Section 4.1 outlines aggregate metrics of a catalogue of photo-z PDFs, and Section 4.2 presents a metric of individual photo-z PDFs in the absence of true photo-z PDFs. Those seeking a connection to previous comparison studies will find metrics of redshift point estimate reductions of photo-z PDFs in Appendix B and metrics of a science-specific summary statistic heuristically derived from photo-z PDFs in Appendix A.

4.1 Metrics of photo-z PDF ensembles

Because LSST’s photo-z PDFs will be used for many scientific applications, some of which require each individual catalogue entry to be accurate, we consider several metrics that probe the population-level performance of the photo-z PDFs. As we have the true redshifts but not true photo-z PDFs for comparison, we remind the reader of the Cumulative Distribution Function (CDF)

$$\text{CDF}[f, q] \equiv \int_{-\infty}^q f(z) dz, \quad (4)$$

of a generic univariate PDF $f(z)$, which is used as the basis for several of our metrics. We describe metrics based on the CDF in Section 4.1.1 and metrics of summary statistics thereof in Section 4.1.2.

4.1.1 CDF-based metrics

A quantile of a distribution is the value q at which the CDF of the distribution is equal to Q ; percentiles and quartiles are familiar examples of linearly spaced sets of 100 and 4 quantiles, respectively. The quantile-quantile (QQ) plot serves as a graphical visualization for comparing two distributions, where the quantiles of one distribution are plotted against the quantiles of the other distribution, providing an intuitive way to qualitatively assess the consistency between an estimated distribution and a true distribution. The closer the QQ plot is to diagonal, the closer the match between the distributions.

The probability integral transform (PIT)

$$\text{PIT} \equiv \text{CDF}[\hat{p}, z_{true}] \quad (5)$$

is the CDF of a photo-z PDF evaluated at its true redshift, and the distribution of PIT values probes the average accuracy of the photo-z PDFs of an ensemble of galaxies. The distribution of PIT values is effectively the derivative of the

QQ plot. A catalogue of accurate photo-z PDFs should have a PIT distribution that is uniform $U(0, 1)$, and deviations from flatness are interpretable: overly broad photo-z PDFs induce underrepresentation of the lowest and highest PIT values, whereas overly narrow photo-z PDFs induce overrepresentation of the lowest and highest PIT values. Catastrophic outliers with a true redshift outside the support of its photo-z PDF have $\text{PIT} \approx 0$ or $\text{PIT} \approx 1$.

The PIT distribution has been used to quantify the performance of photo-z PDF methods in the past (e. g. Bordoloi et al. 2010; Polsterer et al. 2016; Tanaka et al. 2018). Tanaka et al. (2018) use the histogram of PIT values as a diagnostic indicator of overall code performance, while Freeman et al. (2017) independently define the PIT and demonstrate how its individual values may be used both to perform hypothesis testing (via, e. g. the KS, CvM, and AD tests; see below) and to construct QQ plots. Following Kodra & Newman (in prep.) we define the PIT-based catastrophic outlier rate as the fraction of galaxies with $\text{PIT} < 0.0001$ or $\text{PIT} > 0.9999$, which should total 0.0002 for an ideal uniform distribution.

4.1.2 Summary statistics of CDF-based metrics

We evaluate a number of quantitative metrics derived from the visually interpretable QQ plot and PIT histogram, built on the Kolmogorov-Smirnov (KS) statistic

$$\text{KS} \equiv \max_z \left(\left| \text{CDF}[\hat{f}, z] - \text{CDF}[\tilde{f}, z] \right| \right), \quad (6)$$

interpretable as the maximum difference between the CDFs of the empirical distribution of PIT values for the test sample $\hat{f}(z)$ and a reference distribution $\tilde{f}(z)$, in this case $U(0, 1)$, for the ideal distribution of PIT values. We also consider two variants of the KS statistic. A cousin of the KS statistic, the Cramer-von Mises (CvM) statistic

$$\text{CvM}^2 \equiv \int_{-\infty}^{+\infty} (\text{CDF}[\hat{f}, z] - \text{CDF}[\tilde{f}, z])^2 d\text{CDF}[\tilde{f}, z] \quad (7)$$

is the mean-squared difference between the CDFs of an approximate and true PDF. The Anderson-Darling (AD) statistic

$$\text{AD}^2 \equiv N_{tot} \int_{-\infty}^{+\infty} \frac{(\text{CDF}[\hat{f}, z] - \text{CDF}[\tilde{f}, z])^2}{\text{CDF}[\tilde{f}, z](1 - \text{CDF}[\tilde{f}, z])} d\text{CDF}[\tilde{f}, z] \quad (8)$$

is a weighted mean-squared difference featuring enhanced sensitivity to discrepancies in the tails of the distribution. In anticipation of a substantial fraction of galaxies having PIT of 0 or 1, a consequence of catastrophic outliers, we evaluate the AD statistic with modified bounds of integration (0.01, 0.99) to exclude those extremes in the name of numerical stability.

4.2 Conditional Density Estimate (CDE) Loss: a metric of individual photo-z PDFs

The BUZZARD simulation process precludes testing the degree to which samples from our photo-z posteriors reconstruct the space of $p(z, \text{data})$. To the knowledge of the authors, there is only one metric that can be used to evaluate the performance of individual photo-z PDF estimators in the absence of true photo-z posteriors. The conditional density

⁷ See, e. g. the LSST Data Products Definition Document, available at: <https://ls.st/dpdd>

⁸ <http://github.com/aimalz/qp/>

estimation (CDE) loss is an analogue to the familiar root-mean-square-error used in conventional regression, defined as

$$L(f, \hat{f}) \equiv \int \int (f(z|\mathbf{x}) - \hat{f}(z|\mathbf{x}))^2 dz dP(\mathbf{x}), \quad (9)$$

where $f(z|\mathbf{x})$ is the true photo- z PDF that we do not have and $\hat{f}(z|\mathbf{x})$ is an estimate thereof, in terms of the photometry \mathbf{x} . (See Section 3.2.4 for a review of the notation.) We estimate the CDE loss via

$$\hat{L}(f, \hat{f}) = \mathbb{E}_{\mathbf{x}} \left[\int \hat{f}(z | \mathbf{X})^2 dz \right] - 2\mathbb{E}_{\mathbf{x}, Z} \left[\hat{f}(Z | \mathbf{X}) \right] + K_f, \quad (10)$$

where the first term is the expectation value of the photo- z posterior with respect to the marginal distribution of the photometric covariates \mathbf{X} , the second term is the expectation value with respect to the joint distribution of \mathbf{X} and the space Z of all possible redshifts, and the third term K_f is a constant depending only upon the true conditional densities $f(z|\mathbf{x})$. We may estimate these expectations empirically on the test or validation data (Eq. 7 in Izbicki et al. 2017) without knowledge of the true densities.

5 RESULTS

We begin with a demonstrative visual inspection of the photo- z PDFs produced by each code for individual galaxies. Figure 1 shows the photo- z PDFs for four galaxies chosen as examples of photo- z PDF archetypes: a narrow unimodal PDF, a broad unimodal PDF, a bimodal PDF, and a multimodal PDF. We reiterate that under our idealized experimental conditions, differences between the resulting photo- z PDFs are the isolated signature of the implicit prior due to the method by which the photo- z PDFs were derived.

The most striking differences between codes are the small-scale features induced by the interaction between the shared piecewise constant parameterization of 200 bins for $0 < z < 2$ of Section 4 and the smoothing conditions or lack thereof in each algorithm. The $dz = 0.01$ redshift resolution is sufficient to capture the broad peaks of faint galaxies' photo- z PDFs with large photometric errors but is too broad to resolve the narrow peaks for bright galaxies' photo- z PDFs with small photometric errors. This observation is consistent with the findings of Malz et al. (2018) that the piecewise constant form underperforms other parameterizations in the presence of small-scale structures.

However, the shared small-scale features of ANNz2, METAPhoR, CMNN, and SkyNet are a result of various weighted sums of the limited number of training set galaxies with colours similar to those of the test set galaxy in question, with behavior closer to classification than regression in the case of ANNz2. The settings used on GPz in this work forced broadening of the single Gaussian to cover the multimodal redshift solutions of the other codes.

5.1 Performance on photo- z PDF ensembles

The histogram of PIT values, QQ plot, and QQ difference plot relative to the ideal diagonal are provided in Figure 2, showcasing the biases and trends in the average accuracy

Table 2. The catastrophic outlier rate as defined by extreme PIT values. We expect a value of 0.0002 for a proper Uniform distribution. An excess over this small value indicates true redshifts that fall outside the non-zero support of the $p(z)$.

Photo- z Code	fraction $\text{PIT} < 10^{-4}$ or > 0.999
ANNz2	0.0265
BPZ	0.0192
Delight	0.0006
EAZY	0.0154
FlexZBoost	0.0202
GPz	0.0058
LePhare	0.0486
METAPhoR	0.0229
CMNN	0.0034
SkyNet	0.0001
TPZ	0.0130
trainZ	0.0002

of the photo- z PDFs for each code. The high QQ values (i.e. more high than low PIT values) of BPZ, CMNN, Delight, EAZY, and GPz indicate photo- z PDFs biased low, and the low QQ values (more low than high PIT values) of SkyNet and TPZ indicate photo- z PDFs biased high. The gray shaded band marks the 2σ variance in PIT values found using the trainZ algorithm with a bootstrap resampling of the training set and a sample size of 30,000 galaxies, representing a very conservative estimate of the representative training sample size estimated as being required for direct photo- z calibration (Newman et al. 2015), and thus an approximate minimal error significance compared to ideal performance. The existence of deviations in the PIT histograms outside of this gray shaded uncertainty range show that significant biases are present for some codes.

The PIT histograms of Delight, CMNN, SkyNet, and TPZ feature an underrepresentation of extreme values, indicative of overly broad photo- z PDFs, while the overrepresentation of extreme values for METAPhoR indicates overly narrow photo- z PDFs. These five codes in particular have a free parameter for bandwidth, which may be responsible for this vulnerability, in spite of the opportunity for fine-tuning with perfect prior information. FlexZBoost's “sharpening” parameter (described in Section 3.2.4) played a key role in diagonalizing the QQ plot, indicating a common avenue for improvement in the approaches that share this type of parameter. On the other hand, the three purely template-based codes, BPZ, EAZY, and LePhare, do not exhibit much systematic broadening or narrowing, which may indicate that complete template coverage effectively defends from these effects.

Close inspection of the extremes at PIT values of 0 and 1 reveal spikes in the first and last bin of the PIT histogram for some codes in Figure 2, corresponding to catastrophic outliers where the true redshift lies outside of the support of the $p(z)$. The catastrophic outlier rates are provided in Table 2. As expected, trainZ achieves precisely the 0.0002 value expected of an ideal PIT distribution. ANNz2, FlexZBoost, LePhare, and METAPhoR have notably high catastrophic outlier rates > 0.02 , exceeding 100 times the ideal PIT rate, meriting further investigation elsewhere.

Figure 3 highlights the relative values of the KS, CvM,

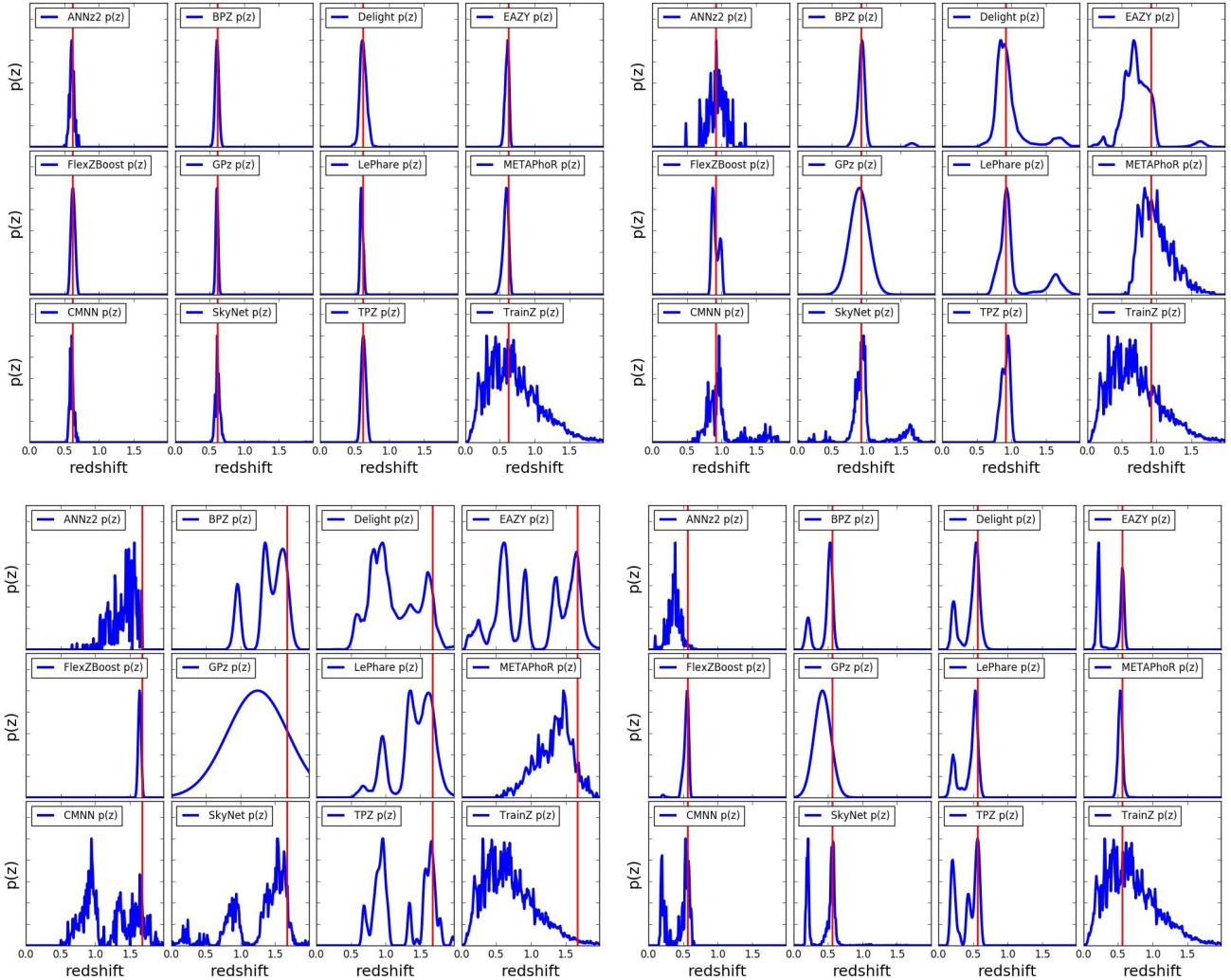


Figure 1. The individual photo- z PDFs (blue) distributions produced by the twelve codes (small panels) on four exemplary galaxies' photometry (large panels) with different true redshifts (red). All photo- z PDFs have been scaled to the same peak value. The photo- z PDFs of all codes share some features for the example galaxies due to physical colour degeneracies and photometric errors: tight unimodal $p(z)$ (upper left), broad unimodal $p(z)$ (upper right), bimodal $p(z)$ (lower right), and complex/multimodal $p(z)$ (lower left). The diverse algorithms and implementations induce differences in small-scale structure and sensitivity to physical systematics.

and AD test statistics calculated by comparing the PIT distribution and a uniform distribution $U(0, 1)$. METAPhOR and LePhare perform well under the AD but poorly under the KS and CvM due to their high catastrophic outlier rates. ANNz2 and FlexZBoost are the top scorers under these metrics of the PIT distribution. ANNz2's strong performance can be attributed to an aspect of the training process in which training set galaxies with PIT values that more closely match the percentiles of the DC1 training set's redshift distribution are upweighted; in effect, these quantile-based metrics were part of the algorithm itself that may or may not serve it well under more realistic experimental conditions. Similar to what was done for the PIT histograms in Figure 2, we create bootstrap training samples of 30,000 galaxies for use with trainZ in order to estimate a conservative statistical floor that we would expect in real data. No code reaches this idealized floor, indicating that all codes suffer some degradation from the ideal when employing their implicit priors,

though ANNz2, FlexZBoost, and GPz are within a factor of two.

5.2 Performance on individual photo- z PDFs

The values of the CDE loss statistic of individual photo- z PDF accuracy are provided in Table 3. It is worth noting that strong performance on the CDE loss, corresponding to lower values of the metric, should imply strong performance on the other metrics, though the inverse is not necessarily true. Thus the CDE loss is the most effective metric for generic science cases.

Of the metrics we were able to consider in this experiment, the **CDE Loss is the only metric that can appropriately penalize the pathological trainZ**. Additionally, it favors CMNN and FlexZBoost, the latter of which is optimized for this metric.

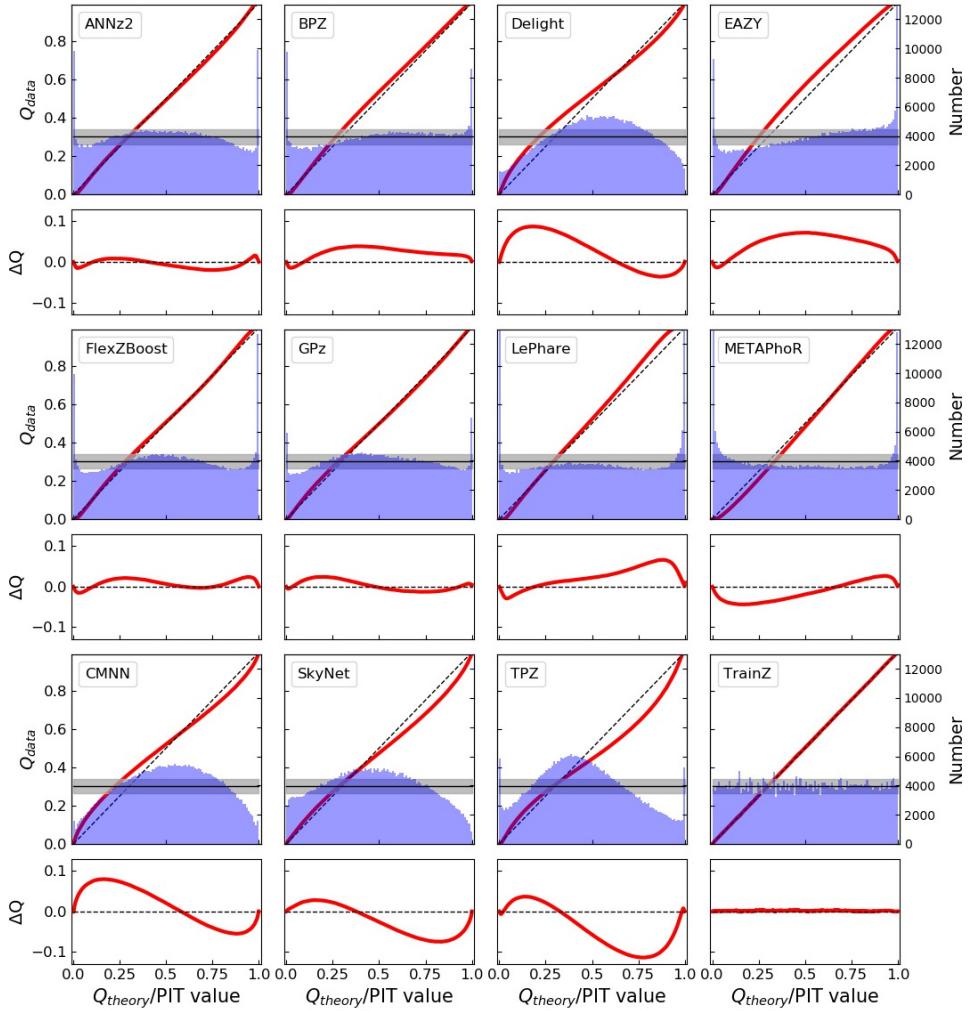


Figure 2. The QQ plot (red) and PIT histogram (blue) of the photo- z PDF codes (panels) along with the ideal QQ (black dashed) and ideal PIT (gray horizontal) curves, as well as a difference plot for the QQ difference from the ideal diagonal (lower inset). The gray shaded region indicates the 2σ range from a bootstrap resampling of the training set with a size of 30,000 galaxies using `trainZ`. The twelve codes exhibit varying degrees of four deviations from perfection: an overabundance of PIT values at the centre of the distribution indicate a catalogue of overly broad photo- z PDFs, an excess of PIT values at the extrema indicates a catalogue of overly narrow photo- z PDFs, catastrophic outliers manifest as overabundances at PIT values of 0 and 1, and asymmetry indicates systematic bias, a form of model misspecification. Values in excess of the 2σ shaded region show that for some codes these errors will be significant given expected training sample sizes.

985 6 DISCUSSION AND FUTURE WORK

986 In contrast with other photo- z PDF comparison papers that 1001
 987 have aimed to identify the “best” code for a given survey, we 1002
 988 have focused on the somewhat more philosophical questions 1003
 989 of how to assess photo- z PDF methods and how to interpret 1004
 990 differences between codes in terms of photo- z PDF perfor- 1005
 991 mance. In Section 6.1, we reframe the strong performance of 1006
 992 our pathological photo- z PDF technique, `trainZ`, as a cau- 1007
 993 tionary tale about the importance of choosing appropriate 1008
 994 comparison metrics. In Section 6.2, we outline the experi- 1009
 995 ments we intend to build upon this study. In Section 6.3, we 1010
 996 discuss the enhancements of the mock data set that will be 1011
 997 necessary to enable the future experiments. 1012

998 6.1 Interpretation of metrics

999 We remind the reader that codes utilized in this study were
 1000 given a goal of obtaining accurate photo- z PDFs, not an
 1001 accurate stacked estimator of the redshift distribution, so
 1002 we do not expect the same codes to necessarily perform well
 1003 for both classes of metrics. Indeed, the codes were optimized
 1004 for their interpretation of our request for “accurate photo-
 1005 z PDFs,” and we expect that the implementations would
 1006 have been adjusted had we requested optimization of the
 1007 traditional metrics of Appendices A and B.

1008 Furthermore, our metrics are not necessarily able to as-
 1009 sess the fidelity of individual photo- z PDFs relative to true
 1010 posteriors: in the absence of a “true PDF” from which
 1011 redshifts are drawn, it is difficult to construct metrics to mea-
 1012 sure performance for individual galaxies rather than ensem-

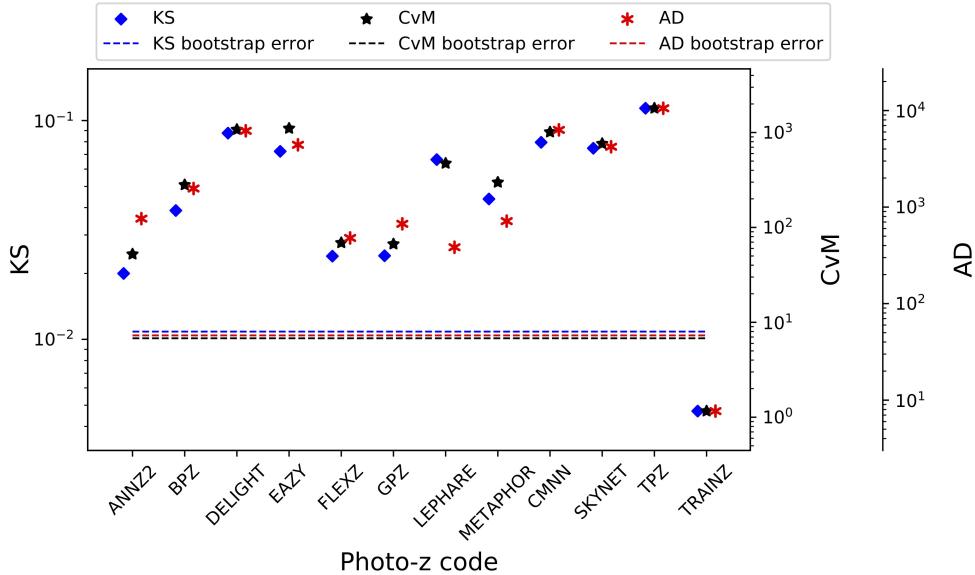


Figure 3. A visualization of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the PIT distributions. There is generally good agreement between these statistics, with differences corresponding to the codes with outstanding catastrophic outlier rates, a reflection of the differences in how each statistic weights the tails of the distribution. Horizontal lines indicate the level of uncertainty found by bootstrapping a training set sample of 30,000 galaxies using `trainZ`; none of the codes reach this conservative ideal floor in expected uncertainty.

Table 3. CDE loss statistic of the individual photo- z PDFs for each code. A lower value of the CDE loss indicates more accurate individual photo- z PDFs, with CMNN and FlexZBoost performing best under this metric.

Photo- z Code	CDE Loss
ANNz2	-6.88
BPZ	-7.82
Delight	-8.33
EAZY	-7.07
FlexZBoost	-10.60
GPz	-9.93
LePhare	-1.66
METAPHOR	-6.28
CMNN	-10.43
SkyNet	-7.89
TPZ	-9.55
<code>trainZ</code>	-0.83

bles. (The CDE Loss metric of section 4.2 is an exception to this rule.) A lack of appropriate metrics more sophisticated than the CDE Loss remains an open issue for science cases requiring accurate individual galaxy PDFs. The metric-specific performance demonstrated in this paper implies that we may need multiple photo- z PDF approaches tuned to each metric in order to maximize returns over all science cases in large upcoming surveys.

The `trainZ` estimator of Section 3.3, which assigns every galaxy a photo- z PDF equal to $N(z)$ of the training set, is introduced as an experimental control or null test to demonstrate this point via *reductio ad absurdum*. Because our training set is perfectly representative of the test set, $N(z)$ should be identical for both sets down to statistical noise. We make the alarming observation that `trainZ`,

the absurd experimental control, outperforms all codes on the CDF-based metrics, and all but one code on the $N(z)$ based statistics. The PIT and other CDF-based metrics upon which modern photo- z PDF comparisons are built (Bordoloi et al. 2010; Polsterer et al. 2016; Tanaka et al. 2018) can be gamed by a trivial estimator that yields only an affirmation of prior knowledge uninformed by the data. In other words, such ensemble metrics are insufficient for the task of selecting photo- z PDF codes for analysis pipelines⁹.

The CDE loss and point estimate metrics appropriately penalize `trainZ`'s naivete. As shown in Appendix B, `trainZ` has identical $ZPEAK$ and $ZWEIGHT$ values for every galaxy, and thus the photo- z point estimates are constant as a function of true redshift, i.e. a horizontal line at the mode and mean of the training set distribution respectively. The explicit dependence on the individual posteriors in the calculation of the CDE loss, described in Section 5.2, distinguishes this metric from those of the photo- z PDF ensemble and stacked estimator of the redshift distribution, despite their prevalence in the photo- z literature.

In summary, context is crucial to defend against deceptively strong performers such as `trainZ`; the best photo-

⁹ That being said, we note that close relatives of `trainZ` have been employed by weak lensing surveys of the past and present (Lima et al. 2008; Hildebrandt et al. 2017; Hoyle et al. 2018) to estimate $N(z)$ by assigning each test set galaxy the redshift distribution of a subset of the training set, where the subset is defined as similar to the test set galaxy in the space of photometric data. The specific science goal naturally guides the choice of metric to focus on $N(z)$ rather than individual photo- z PDFs, and on the basis of that metric such improvements upon `trainZ` are guaranteed to be more robust to training set imbalance.

1051 ***z* PDF method is the one that most effectively** 1110
1052 **achieves the science goals of a particular study**, not
1053 the one that performs best on a metric that does not reflect
1054 those goals. In the absence of a single scientific motivation
1055 or the information necessary for a principled metric defini-
1056 tion, we must consider many metrics and be critical of the
1057 information transmitted by each.

1058 6.2 Extensions to the experimental design

1059 The work presented in this paper is only a first step in as-
1060 sessing photo-*z* PDF approaches and moving toward a pho-
1061 tometric redshift estimator that will be employed for LSST
1062 analyses. Extensions of the experimental design will require
1063 further rounds of analyses, and the authors welcome interest
1064 from those outside LSST-DESC to have their codes assessed
1065 in these future investigations.

1066 This initial paper explores photo-*z* PDF code perfor-
1067 mance in idealized conditions with perfect catalogue-based
1068 photometry and representative training data, but the re-
1069 silience of each code to realistic imperfections in prior in-
1070 formation has not yet been evaluated. A top priority for a
1071 follow-up study is to test realistic forms of incomplete, er-
1072 roneous, and non-representative template libraries and train-
1073 ing sets as well as the impact of other forms of external
1074 priors that must be ingested by the codes, major concerns
1075 in [Newman et al. \(2015\)](#); [Masters et al. \(2017\)](#). Outright red-
1076 shift failures due to emission line misidentification or noise
1077 spikes may be modeled by the inclusion of a small number
1078 of high-confidence yet false redshifts. We plan to perform a
1079 full sensitivity analysis on a realistically incomplete training
1080 set of spectroscopic galaxies, modeling the performance of
1081 spectrographs, emission-line properties, and expected signal-
1082 to-noise to determine which potential training set galaxies
1083 are most likely not to yield a secure redshift.

1084 Appendix A only addresses the stacked estimator of the
1085 redshift distribution of the entire galaxy catalogue rather
1086 than subsets in bins, tomographic or otherwise. The effects
1087 of tomographic binning schemes will be explored in a dedi-
1088 cated future paper, including propagation of redshift uncer-
1089 tainties in a set of fiducial tomographic redshift bins in order
1090 to estimate impact on cosmological parameter estimation.

1091 Sequels to this study will also address some shortcom-
1092 ings of our experimental procedure. The fixed redshift grid
1093 shared between the codes may have unfairly penalized codes
1094 with a different native parameterization, as precision is lost
1095 when converting between formats. Performance on the (ad-
1096mittedly small) population of sharply peaked photo-*z* PDFs
1097 may have been suppressed across all codes due to the insuffi-
1098cient resolution of the redshift grid. In light of the results of
1099 [Malz et al. \(2018\)](#), in future analyses we plan to switch from
1100 a fixed grid to the quantile parameterization or to permit
1101 each code to use its native storage format under a shared
1102 number of parameters.

1103 Section 4 discussed the difficulty in evaluating PDF ac-
1104curacy for individual objects with known (*z*, *d*) information
1105 but without a known $p(z, d)$. In a follow-up study, we will
1106 generate mock data probabilistically, yielding true PDFs in
1107 addition to true redshifts and photometric data. This future
1108 data set will enable tests of PDF accuracy for individual
1109 galaxies rather than solely ensembles.

6.3 Realistic mock data

To make optimal use of the LSST data for cosmological and other astrophysical analyses of the LSST-DESC SRM, future investigations that build upon this one will require a more sophisticated set of galaxy photometry and redshifts. This initial paper explored a data set that was constructed at the catalogue level, with no inclusion of the complications that arise from photometric measurements of imaging data. Future data challenges will move to catalogues constructed from mock images, including the complications of deblending, sensor inefficiencies, and heterogeneous observing conditions, all anticipated to affect the measured colours of LSST’s galaxy sample ([Dawson et al. 2016](#)).

The DC1 galaxy SEDs were linear combinations of just five basis SED templates, and the next generation of data for photo-*z* PDF investigations must include a broader range of physical properties. Though we only considered $z < 2$ here, LSST 10-year data will contain $z > 2$ galaxies, plagued by fainter apparent magnitudes and anomalous colours due to stellar evolution. A subsequent study must also have a data set that includes low-level active galactic nuclei (AGN) features in the SEDs, which perturb colours and other host galaxy properties. An observational degeneracy between the Lyman break of a $z \sim 2 - 3$ galaxy from the Balmer break of a $z \sim 0.2 - 0.3$ galaxy is a known source of catastrophic outliers ([Massarotti et al. 2001](#)) that was not effectively included in this study. To gauge the sensitivity of photo-*z* PDF estimators to catastrophic outliers, our data set must include realistic high-redshift galaxy populations.

7 CONCLUSION

This paper compares twelve photo-*z* PDF codes under controlled experimental conditions of representative and complete prior information to set a baseline for an upcoming sensitivity analysis. This work isolates the impact on metrics of photo-*z* PDF accuracy due to the estimation technique as opposed to the complications of realistic physical systematics of the photometry. Though the mock data set of this investigation did not include true photo-*z* posteriors for comparison, **we interpret deviations from perfect results given perfect prior information as the imprint of the implicit assumptions underlying the estimation approach.**

We evaluate the twelve codes under science-agnostic metrics both established and emerging to stress-test the ensemble properties of photo-*z* PDF catalogues derived by each method. In appendices, we also present metrics of point estimates and a prevalent summary statistic of photo-*z* PDF catalogues used in cosmological analyses to enable the reader to relate this work to studies of similar scope. We observe that no one code dominates in all metrics, and that the standard metrics of photo-*z* PDFs and the stacked estimator of the redshift distribution can be gamed by a very simplistic procedure that asserts the prior over the data. We emphasize to the photo-*z* community that **metrics used to vet photo-*z* PDF methods must be scrutinized to ensure they correspond to the quantities that matter to our science.**

1167 **Acknowledgments**

1168 This paper has undergone internal review in the LSST
 1169 Dark Energy Science Collaboration. The authors acknowl-
 1170 edge feedback from the internal reviewers: Daniel Gruen,
 1171 Markus Rau, and Michael Troxel.

1172 Author contributions are listed below.

1173 S.J. Schmidt: Co-led the project. (conceptualization, data
 1174 curation, formal analysis, investigation, methodology,
 1175 project administration, resources, software, supervision,
 1176 visualization, writing – original draft, writing – review &
 1177 editing).

1178 A.I. Malz: Co-led the project, contributed to choice of
 1179 metrics, implementation in code, and writing. (conceptu-
 1180 alization, methodology, project administration, resources,
 1181 software, visualization, writing – original draft, writing –
 1182 review & editing).

1183 J.Y.H. Soo: Ran ANNz2 and Delight, updated abstract,
 1184 edited sections 1 through 6, added tables in Methods
 1185 and Results, updated references.bib and added references
 1186 throughout the paper.

1187 I.A. Almosallam: vetted the early versions of the data set
 1188 and ran many photo-z codes on it, applied GPz to the final
 1189 version and wrote the GPz subsection.

1190 M. Brescia: Main-ideator of MLPQNA and co-ideator of
 1191 METAPHOR; modification of METAPHOR pipeline to fit
 1192 the LSST data structure and requirements.

1193 S. Cavuoti: Co-ideator of METAPHOR, contributed to
 1194 choice and test of metrics, ran METAPHOR, minor text
 1195 editing.

1196 J. Cohen-Tanugi: contributed to running code, analysis
 1197 discussion, and editing, reviewing the paper.

1198 A.J. Connolly: Developed the colour-matched nearest-
 1199 neighbours photo-z code; participated in discussions of the
 1200 analysis.

1201 J. DeRose: One of the primary developers of the Buzzard-
 1202 highres simulation employed in the analysis.

1203 P.E. Freeman: Contributed to choice of CDE metrics and
 1204 to implementation of FlexZBoost.

1205 M.L. Graham: Ran the colour-matched nearest-neighbours
 1206 photo-z code on the Buzzard catalogue and wrote the
 1207 relevant piece of Section 2; participated in discussions of
 1208 the analysis.

1209 K.G. Iyer: assisted in writing metric functions used to
 1210 evaluate codes.

1211 M.J. Jarvis: Contributed text on AGN to Discussion section
 1212 and portions of GPz work.

1213 J.B. Kalmbach: Worked on preparing the figures for the
 1214 paper.

1215 E. Kovacs: Ran simulations, discussed data format and
 1216 properties for SEDs, dust, and ELG corrections.

1217 A.B. Lee: Co-developed FlexZBoost and the CDE loss
 1218 statistic, wrote text on the work, and supervised the
 1219 development of FlexZBoost software packages.

1220 G. Longo: Scientific advise, test and validation of the
 1221 modified METAPHOR pipeline, text of the METAPHOR
 1222 section.

1223 C.B. Morrison: Managerial support; Discussions with au-
 1224 thors regarding metrics and style; Some coding contribu-
 1225 tion to metric computation.

1226 J.A. Newman: Contributions to overall strategy, design of
 1227 metrics, and supervision of work done by Rongpu Zhou.

1228 E. Nourbakhsh: Ran and optimized TPZ code and wrote a
 1229 subsection of Section 2 for TPZ.

1230 E. Nuss: contributed to running code, analysis discussion,
 1231 and editing, reviewing the paper.

1232 T. Pospisil: Co-developed FlexZBoost software and CDE
 1233 loss calculation code.

1234 H. Tranin: contributed to providing SkyNet results and
 1235 writing the relevant section.

1236 R. Zhou: Optimized and ran EAZY and contributed to the
 1237 draft.

1238 R. Izbicki: Co-developed FlexZBoost and the CDE loss
 1239 statistic, and wrote software for FlexZBoost

The authors express immense gratitude to Alex Abate,
 without whom this paper would not have gotten started.

SJS and EN acknowledge support from DOE grant DE-SC0009999. SJS acknowledges support from NSF/AURA grant N56981C. 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. During the completion of this work, AIM was advised by David W. Hogg and was supported by National Science Foundation grant AST-1517237. JYHS acknowledges financial support from the MyBrainSc Scholarship (Ministry of Education, Malaysia), and the supervision of Ofer Lahav and Benjamin Joachimi. JYHS would also like to thank Antonella Palmese and Boris Leistedt for guidance on the use of the algorithms ANNz2 and Delight respectively. I.A. acknowledges the support of King Abdulaziz City for Science and Technology. MB acknowledges support from the Agreement ASI/INAF 2018-23-HH.0 - Phase D. AJC and JBK acknowledges support from DOE grant DE-SC-0011635. AJC, JBK, MLG, CBM acknowledge support from the DIRAC Institute in the Department of Astronomy at the University of Washington. The DIRAC Institute is supported through generous gifts from the Charles and Lisa Simonyi Fund for Arts and Sciences, and the Washington Research Foundation. PEF acknowledges support from NSF grant 1521786. MJJ acknowledges support from Oxford Hintze Centre for Astrophysical Surveys which is funded through generous support from the Hintze Family Charitable Foundation. ABL and TP acknowledge support from NSF DMS grant 1520786. GL acknowledges partial funding from the EU funded ITN Marie Curie Network SUNDIAL. CBM is supported in part by the National Science Foundation through Cooperative Agreement 1258333 managed by the Association of Universities for Research in Astronomy(AURA), and the Department of Energy under Contract No. DE-AC02-76SF00515 with the SLAC National Accelerator Laboratory. Additional LSST funding comes from private donations, grants to universities, and in-kind support from LSSTC Institutional Members. JAN and RZ were supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics under award number DE-SC0007914. RI acknowledges support from FAPESP grant 2019/11321-9 and CNPq grant 306943/2017-4.

In addition to packages cited in the text, analyses performed in this paper used the following software packages: **Numpy** and **Scipy** ([Oliphant 2007](#)), **Matplotlib** ([Hunter 2007](#)), **Seaborn** ([Waskom et al. 2017](#)), **minFunc** ([Schmidt 2005](#)), **qp** ([Malz & Marshall 2018](#); [Malz et al. 2018](#)),

1290 pySkyNet (Bonnett 2016), and photUtils from the LSST 1339
 1291 simulations package (Connolly et al. 2014). 1340

1292 The DESC acknowledges ongoing support from the In- 1341
 1293 stitut National de Physique Nucléaire et de Physique des 1342
 1294 Particules in France; the Science & Technology Facilities 1343
 1295 Council in the United Kingdom; and the Department of En- 1344
 1296 ergy, the National Science Foundation, and the LSST Cor- 1345
 1297 poration in the United States. DESC uses resources of the 1346
 1298 IN2P3 Computing Center (CC-IN2P3-Lyon/Villeurbanne - 1347
 1299 France) funded by the Centre National de la Recherche Sci- 1348
 1300 entifique; the National Energy Research Scientific Comput- 1349
 1301 ing Center, a DOE Office of Science User Facility supported 1350
 1302 by the Office of Science of the U.S. Department of Energy 1351
 1303 under Contract No. DE-AC02-05CH11231; STFC DiRAC 1352
 1304 HPC Facilities, funded by UK BIS National E-infrastructure 1353
 1305 capital grants; and the UK particle physics grid, supported 1354
 1306 by the GridPP Collaboration. This work was performed in 1355
 1307 part under DOE Contract DE-AC02-76SF00515. 1356

1308 APPENDIX A: EVALUATION OF THE 1357 1309 REDSHIFT DISTRIBUTION 1358

1310 Perhaps the most popular application of photo-z PDFs is 1359
 1311 the estimation of the overall redshift distribution $N(z)$, a 1360
 1312 quantity that enters some cosmological calculations and the 1361
 1313 true value of which is known for the DC1 data set and will be 1362
 1314 denoted as $\tilde{N}(z)$. In terms of the prior information provided 1363
 1315 to each method, the true redshift distribution satisfies the 1364
 1316 tautology $\tilde{N}(z) = p(z|I_D)$ due to our experimental set-up; 1365
 1317 because the DC1 training and template sets are representa- 1366
 1318 tive and complete, I_D represents a prior that is also equal 1367
 1319 to the truth. In this ideal case of complete and represen- 1368
 1320 tative prior information, the method that would give the 1369
 1321 best approximation to $\tilde{N}(z)$ would be one that neglects all 1370
 1322 the information contained in the photometry $\{d_i\}_{N_{tot}}$ and 1371
 1323 gives every galaxy the same photo-z PDF $\hat{p}_i(z) = \tilde{N}(z)$ for 1372
 1324 all i ; the inclusion of any information from the photometry 1373
 1325 would only introduce noise to the optimal result of return- 1374
 1326 ing the prior. This is the exact estimator, **trainZ**, that we 1375
 1327 have described in Section 3.3, and which will serve as an 1376
 1328 experimental control. 1377

1329 A1 Metrics of the stacked estimator of the 1378 1330 redshift distribution 1379

1331 “Stacking” according to 1380

$$1332 \hat{N}^H(z) \equiv \frac{1}{N_{tot}} \sum_i^{N_{tot}} \hat{p}_i^H(z) \quad (A1) \quad 1381$$

1333 is the most widely used method for obtaining $\hat{N}^H(z)$ as an 1382
 1334 estimator of the redshift distribution from photo-z PDFs 1383
 1335 derived by a method H . While the stacked estimator of the 1384
 1336 redshift distribution violates the mathematical definition of 1385
 1337 statistical independence and is thus not formally correct¹⁰, 1386
 1338 we use it as a basis for comparison of photo-z PDF methods 1387
 1339

¹⁰ Malz & Hogg (in prep) shows how the stacking procedure can lead to bias in the estimate of $N(z)$ and presents a principled alternative to this commonly employed method. See <https://github.com/aimalz/chippr> for details.

1340 under the untested assumption that the response of our met- 1341
 1342 rics of $\hat{N}^H(z)$ will be analogous to the same metrics applied 1343
 1344 to a principled estimator of the redshift distribution. 1345

1345 As $N(z)$ is itself a univariate PDF, we apply the met- 1346
 1347 rics of the previous sections to it as well. We additionally 1348
 1349 calculate the first three moments 1350

$$\langle z^m \rangle \equiv \int_{-\infty}^{\infty} z^m N(z) dz \quad (A2)$$

1351 of the estimated redshift distribution $\hat{N}^H(z)$ for each code 1352
 1353 and compare them to the moments of the true redshift distri- 1354
 1355 bution $\tilde{N}(z)$. Under the assumption that the stacked estima- 1356
 1357 tor is unbiased, a superior method minimizes the difference 1358
 1359 between the true and estimated moments. 1360

1361 A2 Performance on the stacked estimator of the 1362 1363 redshift distribution 1364

1365 Figure A1 shows the stacked estimator $\hat{N}(z)$ of the redshift 1366
 1367 distribution for each code compared to the true redshift 1368
 1369 distribution $\tilde{N}(z)$, where the stacked estimator has been 1370
 1371 smoothed for each code in the plot using a kernel density 1372
 1373 estimate (KDE) with a bandwidth chosen by Scott’s Rule 1374
 1375 (Scott 1992) in order to minimize visual differences in small- 1376
 1377 scale features; the quantitative statistics, however, are calcu- 1378
 1379 lated using the empirical CDF which is not smoothed. 1380

1381 Many of the codes, including all the model-fitting ap- 1382
 1383 proaches and ANNz2, GPz, METAPhOr, and SkyNet from the 1384
 1385 data-driven camp, overestimate the redshift density at $z \sim$ 1386
 1387 1.4. This behavior is a consequence of the 4000 Å break 1388
 1389 passing through the gap between the z and y filters, which 1390
 1391 induces a genuine discontinuity in the $z - y$ colour as a func- 1392
 1393 tion of redshift that can sway the photo-z PDF estimates in 1394
 1395 the absence of bluer spectral features. 1396

1397 ANNz2, GPz, and METAPhOr feature exaggerated peaks 1398
 1399 and troughs relative to the training set, a potential sign 1400
 1401 of overtraining. Further investigation on overtraining is 1402
 1403 needed, if present this is an obstacle that may be overcome 1404
 1405 with adjustment of the implementation. 1406

1407 As expected, **trainZ** perfectly recovers the true redshift 1408
 1409 distribution: as the training sample is selected from the same 1410
 1411 underlying distribution as the test set, the redshift distribu- 1412
 1413 tions are identical, up to Poisson fluctuations due to the 1414
 1415 finite number of sample galaxies. CMNN is also in excellent 1416
 1417 agreement for similar reasons: with a representative train- 1418
 1419 ing sample of galaxies spanning the colour-space, the sum 1420
 1421 of the colour-matched neighbour redshifts should return the 1422
 1423 true redshift distribution. FlexZBoost and TPZ also perform 1424
 1425 superb recovery of the true redshift distribution, with only 1426
 1427 a slight deviation at $z \sim 1.4$. Our metrics, however, cannot 1428
 1429 discern whether these four approaches, as well as Delight, 1430
 1431 are spared the $z \sim 1.4$ degeneracy in $\hat{N}(z)$ because they have 1432
 1433 more effectively used information in the data or if the impact 1434
 1435 is simply washed out by the stacked estimator’s effective av- 1436
 1437 erage over the test set galaxy sample. See Appendix B for 1438
 1439 further discussion of the $z \sim 1.4$ issue. 1440

1441 Figure A2 shows the quantitative Kolmogorov-Smirnoff 1442 (KS), Cramer-Von Mises (CvM), and Anderson Darling 1443 (AD) test statistics for each of the codes for the $\hat{N}(z)$ based 1444
 1445 measures. The horizontal lines show the the result of a boot- 1446
 1447 strap resampling of the training set using 30,000 samples for 1448
 1449

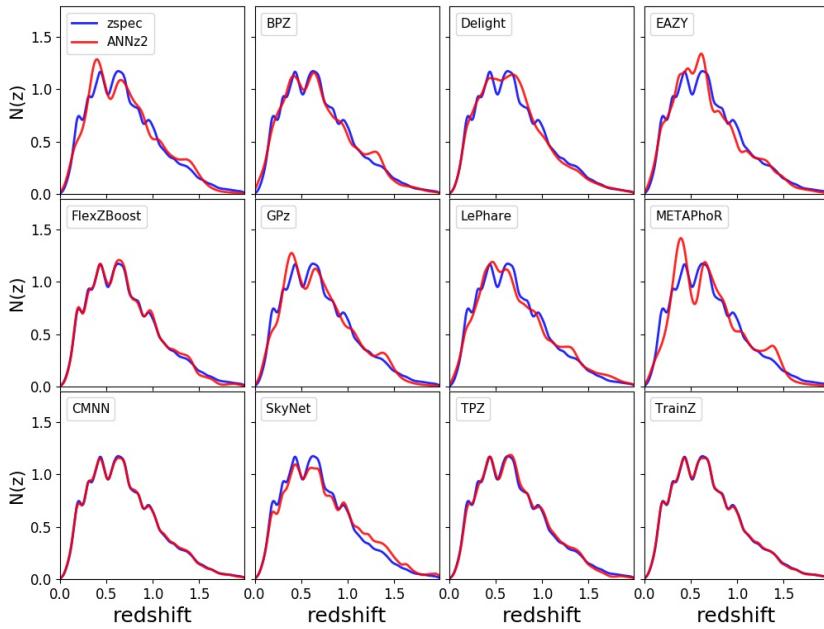


Figure A1. The smoothed stacked estimator $\hat{N}(z)$ of the redshift distribution (red) produced by each code (panels) compared to the true redshift distribution $\tilde{N}(z)$ (blue). Varying levels of agreement are seen among the codes, with the smallest deviations for CMNN, FlexZBoost, TPZ, and trainZ.

trainZ, representing a conservative idealized limit on expected performance for a modest-sized representative training set of galaxies, as mentioned in Section 5.1. The AD bootstrap statistic is elevated due to its sensitivity to the tails of distributions. The stacked estimators of the redshift distribution for CMNN and trainZ best estimate $\tilde{N}(z)$ under these metrics, whereas EAZY, LePhare, METAPhR, and SkyNet underperform; BPZ, GPz, and TPZ are within a factor of two of the conservative limit for all statistics. It is unsurprising that CMNN scores well, as with a nearly complete and representative training set choosing neighbouring points in colour/magnitude space to construct an estimator should lead to excellent agreement in the final $\hat{N}(z)$.

It is, however, surprising that TPZ does well on $\hat{N}(z)$ given its poor performance on the ensemble photo-z PDFs, especially knowing that TPZ was optimized for photo-z PDF ensemble metrics rather than the stacked estimator of the redshift distribution. A possible explanation is the choice of smoothing parameter chosen during validation, which affects photo-z PDF widths as well as overall redshift bias and could be modified to improve performance under the photo-z PDF metrics.

We calculated the first three moments of the stacked $\hat{N}(z)$ distribution of all galaxies and compared it to the moments of the true redshift distribution. Figure A3 shows the residuals of the moments for all codes. Accuracy of the moments varies widely between codes, raising concerns about the propagation to cosmological analyses. The DESC SRD (The LSST Dark Energy Science Collaboration et al. 2018) lists stringent requirements on how well the mean and variance of tomographic redshift bins must be known for each of

the main DESC science cases. We indicate the Year 10 (Y10) requirements assuming our true mean redshift of $z = 0.701$ as dashed lines. In this study with representative training data, ANNz2, CMNN, TPZ, and our pathological trainZ estimator meet the Y10 requirement on the mean redshift. Only ANNz2, CMNN, and trainZ meet both requirements. One should be concerned that many codes fail to meet this ambitious limit under perfect prior information because all codes are anticipated to do no better under realistically imperfect prior information, and indicates that additional calibration to remove these systematic offsets (e.g. Newman 2008) will likely be necessary in order to meet these stringent goals.

SkyNet exhibits redshift bias in Figure A1 and is a clear outlier in the first moment of $\hat{N}(z)$ in Figure A3. The SkyNet algorithm employs a random subsampling of the training set without testing that the subset is representative of the full population, and the implementation used here does not upweight rarer low- and high-redshift galaxies, as in Bonnett (2015), suggesting a possible cause that may be addressed in future work.

APPENDIX B: Photo-z POINT ESTIMATION AND METRICS

While this work assumes that science applications value the information of the full photo-z PDF, we present conventional metrics of photo-z point estimates as a quick and dirty visual diagnostic tool and to facilitate direct comparisons to historical studies.

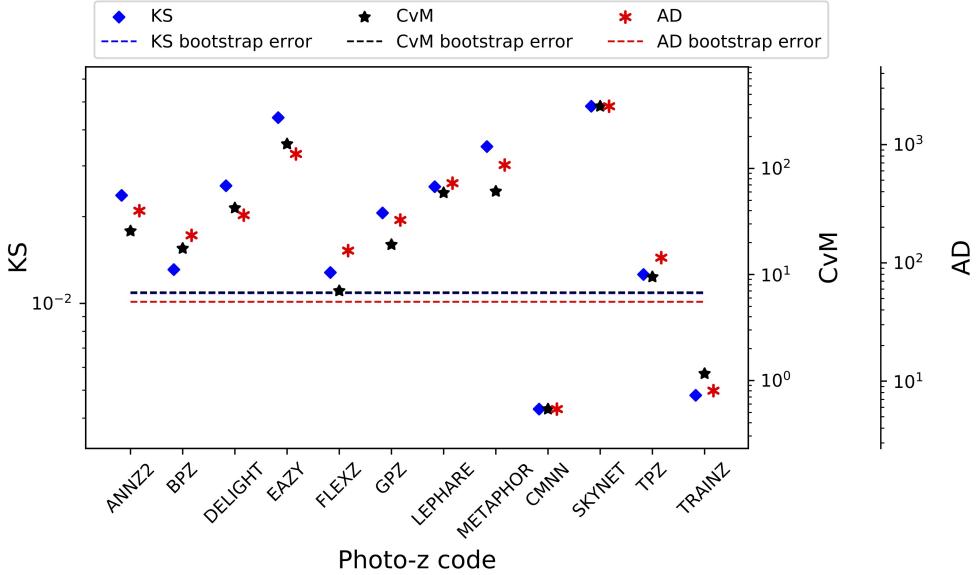


Figure A2. A visualization of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the $\hat{N}(z)$ distributions. Horizontal lines indicate the statistic values (including uncertainty) achieved using `trainZ` via bootstrap resampling a training set containing 30,000 redshifts. We make the reassuring observation that these related statistics do not disagree significantly with one another. `CMNN` outperforms the control case, `trainZ`, and several codes are within a factor of two of this conservative idealized limit. `SkyNet` scores poorly due to an overall bias in its redshift predictions.

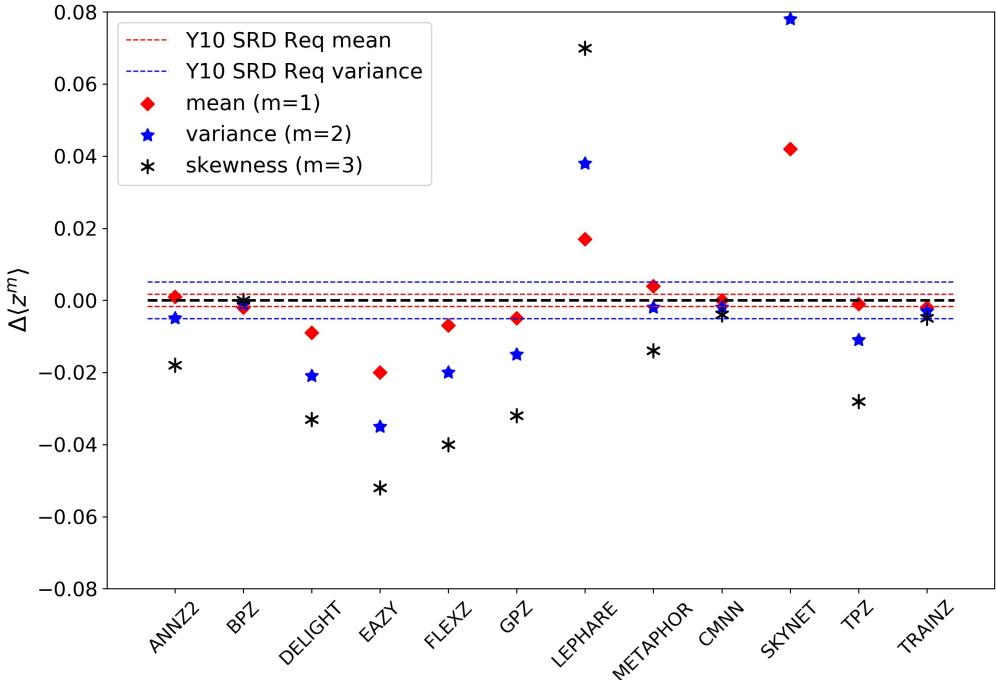


Figure A3. Residuals of the first three moments of the stacked $\hat{N}(z)$ distribution. Red and blue horizontal lines indicate the Year 10 DESC SRD requirements on accuracy of the mean and variance respectively. Only a small number of codes are able to meet these specifications even with perfect training data.

1454 B1 Reduction of photo-z PDFs to point estimates 1459

1455 Though we acknowledge that many of the codes can also 1460
 1456 return a native photo-z point estimate, we put all codes on 1461
 1457 equal footing by considering two generic photo-z point 1462 estimators, the mode z_{PEAK} and main-peak-mean z_{WEIGHT} 1463

(Dahlen et al. 2013), a weighted mean within the bounds of the main peak, as identified by the roots of $p(z) = 0.05 \times z_{PEAK}$. Though z_{WEIGHT} neglects information in a secondary peak of e. g. a bimodal distribution, it avoids the

- 1463 pitfall of reducing the photo- z PDF to a redshift between 1514 Almosallam I. A., Jarvis M. J., Roberts S. J., 2016b, *MNRAS*,
1464 peaks where there is low probability. 1515 462, 726
1516 Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F.,
1517 Fontana A., Giallongo E., 1999, *MNRAS*, 310, 540
1518 Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, *ApJ*, 762, 109
1519 Benítez N., 2000, *ApJ*, 536, 571
1520 Bernstein G., Huterer D., 2010, *MNRAS*, 401, 1399
1521 Blanton M. R., Roweis S., 2007, *AJ*, 133, 734
1522 Blanton M. R., et al., 2005, *AJ*, 129, 2562
1523 Bonnett C., 2015, *MNRAS*, 449, 1043
1524 Bonnett C., 2016, Python wrapper to SkyNet, <https://pyskynet.readthedocs.io/en/latest/>
1525 Bonnett C., et al., 2016, *Phys. Rev. D*, 94, 042005
1526 Bordoloi R., Lilly S. J., Amara A., 2010, *MNRAS*, 406, 881
1527 Brammer G. B., van Dokkum P. G., Coppi P., 2008, *ApJ*, 686,
1528 1503
1529 Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees, Statistics/Probability Series.
1530 Wadsworth Publishing Company, Belmont, California, U.S.A
1531 Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vellucci
1532 C., Longo G., 2018, preprint, (<arXiv:1802.07683>)
1533 Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483
1534 Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, 442, 3380
1535 Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo
1536 G., 2017, *MNRAS*, 465, 1959
1537 Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM
1538 SIGKDD International Conference on Knowledge Discovery
and Data Mining, KDD '16. ACM, New York, NY, USA,
pp 785–794, doi:[10.1145/2939672.2939785](https://doi.acm.org/10.1145/2939672.2939785), <http://doi.acm.org/10.1145/2939672.2939785>
1539 Connolly A. J., et al., 2014, in Angeli G. Z., Dierickx P., eds,
1540 Society of Photo-Optical Instrumentation Engineers (SPIE)
1541 Conference Series Vol. 9150, Modeling, Systems Engineering,
1542 and Project Management for Astronomy VI. p. 14,
doi:[10.1117/12.2054953](https://doi.org/10.1117/12.2054953)
1543 Dahlen T., et al., 2013, *ApJ*, 775, 93
1544 Dalmasso N., Pospisil T., Lee A. B., Izbicki R., Freeman P. E.,
1545 Malz A. I., 2019, arXiv preprint arXiv:1908.11523
1546 Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016,
1547 *ApJ*, 816, 11
1548 DeRose J., et al., 2019, arXiv e-prints, p. <arXiv:1901.02401>
1549 Erben T., et al., 2013, *MNRAS*, 433, 2545
1550 Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, *ApJ*, 513, 34
1551 Firth A. E., Lahav O., Somerville R. S., 2003, *MNRAS*, 339, 1195
1552 Freeman P. E., Izbicki R., Lee A. B., 2017, *MNRAS*, 468, 4556
1553 Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, *MNRAS*, 441,
1554 1741
1555 Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones
1556 R. L., Jurić M., Daniel S. F., Yoachim P., 2018, *AJ*, 155, 1
1557 Green J., et al., 2012, preprint (<arXiv:1208.4012>),
1558 Hildebrandt H., et al., 2010, *A&A*, 523, A31
1559 Hildebrandt H., et al., 2017, *MNRAS*, 465, 1454
1560 Hofmann B., Mathé P., 2018, *Inverse Problems*, 34, 015007
1561 Hoyle B., et al., 2018, *MNRAS*, 478, 592
1562 Hunter J. D., 2007, Matplotlib: A 2D Graphics Environment,
1563 doi:[10.1109/MCSE.2007.55](https://doi.org/10.1109/MCSE.2007.55)
1564 Ilbert O., et al., 2006, *A&A*, 457, 841
1565 Ivezić Ž., et al., 2008, preprint (<arXiv:0805.2366>),
1566 Izbicki R., Lee A. B., 2017, *Electron. J. Statist.*, 11, 2800
1567 Izbicki R., Lee A. B., Freeman P. E., 2017, *Ann. Appl. Stat.*, 11,
1568 698
1569 Laureijs R., et al., 2011, preprint (<1110.3193>),
1570 Leistedt B., Hogg D. W., 2017, *ApJ*, 838, 5
1571 Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon
1572 E. S., 2008, *MNRAS*, 390, 118
1573 Malz A., Marshall P., 2018, qp: Quantile parametrization for
1574 probability distribution functions (ascl:1809.011)
1575 Malz A. I., Marshall P. J., DeRose J., Graham M. L., Schmidt
1576 1577
- 1578 1579 1580 1581
- 1507 REFERENCES
- 1508 Abbott T., et al., 2005, preprint (<arXiv:astro-ph/0510346>)
1509 Abell P. A., et al., 2009, preprint (<arXiv:0912.0201>),
1510 Aihara H., et al., 2018a, *PASJ*, 70, S4
1511 Aihara H., et al., 2018b, *PASJ*, 70, S8
1512 Almosallam I. A., Lindsay S. N., Jarvis M. J., Roberts S. J.,
1513 2016a, *MNRAS*, 455, 2387
- ¹¹ available at: <http://ls.st/srd>

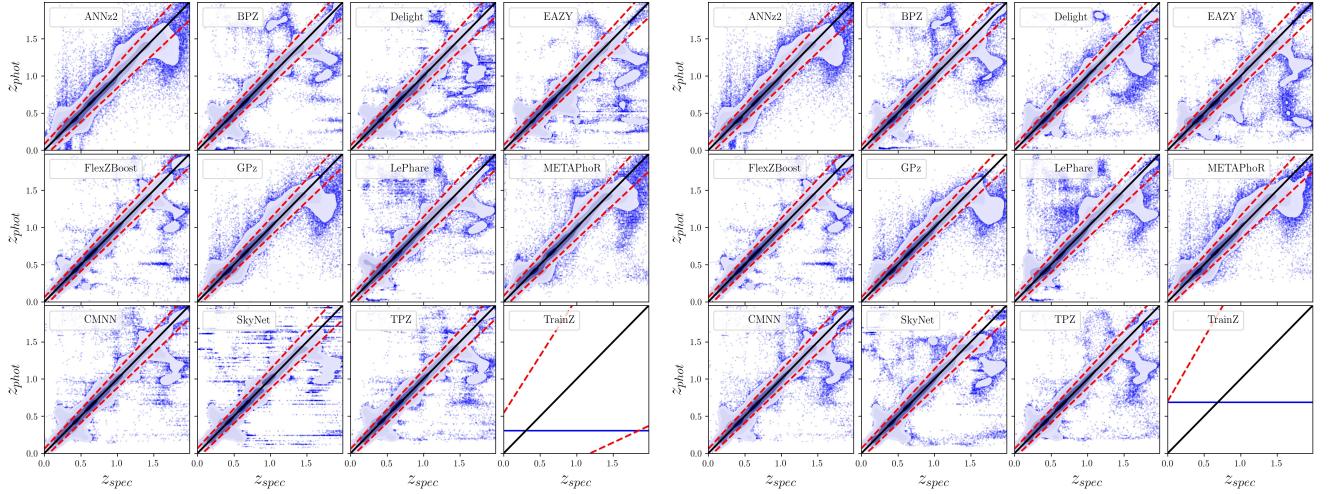


Figure B1. The density of photo- z point estimates (contours) reduced from the photo- z PDFs with outliers (blue) beyond the outlier cutoff (red dashed lines), via the mode (z_{PEAK} , left panel) and main-peak-mean (z_{WEIGHT} , right panel). The **trainZ** estimator (lower right sub-panels) has a shared z_{PEAK} and z_{WEIGHT} for the entire test set galaxy sample.

Table B1. Photo- z point estimate statistics

Photo- z PDF Code	Z_{PEAK}		Z_{WEIGHT}			
	$\frac{\sigma_{IQR}}{(1+z)}$	median	outlier fraction	$\frac{\sigma_{IQR}}{(1+z)}$	median	outlier fraction
ANNz2	0.0270	0.00063	0.044	0.0244	0.000307	0.047
BPZ	0.0215	-0.00175	0.035	0.0215	-0.002005	0.032
Delight	0.0212	-0.00185	0.038	0.0216	-0.002158	0.038
EAZY	0.0225	-0.00218	0.034	0.0226	-0.003765	0.029
FlexZBoost	0.0154	-0.00027	0.020	0.0148	-0.000211	0.017
GPz	0.0197	-0.00000	0.052	0.0195	0.000113	0.051
LePhare	0.0236	-0.00161	0.058	0.0239	-0.002007	0.056
METAPhoR	0.0264	0.00000	0.037	0.0262	0.001333	0.048
CMNN	0.0184	-0.00132	0.035	0.0170	-0.001049	0.034
SkyNet	0.0219	-0.00167	0.036	0.0218	0.000174	0.037
TPZ	0.0161	0.00309	0.033	0.0166	0.003048	0.031
trainZ	0.1808	-0.2086	0.000	0.2335	0.022135	0.000

- 1582 S. J., Wechsler R., (LSST Dark Energy Science Collaboration 1603
1583 2018, *AJ*, **156**, 35 1604 Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013,
1584 Mandelbaum R., et al., 2008, *MNRAS*, **386**, 781 1605 *ApJ*, **771**, 30
1585 Mao Y.-Y., Williamson M., Wechsler R. H., 2015, *ApJ*, **810**, 21 1606 Sadeh I., Abdalla F. B., Lahav O., 2016, *PASP*, **128**, 104502
1586 Massarotti M., Iovino A., Buzzoni A., 2001, *A&A*, **368**, 74 1607 Sánchez C., et al., 2014, *MNRAS*, **445**, 1482
1587 Masters D., et al., 2015, *ApJ*, **813**, 53 1608 Schmidt M., 2005, minFunc: Unconstrained Differentiable Multi-
1588 Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes 1609 variate Optimization in Matlab, <http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>
1589 J. D., Castander F. J., Paltani S., 2017, *ApJ*, **841**, 111 1610 Scott D. W., 1992, Multivariate Density Estimation. Theory,
1590 Newman J. A., 2008, *ApJ*, **684**, 88 1611 Practice, and Visualization. Wiley
1591 Newman J. A., et al., 2015, *Astroparticle Physics*, **63**, 81 1612 Sheldon E. S., Cunha C. E., Mandelbaum R., Brinkmann J.,
1592 Oliphant T., 2007, Python for Scientific Computing, 1613 Weaver B. A., 2012, *The Astrophysical Journal Supplement Series*, **201**, 32
1593 doi:10.1109/MCSE.2007.58 1614 Skrutskie M. F., et al., 2006, *AJ*, **131**, 1163
1594 Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, *ApJ*, 1615 Tanaka M., et al., 2018, *PASJ*, **70**, S9
1595 **689**, 709 1616 The LSST Dark Energy Science Collaboration et al., 2018,
1596 Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint 1617 preprint, ([arXiv:1809.01669](https://arxiv.org/abs/1809.01669))
1597 (arXiv:1608.08016), 1618 Waskom M., et al., 2017, doi:10.5281/zenodo.824567
1598 Rasmussen C., Williams C., 2006, Gaussian Processes for Machine 1619 York D. G., et al., 2000, *AJ*, **120**, 1579
1599 Learning. Adaptative computation and machine learning se- 1620 de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn
1600 ries, MIT Press, Cambridge, MA 1621 E. A., 2013, *Exp. Astron.*, **35**, 25
1601 Rau M. M., Seitz S., Brimioule F., Frank E., Friedrich O., Gruen 1622 de Jong J. T. A., et al., 2017, *A&A*, **604**, A134
1602 D., Hoyle B., 2015, *MNRAS*, **452**, 3710 1623