

Evaluation of probabilistic photometric redshift estimation approaches for LSST

S.J. Schmidt¹, A.I. Malz^{2,3}, J.Y.H. Soo⁴, I.A. Almosallam^{5,6}, M. Brescia⁷, S. Cavaoti^{7,8}, J. Cohen-Tanugi⁹, A.J. Connolly¹⁰, P.E. Freeman¹¹, M.L. Graham¹⁰, K. Iyer¹², R. Izbicki^{13,14}, M.J. Jarvis^{15,16}, J.B. Kalmbach¹⁷, E. Kovacs¹⁸, A.B. Lee¹¹, G. Longo⁸, C. Morrison¹⁰, J. Newman¹⁹, E. Nourbakhsh¹, E. Nuss⁹, T. Pospisil¹¹, H. Tranin⁹

(LSST Dark Energy Science Collaboration)

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

² Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, 10003, USA

³ Department of Physics, New York University, 726 Broadway, New York, 10003, USA

⁴ Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

⁵ King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

⁶ Information Engineering, Parks Road, Oxford, OX1 3PJ, UK

⁷ INAF-Capodimonte Observatory, Salita Moiariello 16, I-80131, Napoli, Italy

⁸ Department of Physics E. Pancini, University Federico II, via Cinthia 6, I-80126, Napoli, Italy

⁹ Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France

¹⁰ Department of Astronomy, University of Washington, Box 351580, U.W., Seattle WA 98195, USA

¹¹ Department of Statistics & Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

¹² Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA

¹³ Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil

¹⁴ External collaborator

¹⁵ Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK

¹⁶ Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa

¹⁷ Department of Physics, University of Washington, Box 351560, Seattle, WA 98195, USA

¹⁸ Argonne National Laboratory, Lemont, IL 60439, USA

¹⁹ Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, 3941 O'Hara St., Pittsburgh, PA 15260, USA

30 March 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 on 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/underbreadth 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; however, as use in precision cosmology measurements became more prevalent there was a realization that point-estimate photo- z ’s were inadequate (Mandelbaum et al. 2008). The nontrivial mapping between broad band fluxes and redshift renders a simplistic point estimate inadequate to quantify the uncertainty landscape by neglecting degenerate redshift solutions. 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, photometric galaxy catalogue data releases, (e. g. Erben et al. (2013), de Jong et al. (2017)), provided 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 yet been established as clearly superior. 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 by its focus on the evaluation criteria for photo- z PDFs and interpretation thereof, as a key project of the Photometric Redshifts working group of the LSST-DESC that aims 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 laid out 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.

2.1 The Buzzard-v1.0 simulation

Our mock catalogue is derived from the BUZZARD-highres-v1.0 of De Rose et al., in prep, Wechsler et al., in prep) catalogue. BUZZARD is built on a dark matter-only N-body simu-

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

lution of 2048^3 particles in a 400 Mpc h^{-1} box. The lightcone was constructed from smoothing and interpolation between a set of time snapshots. Dark matter halos were identified using the `Rockstar` software package (Behroozi et al. 2013) and then populated with galaxies with a stellar mass and absolute r -band magnitude in the SDSS system determined using a sub-halo abundance matching model constrained to match both projected two-point galaxy clustering statistics and an observed conditional stellar mass function (Reddick et al. 2013).

To assign a spectrum to each galaxy, the Adding Density Dependent Spectral Energy Distributions (SEDs) procedure (`ADDSEDS`, deRose in prep.)³ was used. `ADDSEDS` uses a sample of $\sim 5 \times 10^5$ galaxies from the magnitude-limited SDSS Data Release 6 Value Added Galaxy Catalogue (Blanton et al. 2005) to train an empirical relation between absolute r -band magnitude, local galaxy density, and SED. Each SDSS spectrum is parameterized by five weights corresponding to a weighted sum of five basis SED components using the `k-correct v4.3?` software package⁴ (Blanton & Roweis 2007).

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

2.1.1 Caveats

By necessity, BUZZARD does not contain all of the complicating factors present in real data, and here we discuss the most pertinent ways that this limitation affects our experiment. BUZZARD includes only galaxies, not stars of AGN. The catalogue-based construction excludes image-level effects, such as deblending errors, photometric measurement issues, contamination from sky background (Zodiacal light, scattered light, etc.), lensing magnification, and Galactic reddening.

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

While these idealized conditions limit the realism of our mock data, they are irrelevant to the controlled experimental

conditions of this study, if anything assuring that differentiation in the performance of the photo- z PDF codes is due to the inferential techniques rather than nuances in the data.

2.2 LSST-like mock observations

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

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

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

2.3 Shared prior information

For the purpose of performing a controlled experiment that compares photo- z PDF codes on equal footing as a baseline for a future sensitivity analysis, we take care to provide each with maximally optimistic prior information. Redshift estimation approaches built upon physical modeling and machine learning alike have a notion of prior information considered beyond the photometry of the data for which redshift is to be constrained: that information is derived from a template library for a model-based code and a training set for a data-driven code. In this initial study, we seek to set a baseline for a later comparison of the performance of photo- z PDF codes under incomplete and non-representative prior information that will propagate differently in the space of data-driven and model-based algorithms. However, for the baseline case of perfect prior information, physical modeling and machine learning codes can indeed be put on truly equal footing. We outline the equivalent ways of providing all codes perfect prior information below.

³ <https://github.com/vipasu/addsed>

⁴ <http://kcorrect.org>

4 LSST Dark Energy Science Collaboration

220 2.3.1 Training and test set division

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

234 2.3.2 Template library construction

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

254 3 METHODS

255 Here we summarize the twelve photo-z PDF codes com-
 256 pared in this study, summarized in Table 1, which include
 257 both established and emerging approaches in template fit-
 258 ting and machine learning. Though not exhaustive, this sam-
 259 ple represents codes for which there was sufficient exper-
 260 tise within the LSST-DESC Photometric Redshifts Work-
 261 ing Group; the authors welcome interest from those outside
 262 LSST-DESC to have their codes assessed in future investi-
 263 gations that build upon this one.

264 We describe the algorithms and implementations of the
 265 model-based and data-driven codes in Sections 3.1 and 3.2
 266 respectively, with a straw-person approach included in Sec-
 267 tion 3.3.

268 3.1 Template-based Approaches

269 We test three publicly available and commonly used
 270 template-based codes that share the standard physically moti-
 271 vated approach of calculating model fluxes for a set of tem-
 272 plate SEDs on a grid of redshift values and evaluating a χ^2
 273 merit function using the observed and model fluxes.

274 3.1.1 LePhare

275 Photometric Analysis for Redshift Estimate (LePhare⁵,
 276 Arnouts et al. 1999; Ilbert et al. 2006) matches ob-
 277 served colors with those predicted from a template
 278 set, which can be semi-empirical or entirely syn-
 279 synthetic, directly according to the likelihood $\chi^2(z, T, A) \equiv$
 280 $\sum_b^{N_{\text{filt}}} ((F_b^{\text{obs}} - AF_b^{\text{mod}}(T, z)) / \sigma_b^{\text{obs}})^2$ of normalization fac-
 281 tor A , template T , and redshift z . In words, the likelihood
 282 is a sum of observed flux error σ_b^{obs} -weighted squared dif-
 283 ferences between the observed flux F_b^{obs} and the normalized
 284 predicted flux $F_b^{\text{mod}}(T, z)$ in N_{filt} photometric filters b , which
 285 is the LSST *ugrizy* filters in this case. The reported photo-z
 286 PDF is an arbitrary normalization of the likelihood evalua-
 287 ted on the output redshift grid.

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

290 3.1.2 BPZ

291 Bayesian Photometric Redshift (BPZ⁶, Benítez 2000) deter-
 292 mines the likelihood $p(C|z, T)$ of a galaxy’s observed colours
 293 C for a set of SED templates T at redshifts z . The BPZ
 294 likelihood is related to the χ^2 likelihood by $p(C|z, T) \propto$
 295 $\exp[-\chi^2/2]$. Given a Bayesian prior $p(z, T|m_0)$ over appar-
 296 ent magnitude m_0 and assuming that the SED templates
 297 are spanning and exclusive, BPZ constructs the redshift pos-
 298 terior $p(z|C, m_0)$ by marginalizing over all SED templates
 299 as in (Eq. 3 from Benítez 2000), corresponding to setting
 300 the parameter PROBS_LITE=TRUE in the BPZ parameter file.
 301 The BPZ prior is the product of an SED template propor-
 302 tion that varies with apparent magnitude $p(T|m_0)$ and a
 303 prior $p(z|T, m_0)$ over the expected redshift as a function of
 304 apparent magnitude and SED template.

305 Here we test BPZ-v 1.99.3 with the DC1 template set of
 306 Section 2.3.2. To keep the number of free parameters man-
 307 ageable, the DC1 template set is pre-sorted by the rest-frame
 308 $u - g$ colour and split into three broad classes of SED tem-
 309 plate, equivalent to the E, Sp and Im/SB types in . The
 310 Bayesian prior term $p(T|m_0)$ was derived directly from the
 311 DC1 training set, and the other term $p(z|T, m_0)$ was chosen
 312 to be the best fit for the eleven free parameters of the func-
 313 tional form of Benítez (2000). Prior to running the code,
 314 the non-detection placeholder magnitude was replaced with
 315 an estimate of the one- σ detection limit for the undetected
 316 band as a proxy for a value close to the estimated sky noise
 317 threshold.

318 3.1.3 EAZY

319 Easy and Accurate Photometric Redshifts from Yale (EAZY⁷,
 320 Brammer et al. 2008) extends the basic χ^2 fit procedure that
 321 defines template-fitting approaches. The algorithm models
 322 the observed photometry with a linear combination of tem-
 323 plate SEDs at each redshift. The best-fit SED is found by
 324 simultaneously fitting one, two, or all of the templates via χ^2
 325 minimization, which is distinct from marginalizing across all

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

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

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

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/tospiszi/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.ds.unina.it
CMNN (Graham et al. 2018)	machine learning	N/A
SkyNet (Graff et al. 2014)	machine learning	http://ccpforge.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

326 templates. The minimized χ^2 likelihood at each redshift is
 327 then combined with an apparent magnitude prior to obtain
 328 the redshift posterior PDF. We note that the utilization of
 329 the best-fit SED rather than a proper marginalization does
 330 not lead to the correct posterior distribution, an implemen-
 331 tation issue that has now been identified and will be ad-
 332 dressed by the developers in the future. EAZY can account
 333 for uncertainty in the template set by adding in quadrature
 334 to the flux errors an empirically derived template error as a
 335 function of redshift.

336 The SED-independent apparent magnitude prior was
 337 derived empirically from the DC1 training set. The EAZY ar-
 338 chitecture cannot accept a template set other than the same
 339 five basis templates employed by **k-correct** when construct-
 340 ing the DC1 catalogue. However, EAZY does feature a flexible
 341 **all-templates** mode, which fits the photometric data with
 342 a linear combination of the five basis templates. We set the
 343 template error to zero since the same templates were in fact
 344 used to produce the DC1 photometry.

3.2 Training-based Approaches

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

3.2.1 ANNz2

352 ANNz2⁸ (Sadeh et al. 2016) employs several machine learn-
 353 ing algorithms, including artificial neural networks (ANN),
 354 boosted decision tree, and k-nearest neighbour (KNN) re-
 355 gression. In addition to accounting for errors on the input
 356 photometry, ANNz2 uses the KNN-uncertainty estimate of
 357 Oyaizu et al. (2008) to quantify uncertainty in the choice of
 358 method over multiple runs. Using the Toolkit for Multivari-
 359 ate Data Analysis with ROOT⁹, it can return the results
 360 of running a single machine learning algorithm, a “best”
 361 choice of the results from simultaneously running multiple

363 algorithms, or a combination of the results of multiple al-
 364 gorithms weighted by their method uncertainties averaged
 365 over multiple runs.

366 In this study, we used ANNz2-v.2.0.4 to output only the
 367 result of the ANN algorithm. Photo-z PDFs were produced
 368 by running an ensemble of 5 ANNs with a 6 : 12 : 12 : 1
 369 architecture corresponding to the 6 *ugrizy* inputs, 2 hidden
 370 layers with 12 nodes each, and 1 output of redshift. Each
 371 of the five ANNs was trained with different random seeds
 372 for the initialization of input parameters. Additionally, all
 373 ANNs were trained on only a $i \leq 25.3$ subsample of the DC1
 374 training set, and half of the training set was reserved for
 375 validation to prevent overfitting. Undetected galaxies were
 376 excluded from the training set, and per-band non-detections
 377 in the test set were replaced with the mean magnitude in
 378 that band within the entire test set.

3.2.2 Colour-Matched Nearest-Neighbours

380 The nearest-neighbours colour-matching photometric red-
 381 shift estimator (CMNN, Graham et al. 2018) uses a training
 382 set of galaxies with known redshifts that has equivalent or
 383 better photometry than the test set in terms of quality and
 384 filter coverage. For each galaxy in the test set, CMNN identifies
 385 a colour-matched subset of training galaxies using a thresh-
 386 old in the Mahalanobis distance $D_M = \sum_j^{N_{\text{colours}}} (c_j^{\text{train}} -$
 387 $c_j^{\text{test}})^2 / \delta c_{\text{test}}^2$ in the space of available colours c , with colour
 388 measurement errors δc_{test} and $N_{\text{colours}} = 5$ colors j defined
 389 by the *ugrizy* filters, which defines the set of colour-matched
 390 neighbours based on a value of the percent point function
 391 (PPF). As an example, for $N_{\text{filt}} = 5$ with PPF= 0.95, 95%
 392 of all training galaxies consistent with the test galaxy will
 393 have $D_M < 11.07$. Undetected bands are dropped, thereby
 394 reducing the effective N_{filt} for that galaxy. The photo-z PDF
 395 of a given test set galaxy is the normalized distribution of
 396 redshifts of its colour-matched subset of training set galax-
 397 ies.

398 Here, we make two modifications to the implementation
 399 of Graham et al. (2018) to comply with the controlled exper-
 400 imental conditions. First, we do not impose nondetections
 401 on galaxies fainter than the expected LSST 10-year limit-
 402 ing magnitude or bright enough to saturate with LSST’s
 403 CCDs, instead using all of the photometry for the DC1 test
 404 and training sets. Second, we apply the initial colour cut
 405 to the training set before calculating the Mahalanobis dis-
 406 tance in order to accelerate processing and use a magnitude

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

⁹ <http://tmva.sourceforge.net/>

6 LSST Dark Energy Science Collaboration

407 pseudo-prior as in [Graham et al. \(2018\)](#), but for both we use
 408 cut-off values corresponding to the DC1 training set galaxies’
 409 colours and magnitudes.

410 We make an additional adaptation to enable the CMNN
 411 algorithm to yield accurate photo- z PDFs for all galaxies,
 412 as the original [Graham et al. \(2018\)](#) algorithm is optimized
 413 for photo- z point estimates and is susceptible to less ac-
 414 curate photo- z PDFs for bright galaxies or those with few
 415 matches in colour-space. We use PPF = 0.95 rather than
 416 PPF = 0.68 to generate the subset of colour-matched train-
 417 ing galaxies, whose redshifts are weighted by their inverse
 418 Mahalanobis distances of the when composing the photo-
 419 z PDF rather than weighting all colour-matched training
 420 galaxies equally. Additionally, when the number of colour-
 421 matched training set galaxies is less than 20, the nearest 20
 422 neighbours in color-space are used instead, and the output
 423 photo- z PDF is convolved with a Gaussian kernel of variance
 424 $\sigma_{\text{train}}^2 (\text{PPF}_{20}/0.95)^2 - 1$ to account for the corresponding
 425 growth of the effective PPF to include 20 neighbors.

460 sian Processes ([Rasmussen & Williams 2006](#)), that pro-
 461 duces input-dependent variance estimates corresponding to
 462 heteroscedastic noise. The model assumes a Gaussian pos-
 463 terior probability $p(z|\mathbf{x}) = \mathcal{N}(z|\mu(\mathbf{x}), \sigma(\mathbf{x})^2)$ of the out-
 464 put redshift z given the input photometry \mathbf{x} . The mean
 465 $\mu(\mathbf{x})$ and the variance $\sigma(\mathbf{x})^2$ are modeled as functions
 466 $f(\mathbf{x}) = \sum_{i=1}^m w_i \phi_i(\mathbf{x})$ linear combinations of m basis func-
 467 tions $\{\phi_i(\mathbf{x})\}_{i=1}^m$ with associated weights $\{w_i\}_{i=1}^m$. The de-
 468 tails on how to learn the parameters of the model and the
 469 hyper-parameters of the basis functions are described in [Al-](#)
 470 [mosallam et al. \(2016b\)](#). GPz’s variance estimate is composed
 471 of a model uncertainty term corresponding to sparsity of the
 472 training set photometry and a noise uncertainty term en-
 473 compassing noisy photometric observations, enabling quan-
 474 tification of any need for more representative or more precise
 475 training samples. GPz may also weight training set samples
 476 by importance according to $|z_{\text{spec}} - z_{\text{phot}}|/(1+z_{\text{spec}})$ to mini-
 477 mize the normalized photo- z point estimate error, however,
 478 this function may be adapted to photo- z PDFs, pressuring
 479 the model to dedicate more resources to test set galaxies
 480 that are not well-represented in the training set.

481 To smooth the long tail in the distribution of magni-
 482 tude errors, we use the log of the magnitude errors, im-
 483 proving numerical stability and eliminating the need for
 484 constraints on the optimization process. Unobserved mag-
 485 nitudes $x_u = \mu_u + \Sigma_{uo} \Sigma_{oo}^{-1} (x_o - \mu_o)$ were imputed from
 486 observed magnitudes x_o and the training set mean μ and
 487 covariance Σ using a linear model. This is the optimal ex-
 488 pected value of the unobserved variables given the observed
 489 ones under the assumption that the distribution is jointly
 490 Gaussian; note that this reduces to a simple average if the
 491 covariates are independent with $\Sigma_{uo} = 0$. We reserved for
 492 validation 20% of the training set and used the Variable Co-
 493 variance option in GPz with 200 basis functions, neglecting
 494 to apply cost-sensitive learning options.

3.2.3 FlexZBoost

426 **FlexZBoost**¹⁰ ([Izbicki & Lee 2017](#)) is built on **FlexCode**, a
 427 general-purpose methodology for converting any conditional
 428 mean point estimator of z to a conditional density estima-
 429 tor $p(z|\mathbf{x}) \equiv f(z|\mathbf{x})$, where \mathbf{x} here represents our photomet-
 430 ric covariates and errors. **FlexZBoost** expands the unknown
 431 function $f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x}) \phi_i(z)$ using an orthonormal ba-
 432 sis $\{\phi_i(z)\}_i$. By the orthogonality property, the expansion
 433 coefficients $\beta_i(\mathbf{x}) = \mathbb{E}[\phi_i(z)|\mathbf{x}] \equiv \int f(z|\mathbf{x}) \phi_i(z) dz$ are thus
 434 conditional means. The expectation value $\mathbb{E}[\phi_i(z)|\mathbf{x}]$ of the
 435 expansion coefficients conditioned on the data is equivalent
 436 to the regression of the space of possible redshifts on the
 437 space of possible photometry. Thus the expansion coeffi-
 438 cients $\beta_i(\mathbf{x})$ can be estimated from the data via regression
 439 to yield the conditional density estimate $\hat{f}(z|\mathbf{x})$.

440 In this paper, we used **xgboost** ([Chen & Guestrin
 441 2016](#)) for the regression; it should however be noted that
 442 **FlexCode-RF**¹⁰, based on Random Forests, generally per-
 443 forms better for smaller datasets. As our basis $\phi_i(z)$, we
 444 choose a standard Fourier basis. The two tuning parame-
 445 ters in our photo- z PDF estimate are the number I of terms
 446 in the series expansion and an exponent α that we use to
 447 sharpen the computed density estimates $\tilde{f}(z|\mathbf{x}) \propto \hat{f}(z|\mathbf{x})^\alpha$.
 448 Both I and α were chosen in an automated way by mini-
 449 mizing the weighted L_2 -loss function (Eq. 5 in [Izbicki & Lee
 450 2017](#)) on a validation set comprised of a randomly selected
 451 15% of the DC1 training set. While **FlexCode**’s lossless na-
 452 tive encoding stores each photo- z PDF using the basis co-
 453 efficients $\beta_i(\mathbf{x})$, we discretized the final estimates into 200
 454 linearly-spaced redshift bins $0 < z < 2$ to match the consis-
 455 tent output format of the experimental conditions.

3.2.4 GPz

457 **GPz**¹¹ ([Almosallam et al. 2016a,b](#)) is a sparse Gaussian pro-
 458 cess based code, a scalable approximation of full Gaus-

495 **METAPhOr**

496 Machine-learning Estimation Tool for Accurate Photomet-
 497 ric Redshifts (**METAPhOr**¹², [Cavuoti et al. 2017](#)) is based on
 498 the Multi Layer Perceptron with Quasi Newton Algorithm
 499 (MLPQNA) with the least square error model and Tikhonov
 500 L_2 -norm regularization ([Hofmann & Mathé 2018](#)). Photo- z
 501 PDFs are generated by running N trainings on the same
 502 training set, or M trainings on M different random sam-
 503 plings of the training set. Upon regression of the test set,
 504 the photometry m_{ij} of each test set galaxy j in filter i is
 505 perturbed according to $m'_{ij} = m_{ij} + \alpha_i F_{ij} \epsilon$ in terms of
 506 the standard normal random variable $\epsilon \sim \mathcal{N}(0, 1)$, a mul-
 507 tiplicative constant α_i permitting accommodation of multi-
 508 survey photometry, and a bimodal function F_{ij} composed of
 509 a polynomial fit of the mean magnitude errors on the binned
 510 bands plus a constant term representing the threshold be-
 511 low which the polynomial’s noise contribution is negligible
 512 ([Brescia et al. 2018](#)).

513 In this work, we used a hierarchical KNN to replace non-
 514 detections with values based on their neighbors. The usual
 515 cross-validation step was also omitted for this study.

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

¹¹ <https://github.com/OxfordML/GPz>

¹² <http://dame.dsfa.unina.it>

516 3.2.6 *SkyNet*

517 **SkyNet**¹³ (Graff et al. 2014) employs a neural network based
 518 on a second order conjugate gradient optimization scheme
 519 (see Graff et al. 2014, for further details). The neural net-
 520 work is configured as a standard multilayer perceptron with
 521 three hidden layers and one input layer with 12 nodes cor-
 522 responding to the 6 photometric magnitudes and their mea-
 523 surement errors. We use **SkyNet** as a regressor for photo- z
 524 point estimation and as a classifier for photo- z PDF estima-
 525 tion.

526 The regressor used a standard χ^2 error function with a
 527 single linear node as the output layer and 10 nodes with a
 528 tanh activation function for each hidden layer. The classifier
 529 used a cross-entropy error function with a 20:40:40 node (all
 530 rectified linear units) architecture for each hidden layer and
 531 an output layer of 200 nodes corresponding to 200 bins for
 532 the PDF, with a softmax activation function to enforce the
 533 normalization condition that the probabilities sum to unity.
 534 While previous implementations of the code (see Appendix
 535 C.3 of Sánchez et al. 2014; Bonnett 2015) implement a
 536 sliding bin smoothing, no such procedure was used in this
 537 study.

538 We pre-whitened the data by pegging the magnitudes
 539 to (45,45,40,35,42,42) and errors to (20,20,10,5,15,15) for
 540 *ugrizy* filters, respectively. To avoid over-fitting, 30% of the
 541 training set was reserved for validation, and training was
 542 halted as soon as the error rate began to increase on the
 543 validation set. The weights were randomly initialized based
 544 on normal sampling.

545 3.2.7 *TPZ*

546 Trees for Photo- z (TPZ¹⁴, Carrasco Kind & Brunner 2013;
 547 Carrasco Kind & Brunner 2014) uses prediction trees and
 548 random forest techniques to estimate photo- z PDFs. TPZ re-
 549 cursively splits the training set into branch pairs based on
 550 maximizing information gain among a random subsample of
 551 features, to minimize correlation between the trees, termi-
 552 nating only when a newly created leaf meets a criterion, such
 553 as a leaf size minimum or a variance threshold. The regions
 554 in each terminal leaf node correspond to a subsample of the
 555 training set with similar properties. Bootstrap samples from
 556 the training set photometry and errors are used to build a
 557 set of prediction trees.

558 To run TPZ, we replaced nondetections with an approxi-
 559 mation of the 1σ detection threshold based on the error fore-
 560 cast of the 10-year LSST data, i. e. $dm = 2.5 \log(1 + N/S)$
 561 where $dm \sim 0.7526$ mag for $N/S = 1$. We calibrated TPZ
 562 with the Out-of-Bag cross-validation technique (Breiman
 563 et al. 1984; Carrasco Kind & Brunner 2013) to evaluate
 564 its predictive validity and determine the relative impor-
 565 tance of the different input attributes. We grew 100 trees
 566 to a minimum leaf size of 5 using the *ugri* magnitudes, all
 567 $u - g, g - r, r - i, i - z, z - y$ colours, and the associated
 568 errors, as the z and y magnitudes did not show significant
 569 correlation with the redshift in our cross-validation. We par-
 570 titioned our redshift space into 200 bins and smoothed each
 571 individual PDF with a smoothing scale of twice the bin size.

572 3.2.8 *Delight*

573 **Delight**¹⁵ (Leistedt & Hogg 2017) is a hybrid technique that
 574 infers photo- z s with a data-driven model of latent SEDs and
 575 a physical model of photometric fluxes as a function of red-
 576 shift. Generally, machine learning methods rely on represen-
 577 tative training data with shared photometric filters, while
 578 template based methods rely on a complete library of tem-
 579 plates based on physical models constructed. **Delight** aims
 580 to take the best aspects of both approaches by construct-
 581 ing a large collection of latent SED templates (or physical
 582 flux-redshift models) from training data, with a template
 583 SED library as a guide to the learning of the model, thereby
 584 circumventing the machine learning prerequisite of represen-
 585 tative training data in the same photometric bands and the
 586 template fitting requirement of detailed galaxy SED models.
 587 It models noisy observed flux $\hat{\mathbf{F}} = \mathbf{F} + F_b$ as a sum of a noise-
 588 less flux plus a Gaussian processes $F_b \sim \mathcal{GP}(\mu^F, k^F)$ with
 589 zero mean function μ^F and a physically motivated kernel k^F
 590 that induces realistic correlations in flux-redshift space.

591 From a template-fitting perspective, each test set galaxy
 592 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-
 593 shift z conditioned on noisy flux $\hat{\mathbf{F}}$, where $p(z|T_i)p(T_i)$ cap-
 594 tures prior information about the redshift distributions and
 595 abundances of the galaxy templates T_i . As in traditional
 596 template fitting, each likelihood $p(\hat{\mathbf{F}}|\mathbf{F})$ relates the noisy flux
 597 $\hat{\mathbf{F}}$ with the noiseless flux \mathbf{F} predicted by the model of a linear
 598 combination of templates, carefully constructed to account
 599 for model uncertainties and different normalization of the
 600 same SED, plus the Gaussian process term.

601 The machine learning approach appears in the inclu-
 602 sion of a pairwise comparison term $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$ for the
 603 prediction of model flux \mathbf{F} at a model redshift z with re-
 604 spect to training set galaxy j with redshift z_j and ob-
 605 served flux $\hat{\mathbf{F}}_j$. Thus the photo- z posterior $p(\hat{\mathbf{F}}|z, T_i) =$
 606 $\int p(\hat{\mathbf{F}}|\mathbf{F})p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)d\mathbf{F}$ may be interpreted as the proba-
 607 bility that the training and the target galaxies have the same
 608 SED at different redshifts. The flux prediction $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$
 609 of the training galaxy at redshift z is modeled via the Gaus-
 610 sian process described above; more detail is provided in
 611 Leistedt & Hogg (2017).

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

620 3.3 **trainZ**: a pathological photo- z estimator

621 We also consider a pathological photo- z PDF estimation
 622 method, dubbed **trainZ**, which assigns each test set galaxy
 623 a photo- z PDF equal to the normalized redshift distribution
 624 $N(z)$ of the training set, according to

$$625 p(z|\{z_j\}) \equiv \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \begin{cases} 1 & \text{if } z_k \leq z_i < z_{k+1} \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

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

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 will dedicate some attention to breaking it apart. 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 the true photo- z likelihoods, a mock catalogue generation procedure that has not yet appeared in the literature and was not performed for this work.

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

$$\hat{p}_i^H(z) \equiv p(z|d_i, I_D, I_H), \quad (2)$$

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 metrics of univariate PDFs. We present the metrics of photo- z PDFs that address our goals in the sections below. Section 4.1 outlines aggregate metrics of a catalogue of photo- z PDFs, and Section 4.2 presents a metric of individual photo- z PDFs in the absence of true photo- z PDFs. 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 an ensemble of photo- z PDFs

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)

$$\text{CDF}[f, q] \equiv \int_{-\infty}^q f(z) dz, \quad (3)$$

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 easy way to qualitatively assess the consistency between an estimating 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)

$$\text{PIT} \equiv \text{CDF}[\hat{p}, z_{\text{true}}] \quad (4)$$

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

¹⁷ <http://github.com/aimalz/qp/>

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. Borodolo 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

$$\text{KS} \equiv \max_z \left(|CDF[\hat{f}, z] - CDF[\tilde{f}, z]| \right), \quad (5)$$

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

$$\text{CvM}^2 \equiv \int_{-\infty}^{+\infty} (CDF[\hat{f}, z] - CDF[\tilde{f}, z])^2 dCDF[\tilde{f}, z] \quad (6)$$

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

$$\text{AD}^2 \equiv N_{\text{tot}} \int_{-\infty}^{+\infty} \frac{(CDF[\hat{f}, z] - CDF[\tilde{f}, z])^2}{CDF[\tilde{f}, z](1 - CDF[\tilde{f}, z])} dCDF[\tilde{f}, z] \quad (7)$$

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

$$L(f, \hat{f}) \equiv \int \int (f(z|\mathbf{x}) - \hat{f}(z|\mathbf{x}))^2 dz dP(\mathbf{x}) \quad (8)$$

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

$$\hat{L}(f, \hat{f}) = \mathbb{E}_{\mathbf{x}} \left[\int \hat{f}(z | \mathbf{X})^2 dz \right] - 2\mathbb{E}_{\mathbf{x}, Z} \left[\hat{f}(Z | \mathbf{X}) \right] + K_f, \quad (9)$$

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

5 RESULTS

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

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

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

5.1 Performance on photo- z PDF ensembles

The histogram of PIT values, QQ plot, and QQ difference plot relative to the ideal diagonal are provided in Figure 2, showcasing the biases and trends in the average accuracy of the photo- z PDFs for each code. The high QQ values (more high than low PIT values) of BPz, CMNN, Delight, EAZY, and GPz indicate photo- z PDFs biased low, and the

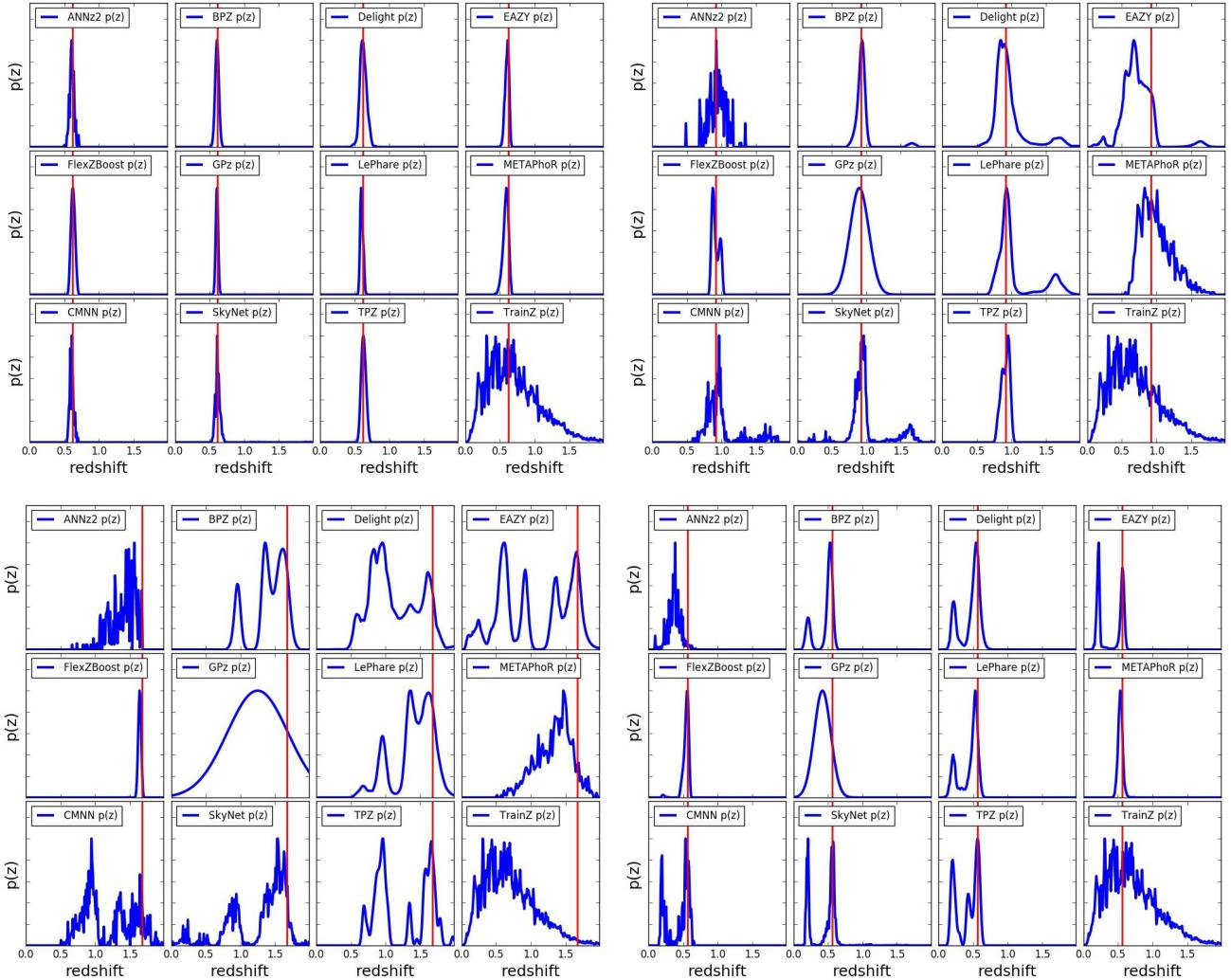


Figure 1. The individual photo- z PDFs (blue) distributions produced by the twelve codes (small panels) on four exemplary galaxies' photometry (large panels) with different true redshifts (red). The photo- z PDFs of all codes share some features for the example galaxies due to physical color degeneracies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

845 low QQ values (more low than high PIT values) of SkyNet
 846 and TPZ indicate photo- z PDFs biased high.

847 The PIT histograms of Delight, CMNN, SkyNet, and TPZ
 848 feature an underrepresentation of extreme values, indicative
 849 of overly broad photo- z PDFs, while the overrepresentation
 850 of extreme values for for METAPhoR indicate overly narrow
 851 photo- z PDFs. These five codes in particular have a free
 852 parameter for bandwidth, which may be responsible for this
 853 vulnerability, in spite of the opportunity for fine-tuning with
 854 perfect prior information. FlexZBoost's "sharpening" para-
 855 meter (described in Section 3.2.3) played a key role in
 856 diagonalizing the QQ plot, indicating a common avenue for
 857 improvement in the approaches that share this type of para-
 858 meter. On the other hand, the three purely template-based
 859 codes, BPz, EAZY, and LePhare, do not exhibit much sys-
 860 tematic broadening or narrowing, which may indicate that
 861 complete template coverage effectively defends from these
 862 effects.

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 > 0.9999
ANNz2	0.0265
BPz	0.0192
Delight	0.0006
EAZY	0.0154
FlexZBoost	0.0202
GPz	0.0058
LePhare	0.0486
METAPhoR	0.0229
CMNN	0.0034
SkyNet	0.0001
TPZ	0.0130
trainZ	0.0002

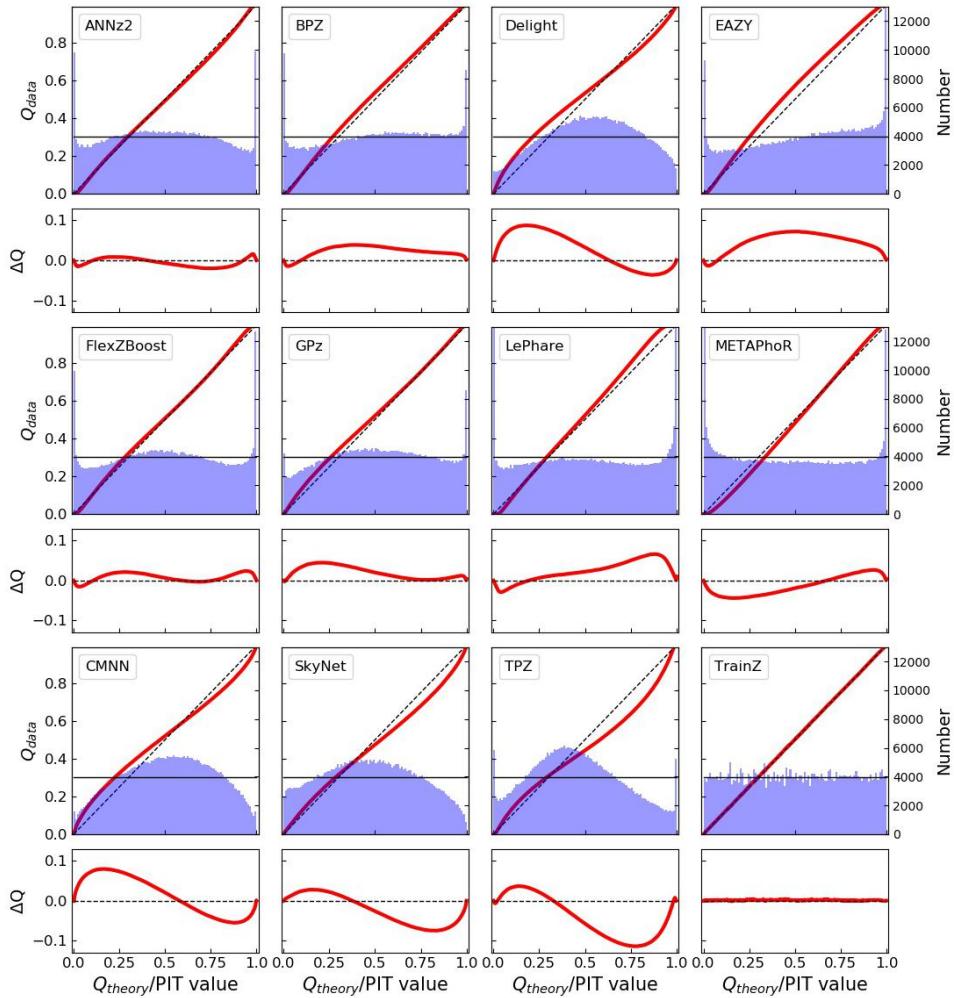


Figure 2. The QQ plot (red) and PIT histogram (blue) of the photo- z PDF codes (panels) along with the ideal QQ (black dashed diagonal) and ideal PIT (gray horizontal) curves, as well as a difference plot for the QQ difference from the ideal diagonal (lower inset). The twelve codes exhibit varying degrees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

Though the spikes in the first and last bin of the PIT histogram were cut off in Figure 2 for visualization, the catastrophic outlier rates are provided in Table 2. As expected, **trainZ** achieves precisely the 0.0002 value expected of an ideal PIT distribution. **ANNz2**, **FlexZBoost**, **LePhare**, and **METAPhoR** have notably high catastrophic outlier rates > 0.02 , exceeding 100 times the ideal PIT rate, meriting further investigation.

Figure 3 displays the values of the KS, CvM, and AD test statistics between the PIT distribution and a uniform distribution $U(0, 1)$, highlighting the relative rather than absolute numbers. **METAPhoR** and **LePhare** perform well under the AD but poorly under the KS and CvM due to their high catastrophic outlier rates. **ANNz2** and **FlexZBoost** are the top scorers under these metrics of the PIT distribution. **ANNz2**'s strong performance can be attributed to an aspect of the training process in which training set galaxies with a

PIT that more closely matches the percentiles of the DC1 training set's redshift distribution are upweighted; in effect, these quantile-based metrics were part of the algorithm itself that may or may not serve it well under more realistic experimental conditions.

5.2 Performance on individual photo- z PDFs

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

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.

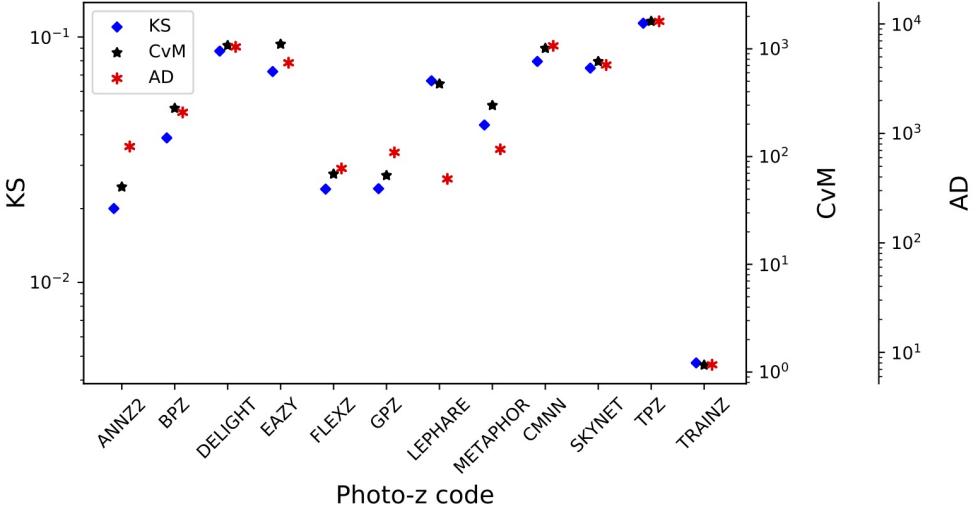


Figure 3. A visualization of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the PIT distributions. There is generally good agreement between these statistics, with differences corresponding to the codes with outstanding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

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
METAPhR	-6.28
CMNN	-10.43
SkyNet	-7.89
TPz	-9.55
trainZ	-0.83

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 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 ??.

Furthermore, our metrics are not necessarily able to assess the fidelity of individual photo-z PDFs relative to true posteriors. Metric-specific performance implies that we may need multiple photo-z PDF approaches tuned to each metric in order to maximize returns over all science cases in large upcoming surveys.

The `trainZ` estimator of Section 3.3, which assigns every galaxy a photo-z PDF equal to $N(z)$ of the training set, is introduced as an experimental control or null test to demonstrate this point via *reductio ad absurdum*. Because our training set is perfectly representative of the test set, $N(z)$ should be identical for both sets down to statistical noise. We make the alarming observation that `trainZ` 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). 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. In addition to outright redshift failures we will model the inclusion of a small number of high-confidence yet false redshifts due to emission line misidentification or noise spikes.

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.

Sequel 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.

6.3 Realistic mock data

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

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

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

7 CONCLUSION

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

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

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
 1124
 1125 J.Y.H. Soo: Ran ANNz2 and Delight, updated abstract,
 1126 edited sections 1 through 6, added tables in Methods
 1127 and Results, updated references.bib and added references
 1128 throughout the paper
 1129
 1130 I.A. Almosallam: vetted the early versions of the data set
 1131 and ran many photo-z codes on it, applied GPz to the final
 1132 version and wrote the GPz subsection
 1133
 1134 M. Brescia: main ideator of METAPHOR and of MLPQNA;
 1135 modification of METAPHOR pipeline to fit the LSST data
 1136 structure and requirements
 1137
 1138 S. Cavaoti: Contributed to choice and test of metrics, ran
 1139 METAPHOR, minor text editing
 1140
 1141 J. Cohen-Tanugi: contributed to running code, analysis
 1142 discussion, and editing, reviewing the paper
 1143
 1144 A.J. Connolly: Developed the colour-matched nearest-
 1145 neighbours photo-z code; participated in discussions of the
 1146 analysis.
 1147
 1148 P.E. Freeman: Contributed to choice of CDE metrics and
 1149 to implementation of FlexZBoost
 1150
 1151 M.L. Graham: Ran the colour-matched nearest-neighbours
 1152 photo-z code on the Buzzard catalog and wrote the relevant
 1153 piece of Section 2; participated in discussions of the analysis.
 1154
 1155 K. Iyer: assisted in writing metric functions used to evaluate
 1156 codes
 1157
 1158 R. Izbicki: Co-developed FlexZBoost and the CDE loss
 1159 statistic, and wrote software for FlexZBoost
 1160
 1161 M.J. Jarvis: Contributed text on AGN to Discussion section
 1162 and portions of GPz work
 1163
 1164 J.B. Kalmbach: Worked on preparing the figures for the
 1165 paper.
 1166
 1167 E. Kovacs: Ran simulations, discussed data format and
 1168 properties for SEDs, dust, and ELG corrections
 1169
 1170 A.B. Lee: Co-developed FlexZBoost and the CDE loss
 1171 statistic, wrote text on the work, and supervised the
 1172 development of FlexZBoost software packages
 1173
 1174 G. Longo: Scientific advise, test and validation of the
 1175 modified METAPHOR pipeline, text of the METAPHOR
 1176 section
 1177
 1178 C. Morrison: Managerial support; Discussions with authors
 1179 regarding metrics and style; Some coding contribution to
 1180 metric computation.
 1181
 1182 J. Newman: Contributions to overall strategy, design of
 1183 metrics, and supervision of work done by Rongpu Zhou
 1184
 1185 E. Nourbakhsh: Ran and optimized TPZ code on the
 1186 Buzzard catalog and wrote a subsection of Section 2 for
 1187 that
 1188
 1189 E. Nuss: contributed to running code, analysis discussion,
 1190 and editing, reviewing the paper
 1191
 1192 T. Pospisil: Co-developed FlexZBoost software and CDE
 1193 loss calculation code
 1194
 1195 H. Tranin: contributed to providing SkyNet results and
 1196 writing the relevant section
 1197
 1198

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 is 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` ([Bennett 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 $\hat{N}(z)$. In terms of the prior information provided to each method, the true redshift distribution satisfies the tautology $\hat{N}(z) = p(z|I_D)$ 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 $\hat{N}(z)$ 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 $\hat{p}_i(z) = \hat{N}(z)$ for all i ; the inclusion of any information from the photometry would only introduce noise to the optimal result of returning 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 we do not endorse the use of the stacked estimator of the redshift distribution, 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 ANNz2, 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.

ANNz2, GPz, and METAPhR show signs of overtraining, estimating enhanced peaks and diminished troughs relative to the training set, 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 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-Smirnov (KS), Cramer-Von Mises (CvM), and Anderson Darling (AD) test statistics for each of the codes for the $\hat{N}(z)$ based measures. The stacked estimators of the redshift distribution for CMNN and trainZ best estimate $\tilde{N}(z)$ under these metrics, and the only codes that do especially poorly are EAZY, LePhare, METAPhR, and SkyNet. It is unsurprising that CMNN scores well, as with a near perfectly representative training set means that choosing neighbouring points in

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 $\hat{N}(z)$			
Estimator	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
trainZ	0.699	0.627	0.666

color/magnitude space should lead to excellent agreement in the final $\hat{N}(z)$ estimate.

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.

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} .

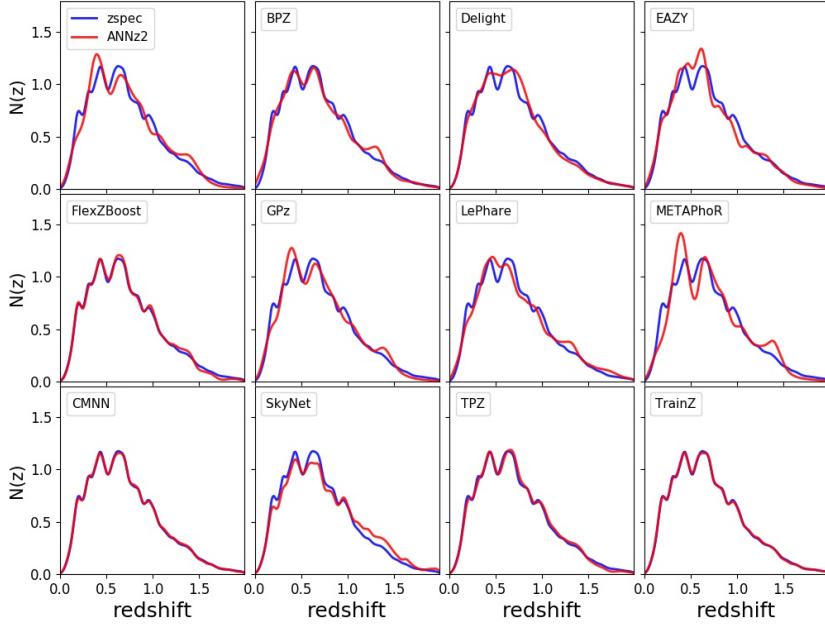


Figure A1. The smoothed stacked estimator $\hat{N}(z)$ of the redshift distribution (red) produced by each code (panels) compared to the true redshift distribution $\tilde{N}(z)$ (blue). Varying levels of agreement are seen among the codes, with the smallest deviations for CMNN, FlexZBoost, TPZ, and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

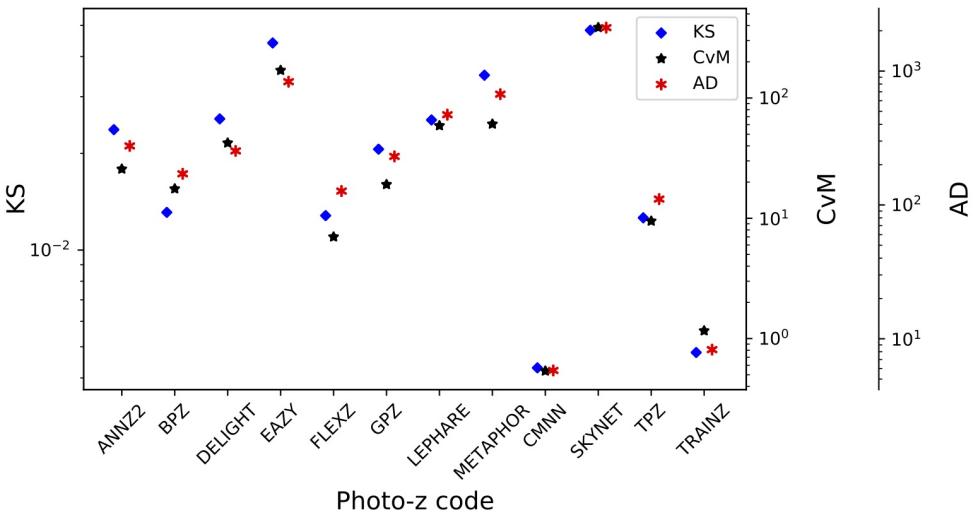


Figure A2. A visualization of the Kolmogorov-Smirnoff (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the $\hat{N}(z)$ distributions. We make the reassuring observation that these related statistics do not disagree significantly with one another. CMNN performs comparably well to trainZ, the control case, while SkyNet scores poorly due to an overall bias in its redshift predictions.

(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.

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 scat-

¹²⁷¹ (Dahlen et al. 2013), a weighted mean within the bounds of ¹²⁷⁷
¹²⁷² the main peak, as identified by the roots of $p(z) - 0.05 \times$

¹²⁷³ z_{PEAK} . Though z_{WEIGHT} neglects information in a sec- ¹²⁷⁸
¹²⁷⁴ ondary peak of a, say, bimodal distribution, it avoids the ¹²⁷⁹
¹²⁷⁵ pitfall of reducing the photo-z PDF to a redshift between ¹²⁸⁰
¹²⁷⁶ peaks where there is low probability. ¹²⁸¹
¹²⁷⁷ $z_{PEAK} - z_{true})/(1 + z_{true})$. Because the standard deviation of the ¹²⁸²
¹²⁷⁸ photo-z residuals is sensitive to outliers, we define the scat-

ter in terms of the Interquartile Range (IQR), the difference between the 75th and 25th percentiles of the distribution of e_z , imposing the scaling $\sigma_{\text{IQR}} = \text{IQR}/1.349$ 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 $\max(3\sigma_{\text{IQR}}, 0.06)$.

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

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 $z_{\text{phot}} = z_{\text{spec}}$ 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 $z_{\text{phot}} = z_{\text{spec}}$ line are due to the 4000Å break passing through the gaps between adjacent filters. Even the strongest codes feature populations far from the $z_{\text{phot}} = z_{\text{spec}}$ line representing a degeneracy in the space of colours and redshifts.

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 $i < 25$.

REFERENCES

- Abbott T., et al., 2005, preprint (arXiv:astro-ph/0510346)
 Abell P. A., et al., 2009, preprint (arXiv:0912.0201),
 Aihara H., et al., 2018a, *PASJ*, **70**, S4
 Aihara H., et al., 2018b, *PASJ*, **70**, S8
 Almosallam I. A., Lindsay S. N., Jarvis M. J., Roberts S. J., 2016a, *MNRAS*, **455**, 2387
 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., Cavaudi S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2018, preprint, (<arXiv:1802.07683>)
 Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, **432**, 1483
 Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, **442**, 3380
 Cavaudi S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo G., 2017, *MNRAS*, **465**, 1959
 Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. ACM, New York, NY, USA, pp 785–794, doi:10.1145/2939672.2939785, <http://doi.acm.org/10.1145/2939672.2939785>
 Connolly A. J., et al., 2014, in Angeli G. Z., Dierickx P., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9150, Modeling, Systems Engineering, and Project Management for Astronomy VI. p. 14, doi:10.1117/12.2054953
 Dahlen T., et al., 2013, *ApJ*, **775**, 93
 Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016, *ApJ*, **816**, 11
 Erben T., et al., 2013, *MNRAS*, **433**, 2545
 Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, *ApJ*, **513**, 34
 Firth A. E., Lahav O., Somerville R. S., 2003, *MNRAS*, **339**, 1195
 Freeman P. E., Izbicki R., Lee A. B., 2017, *MNRAS*, **468**, 4556
 Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, *MNRAS*, **441**, 1741
 Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones R. L., Jurić M., Daniel S. F., Yoachim P., 2018, *AJ*, **155**, 1
 Green J., et al., 2012, preprint (arXiv:1208.4012),
 Hildebrandt H., et al., 2010, *A&A*, **523**, A31
 Hofmann B., Mathé P., 2018, *Inverse Problems*, **34**, 015007
 Hunter J. D., 2007, Matplotlib: A 2D Graphics Environment, doi:10.1109/MCSE.2007.55
 Ilbert O., et al., 2006, *A&A*, **457**, 841
 Ivezić Ž., et al., 2008, preprint (arXiv:0805.2366),
 Izbicki R., Lee A. B., 2017, *Electron. J. Statist.*, **11**, 2800
 Izbicki R., Lee A. B., Freeman P. E., 2017, *Ann. Appl. Stat.*, **11**, 698
 Laureijs R., et al., 2011, preprint (1110.3193),
 Leistedt B., Hogg D. W., 2017, *ApJ*, **838**, 5
 Malz A., Hogg D., in prep., CHIPPR, chippr
 Malz A., Marshall P., DeRose J., Graham M., Schmidt S., Wechsler R., 2018, AJ, Accepted,
 Mandelbaum R., et al., 2008, *MNRAS*, **386**, 781
 Massarotti M., Iovino A., Buzzoni A., 2001, *A&A*, **368**, 74
 Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltoni S., 2017, *ApJ*, **841**, 111
 Newman J. A., et al., 2015, *Astroparticle Physics*, **63**, 81
 Oliphant T., 2007, Python for Scientific Computing, doi:10.1109/MCSE.2007.58
 Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, *ApJ*, **689**, 709
 Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint (arXiv:1608.08016),
 Rasmussen C., Williams C., 2006, Gaussian Processes for Machine Learning, Adapative computation and machine learning series, MIT Press, Cambridge, MA
 Rau M. M., Seitz S., Brimouille F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, *MNRAS*, **452**, 3710
 Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, *ApJ*, **771**, 30
 Sadeh I., Abdalla F. B., Lahav O., 2016, *PASP*, **128**, 104502
 Sánchez C., et al., 2014, *MNRAS*, **445**, 1482

¹⁸ available at: <http://ls.st/srd>

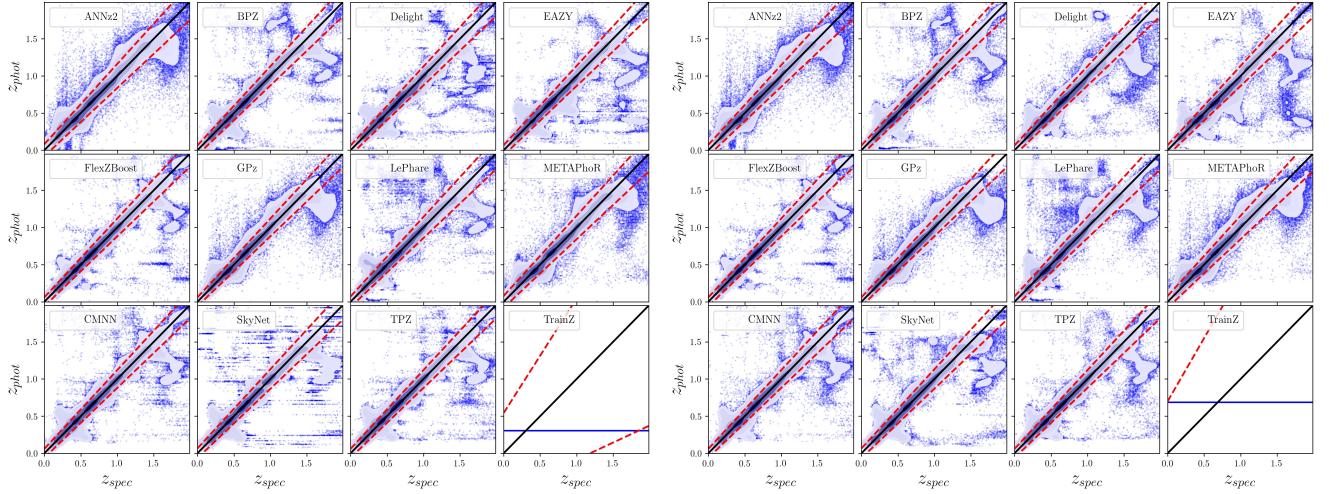


Figure B1. The density of photo- z point estimates (contours) reduced from the photo- z PDFs with outliers (blue) beyond the outlier cutoff (red dashed lines), via the mode (z_{PEAK} , left panel) and main-peak-mean (z_{WEIGHT} , right panel). The **trainZ** estimator (lower right sub-panels) has a shared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

Table B1. Photo- z point estimate statistics

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

- 1406 Schmidt M., 2005, minFunc: Unconstrained Differentiable Multi-
 1407 variate Optimization in Matlab, <http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>
 1408
 1409 Scott D. W., 1992, Multivariate Density Estimation. Theory,
 Practice, and Visualization. Wiley
 1410 Skrutskie M. F., et al., 2006, AJ, 131, 1163
 1411 Tanaka M., et al., 2018, PASJ, 70, S9
 1412 The LSST Dark Energy Science Collaboration et al., 2018,
 preprint, ([arXiv:1809.01669](https://arxiv.org/abs/1809.01669))
 1413 Waskom M., et al., 2017, doi:10.5281/zenodo.824567
 1414 York D. G., et al., 2000, AJ, 120, 1579
 1415 de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn
 1416 E. A., 2013, Exp. Astron., 35, 25
 1417 de Jong J. T. A., et al., 2017, A&A, 604, A134