

Evaluation of probabilistic photometric redshift estimation approaches for LSST

S.J. Schmidt¹, A.I. Malz^{2,3,4}, J.Y.H. Soo⁵, I.A. Almosallam^{6,7}, M. Brescia⁸, S. Cavaudi^{8,9}, J. Cohen-Tanugi¹⁰, A.J. Connolly¹¹, P.E. Freeman¹², M.L. Graham¹¹, K. Iyer¹³, M.J. Jarvis^{14,15}, J.B. Kalmbach¹⁶, E. Kovacs¹⁷, A.B. Lee¹², G. Longo⁹, C. B. Morrison¹¹, J. Newman¹⁸, E. Nourbakhsh¹, E. Nuss¹⁰, T. Pospisil¹², H. Tranin¹⁰, R. Zhou¹⁸, R. Izbicki^{19,20}

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

⁶ King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

⁷ Information Engineering, Parks Road, Oxford, OX1 3PJ, UK

⁸ INAF-Capodimonte Observatory, Salita Moiariello 16, I-80131, Napoli, Italy

⁹ Department of Physics E. Pancini, University Federico II, via Cinthia 6, I-80126, Napoli, Italy

¹⁰ Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France

¹¹ Department of Astronomy, University of Washington, Box 351580, U.W., Seattle WA 98195, USA

¹² Department of Statistics & Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

¹³ Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA

¹⁴ Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK

¹⁵ Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa

¹⁶ Department of Physics, University of Washington, Box 351560, Seattle, WA 98195, USA

¹⁷ Argonne National Laboratory, Lemont, IL 60439, USA

¹⁸ Department of Physics and Astronomy and the Pittsburgh Particle Physics, Astrophysics and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA 15260, USA

¹⁹ Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil

²⁰ External collaborator

4 September 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.

Key words: galaxies: distances and redshifts – galaxies: statistics – methods: statistical

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), present a paradigm shift from spectroscopic to photometric 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 $\sim 10^7$ 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%, 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 subsamples populations 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, this degeneracy is a real consequence 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, later photometric galaxy catalogue data releases, (e. g. 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 observed 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. In fact, 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). This caveat aside, quantitative 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. We aim to perform a comprehensive sensitivity analysis of photo- z PDF techniques in order to ultimately select those that will become part of the LSST pipelines, as part of a key project of the Photometric Redshifts working group of the LSST-DESC, described in the Science Roadmap (SRM)². In this initial study, we focus on evaluating the performance of photo- z PDF codes using PDF-specific performance metrics in a 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.

2 DATA

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

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

² Available at: http://lsst-desc.org/sites/default/files/DESC_SRMs_V1_1.pdf

112 2.1 The Buzzard-v1.0 simulation

113 Our mock catalogue is derived from the BUZZARD-highres-
 114 v1.0 catalogue (DeRose et al. 2019, Wechsler et al., in prep).
 115 BUZZARD is built on a dark matter-only N-body simulation
 116 of 2048^3 particles in a 400 Mpc h^{-1} box. The lightcone was
 117 constructed from smoothing and interpolation between a set
 118 of time snapshots. Dark matter halos were identified using
 119 the Rockstar software package (Behroozi et al. 2013) and
 120 then populated with galaxies with a stellar mass and ab-
 121 solute r -band magnitude in the SDSS system determined
 122 using a sub-halo abundance matching model constrained to
 123 match both projected two-point galaxy clustering statistics
 124 and an observed conditional stellar mass function (Reddick
 125 et al. 2013).

126 To assign a spectrum to each galaxy, the Adding Den-
 127 sity Dependent Spectral Energy Distributions (SEDs) pro-
 128 cedure (ADDSEDS, deRose in prep.)³ was used. ADDSEDS uses
 129 a sample of $\sim 5 \times 10^5$ galaxies from the magnitude-limited
 130 SDSS Data Release 6 Value Added Galaxy Catalogue (Blan-
 131 ton et al. 2005) to train an empirical relation between abso-
 132 lute r -band magnitude, local galaxy density, and SED. Each
 133 SDSS spectrum is parameterized by five weights correspond-
 134 ing to a weighted sum of five basis SED components using
 135 the k-correct software package⁴ (Blanton & Roweis 2007).

136 Correlations between SED and galaxy environment
 137 were included so as to preserve the colour-density relation of
 138 galaxy environment. The distance to the spatially projected
 139 fifth-nearest neighbour was used as a proxy for local density
 140 in the SDSS training sample. For each simulated galaxy,
 141 a galaxy with similar absolute r -band magnitude and local
 142 galaxy density was chosen from the training set, and that
 143 training galaxy's SED was assigned to the simulated galaxy.
 144 In Section 2.1.1, we critique the realism of this mock data.

145 2.1.1 Caveats

146 By necessity, BUZZARD does not contain all of the compli-
 147 cating factors present in real data, and here we discuss the
 148 most pertinent ways that this limitation affects our exper-
 149 iment. BUZZARD includes only galaxies, not stars or AGN.
 150 The catalogue-based construction excludes image-level ef-
 151 fects, such as deblending errors, photometric measurement
 152 issues, contamination from sky background (Zodiacal light,
 153 scattered light, etc.), lensing magnification, and Galactic
 154 reddening.

155 The SEDs are five-component linear combinations of
 156 $\sim 5 \times 10^5$ SDSS galaxies, so the sample contains only galax-
 157 ies that resemble linear combinations of those for which
 158 SDSS obtained spectra. The linear combination SEDs also
 159 restrict the properties of the galaxy population to linear
 160 combinations of the properties corresponding to five basis
 161 templates, precluding the modeling of non-linear features
 162 such as the full range of emission line fluxes relative to the
 163 continuum. The only form of intrinsic dust reddening comes
 164 from what is already present in the five basis SEDs via the
 165 training set used to create the basis templates, and linear
 166 combinations thereof do not span the full range of realistic
 167 dust extinction observed in galaxy populations.

³ <https://github.com/vipasu/addsed>

⁴ <http://kcorrect.org>

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

173 2.2 LSST-like mock observations

174 Given the SED, absolute r -band magnitude, and redshift,
 175 we computed apparent magnitudes in the six LSST filter
 176 passbands, $ugrizy$. We assigned magnitude errors in the six
 177 bands using the simple model of Ivezić et al. (2008), assum-
 178 ing achievement of the full 10-year depth, with a modifica-
 179 tion of fiducial LSST total numbers of 30-second visits for
 180 photometric error generation: we assume 60 visits in u -band,
 181 80 visits in g -band, 180 visits in r -band, 180 visits in i -band,
 182 160 visits in z -band, and 160 visits in y -band.

183 As a consequence of adding Gaussian-distributed pho-
 184 tometric errors, 2.0% of our galaxies exhibit a negative flux
 185 in one or more bands, the vast majority of which are in the
 186 u -band. We deem such negative fluxes *non-detections* and
 187 assign a placeholder magnitude of 99.0 in the catalogue to
 188 indicate to the photo-z PDF codes that such galaxies would
 189 be “looked at but not seen” in multi-band forced photome-
 190 try.

191 The full dataset thus covers 400 square degrees and con-
 192 tains 238 million galaxies of redshift $0 < z \leq 8.7$ down
 193 to $r = 29$. Systematic inconsistencies with galaxy colors
 194 at $z > 2$ were observed, so the catalogue was limited to
 195 $0 < z \leq 2.0$. To obtain a catalogue matching the LSST
 196 Gold Sample, we imposed an cut of $i < 25.3$, which gives a
 197 signal-to-noise ratio ~ 30 for most galaxies. In order for sta-
 198 tistical errors to be subdominant to the systematic errors we
 199 aim to probe, we further reduced the sample size to $< 10^7$
 200 galaxies by isolating ~ 16.8 square degrees selected from five
 201 separate spatial regions of the simulation. We refer to this
 202 final set of galaxies as DC1, for the first LSST-DESC Data
 203 Challenge.

204 2.3 Shared prior information

205 For the purpose of performing a controlled experiment that
 206 compares photo-z PDF codes on equal footing as a baseline
 207 for a future sensitivity analysis, we take care to provide each
 208 with maximally optimistic prior information. Redshift esti-
 209 mation approaches built upon physical modeling and ma-
 210 chine learning alike have a notion of prior information con-
 211 sidered beyond the photometry of the data for which redshift
 212 is to be constrained: that information is derived from a tem-
 213 plate library for a model-based code and a training set for
 214 a data-driven code. In this initial study, we seek to set a
 215 baseline for a later comparison of the performance of photo-
 216 z PDF codes under incomplete and non-representative prior
 217 information that will propagate differently in the space of
 218 data-driven and model-based algorithms. However, for the
 219 baseline case of perfect prior information, physical model-
 220 ing and machine learning codes can indeed be put on truly
 221 equal footing. We outline the equivalent ways of providing
 222 all codes perfect prior information below.

4 LSST Dark Energy Science Collaboration

223 2.3.1 Training and test set division

224 Following the findings of Bernstein & Huterer (2010), Masters et al. (2017) that only 10^4 spectra are necessary to
 225 calibrate photo- z s to Stage IV requirements, we aimed to
 226 set aside a randomly selected training set of $3 - 5 \times 10^4$
 227 galaxies, $\sim 10\%$ of the full sample. After all cuts described
 228 above, we designated the *DC1 training set* of 44 404 galaxies
 229 for which observed photometry, true SEDs, and true red-
 230 shifts would be provided to all codes and the blinded
 231 *DC1 test set* of 399 356 galaxies for which photometry alone
 232 would be provided to all codes and photo- z PDFs would be
 233 requested. The LSST photometric filter transmission curves
 234 were also considered public information that could be used
 235 by any code.

237 2.3.2 Template library construction

238 We aimed to provide template-fitting codes with complete
 239 yet manageable library of templates spanning the space of
 240 SEDs of the DC1 galaxies. We constructed $K = 100$ repre-
 241 sentative templates from the $\sim 5 \times 10^5$ SEDs of the SDSS
 242 DR6 NYU-VAGC by using the five-dimensional vectors of
 243 SED weight coefficients described above. After regularizing
 244 the SED weight coefficients $\in [0, 1]$, we ran a simple K-
 245 means clustering algorithm on the five-dimensional space
 246 of regularized SED weight coefficients of the SDSS galaxy
 247 sample. The resulting clusters were used to define Voronoi
 248 cells in the space of weight coefficients, with centre pos-
 249 tions corresponding to weights for the k-correct SED com-
 250 ponents, yielding the 100 *DC1 template set* to be provided
 251 to all template-based codes. We did not, however, exclude
 252 from consideration template-based codes that made modi-
 253 fications in their use of these templates due to architecture
 254 limitations (as opposed to knowledge of the experimental
 255 conditions that could artificially boost the code’s apparent
 256 performance), with deviations noted in Section 3.

257 3 METHODS

258 Here we summarize the twelve photo- z PDF codes compared
 259 in this study, also in Table 1, which include both estab-
 260 lished and emerging approaches in template fitting and
 261 machine learning. Though not exhaustive, this sample rep-
 262 presents codes for which there was sufficient expertise within
 263 the LSST-DESC Photometric Redshifts Working Group.
 264 Some aspects of data treatment were left to the individual
 265 code runners, for example, whether/how to split the avail-
 266 able data with known redshifts into separate training and
 267 validation sets.

268 Another key difference is the treatment of non-
 269 detections in one or more bands: some codes ignore incom-
 270 plete bands, while others replace the value with either an
 271 estimate for the detection limit, the mean of other values in
 272 the training set, or another default value. There are varying
 273 conventions among machine learning based codes for treat-
 274 ment of non-detections, and no one prescription dominates
 275 in the photo- z literature. The specific choices for each code
 276 affect the results and contribute to the implicit prior influ-
 277 encing their output. However, we remind the reader that
 278 only 2.0 per cent of our sample has non-detections, almost

279 exclusively in the u -band, and thus should not dominate the
 280 code performance differences. The authors welcome interest
 281 from those outside LSST-DESC to have their codes assessed
 282 in future investigations that build upon this one.

283 We describe the algorithms and implementations of the
 284 model-based and data-driven codes in Sections 3.1 and 3.2
 285 respectively, with a straw-person approach included in Sec-
 286 tion 3.3.

287 3.1 Template-based Approaches

288 We test three publicly available and commonly used
 289 template-based codes that share the standard physically moti-
 290 vated approach of calculating model fluxes for a set of tem-
 291 plate SEDs on a grid of redshift values and evaluating a χ^2
 292 merit function using the observed and model fluxes:

$$293 \chi^2(z, T, A) = \sum_i^{N_{\text{filt}}} \left(\frac{F_{\text{obs}}^i - A F_{\text{pred}}^i(T, z)}{\sigma_{\text{obs}}^i} \right)^2 \quad (1)$$

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

3.1.1 LePhare

304 Photometric Analysis for Redshift Estimate (LePhare⁵,
 305 Arnouts et al. 1999; Ilbert et al. 2006) uses the χ^2 of Equa-
 306 tion 1 to match observed colors with those predicted from
 307 a template set, which can be semi-empirical or entirely syn-
 308 synthetic, directly according to the The reported photo- z PDF
 309 is an arbitrary normalization of the likelihood evaluated on
 310 the output redshift grid.

311 Here we use LePhare-v 2.2 with the DC1 template set
 312 of Section 2.3.2.

314 3.1.2 BPZ

315 Bayesian Photometric Redshift (BPZ⁶, Benítez 2000) de-
 316 termines the likelihood $p(C|z, T)$ of a galaxy’s observed
 317 colours C for a set of SED templates T at redshifts
 318 z . The BPZ likelihood is related to the χ^2 likelihood by
 319 $p(C|z, T) \propto \exp[-\chi^2/2]$. Given a Bayesian prior $p(z, T|m_0)$
 320 over apparent magnitude m_0 and type T , and assuming
 321 that the SED templates are spanning and exclusive, BPZ
 322 constructs the redshift posterior $p(z|C, m_0)$ by marginaliz-
 323 ing over all SED templates with the form $p(z|C, m_0) \propto$
 324 $\sum_T p(C|z, T) p(z, T|m_0)$ (Eq. 3 from Benítez 2000), corre-
 325 sponding to setting the parameter PROBS_LITE=TRUE in the
 326 BPZ parameter file. The BPZ prior is the product of an SED
 327 template proportion that varies with apparent magnitude

5 <http://www.cfht.hawaii.edu/~arnouts/lephare.html>

6 <http://www.stsci.edu/~dcoe/BPZ/>

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

Published code	Type	Public source code
LePhare (Arnouts et al. 1999)	template fitting	http://www.cfht.hawaii.edu/~arnouts/lephare.html
BPZ (Benítez 2000)	template fitting	http://www.stsci.edu/~dcoe/BPZ/
EAZY (Brammer et al. 2008)	template fitting	https://github.com/gbrammer/eazy-photoz
ANNz2 (Sadeh et al. 2016)	machine learning	https://github.com/IftachSadeh/ANNZ
FlexZBoost (Izbicki & Lee 2017)	machine learning	https://github.com/tospis/i/flexcode ; https://github.com/rizbicki/FlexCoDE
GPz (Almosallam et al. 2016b)	machine learning	https://github.com/OxfordML/GPz
METAPhoR (Cavuoti et al. 2017)	machine learning	http://dame.dsf.unina.it
CMNN (Graham et al. 2018)	machine learning	N/A
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
Delight (Leistedt & Hogg 2017)	hybrid	https://github.com/ixkael/Delight
trainZ	machine learning	See Section 3.3

³²⁸ $p(T|m_0)$ and a prior $p(z|T, m_0)$ over the expected redshift
³²⁹ as a function of apparent magnitude and SED template.

³³⁰ Here we test BPZ-v 1.99.3 with the DC1 template set
³³¹ of Section 2.3.2. To keep the number of free parameters
³³² manageable, the DC1 template set is pre-sorted by the rest-
³³³ frame $u - g$ colour and split into three broad classes of SED
³³⁴ template, equivalent to the E, Sp and Im/SB types. The
³³⁵ Bayesian prior term $p(T|m_0)$ was derived directly from the
³³⁶ DC1 training set, and the other term $p(z|T, m_0)$ was cho-
³³⁷ sen to be the best fit for the eleven free parameters of the
³³⁸ functional form of Benítez (2000). We use template inter-
³³⁹ polation, creating two linearly interpolated templates between
³⁴⁰ each basis SED (sorted by rest-frame $u - g$ colour) by set-
³⁴¹ ting the parameter INTERP=2. Prior to running the code,
³⁴² the non-detection placeholder magnitude was replaced with
³⁴³ an estimate of the one- σ detection limit for the undetected
³⁴⁴ band as a proxy for a value close to the estimated sky noise
³⁴⁵ threshold.

³⁴⁶ 3.1.3 EAZY

³⁴⁷ Easy and Accurate Photometric Redshifts from Yale (EAZY⁷,
³⁴⁸ Brammer et al. 2008) extends the basic χ^2 fit procedure that
³⁴⁹ defines template-fitting approaches. The algorithm models
³⁵⁰ the observed photometry with a linear combination of tem-
³⁵¹ plate SEDs at each redshift. The best-fit SED is found by
³⁵² simultaneously fitting one, two, or all of the templates via χ^2
³⁵³ minimization, which is distinct from marginalizing across all
³⁵⁴ templates. The minimized χ^2 likelihood at each redshift is
³⁵⁵ then combined with an apparent magnitude prior to obtain
³⁵⁶ the redshift posterior PDF. We note that the utilization of
³⁵⁷ the best-fit SED rather than a proper marginalization does
³⁵⁸ not lead to the correct posterior distribution, an implemen-
³⁵⁹ tation issue that has now been identified and will be ad-
³⁶⁰ dressed by the developers in the future. EAZY can account
³⁶¹ for uncertainty in the template set by adding in quadrature
³⁶² to the flux errors an empirically derived template error as a
³⁶³ function of redshift.

³⁶⁴ The SED-independent apparent magnitude prior was
³⁶⁵ derived empirically from the DC1 training set. The EAZY ar-
³⁶⁶ chitecture cannot accept a template set other than the same
³⁶⁷ five basis templates employed by k-correct when construct-
³⁶⁸ ing the DC1 catalogue. However, EAZY does feature a flexible
³⁶⁹ all-templates mode, which fits the photometric data with

³⁷⁰ a linear combination of the five basis templates. We set the
³⁷¹ template error to zero since the same templates were in fact
³⁷² used to produce the DC1 photometry.

³⁷³ 3.2 Machine Learning-based Approaches

³⁷⁴ We compared nine data-driven photo- z estimation ap-
³⁷⁵ proaches, eight of which are described in this section and one
³⁷⁶ of which is discussed in Section 3.3. Because the algorithms
³⁷⁷ differ more from one another and the techniques are rela-
³⁷⁸ tive newcomers to the astronomical literature, we provide
³⁷⁹ somewhat more detail about the implementations below.

³⁸⁰ 3.2.1 ANNz2

³⁸¹ ANNz2⁸ (Sadeh et al. 2016) employs several machine learn-
³⁸² ing algorithms, including artificial neural networks (ANN),
³⁸³ boosted decision tree, and k-nearest neighbour (KNN) re-
³⁸⁴ gression. In addition to accounting for errors on the input
³⁸⁵ photometry, ANNz2 uses the KNN-uncertainty estimate of
³⁸⁶ Oyaizu et al. (2008) to quantify uncertainty in the choice of
³⁸⁷ method over multiple runs. Using the Toolkit for Multivariate
³⁸⁸ Data Analysis with ROOT⁹, ANNz2 can return the results
³⁸⁹ of running a single machine learning algorithm, a “best”
³⁹⁰ choice of the results from simultaneously running multiple
³⁹¹ algorithms (based on evaluation the cumulative distribution
³⁹² functions of validation set objects), or a combination of the
³⁹³ results of multiple algorithms weighted by their method un-
³⁹⁴ certainty averaged over multiple runs.

³⁹⁵ In this study, we used ANNz2-v.2.0.4 to output only the
³⁹⁶ result of the ANN algorithm. Photo- z PDFs were produced
³⁹⁷ by running an ensemble of 5 ANNs with a 6 : 12 : 12 : 1
³⁹⁸ architecture corresponding to the 6 $ugrizy$ inputs, 2 hidden
³⁹⁹ layers with 12 nodes each, and 1 output of redshift. Each
⁴⁰⁰ of the five ANNs was trained with different random seeds
⁴⁰¹ for the initialization of input parameters. Additionally, all
⁴⁰² ANNs were trained on only a $i \leq 25.3$ subsample of the DC1
⁴⁰³ training set, and half of the training set was reserved for
⁴⁰⁴ validation to prevent overfitting. Undetected galaxies were
⁴⁰⁵ excluded from the training set, and per-band non-detections
⁴⁰⁶ in the test set were replaced with the mean magnitude in
⁴⁰⁷ that band within the entire test set.

⁷ <https://github.com/gbrammer/eazy-photoz>

⁸ <https://github.com/IftachSadeh/ANNZ>

⁹ <http://tmva.sourceforge.net/>

6 LSST Dark Energy Science Collaboration

408 3.2.2 Colour-Matched Nearest-Neighbours

409 The nearest-neighbours colour-matching photometric red-
 410 shift estimator (CMNN, Graham et al. 2018) uses a training
 411 set of galaxies with known redshifts that has equivalent or
 412 better photometry than the test set in terms of quality and
 413 filter coverage. For each galaxy in the test set, CMNN identifies
 414 a colour-matched subset of training galaxies using a thresh-
 415 old in the Mahalanobis distance $D_M = \sum_j^{N_{\text{colours}}} (c_j^{\text{train}} -$
 416 $c_j^{\text{test}})^2 / \delta c_{\text{test}}^2$ in the space of available colours c , with colour
 417 measurement errors δc_{test} and $N_{\text{colours}} = 5$ colors j defined
 418 by the $ugrizy$ filters, which defines the set of colour-matched
 419 neighbours based on a value of the percent point function
 420 (PPF). As an example, for $N_{\text{fit}} = 5$ with PPF= 0.95, 95%
 421 of all training galaxies consistent with the test galaxy will
 422 have $D_M < 11.07$. Undetected bands are dropped, thereby
 423 reducing the effective N_{fit} for that galaxy. The photo- z PDF
 424 of a given test set galaxy is the normalized distribution of
 425 redshifts of its colour-matched subset of training set galax-
 426 ies.

427 Here, we make two modifications to the implementation
 428 of Graham et al. (2018) to comply with the controlled exper-
 429 imental conditions. First, we do not impose non-detections
 430 on galaxies fainter than the expected LSST 10-year limit-
 431 ing magnitude nor galaxies bright enough to saturate with
 432 LSST’s CCDs, instead using all of the photometry for the
 433 DC1 test and training sets. Second, we apply the initial
 434 colour cut to the training set before calculating the Maha-
 435 lanobis distance in order to accelerate processing and use a
 436 magnitude pseudo-prior as in Graham et al. (2018), but for
 437 both we use cut-off values corresponding to the DC1 training
 438 set galaxies’ colours and magnitudes.

439 We make an additional adaptation to enable the CMNN
 440 algorithm to yield accurate photo- z PDFs for all galaxies,
 441 as the original Graham et al. (2018) algorithm is optimized
 442 for photo- z point estimates and is susceptible to less ac-
 443 curate photo- z PDFs for bright galaxies or those with few
 444 matches in colour-space. We use PPF= 0.95 rather than
 445 PPF= 0.68 to generate the subset of colour-matched train-
 446 ing galaxies, whose redshifts are weighted by their inverse
 447 Mahalanobis distances of the when composing the photo-
 448 z PDF rather than weighting all colour-matched training
 449 galaxies equally. Additionally, when the number of colour-
 450 matched training set galaxies is less than 20, the nearest 20
 451 neighbours in color-space are used instead, and the output
 452 photo- z PDF is convolved with a Gaussian kernel of variance
 453 $\sigma_{\text{train}}^2 (\text{PPF}_{20}/0.95)^2 - 1$ to account for the correspond-
 454 ing growth of the effective PPF to include 20 neighbors.

455 3.2.3 FlexZBoost

456 **FlexZBoost**¹⁰ (Izbicki & Lee 2017) is built on **FlexCode**, a
 457 general-purpose methodology for converting any conditional
 458 mean point estimator of z to a conditional density estima-
 459 tor $p(z|\mathbf{x}) \equiv f(z|\mathbf{x})$, where \mathbf{x} here represents our photomet-
 460 ric covariates and errors. **FlexZBoost** expands the unknown
 461 function $f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x})\phi_i(z)$ using an orthonormal ba-
 462 sis $\{\phi_i(z)\}_i$. By the orthogonality property, the expansion

463 coefficients $\beta_i(\mathbf{x}) = \mathbb{E}[\phi_i(z)|\mathbf{x}] \equiv \int f(z|\mathbf{x})\phi_i(z)dz$ are thus
 464 conditional means. The expectation value $\mathbb{E}[\phi_i(z)|\mathbf{x}]$ of the
 465 expansion coefficients conditioned on the data is equivalent
 466 to the regression of the space of possible redshifts on the
 467 space of possible photometry. Thus the expansion coeffi-
 468 cients $\beta_i(\mathbf{x})$ can be estimated from the data via regression
 469 to yield the conditional density estimate $\hat{f}(z|\mathbf{x})$.

470 In this paper, we used **xgboost** (Chen & Guestrin
 471 2016) for the regression; it should however be noted that
 472 **FlexCode-RF**¹⁰, based on Random Forests, generally per-
 473 forms better for smaller datasets. As our basis $\phi_i(z)$, we
 474 choose a standard Fourier basis. The two tuning parame-
 475 ters in our photo- z PDF estimate are the number I of terms
 476 in the series expansion and an exponent α that we use to
 477 sharpen the computed density estimates $\tilde{f}(z|\mathbf{x}) \propto \hat{f}(z|\mathbf{x})^\alpha$.
 478 Both I and α were chosen in an automated way by mini-
 479 mizing the weighted L_2 -loss function (Eq. 5 in Izbicki & Lee
 480 2017) on a validation set comprised of a randomly selected
 481 15% of the DC1 training set. While **FlexCode**’s lossless na-
 482 tive encoding stores each photo- z PDF using the basis co-
 483 efficients $\beta_i(\mathbf{x})$, we discretized the final estimates into 200
 484 linearly-spaced redshift bins $0 < z < 2$ to match the consis-
 485 tent output format of the experimental conditions.

486 3.2.4 GPz

487 **GPz**¹¹ (Almosallam et al. 2016a,b) is a sparse Gaussian pro-
 488 cess based code, a scalable approximation of full Gaus-
 489 sian Processes (Rasmussen & Williams 2006), that pro-
 490 duces input-dependent variance estimates corresponding to
 491 heteroscedastic noise. The model assumes a Gaussian pos-
 492 terior probability $p(z|\mathbf{x}) = \mathcal{N}(z|\mu(\mathbf{x}), \sigma(\mathbf{x})^2)$ of the out-
 493 put redshift z given the input photometry \mathbf{x} . The mean
 494 $\mu(\mathbf{x})$ and the variance $\sigma(\mathbf{x})^2$ are modeled as functions
 495 $f(\mathbf{x}) = \sum_{i=1}^m w_i \phi_i(\mathbf{x})$ linear combinations of m basis func-
 496 tions $\{\phi_i(\mathbf{x})\}_{i=1}^m$ with associated weights $\{w_i\}_{i=1}^m$. The de-
 497 tails on how to learn the parameters of the model and the
 498 hyper-parameters of the basis functions are described in Al-
 499 mosallam et al. (2016b). **GPz**’s variance estimate is composed
 500 of a model uncertainty term corresponding to sparsity of the
 501 training set photometry and a noise uncertainty term en-
 502 compassing noisy photometric observations, enabling quan-
 503 tification of any need for more representative or more precise
 504 training samples. **GPz** may also weight training set samples
 505 by importance according to $|z_{\text{spec}} - z_{\text{phot}}|/(1+z_{\text{spec}})$ to min-
 506 imize the normalized photo- z point estimate error, however,
 507 this function may be adapted to photo- z PDFs, pressuring
 508 the model to dedicate more resources to test set galaxies
 509 that are not well-represented in the training set.

510 To smooth the long tail in the distribution of magni-
 511 tude errors, we use the log of the magnitude errors, im-
 512 proving numerical stability and eliminating the need for
 513 constraints on the optimization process. Unobserved mag-
 514 nitudes $x_u = \mu_u + \Sigma_{uo}\Sigma_{oo}^{-1}(x_o - \mu_o)$ were imputed from
 515 observed magnitudes x_o and the training set mean μ and
 516 covariance Σ using a linear model. This is the optimal ex-
 517 pected value of the unobserved variables given the observed
 518 ones under the assumption that the distribution is jointly
 519 Gaussian; note that this reduces to a simple average if the

10 <https://github.com/tpospisi/flexcode>;
<https://github.com/rizbicki/FlexCoDE>

11 <https://github.com/OxfordML/GPz>

520 covariates are independent with $\Sigma_{\text{uo}} = 0$. We reserved for
 521 validation 20% of the training set and used the Variable
 522 Covariance option in GPz with 200 basis functions (see Al-
 523 mosallam et al. (2016b) for details), neglecting to apply cost-
 524 sensitive learning options.

525 3.2.5 METAPhOr

526 Machine-learning Estimation Tool for Accurate Photometric
 527 Redshifts (METAPhOr¹², Cavuoti et al. 2017) is based on
 528 the Multi Layer Perceptron with Quasi Newton Algorithm
 529 (MLPQNA) with the least square error model and Tikhonov
 530 L_2 -norm regularization (Hofmann & Mathé 2018). Photo-z
 531 PDFs are generated by running N trainings on the same
 532 training set, or M trainings on M different random sam-
 533 plings of the training set. Upon regression of the test set,
 534 the photometry m_{ij} of each test set galaxy j in filter i is
 535 perturbed according to $m'_{ij} = m_{ij} + \alpha_i F_{ij} \epsilon$ in terms of
 536 the standard normal random variable $\epsilon \sim \mathcal{N}(0, 1)$, a mul-
 537 tiplicative constant α_i permitting accommodation of multi-
 538 survey photometry, and a bimodal function F_{ij} composed of
 539 a polynomial fit of the mean magnitude errors on the binned
 540 bands plus a constant term representing the threshold be-
 541 low which the polynomial's noise contribution is negligible
 542 (Brescia et al. 2018).

543 In this work, we used a hierarchical KNN to replace
 544 non-detections with values based on their neighbors. The
 545 usual cross-validation of redshift estimates and PDFs was
 546 also omitted for this study.

547 3.2.6 SkyNet

548 SkyNet¹³ (Graff et al. 2014) employs a neural network based
 549 on a second order conjugate gradient optimization scheme
 550 (see Graff et al. 2014, for further details). The neural net-
 551 work is configured as a standard multilayer perceptron with
 552 three hidden layers and one input layer with 12 nodes cor-
 553 responding to the 6 photometric magnitudes and their mea-
 554 surement errors. We use SkyNet as a regressor for photo-z
 555 point estimation and as a classifier for photo-z PDF estima-
 556 tion.

557 The regressor used a standard χ^2 error function with a
 558 single linear node as the output layer and 10 nodes with a
 559 tanh activation function for each hidden layer. The classifier
 560 used a cross-entropy error function with a 20:40:40 node (all
 561 rectified linear units) architecture for each hidden layer and
 562 an output layer of 200 nodes corresponding to 200 bins for
 563 the PDF, with a softmax activation function to enforce the
 564 normalization condition that the probabilities sum to unity.
 565 While previous implementations of the code (see Appendix
 566 C.3 of Sánchez et al. 2014; Bonnett 2015) implement a
 567 sliding bin smoothing, no such procedure was used in this
 568 study.

569 We pre-whitened the data by pegging the magnitudes
 570 to (45,45,40,35,42,42) and errors to (20,20,10,5,15,15) for
 571 *ugrizy* filters, respectively. To avoid over-fitting, 30% of the
 572 training set was reserved for validation, and training was
 573 halted as soon as the error rate began to increase on the

574 validation set. The weights were randomly initialized based
 575 on normal sampling.

576 3.2.7 TPZ

577 Trees for Photo-z(TPZ¹⁴, Carrasco Kind & Brunner 2013;
 578 Carrasco Kind & Brunner 2014) uses prediction trees and
 579 random forest techniques to estimate photo-z PDFs. TPZ re-
 580 cursively splits the training set into branch pairs based on
 581 maximizing information gain among a random subsample of
 582 features, to minimize correlation between the trees, termi-
 583 nating only when a newly created leaf meets a criterion, such
 584 as a leaf size minimum or a variance threshold. The regions
 585 in each terminal leaf node correspond to a subsample of the
 586 training set with similar properties. Bootstrap samples from
 587 the training set photometry and errors are used to build a
 588 set of prediction trees.

589 To run TPZ, we replaced non-detections with an approxi-
 590 mation of the 1σ detection threshold based on the error fore-
 591 cast of the 10-year LSST data, i. e. $dm = 2.5 \log(1 + N/S)$
 592 where $dm \sim 0.7526$ magnitudes for $N/S = 1$. We cali-
 593 brated TPZ with the Out-of-Bag cross-validation technique
 594 (Breiman et al. 1984; Carrasco Kind & Brunner 2013) to
 595 evaluate its predictive validity and determine the relative
 596 importance of the different input attributes. We grew 100
 597 trees to a minimum leaf size of 5 using the *ugri* magnitudes,
 598 all $u - g, g - r, r - i, i - z, z - y$ colours, and the associated
 599 errors, as the z and y magnitudes did not show significant
 600 correlation with the redshift in our cross-validation. We par-
 601 titioned our redshift space into 200 bins and smoothed each
 602 individual PDF with a smoothing scale of twice the bin size.

603 3.2.8 Delight

604 Delight¹⁵ (Leistedt & Hogg 2017) is a hybrid technique that
 605 infers photo-zs with a data-driven model of latent SEDs and
 606 a physical model of photometric fluxes as a function of red-
 607 shift. Generally, machine learning methods rely on represen-
 608 tative training data with shared photometric filters, while
 609 template based methods rely on a complete library of tem-
 610 plates based on physical models constructed. Delight aims
 611 to take the best aspects of both approaches by construct-
 612 ing a large collection of latent SED templates (or physical
 613 flux-redshift models) from training data, with a template
 614 SED library as a guide to the learning of the model, thereby
 615 circumventing the machine learning prerequisite of represen-
 616 tative training data in the same photometric bands and the
 617 template fitting requirement of detailed galaxy SED models.
 618 It models noisy observed flux $\hat{\mathbf{F}} = \mathbf{F} + \mathbf{F}_b$ as a sum of a noise-
 619 less flux plus a Gaussian processes $\mathbf{F}_b \sim \mathcal{GP}(\mu^F, k^F)$ with
 620 zero mean function μ^F and a physically motivated kernel k^F
 621 that induces realistic correlations in flux-redshift space.

622 From a template-fitting perspective, each test set galaxy
 623 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-
 624 shift z conditioned on noisy flux $\hat{\mathbf{F}}$, where $p(z|T_i)p(T_i)$ cap-
 625 tures prior information about the redshift distributions and
 626 abundances of the galaxy templates T_i . As in traditional
 627 template fitting, each likelihood $p(\hat{\mathbf{F}}|\mathbf{F})$ relates the noisy flux

12 <http://dame.dsfs.unina.it>

13 <http://ccpforge.cse.rl.ac.uk/gf/project/skynet/>

14 <https://github.com/mgckind/MLZ>

15 <https://github.com/ixkael/Delight>

8 LSST Dark Energy Science Collaboration

$\hat{\mathbf{F}}$ with the noiseless flux \mathbf{F} predicted by the model of a linear combination of templates, carefully constructed to account for model uncertainties and different normalization of the same SED, plus the Gaussian process term.

The machine learning approach appears in the inclusion of a pairwise comparison term $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$ for the prediction of model flux \mathbf{F} at a model redshift z with respect to training set galaxy j with redshift z_j and observed flux $\hat{\mathbf{F}}_j$. Thus the photo- z posterior $p(\hat{\mathbf{F}}|z, T_i) = \int p(\hat{\mathbf{F}}|\mathbf{F})p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)d\mathbf{F}$ may be interpreted as the probability that the training and the target galaxies have the same SED at different redshifts. The flux prediction $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$ of the training galaxy at redshift z is modeled via the Gaussian process described above; more detail is provided in Leistedt & Hogg (2017).

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

3.3 trainZ: a pathological photo- z 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 $k = N_{\text{train}}$.

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. As a result, each observed galaxy i has a *true posterior photo- z PDF* $p(z|d_i)$ as well as a true likelihood $p(d|z_i)$. 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 is in general not accessible unless the photometry is in fact drawn from a ground

truth for the joint probability density of redshift and photometry $p(z, d)$. In contrast, existing mock catalogs 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 a number of things that will differ depending on the method H used to produce it, namely the often implicit assumptions I_H necessary for the method to be valid and any inputs I_D it takes as prior information, such as a template library or training set. Because of this, direct comparison of photo- z PDFs produced by different methods is in some sense 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. We call I_H the *implicit prior* specific to the method, though some aspects of its nature may be discerned.

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 will 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.¹⁶ We will discuss the choice of photo- z PDF parameterization further in Section 6.

This analysis is conducted using the `qp`¹⁷ software package (Malz et al. 2018) for manipulating and calculating met-

¹⁶ See, e. g. the LSST Data Products Definition Document, available at: <https://ls.st/dpdd>

¹⁷ <http://github.com/aimalz/qp/>

rics 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. Though the outmoded practices should not be encouraged, 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 statistics 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 accuracy of each individual catalog entry, we consider several metrics that probe the population-level performance of the photo- z PDFs. Because 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.

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)

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

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. Borodoloi 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 quantile-quantile 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.

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

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

interpretable as the maximum difference between the CDFs of an approximating univariate distribution $\hat{f}(z)$ and a reference distribution $\tilde{f}(z)$. 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

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

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. Using the notation introduced in Section 3.2.3, the conditional density estimation (CDE) loss is defined as

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

in terms of the photometry \mathbf{x} , an analogue to the familiar root-mean-square-error used in conventional regression. 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.

845 **5 RESULTS**

846 We begin with a demonstrative visual inspection of the
 847 photo- z PDFs produced by each code for individual galaxies.
 848 Figure 1 shows the photo- z PDFs for four galaxies chosen
 849 as examples of photo- z PDF archetypes: a narrow unimodal
 850 PDF, a broad unimodal PDF, a bimodal PDF, and a multi-
 851 modal PDF. We reiterate that under our idealized experi-
 852 mental conditions, differences between codes are the isolated
 853 signature of the implicit prior due to the method by which
 854 the photo- z PDFs were derived.

855 The most striking differences between codes are due
 856 to small-scale features induced by the interaction between
 857 the shared piecewise constant parameterization of 200 bins
 858 $0 < z < 2$ of Section 4 and the smoothing conditions or
 859 lack thereof in each algorithm. The $\text{dz} = 0.01$ redshift reso-
 860 lution is sufficient to capture the broad peaks of faint galax-
 861 ies’ photo- z PDFs with large photometric errors but is too
 862 broad to resolve the narrow peaks for bright galaxies’ photo-
 863 z PDFs with small photometric errors. This observation is
 864 consistent with the findings of Malz et al. (2018) that the
 865 piecewise constant parameterization underperforms in the
 866 presence of small-scale structures.

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

875 **5.1 Performance on photo- z PDF ensembles**

876 The histogram of PIT values, QQ plot, and QQ difference
 877 plot relative to the ideal diagonal are provided in Figure 2,
 878 showcasing the biases and trends in the average accuracy
 879 of the photo- z PDFs for each code. The high QQ values
 880 (i.e. more high than low PIT values) of BPZ, CMNN, Delight,
 881 EAZY, and GPz indicate photo- z PDFs biased low, and the
 882 low QQ values (more low than high PIT values) of SkyNet
 883 and TPZ indicate photo- z PDFs biased high. The gray shaded
 884 band marks the 2σ variance in PIT values found using the
 885 trainZ algorithm with a bootstrap resampling of the train-
 886 ing set and a sample size of 30,000 galaxies, representing
 887 a very conservative estimate of the representative training
 888 sample size, and thus an approximate minimal error signifi-
 889 cance compared to ideal performance. Deviations in the PIT
 890 histograms outside of this range show that significant biases
 891 are present for some codes.

892 The PIT histograms of Delight, CMNN, SkyNet, and TPZ
 893 feature an underrepresentation of extreme values, indicative
 894 of overly broad photo- z PDFs, while the overrepresentation
 895 of extreme values for METAPhR indicate overly narrow photo-
 896 z PDFs. These five codes in particular have a free parameter
 897 for bandwidth, which may be responsible for this vulnera-
 898 bility, in spite of the opportunity for fine-tuning with per-
 899 fect prior information. FlexZBoost’s “sharpening” parame-
 900 ter (described in Section 3.2.3) played a key role in diagonal-
 901 izing the QQ plot, indicating a common avenue for improve-
 902 ment in the approaches that share this type of parameter.
 903 On the other hand, the three purely template-based codes,

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
METAPhR	0.0229
CMNN	0.0034
SkyNet	0.0001
TPZ	0.0130
trainZ	0.0002

904 BPZ, EAZY, and LePhare, do not exhibit much systematic
 905 broadening or narrowing, which may indicate that complete
 906 template coverage effectively defends from these effects.

907 Close inspection of the extremes at PIT values of 0 and 1
 908 reveal spikes in the first and last bin of the PIT histogram for
 909 some codes in Figure 2, corresponding to catastrophic out-
 910 liers where the true redshift lies outside of the support of the
 911 $p(z)$. The catastrophic outlier rates are provided in Table 2.
 912 As expected, trainZ achieves precisely the 0.0002 value ex-
 913 pected of an ideal PIT distribution. ANNz2, FlexZBoost,
 914 LePhare, and METAPhR have notably high catastrophic out-
 915 lier rates > 0.02 , exceeding 100 times the ideal PIT rate,
 916 meriting further investigation.

917 Figure 3 highlights the relative values of the KS, CvM,
 918 and AD test statistics calculated by comparing the PIT dis-
 919 tribution and a uniform distribution $U(0, 1)$. METAPhR and
 920 LePhare perform well under the AD but poorly under the
 921 KS and CvM due to their high catastrophic outlier rates.
 922 ANNz2 and FlexZBoost are the top scorers under these met-
 923 rics of the PIT distribution. ANNz2’s strong performance can
 924 be attributed to an aspect of the training process in which
 925 training set galaxies with a PIT that more closely matches
 926 the percentiles of the DC1 training set’s redshift distribu-
 927 tion are upweighted; in effect, these quantile-based metrics
 928 were part of the algorithm itself that may or may not serve
 929 it well under more realistic experimental conditions. Similar
 930 to what was done for the PIT histograms in Figure 2, we
 931 create bootstrap training samples of 30,000 galaxies for use
 932 with trainZ in order to estimate a conservative statistical
 933 floor that we would expect in real data. No code reaches this
 934 idealized floor, indicating that all codes suffer some degra-
 935 dation from the ideal when employing their implicit priors,
 936 though ANNz2, FlexZBoost, and GPz are within a factor of
 937 two.

938 **5.2 Performance on individual photo- z PDFs**

939 The values of the CDE loss statistic of individual photo-
 940 z PDF accuracy are provided in Table 3. It is worth noting
 941 that strong performance on the CDE loss should imply
 942 strong performance on the other metrics, though the inverse

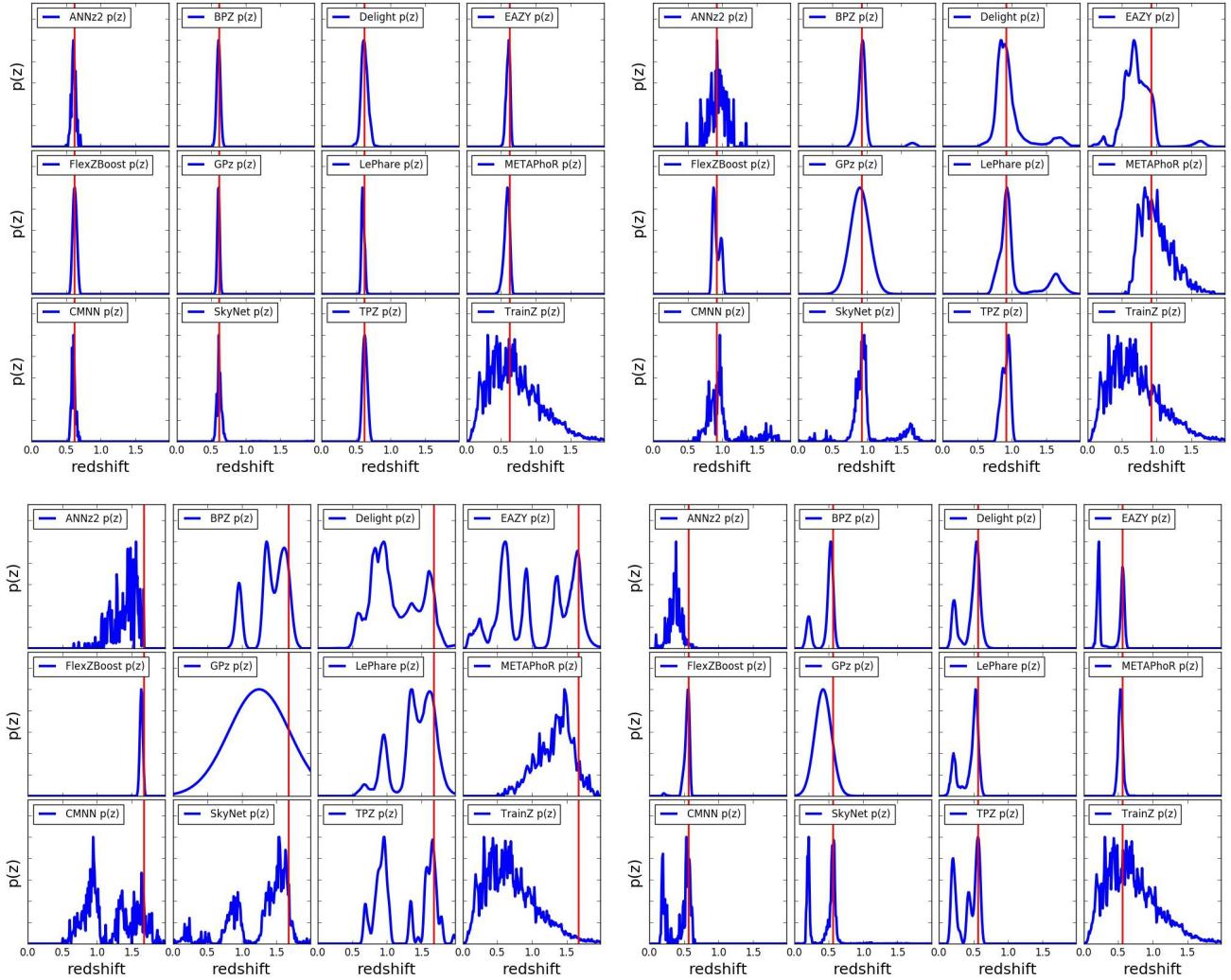


Figure 1. The individual photo- z PDFs (blue) distributions produced by the twelve codes (small panels) on four exemplary galaxies' photometry (large panels) with different true redshifts (red). The photo- z PDFs of all codes share some features for the example galaxies due to physical color degeneracies 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.

is not necessarily true. Thus the CDE loss is the most effective metric for generic science cases.

This metric is the only one that can appropriately penalize `trainZ` and indicates strong performance for `CMNN` and `FlexZBoost`, the latter of which is optimized for this metric.

6 DISCUSSION AND FUTURE WORK

In contrast with other photo- z PDF comparison papers that have aimed to identify the “best” code for a given survey, we have focused on the somewhat more philosophical questions of how to assess photo- z PDF methods and how to interpret differences between codes in terms of photo- z PDF performance. In Section 6.1, we reframe the strong performance of our pathological photo- z PDF technique, `trainZ`, as a cautionary tale about the importance of choosing appropriate comparison metrics. In Section 6.2, we outline the experiments we intend to build upon this study. In Section 6.3, we

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

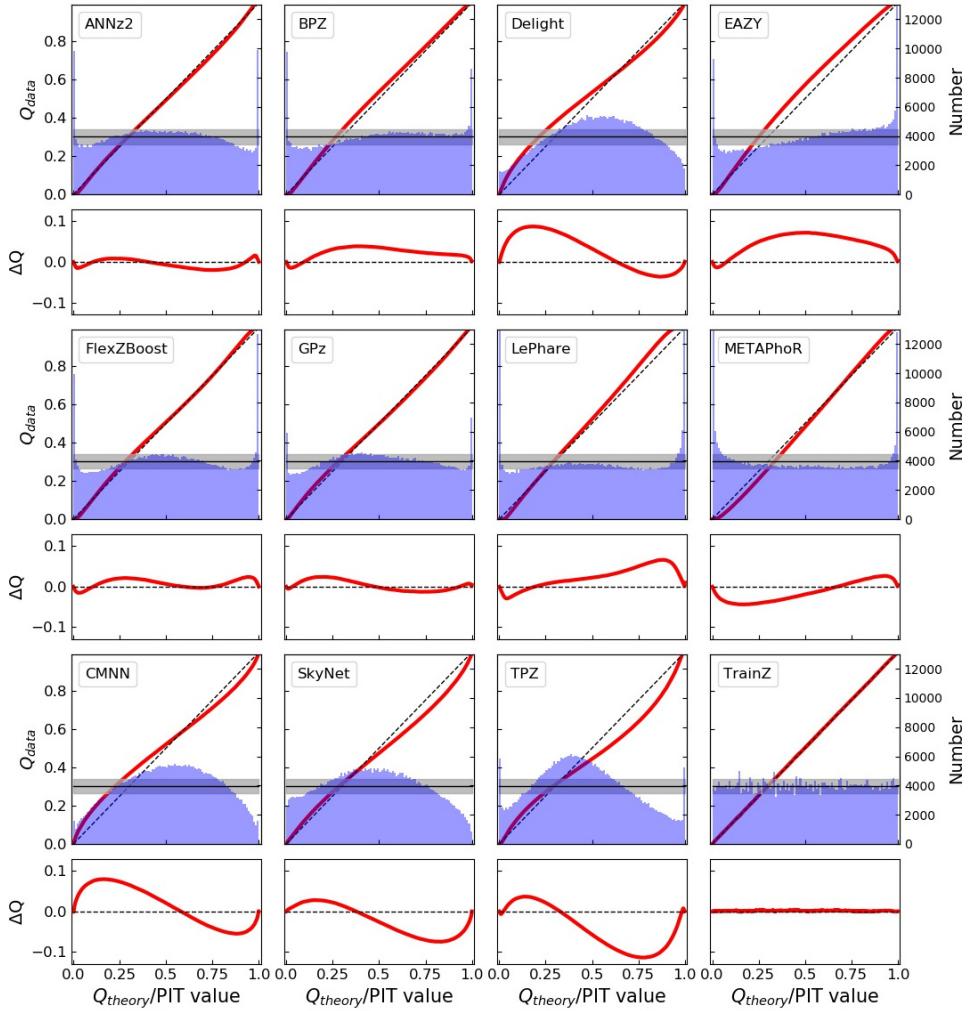


Figure 2. The QQ plot (red) and PIT histogram (blue) of the photo- z PDF codes (panels) along with the ideal QQ (black dashed) and ideal PIT (gray horizontal) curves, as well as a difference plot for the QQ difference from the ideal diagonal (lower inset). The gray shaded region indicates the 2σ range 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.

discuss the enhancements of the mock data set that will be necessary to enable the future experiments.

6.1 Interpretation of metrics

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

Furthermore, our metrics are not necessarily able to as-

sess the fidelity of individual photo- z PDFs relative to true posteriors: in the absence of a “true PDF” from which redshifts are drawn, it is difficult to construct metrics to measure performance for individual galaxies rather than ensembles. (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

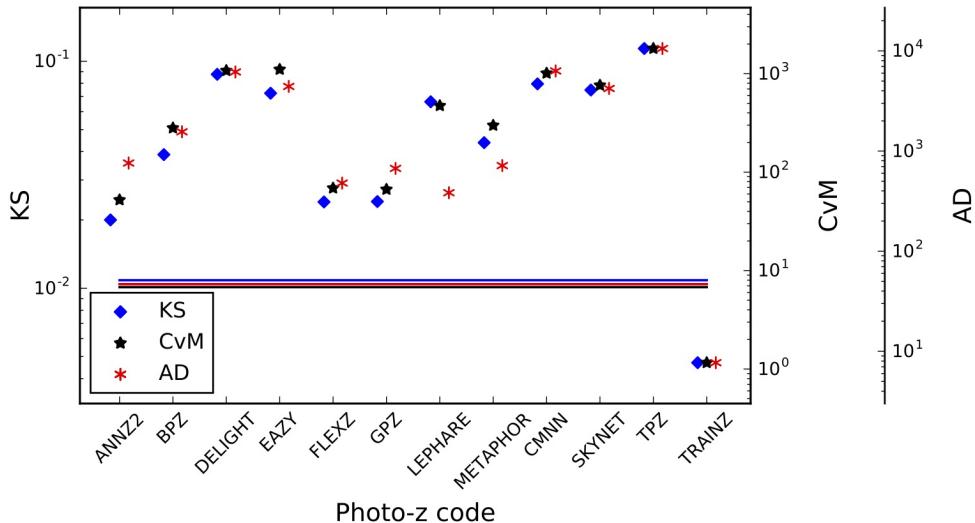


Figure 3. A visualization of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the PIT distributions. There is generally good agreement between these statistics, with differences corresponding to the codes with outstanding catastrophic outlier rates, a reflection in 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.

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` outperforms all codes on the PIT-based metrics, and all but one code on the $N(z)$ based statistics.

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 defending against deceptively strong performers such as `trainZ`; the best photo- z PDF method is the one that most effectively achieves our science goals, not the one that performs best on a metric that does not reflect those goals. In the absence of a single scientific motivation or the information necessary for a principled metric definition, we must consider many metrics and be critical of the information transmitted by each.

6.2 Extensions to the experimental design

The work presented in this paper is only a first step in assessing photo- z PDF approaches and moving toward an improved photometric redshift estimator. Here we discuss the next steps for subsequent investigations.

This initial paper explored code performance in idealized conditions with perfect catalog-based photometry and representative training data. A top priority for a follow-up

study is to test realistic forms of incomplete, erroneous, and non-representative template libraries and training sets as well as the impact of other forms of external priors that must be ingested by the codes, major concerns in Newman et al. (2015); Masters et al. (2017). Outright redshift failures due to emission line misidentification or noise spikes may be modeled by the inclusion of a small number of high-confidence yet false redshifts. We plan to perform a full sensitivity analysis on a realistically incomplete training set of spectroscopic galaxies, modeling the performance of spectrographs, emission-line properties, and expected signal-to-noise to determine which potential training set galaxies are most likely to be excluded.

Appendix A only addresses the stacked estimator of the redshift distribution of the entire galaxy catalogue rather than subsets in bins, tomographic or otherwise. The effects of tomographic binning scheme will be explored in a dedicated future paper, including propagation of redshift uncertainties in a set of fiducial tomographic redshift bins in order to estimate impact on cosmological parameter estimation.

Sequels to this study will also address some shortcomings of our experimental procedure. The fixed redshift grid shared between the codes may have unfairly penalized codes with a different native parameterization, as precision is lost when converting between formats. Performance on sharply peaked photo- z PDFs may have been suppressed across all codes due to the insufficient resolution of the redshift grid. In light of the results of Malz et al. (2018), in future analyses we plan to switch from a fixed grid to the quantile parameterization or to permit each code to use its native storage format under a shared number of parameters.

Section 4 discussed the difficulty in evaluating PDF accuracy for individual objects. In a follow-up study, we will generate “true PDF” distributions, yielding a dataset that enables a test of PDF accuracy for individual galaxies rather than solely ensembles.

1057 6.3 Realistic mock data

1058 To make optimal use of the LSST data for cosmological
 1059 and other astrophysical analyses of the Science Roadmap,
 1060 future investigations that build upon this one will require a
 1061 more sophisticated set of galaxy photometry and redshifts.
 1062 This initial paper explored a data set that was constructed
 1063 at the catalog level, with no inclusion of the complications
 1064 that come from measuring photometry from images. Future
 1065 data challenges will move to catalogs constructed from mock
 1066 images, including the complications of deblending, sensor in-
 1067 efficiencies, and heterogeneous observing conditions, all an-
 1068 ticipated to affect the measured colours of LSST’s galaxy
 1069 sample (Dawson et al. 2016).

1070 The DC1 galaxy SEDs were linear combinations of just
 1071 five basis SED templates, but a next generation of data for
 1072 photo-z PDF investigations must include a broader range of
 1073 physical properties. Though we only considered $z < 2$ here,
 1074 LSST 10-year data will contain $z > 2$ galaxies, plagued by
 1075 fainter apparent magnitudes and anomalous colours due to
 1076 stellar evolution. A subsequent study must also have a data
 1077 set that includes low-level active galactic nuclei (AGN) fea-
 1078 tures in the SEDs, which perturb colours and other host
 1079 galaxy properties. An observational degeneracy between the
 1080 Lyman break of a $z \sim 2 - 3$ galaxy from the Balmer break
 1081 of a $z \sim 0.2 - 0.3$ galaxy is a known source of catastrophic
 1082 outliers (Massarotti et al. 2001) that was not effectively in-
 1083 cluded in this study. To gauge the sensitivity of photo-z
 1084 PDF estimators to catastrophic outliers, our data set must
 1085 include realistic high-redshift galaxy populations.

1086 The overarching plan describing everything laid out in
 1087 this section is described in more detail in the LSST-DESC
 1088 Science Roadmap (see Footnote in Section 1).

1115 they correspond to the quantities that matter to our
 1116 science.

Acknowledgments

Author contributions are listed below.

S.J. Schmidt: Led the project. (conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, visualization, writing – original draft, writing – review & editing)

A.I. Malz: Contributed to choice of metrics, implementation in code, and writing. (conceptualization, methodology, project administration, resources, software, visualization, writing – original draft, writing – review & editing)

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

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

M. Brescia: main ideator of METAPHOR and of MLPQNA; modification of METAPHOR pipeline to fit the LSST data structure and requirements

S. Cavaudi: Contributed to choice and test of metrics, ran METAPHOR, minor text editing

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

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

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

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

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

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

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

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

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

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

C. B. Morrison: Managerial support; Discussions with authors regarding metrics and style; Some coding contribution to metric computation.

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

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

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

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

1089 7 CONCLUSION

1090 This paper compares twelve photo-z PDF codes under con-
 1091 trolled experimental conditions of representative and com-
 1092 plete prior information to set a baseline for an upcoming
 1093 sensitivity analysis. This work isolates the impact on met-
 1094 rics of photo-z PDF accuracy due to the estimation tech-
 1095 nique as opposed to the complications of realistic physical
 1096 systematics of the photometry. Though the mock data set of
 1097 this investigation did not include true photo-z posteriors for
 1098 comparison, we interpret deviations from perfect re-
 1099 sults given perfect prior information as the imprint
 1100 of the implicit assumptions underlying the estima-
 1101 tion approach.

1102 We evaluate the twelve codes under science-agnostic
 1103 metrics both established and emerging to stress-test the
 1104 ensemble properties of photo-z PDF catalogues derived by
 1105 each method. In appendices, we also present metrics of point
 1106 estimates and a prevalent summary statistic of photo-z PDF
 1107 catalogues used in cosmological analyses to enable the reader
 1108 to relate this work to studies of similar scope. We observe
 1109 that no one code dominates in all metrics, and that the stan-
 1110 dard metrics of photo-z PDFs and the stacked estimator of
 1111 the redshift distribution can be gamed by a very simplistic
 1112 procedure that asserts the prior over the data. We empha-
 1113 size to the photo-z community that metrics used to vet
 1114 photo-z PDF methods must be scrutinized to ensure

H. Tranin: contributed to providing SkyNet results and writing the relevant section
R. Zhou: Optimized and ran EAZY and contributed to the draft
R. Izbicki: Co-developed FlexZBoost and the CDE loss statistic, and wrote software for FlexZBoost

The authors would like to thank their LSST-DESC publication review committee for comments that improved the paper draft.

personal funding sources S. Schmidt acknowledges support from DOE grant DE-SC0009999 and 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.

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), **pySkyNet** (Bonnell 2016), and **photUtils** from the LSST simulations package (Connolly et al. 2014).

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne – France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.

APPENDIX A: EVALUATION OF THE REDSHIFT DISTRIBUTION

Perhaps the most popular application of photo- z PDFs is the estimation of the overall redshift distribution $N(z)$, a quantity that enters some cosmological calculations and the true value of which is known for the DC1 data set and will be denoted as $\tilde{N}(z)$. In terms of the prior information provided to each method, the true redshift distribution satisfies the tautology $\tilde{N}(z) = p(z|I_D)$ due to our experimental set-up; because the DC1 training and template sets are representative and complete, I_D represents a prior that is also equal to the truth. In this ideal case of complete and representative prior information, the method that would give the best approximation to $\tilde{N}(z)$ would be one that neglects all the information contained in the photometry $\{d_i\}_{N_{tot}}$ and gives every galaxy the same photo- z PDF $\hat{p}_i(z) = \tilde{N}(z)$ for all i ; the inclusion of any information from the photometry would only introduce noise to the optimal result of return-

ing the prior. This is the exact estimator, **trainZ**, that we have described in Section 3.3, and which will serve as an experimental control.

A1 Metrics of the stacked estimator of the redshift distribution

Though alternatives exist (Malz & Hogg prep), “stacking” according to

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

is the most widely accepted method for obtaining $\hat{N}^H(z)$ as an estimator of the redshift distribution from photo- z PDFs derived by a method H . Though the use of the stacked estimator of the redshift distribution is not formally correct, we use it under the untested assumption that the response of our metrics of $\hat{N}^H(z)$ will be analogous to the same metrics applied to a principled estimator of the redshift distribution.

As $N(z)$ is itself a univariate PDF, we apply the metrics of the previous sections to it as well. We additionally calculate the first three moments

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

of the estimated redshift distribution $\hat{N}^H(z)$ for each code and compare them to the moments of the true redshift distribution $\tilde{N}(z)$. Under the assumption that the stacked estimator is unbiased, a superior method minimizes the difference between the true and estimated moments.

A2 Performance on the stacked estimator of the redshift distribution

Figure A1 shows the stacked estimator $\hat{N}(z)$ of the redshift distribution for each code compared to the true redshift distribution $\tilde{N}(z)$, where the stacked estimator has been smoothed for each code in the plot using a kernel density estimate (KDE) with a bandwidth chosen by Scott’s Rule (Scott 1992) in order to minimize visual differences in small-scale features; the quantitative statistics, however, are calculated using the empirical CDF which is not smoothed.

Many of the codes, including all the model-fitting approaches and **NNz2**, **GPz**, **METAPhR**, and **SkyNet** from the data-driven camp, overestimate the redshift density at $z \sim 1.4$. This behavior is a consequence of the 4000 Å break passing through the gap between the z and y filters, which induces a genuine discontinuity in the $z - y$ colour as a function of redshift that can sway the photo- z PDF estimates in the absence of bluer spectral features.

NNz2, **GPz**, and **METAPhR** feature exaggerated peaks and troughs relative to the training set, a potential sign of overtraining. Further investigation on overtraining is needed, if present this is an obstacle that may be overcome with adjustment of the implementation.

As expected, **trainZ** perfectly recovers the true redshift distribution: as the training sample is selected from the same underlying distribution as the test set, the redshift distributions are identical, up to Poisson fluctuations due to the finite number of sample galaxies. **CMNN** is also in excellent

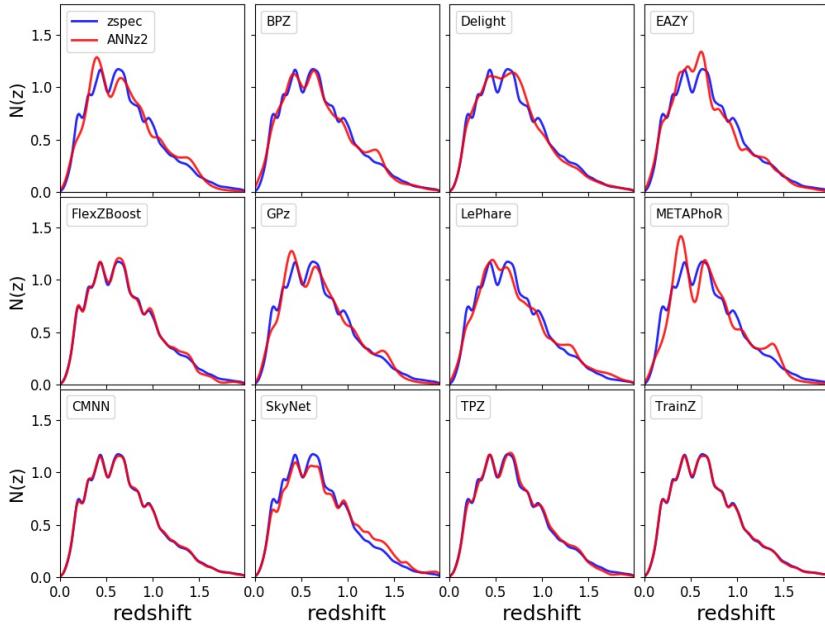


Figure A1. The smoothed stacked estimator $\hat{N}(z)$ of the redshift distribution (red) produced by each code (panels) compared to the true redshift distribution $\tilde{N}(z)$ (blue). Varying levels of agreement are seen among the codes, with the smallest deviations for CMNN, FlexZBoost, TPZ, and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

agreement for similar reasons: with a representative training sample of galaxies spanning the colour-space, the sum of the colour-matched neighbour redshifts should return the true redshift distribution. FlexZBoost and TPZ also perform superb recovery of the true redshift distribution, with only a slight deviation at $z \sim 1.4$. Our metrics, however, cannot discern whether these four approaches, as well as Delight, are spared the $z \sim 1.4$ degeneracy in $\hat{N}(z)$ because they have more effectively used information in the data or if the impact is simply washed out by the stacked estimator's effective average over the test set galaxy sample. See Appendix B for further discussion of the $z \sim 1.4$ issue.

Figure A2 shows the quantitative Kolmogorov-Smirnoff (KS), Cramer-Von Mises (CvM), and Anderson Darling (AD) test statistics for each of the codes for the $\hat{N}(z)$ based measures. The horizontal lines show the result of a bootstrap resampling of the training set using 30,000 samples for 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 $\hat{N}(z)$ 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 color/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.

The first three moments of the stacked $\hat{N}(z)$ distribution relative to the empirical estimate of the truth distribution are given in Table A1. Accuracy of the moments varies widely between codes, raising concerns about the propagation to cosmological analyses.

SkyNet exhibits redshift bias in Figure A1 and is a clear outlier in the first moment of $\hat{N}(z)$ in Table A1. 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 diagnostic tool and to facilitate direct comparisons to historical studies.

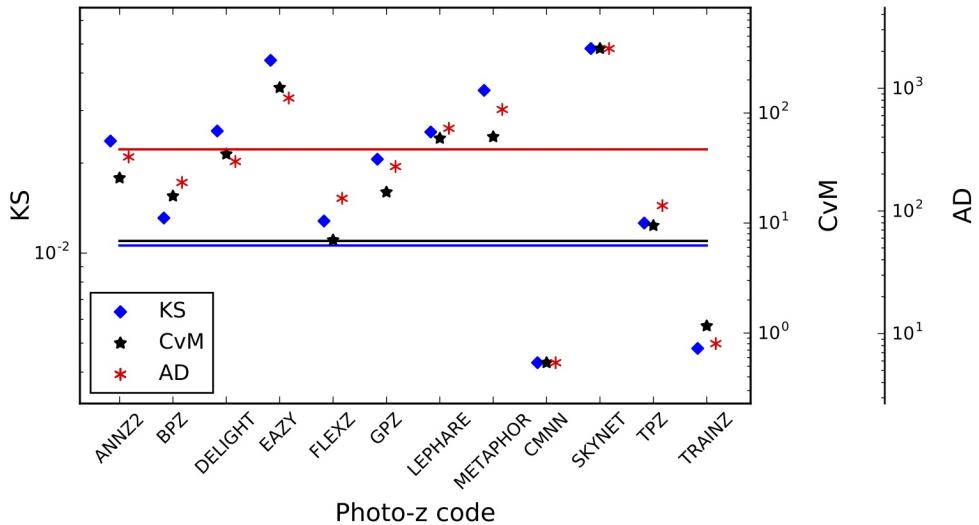


Figure A2. A visualization of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the $\hat{N}(z)$ distributions. Horizontal lines indicate the statistic values (including uncertainty) achieved using `trainZ` via bootstrap 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.

Table A1. Moments of the stacked estimator $\hat{N}(z)$ of the redshift distribution. Most of the codes considered recover the moments of $\tilde{N}(z)$

Moments of $\hat{N}(z)$	mean	variance	skewness
Empirical “truth”	0.701	0.630	0.671
ANNz2	0.702	0.625	0.653
BPZ	0.699	0.629	0.671
Delight	0.692	0.609	0.638
EAZY	0.681	0.595	0.619
FlexZBoost	0.694	0.610	0.631
GPz	0.696	0.615	0.639
LePhare	0.718	0.668	0.741
METAPhR	0.705	0.628	0.657
CMNN	0.701	0.628	0.667
SkyNet	0.743	0.708	0.797
TPZ	0.700	0.619	0.643
<code>trainZ</code>	0.699	0.627	0.666

B2 Metrics of photo-z point estimates

We calculate the commonly used point estimate metrics of the overall intrinsic scatter, bias, and catastrophic outlier rate, defined in terms of the standard error $e_z \equiv (z_{PEAK} - z_{true})/(1 + z_{true})$. Because the standard deviation of the photo-z residuals is sensitive to outliers, we define the scatter in terms of the Interquartile Range (IQR), the difference between the 75th and 25th percentiles of the distribution of e_z , imposing the scaling $\sigma_{IQR} = IQR/1.349$ to ensure that the area within σ_{IQR} is the same as that within one standard deviation from a standard Normal distribution. We also resist the effect of catastrophic outliers by defining the bias b_z as the median rather than mean value of e_z . The catastrophic outlier rate f_{out} is defined as the fraction of galaxies with e_z greater than $\max(3\sigma_{IQR}, 0.06)$.

For reference, Section 3.8 of the LSST Science Book (Abell et al. 2009) uses the standard definitions of these parameters in requiring

- RMS scatter $\sigma < 0.02(1 + z_{true})$
- bias $b_z < 0.003$
- catastrophic outlier rate $f_{out} < 10\%$.

B3 Comparison of photo-z point estimate metrics

Figure B1 shows both point estimates for all codes both z_{PEAK} and z_{WEIGHT} . Point density is shown with mixed contours to emphasize that most of the galaxies do fall close to the $z_{phot} = z_{spec}$ line, while points trace the details of the catastrophic outlier populations.

The finite grid spacing of the photo-z PDFs induces some discretization in z_{PEAK} . The features perpendicular to the $z_{phot} = z_{spec}$ line are due to the 4000Å break passing through the gaps between adjacent filters. Even the strongest codes feature populations far from the $z_{phot} = z_{spec}$ line representing a degeneracy in the space of colours and redshifts.

B1 Reduction of photo-z PDFs to point estimates

Though we acknowledge that many of the codes can also return a native photo-z point estimate, we put all codes on equal footing by considering two generic photo-z point estimators, the mode z_{PEAK} and main-peak-mean z_{WEIGHT} (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 a, say, bimodal distribution, it avoids the pitfall of reducing the photo-z PDF to a redshift between peaks where there is low probability.

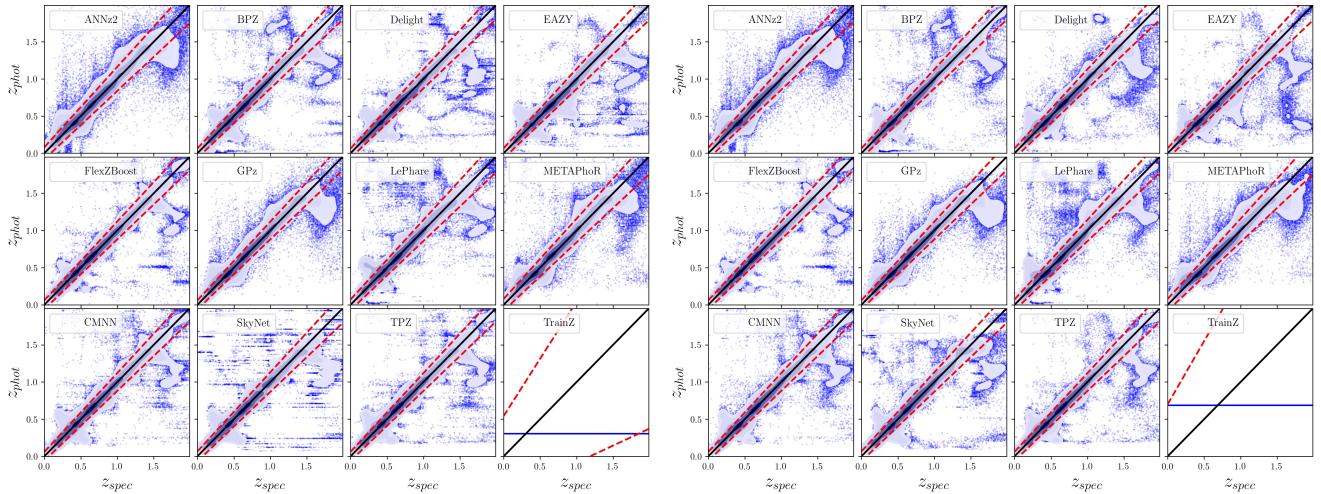


Figure B1. The density of photo- z point estimates (contours) reduced from the photo- z PDFs with outliers (blue) beyond the outlier cutoff (red dashed lines), via the mode (z_{PEAK} , left panel) and main-peak-mean (z_{WEIGHT} , right panel). The `trainZ` estimator (lower right sub-panels) has a shared 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}			
	$\sigma_{IQR}/(1+z)$	median	outlier fraction	$\sigma_{IQR}/(1+z)$	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
<code>trainZ</code>	0.1808	-0.2086	0.000	0.2335	0.022135	0.000

The intrinsic scatter, bias, and catastrophic outlier rate are given in Table B1. Perhaps unsurprisingly, performance under these metrics largely tracks that of the metrics of Section 4 of the photo- z PDFs from which the point estimates were derived. All twelve codes perform at or near the goals of the LSST Science Requirements Document¹⁸ and Graham et al. (2018), which is encouraging if not unexpected for $i < 25$.

- Almosallam I. A., Jarvis M. J., Roberts S. J., 2016b, *MNRAS*, **462**, 726
 Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, *MNRAS*, **310**, 540
 Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, *ApJ*, **762**, 109
 Benítez N., 2000, *ApJ*, **536**, 571
 Bernstein G., Huterer D., 2010, *MNRAS*, **401**, 1399
 Blanton M. R., Roweis S., 2007, *AJ*, **133**, 734
 Blanton M. R., et al., 2005, *AJ*, **129**, 2562
 Bonnett C., 2015, *MNRAS*, **449**, 1043
 Bonnett C., 2016, Python wrapper to SkyNet, <https://pyskynet.readthedocs.io/en/latest/>
 Bonnett C., et al., 2016, *Phys. Rev. D*, **94**, 042005
 Bordoloi R., Lilly S. J., Amara A., 2010, *MNRAS*, **406**, 881
 Brammer G. B., van Dokkum P. G., Coppi P., 2008, *ApJ*, **686**, 1503
 Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees, Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A
 Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2018, preprint, ([arXiv:1802.07683](https://arxiv.org/abs/1802.07683))

¹⁸ available at: <http://ls.st/srd>

- 1425 Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483 1493
 1426 Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, 442, 3380 1494
 1427 Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo 1495
 G., 2017, *MNRAS*, 465, 1959 1496
 1428 Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM 1497
 SIGKDD International Conference on Knowledge Discovery 1498
 and Data Mining, KDD '16. ACM, New York, NY, USA, 1499
 pp 785–794, doi:10.1145/2939672.2939785, <http://doi.acm.org/10.1145/2939672.2939785>
 1429 Connolly A. J., et al., 2014, in Angeli G. Z., Dierickx P., eds, 1499
 Society of Photo-Optical Instrumentation Engineers (SPIE) 1499
 Conference Series Vol. 9150, Modeling, Systems Engineering- 1499
 and Project Management for Astronomy VI. p. 14, 1499
 doi:10.1117/12.2054953
 1430 Dahlen T., et al., 2013, *ApJ*, 775, 93
 1431 Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016, 1498
 ApJ, 816, 11
 1432 DeRose J., et al., 2019, arXiv e-prints, p. arXiv:1901.02401
 1433 Erben T., et al., 2013, *MNRAS*, 433, 2545
 1434 Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, *ApJ*, 513, 34
 1435 Firth A. E., Lahav O., Somerville R. S., 2003, *MNRAS*, 339, 1195
 1436 Freeman P. E., Izbicki R., Lee A. B., 2017, *MNRAS*, 468, 4556
 1437 Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, *MNRAS*, 441, 1448
 1741
 1438 Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones 1449
 R. L., Jurić M., Daniel S. F., Yoachim P., 2018, *AJ*, 155, 1
 1440 Green J., et al., 2012, preprint (arXiv:1208.4012),
 1441 Hildebrandt H., et al., 2010, *A&A*, 523, A31
 1442 Hofmann B., Mathé P., 2018, *Inverse Problems*, 34, 015007
 1443 Hunter J. D., 2007, Matplotlib: A 2D Graphics Environment, 1444
 doi:10.1109/MCSE.2007.55
 1444 Ilbert O., et al., 2006, *A&A*, 457, 841
 1445 Ivezić Ž., et al., 2008, preprint (arXiv:0805.2366),
 1446 Izbicki R., Lee A. B., 2017, *Electron. J. Statist.*, 11, 2800
 1447 Izbicki R., Lee A. B., Freeman P. E., 2017, *Ann. Appl. Stat.*, 11, 1448
 698
 1448 Laureijs R., et al., 2011, preprint (1110.3193),
 1449 Leistedt B., Hogg D. W., 2017, *ApJ*, 838, 5
 1450 Malz A., Hogg D., in prep., CHIPPR, chippr
 1451 Malz A., Marshall P., DeRose J., Graham M., Schmidt S., Wechsler R., 2018, AJ, Accepted,
 1452 Mandelbaum R., et al., 2008, *MNRAS*, 386, 781
 1453 Massarotti M., Iovino A., Buzzoni A., 2001, *A&A*, 368, 74
 1454 Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes 1455
 J. D., Castander F. J., Paltani S., 2017, *ApJ*, 841, 111
 1455 Newman J. A., et al., 2015, *Astroparticle Physics*, 63, 81
 1456 Oliphant T., 2007, Python for Scientific Computing, 1456
 doi:10.1109/MCSE.2007.58
 1457 Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, *ApJ*, 1457
 689, 709
 1458 Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint 1458
 (arXiv:1608.08016),
 1459 Rasmussen C., Williams C., 2006, Gaussian Processes for Machine 1459
 Learning. Adaptative computation and machine learning series, MIT Press, Cambridge, MA
 1460 Rau M. M., Seitz S., Brimioule F., Frank E., Friedrich O., Gruen 1460
 D., Hoyle B., 2015, *MNRAS*, 452, 3710
 1461 Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, 1461
 ApJ, 771, 30
 1462 Sadeh I., Abdalla F. B., Lahav O., 2016, *PASP*, 128, 104502
 1463 Sánchez C., et al., 2014, *MNRAS*, 445, 1482
 1464 Schmidt M., 2005, minFunc: Unconstrained Differentiable Mul- 1464
 tivariate Optimization in Matlab, <http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>
 1465 Scott D. W., 1992, Multivariate Density Estimation. Theory, 1465
 Practice, and Visualization. Wiley
 1466 Skrutskie M. F., et al., 2006, *AJ*, 131, 1163
 1467 Tanaka M., et al., 2018, *PASJ*, 70, S9
 1468 The LSST Dark Energy Science Collaboration et al., 2018, 1468
 preprint, (arXiv:1809.01669)
 1469 Waskom M., et al., 2017, doi:10.5281/zenodo.824567
 1470 York D. G., et al., 2000, *AJ*, 120, 1579
 1471 de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn 1471
 E. A., 2013, *Exp. Astron.*, 35, 25
 1472 de Jong J. T. A., et al., 2017, *A&A*, 604, A134