Beyond Point Photo-z Estimates: Normalizing Flows for Tomographic $N(z)$ Reconstruction with DESI DR1

Zhuo Chen,[1,2], Antoine Rocher[2], Kwan Chuen Chan,[1,3]★, Jean-Paul Kneib[2]
[1]School of Physics and Astronomy, Sun Yat-Sen University, 2 Daxue Road, Tangjia, Zhuhai 519082, China
[2] Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland
[3] CSST Science Center for the Guangdong-Hongkong-Macau Greater Bay Area, SYSU, Zhuhai 519082, China

arXiv:2611.00121v1 [astro-ph.CO] 29 Oct 2026

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Accurate calibration of tomographic redshift distributions, $N(z)$ is essential for cosmological analysis of weak lensing data. We investigate conditional normalizing flows (NFs) and compare their performance with random forest (RF) and quantile random forest (QRF) models. We consider DESI galaxy and quasar targets from DESI DR1 spectroscopic data and with the photometry from Legacy Surveys DR10. We train on 1.1 million sources and test on 2.0 million sources in a non-overlapping sky region ($0 < z_{\rm spec} < 4.5$). For point estimation, the NF achieves the lowest scatter and outlier fraction at $0 < z_{\rm spec} < 1.6$ and outperforms both baselines across all three metrics at $z_{\rm spec} > 1.6$. By introducing distribution-level metrics like Wasserstein distance and JS divergence, we find that stacking NF posterior PDFs provides more accurate tomographic $N(z)$ reconstruction than QRF, particularly at high redshifts. For $z_{\rm median} > 2.5$, the NF achieves a smaller absolute mean-redshift bias (0.060 ± 0.031 versus 0.318) and higher probability overlap (0.905 ± 0.008 versus 0.787) than QRF. We apply redshift-dependent post-hoc PDF calibration derived from a held-out validation sample. After calibration, the absolute mean-redshift biases in four bins spanning $0.2 < z_{\rm median} < 1.0$ range from 0.0016 to 0.0019, with central values below referring benchmark $0.002(1+z)$. In the highest-redshift bin, calibration further reduces the bias to 0.016 ± 0.011. Our results demonstrate the potential of NF PDFs for tomographic $N(z)$ reconstruction while highlighting challenges in transferring models across galaxy populations.

Key words: Galaxy photometry – Neural network – Sky surveys – Photometric Redshift estimation – Cosmology

1 INTRODUCTION

Wide-field imaging surveys rely heavily on photometric redshift (photo-z) estimation to infer redshift information for millions to billions of galaxies. Photo-zs are estimated from broadband photometry and therefore have substantially larger uncertainties than spectroscopic measurements. In particular, photo-z techniques will enable the construction of enormous galaxy samples that are essential for Stage IV cosmological surveys such as Euclid (Laureijs et al. 2011), the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) (Ivezić et al. 2019), and the Chinese Space Station Survey Telescope (CSST) (CSST Collaboration et al. 2026). In these surveys, the accuracy of cosmological analyses depends not only on the precision of individual galaxy redshift estimates but also on the accurate reconstruction of the galaxy true redshift distribution, $N(z)$ of the sample. Weak lensing tomography and galaxy clustering analyses require precise measurement of $N(z)$ within each photometric bin. For example, the Euclid weak lensing survey requires the uncertainty in the mean redshift of each tomographic bin to be controlled at the level of $|\Delta z| \lesssim 0.002(1+z)$ (Euclid Collaboration 2020, 2021). Achieving this level of precision represents a major challenge for photometric redshift methodologies.

In recent years, machine learning (ML) methods have become widely adopted for photometric redshift estimation (Salvato et al. 2019; Newman & Gruen 2022). ML methods can learn complex nonlinear relationships between photometric observables and galaxy redshifts, often outperforming traditional template-fitting approaches in terms of prediction accuracy and computational efficiency. Various algorithms have been explored including Multi-Layer Perceptron (Collister & Lahav 2004; Sadeh et al. 2016), Nearest Neighbor Fitting (De Vicente et al. 2016), Random Forest (Carliles et al. 2010; Zhou et al. 2021; Lu et al. 2024), (Boosted) Decision Tree (Gerdes et al. 2010; Carrasco Kind & Brunner 2013), Sparse Gaussian Process (Almosallam et al. 2016), Convolutional Neural Network (Hoyle 2016; D'Isanto & Polsterer 2018; Schuldt et al. 2021; Li et al. 2022; Zhou et al. 2022), and Normalizing Flow (NF) (Crenshaw et al. 2024; Sun et al. 2023; Ren et al. 2026).

Traditional ML approaches often focus on predicting a single redshift point estimate. Beyond point estimates such as the mean or median redshift, it is essential to characterize the uncertainty in redshift predictions, which can be represented by the full redshift probability density function (PDF). Normalizing Flows (NFs) (Tabak & Turner 2013; Rezende & Mohamed 2015) provide a flexible framework for

★ E-mail:chankc@mail.sysu.edu.cn (SYSU)

© 2026 The Authors

modelling complex probability distributions through a sequence of invertible transformations from a simple base distribution to a target distribution. Their ability to perform exact likelihood evaluation, efficient sampling, and expressive modelling of non-Gaussian and multimodal densities makes them well suited to photo-z estimation, where redshift PDFs can exhibit asymmetric shapes, extended tails, and multiple peaks due to photometric degeneracies.

Accurate reconstruction and calibration of tomographic redshift distributions, $N(z)$, are essential for reliable cosmological inference, as systematic errors in these distributions can propagate into cosmological parameter estimates. Existing approaches to redshift-distribution calibration include self-organizing maps (SOM) (Carrasco Kind & Brunner 2014; Masters et al. 2015; Buchs et al. 2019; Wright et al. 2020) and clustering-based methods. The latter can be divided into clustering-redshift approaches using external spectroscopic reference samples (Newman 2008; McQuinn & White 2013; Ménard et al. 2013; Schmidt et al. 2013), self-calibration methods using photometric clustering alone (Schneider et al. 2006; Zhang et al. 2010, 2017; Peng et al. 2022; Xu et al. 2023), and combinations of these approaches (Zheng et al. 2024). Clustering-redshift techniques have been widely applied in different surveys (Gatti et al. 2018; Cawthon et al. 2022; Hildebrandt et al. 2021; Rau et al. 2023; Choppin de Janvry et al. 2026). However, recovering the detailed shape of $N(z)$ can be challenging due to degeneracies with galaxy bias and other observational effects. In this context, redshift distributions reconstructed from photometric redshift PDFs can provide complementary information that may be combined with clustering-based constraints.

Stacking individual photo-z PDFs has been widely used to reconstruct redshift distributions, particularly for weak-lensing tomography (Bordoloi et al. 2010; Bonnett 2015; Euclid Collaboration 2021). More recently, Ren et al. (2026) applied conditional NFs to photo-z estimation, focusing primarily on point-estimate accuracy, while Tian et al. (2026) developed a discretized photo-z PDF estimator using spectroscopic samples including DESI DR1 and evaluated the overall stacked $N(z)$. Building on these studies, we systematically quantify the performance of continuous NF PDFs in tomographic $N(z)$ reconstruction across photometric-redshift bins, compare their performance with QRF, and assess the impact of post-hoc PDF calibration, including at high redshifts.

To investigate the potential of NFs for applications in large galaxy surveys, we construct million-scale datasets by combining spectroscopic redshifts from DESI Data Release 1 (DR1) (DESI Collaboration 2026) with photometric measurements from the DESI Legacy Surveys (Dey et al. 2019). We evaluate the models using independent spectroscopic testing samples from non-overlapping sky regions, allowing us to assess probabilistic photo-z estimation and $N(z)$ reconstruction within a controlled spectroscopic sample. We also compare the NF with Random Forest (RF) and Quantile Random Forest (QRF) baselines.

Importantly, the NF framework is not inherently specific to DESI, but provides a flexible methodology for probabilistic photo-z estimation that can be adapted to different photometric surveys. The conditioning features can be adjusted according to the available photometric measurements, allowing the framework to be retrained and calibrated for different galaxy populations and observational datasets, including those from Euclid and LSST. However, the performance obtained from the DESI-selected spectroscopic samples should not be interpreted as the expected accuracy for the broader photometric populations of these surveys. Extending the methodology to such populations will require careful assessment of spectroscopic representativeness and selection effects.

This paper is organized as follows: Section 2 describes the DESI DR1 spectroscopic sample and the Legacy Surveys DR10 photometric data, while Section 3 introduces the NF, RF, and QRF methods and the $N(z)$ reconstruction metrics. Section 4 presents the main results, including point-estimation performance, tomographic reconstruction, PDF calibration, and high-redshift tests. We discuss the implications and limitations of the NF approach in Section 5 and summarize our conclusions in Section 6.

2 DATASETS

2.1 Spectroscopic data

In this work, we rely on the spectroscopic measurement from the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration (2022)), an eight-years program that aims to observe over 60 million galaxies and quasars across approximately 17 000 deg^2 of the extragalactic sky. Its first public Data Release 1 (DR1; DESI Collaboration (2026)) contains approximately 18.7 million high-quality spectra from the Main Survey, including about 13.0 million galaxies and 1.6 million quasars with well-characterized uncertainties and quality flags. The large sample size and broad redshift coverage of DESI DR1 provide a valuable dataset for training and evaluating probabilistic photo-z models, with spectroscopic redshift distributions, $N_{\rm spec}(z)$, serving as references for assessing the accuracy of reconstructed $N(z)$.

2.2 Photometric features

The DESI target selection is primarily based on colour and magnitude information from the DESI Legacy Imaging Surveys DR9 (DESI Collaboration 2022), which combine optical imaging from the Dark Energy Camera Legacy Survey (DECaLS), the Beijing–Arizona Sky Survey (BASS), and the Mayall z-band Legacy Survey (MzLS) with infrared observations from WISE (Wright et al. 2010; Dey et al. 2019).

For our analysis, we use the Legacy Surveys DR10.1 release[1], which additionally includes DECam i-band imaging (Flaugher et al. 2015), providing extended wavelength coverage for photometric redshift estimation.

The DESI spectroscopic sample comprises several target classes with different selection criteria. The main magnitude limits include $r < 19.5$ for BGS Bright, $19.5 < r < 20.175$ for BGS Faint, $z_{\rm fiber} < 21.6$ for LRGs, $g_{\rm fiber} < 24.1$ for ELGs, and $16.5 < r < 23.0$ for QSOs (Hahn et al. 2023; Zhou et al. 2023; Raichoor et al. 2023; Chaussidon et al. 2022). These magnitude limits, together with the additional colour and quality selections, imply that the DESI spectroscopic sample is not necessarily representative of the broader photometric galaxy population, as it preferentially samples specific regions of colour–magnitude space.

As input features to train the photometric redshift calibration, we use fluxes and inverse variances (IVAR) in the g, r, i, z, $W1$, and $W2$ bands. The fluxes are converted into extinction-corrected magnitudes as

$$m_{\rm band} = 22.5 - 2.5\log_{10}\left(\frac{F_{\rm band}}{T_{\rm MW,band}}\right), \quad (1)$$

where $T_{\rm MW,band}$ accounts for Galactic extinction using the extinction coefficients from Schlegel et al. (1998); Schlafly & Finkbeiner

[1] At the beginning of this analysis, DR11 was not yet available.

(2011). Additionally, we include the colours $g-r$, $r-i$, $i-z$, $z-W1$ and $W1-W2$, which provide complementary information. Finally we incorporate galaxy properties beyond photometric information using morphological and observational features including the half-light radius (`shape_r`), Sérsic index (`sersic`), and PSF depths in the g, r, i, and z bands.

2.3 Training and testing datasets

For this analysis, we construct an initial parent sample of approximately 3×10^7 sources from the DESI Legacy Imaging Surveys DR10 photometric catalogue (Dey et al. 2019). We identify sources with corresponding spectroscopic information from DESI DR1 to construct the dataset for photometric-redshift estimation. We then apply the following quality selections to ensure reliable photometric and spectroscopic measurements:

(i) sources are DESI galaxy or QSO targets
(ii) Positive flux measurements in all bands:

$\mathrm{Flux}_{\mathrm{all}} > 0.$

(iii) Reliable spectroscopic redshifts:

ZCAT_PRIMARY = True, ZWARN = 0.

(iv) Signal-to-noise ratio greater than 3 in all bands:

$\mathrm{SNR}_{\mathrm{Flux}_{\mathrm{all}}} > 3.$

After these selections, about 6×10^6 objects remain, including stars, galaxies and quasars with redshift ranging around $0 < z < 4.5$. Instead of training separate models for different galaxy populations or source classes (e.g. Bright Galaxies, Luminous Red Galaxies, and quasars), we train a single model on the full sample. This allows the NF to learn a unified mapping between photometric observables and redshift without introducing additional selection effects associated with population-specific pre-selection. We also do not remove stars or other contaminants using spectroscopic information, such as z_{spec} or spectroscopic classifications. This choice is intended to ensure that the pipeline can be applied directly to realistic imaging surveys, where such spectroscopic information is not available for pre-selection.

To evaluate the transferability of the model across different sky regions, we divide the dataset into two spatially separated areas within the DESI footprint. The first region ($80° < \mathrm{RA} < 160°$, $-15° < \mathrm{DEC} < 40°$) is used for training and validation, while the second region ($160° < \mathrm{RA} < 240°$, $-15° < \mathrm{DEC} < 40°$) is reserved as an independent test set, as shown in Fig. 1. The normalised distributions of photometric features and observational features for these sources in two patches are shown in the Fig. A2 in Appendix.

Within the training region, 10% of the objects are randomly selected to form a validation set, while the remaining 90% are used for training. The final dataset comprises 1,094,347 training sources, 121,594 validation sources, and the testing set has 2,028,371 sources (including 252,719 quasars and 25,511 stars). The spectroscopic redshift distributions of the training and testing samples are shown in Fig. 2. Both samples exhibit a pronounced excess at very low redshift ($z_{\mathrm{spec}} < 0.005$), which is mainly due to stellar contamination.

3 METHODS

In this section, we first describe the three methods used for photo-z estimation: NF, random forest and quantile random forest. We then introduce the photometric redshift distributions by stacking PDFs and the metrics used to quantify its accuracy relative to the ground spectroscopic redshift distribution.

3.1 Normalizing flows

Below we briefly review the method used for photo-z estimation and refer the reader to Ren et al. (2026) for more details. NFs are generative probabilistic models that construct complex probability distributions by applying a sequence of invertible and differentiable transformations to a simple base distribution (Dinh et al. 2015). Their tractable change-of-variables formulation allows both exact likelihood evaluation and efficient sampling.

For photo-z estimation, we use a conditional NF to model the redshift distribution $p(z \mid \mathbf{x}_{\mathrm{phot}})$, where $\mathbf{x}_{\mathrm{phot}}$ denotes the input photometric features. Let u denote a latent variable drawn from a simple base distribution, $p_U(u)$, which is taken to be Gaussian. The conditional flow defines an invertible transformation $z = g(u; \mathbf{x}_{\mathrm{phot}})$ with inverse $u = f(z; \mathbf{x}_{\mathrm{phot}})$. The conditional density is then obtained from the change-of-variables formula,

$$p(z \mid \mathbf{x}_{\mathrm{phot}}) = p_U\ \left(f(z; \mathbf{x}_{\mathrm{phot}})\right) \left|\frac{df}{dz}\right|. \tag{2}$$

The flow is composed of a sequence of invertible transformations, and the model parameters are optimized by maximizing the mean conditional log-likelihood of the training sample,

$$\mathcal{L} = \frac{1}{N} \sum_{j=1}^{N} \log p\left(z_j \mid \mathbf{x}_{\mathrm{phot},j}\right), \tag{3}$$

where N is the number of training targets.

We adopt the same network implementation and hyperparameters as Ren et al. (2026). Following Ren et al. (2026), we use the Neural Spline Flows (NSFs) (Durkan et al. 2019) to implement the flow transformation for its capability to express complex transformation and hence superior performance in redshift estimation. In practice, the NF network uses a standard normal base distribution and three conditional neural spline layers, each parameterized by an MLP with three hidden layers of 32 neurons and a spline bound of 6.0. The model is optimized using the Adam optimizer with a weight decay of 10^{-4}, with gradient clipping at 10.

Before training, the spectroscopic redshifts and input photometric features are standardised using z-score normalisation to have zero mean and unit variance. The model is trained for 75 epochs, with the learning rate reduced by a factor of 10 after 50 epochs. Training is performed by maximising the mean log-likelihood defined in Eq. 3, or equivalently, minimising its negative value. The validation loss is monitored throughout training to assess potential overfitting. Towards the end of training, both the training and validation losses fluctuated around approximately stable values without sustained improvement and remained close to each other, motivating the adopted training length of 75 epochs.

Training the model for 75 epochs with approximately 1.1 million sources requires about 6 hours on an NVIDIA RTX 4060 GPU. To assess training consistency, we evaluate the final loss values of five independently trained NF models on the same training and validation samples. In this case, the resulting mean training and validation losses are -1.1499 ± 0.0234 and -1.1488 ± 0.0238, respectively, where the uncertainties represent the standard deviations across the five realisations with different random seeds. The close agreement between the training and validation losses suggests no substantial generalisation gap in the final models.

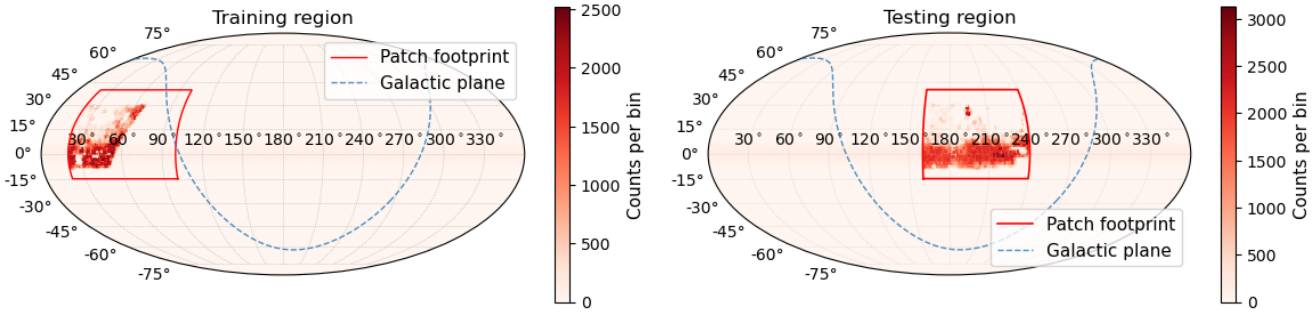


Figure 1. Spatial footprints and number densities of the training and testing patches in equatorial coordinates using a Mollweide projection. The red lines indicate the adopted patch boundaries, while the blue dashed line shows the Galactic plane.

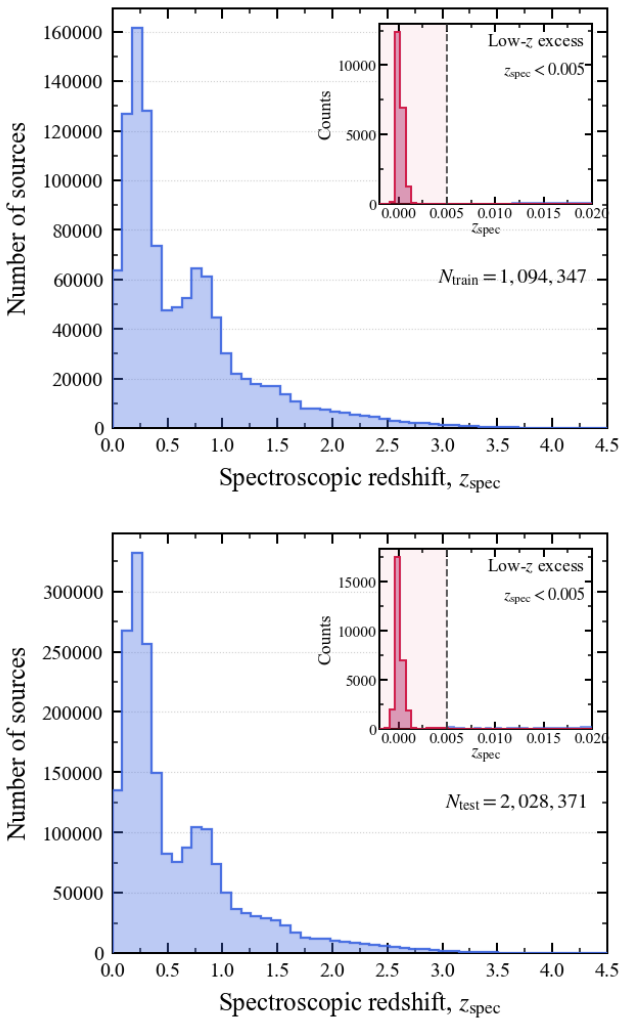


Figure 2. Upper panel: Spectroscopic redshift distribution of the training sample ($N_{\rm train}$ = 1,094,347). Lower panel: Spectroscopic redshift distribution of the testing sample $N_{\rm test}$ = 2,028,371. The inset panels enlarge the very low-redshift range, $-0.002 < z_{\rm spec} < 0.02$, highlighting the pronounced excess at $z_{\rm spec} < 0.005$, which is primarily due to stellar contamination.

For a galaxy with photometric features $\mathbf{x}_{\rm phot}$, the trained model predicts the conditional redshift PDF, $p(z \mid \mathbf{x}_{\rm phot})$. The full predicted PDF can be retained for probabilistic analyses or summarised using point estimates. We adopt the posterior median, $z_{\rm median}$, as the primary point estimate because of its robustness to asymmetric and multimodal distributions. For each object in the test sample, we draw $N_{\rm sample}$ = 200 Monte Carlo samples from its predicted conditional distribution:

$$z \sim p(z \mid \mathbf{x}_{\rm phot}). \tag{4}$$

We adopt fewer samples than the $N_{\rm sample}$ = 2000 used by Ren et al. (2026) to reduce the computational cost for the large testing dataset.

3.2 Random Forest and Quantile Random Forest

Additionally, we use Random Forest (RF; Breiman (2001)) and Quantile Random Forest (QRF; Meinshausen (2006)) as baseline methods because of their computational efficiency, robustness, and widespread use in astronomical applications (Salvato et al. 2019; Newman & Gruen 2022). RF is an ensemble-learning method that combines the predictions of multiple decision trees. For photo-z estimation, each tree is trained using photometric features and spectroscopic redshifts, and the final point estimate is obtained by averaging the predictions over the ensemble. QRF extends RF by estimating conditional quantiles of the target variable rather than only its conditional mean. These quantiles provide a discrete representation of the conditional redshift distribution. Increasing the number of quantiles can improve the resolution of this representation, although estimates at extreme or closely spaced quantile levels may become noisy in sparsely sampled regions of the photometric feature space. Consequently, QRF generally provides less smooth PDFs than those produced by the NF model.

The RF is implemented using the `RandomForestRegressor` class from `scikit-learn`[2]. To train the RF, we use 500 decision trees with `max_depth=None`, `min_samples_leaf=5`, and `max_features="sqrt"`. Missing input values are handled using median imputation. The QRF is implemented using the `quantile_forest` package[3]. The QRF uses 500 decision trees with `max_depth=20`, `min_samples_leaf=5`, and `max_features="sqrt"`. Its conditional redshift distribution is evaluated at 99 uniformly spaced quantile levels between 0.01 and 0.99. Both models use the same training and testing samples as the NF model (see Sec. 2.3), ensuring a consistent comparison.

[2] https://scikit-learn.org/stable/
[3] https://zillow.github.io/quantile-forest/

3.3 Redshift-distribution reconstruction and evaluation metrics

The photometric redshift distribution is reconstructed by stacking the normalised PDFs of individual source (Bordoloi et al. 2010; Bonnett 2015):

$$N_{\rm phot}(z) = \sum_{i=1}^{N_{\rm gal}} P_i(z), \tag{5}$$

where $P_i(z)$ is the normalised redshift PDF of the $i^{\rm th}$ source. For methods that provide only point estimates, such as RF, the distribution is replaced by the Dirac delta function distribution. In practice, the point estimates are binned using the same redshift bins as the PDF-based reconstruction.

To evaluate the performance metrics of the photo-z point estimation, we use the normalised redshift error

$$\Delta z = \frac{z_{\rm phot} - z_{\rm spec}}{1 + z_{\rm spec}}, \tag{6}$$

and three commonly used summary statistics (Salvato et al. 2019). The scatter is quantified by the normalised median absolute deviation,

$$\sigma_{\rm NMAD} = 1.4868\,{\rm median}\left(|\Delta z - {\rm median}(\Delta z)|\right), \tag{7}$$

while the absolute median bias is defined as

$$|{\rm bias}| = |{\rm median}(\Delta z)|\,. \tag{8}$$

We further define catastrophic outliers as sources satisfying

$$|\Delta z| > 0.15. \tag{9}$$

The point estimates of the NF model are obtained from the predicted posterior $p(z \mid \mathbf{x})$ using its mean, median, or mode ($z_{\rm mean}$, $z_{\rm median}$, and $z_{\rm maxp}$, respectively) (Crenshaw et al. 2024; Ren et al. 2026). We adopt $z_{\rm median}$ as the primary estimator because of its robustness to asymmetric and multimodal PDFs as Ren et al. (2026).

Furthermore, we introduce several performance metrics to compare the reconstructed photometric redshift distribution, $N_{\rm phot}(z)$ and the spectroscopic reference, $N_{\rm spec}(z)$. Both distributions are normalised to unit integral before these metrics are evaluated. The mean-redshift bias is defined as

$$\Delta z_{\rm mean} = \langle z\rangle_{\rm phot} - \langle z\rangle_{\rm spec}, \tag{10}$$

where $\langle z\rangle_{\rm phot}$ and $\langle z\rangle_{\rm spec}$ are respectively the mean redshifts of photometric redshift distribution $N_{\rm phot}(z)$ and spectroscopic redshift distribution $N_{\rm spec}(z)$. This quantifies the displacement of the reconstructed distribution relative to the spectroscopic reference. We also measure the difference of their widths using

$$\Delta\sigma = \sigma_{\rm phot} - \sigma_{\rm spec}, \tag{11}$$

where $\sigma_{\rm phot}$ and $\sigma_{\rm spec}$ are the standard deviation of both distributions. Differences in the overall distribution shape are quantified using the first-order Wasserstein distance (Villani 2009):

$$W_1 = \int \left|F_{\rm phot}(z) - F_{\rm spec}(z)\right| {\rm d}z, \tag{12}$$

where $F_{\rm phot}$ ($F_{\rm spec}$) denotes the Cumulative Distribution Function (CDF) derived from $N_{\rm phot}$ ($N_{\rm spec}$). This metric is sensitive to shifts and broad distortions of the reconstructed distribution. Additionally, we use the Jensen–Shannon (JS) divergence (Lin 1991),

$${\rm JS}(N_{\rm phot} \parallel N_{\rm spec}) = \frac{1}{2}D_{\rm KL}(N_{\rm phot} \parallel M) + \frac{1}{2}D_{\rm KL}(N_{\rm spec} \parallel M), \tag{13}$$

where $M = \frac{1}{2}(N_{\rm phot} + N_{\rm spec})$ and $D_{\rm KL}$ is the Kullback-Leibler divergence (Kullback & Leibler 1951). The JS divergence is symmetric and measures differences in both the central and tail regions of the distributions. Finally, the shared probability mass is quantified using the overlap coefficient

$${\rm Overlap} = \int \min\left[N_{\rm phot}(z), N_{\rm spec}(z)\right] {\rm d}z. \tag{14}$$

For distributions normalised to unit integral, the overlap ranges from 0 to 1, with larger values indicating closer agreement. Together, these metrics characterize differences in the location, width, shape, and overall agreement between the reconstructed redshift distribution, $N_{\rm phot}(z)$, and the spectroscopic reference distribution, $N_{\rm spec}(z)$.

4 RESULTS

4.1 Point-estimation performance

The 3 models are trained with $\sim$ 1.1 million sources, using the photometric magnitudes and colours ($g, r, i, z, W1, W2, g-r, r-i, i-z, z-W1, W1-W2$), together with the morphological and observational features described in Section 2.2. The influence of the individual photometric features for NF model is presented in Appendix C, where the colours like $g-r$, $r-i$ and $i-z$ show strongest sensitivity in NF model training based on Fig. C1. We use $z_{\rm median}$ as the point estimate for both the NF and QRF models, while the RF directly provides a single redshift prediction.

The quantitative comparison of point-estimation performance is summarized in Table 1 for two redshift ranges, $0 < z_{\rm spec} < 1.6$ dominated by galaxies, and $z_{\rm spec} > 1.6$ dominated by quasars. At low redshift, the NF achieves the lowest scatter, $\sigma_{\rm NMAD} = 0.0223 \pm 0.0002$, and outlier fraction, $(3.79 \pm 0.03)\%$, while maintaining a competitive absolute median bias of $(6.54 \pm 5.40) \times 10^{-4}$. The QRF yields the smallest median bias, 4.8×10^{-5}. At high redshift, the NF performs best across all three metrics, with an absolute median bias of 0.0430 ± 0.0034, $\sigma_{\rm NMAD} = 0.1015 \pm 0.0011$, and an outlier fraction of $(25.74 \pm 0.51)\%$, indicating a clearer advantage in the more challenging high-redshift regime. The comparison between photometric and spectroscopic redshifts for the three methods is shown in Fig. A1 in the Appendix. The reported uncertainties for NF are estimated as the standard deviations across five independent training realisations with different random seeds and data shuffling. The uncertainties for RF and QRF are not reported because the corresponding seed-to-seed variations are below the displayed precision. To summarise, the NF is competitive for point estimation, with a clearer advantage at high redshift, but it does not consistently outperform RF or QRF in the low-redshift regime. Our results agree with the findings in Ren et al. (2026), which applied the conditional NF on different datasets and concluded that NF excels when the data are noisy and sparse by comparing with various photo-z estimation methods.

The main strength of NF becomes more apparent when the full predicted PDFs are retained, as they preserve information on the uncertainty, width, asymmetry, and extended tails of the redshift distribution.

.

4.2 Tomographic N(z) reconstruction

In this subsection, we present the reconstruction results for the overall and tomographic redshift distributions, $N(z)$, obtained by stacking the predicted photo-z PDFs. We compare the performance of the NF and QRF models using the metrics defined in Section 3.3, with the spectroscopic redshift distributions from DESI DR1 as the reference.

Firstly, we consider reconstructing the overall redshift distributions

Table 1. Summary of the point-estimation performance metrics for the Normalizing Flow (NF), quantile random forest (QRF), and random forest (RF) models. For NF and QRF, the median of the predicted redshift distribution, $z_{\rm median}$, is adopted as the point estimate, while for RF we use the default ensemble-averaged prediction. We report the metrics in two redshift ranges: $0 < z_{\rm spec} < 1.6$, containing 1,880,915 sources, and $z_{\rm spec} > 1.6$, containing 135,522 sources. The uncertainties for NF are estimated as the standard deviations across five independent training realisations with different random seeds and data shuffling. Some uncertainties are not reported because the corresponding seed-to-seed variations are below the displayed precision.

Redshift	Model	$\lvert\text{bias}\rvert$ (10^{-4})	$\sigma_{\rm NMAD}$ (10^{-2})	$f_{\rm out}$ (%)
$0 < z_{\rm spec} < 1.6$	NF	6.54 ± 5.40	2.23 ± 0.02	3.79 ± 0.03
	RF	16.47	2.72	4.78
	QRF	0.48	2.69	3.99
$z_{\rm spec} > 1.6$	NF	430 ± 34	10.15 ± 0.11	25.74 ± 0.51
	RF	794	11.89	31.49
	QRF	631	11.22	28.88

of testing sources, $N(z)$, by stacking the predicted photo-z PDFs over the redshift range $0 < z < 4.5$, using 200 uniformly spaced bins. We compare the NF and QRF reconstructions with the reference redshift distribution derived from DESI DR1 spectroscopic redshifts. We also display the redshift distribution constructed from the NF posterior-median estimates to illustrate the differences between PDF stacking and point-estimate reconstruction. We do not include the RF results in the $N(z)$ comparison, as RF provides only point estimates and does not support PDF stacking. 200 Monte Carlo samples are used to estimate the PDF for each sources with the NF models, whereas the QRF distribution is represented by 99 quantiles uniformly spaced from 0.01 to 0.99. For NF, we checked the variance of the 200 Monte Carlo samples and it was negligible compare to the run-to-run case. For comparison, we construct $N(z_{\rm median})$ using the posterior median of each galaxy, shown in Fig. 3.

For tomographic analysis, we compare the tomographic $N(z)$ reconstructions obtained by stacking the NF and QRF redshift distributions. The test samples are divided according to the predicted posterior-median redshift with boundaries [0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.6, 2.5, 4.5] in 8 tomographic bins. For the first 5 bins, the tomographic redshift distributions $N(z)$ are reconstructed on the redshift range $0.0 < z < 1.2$ with 200 uniform bins. For the remaining 3 bins, the tomographic redshift distributions $N(z)$ are reconstructed on the redshift range $0.0 < z < 5.0$ with 250 uniform bins. In practice, the tomographic bins are defined using the NF $z_{\rm median}$ estimates for NF model, while the QRF reconstruction uses the corresponding QRF median estimates. This provides an end-to-end comparison of the two pipelines, in which each method defines its own tomographic sample using its predicted median redshift, more closely reflecting a realistic photo-z tomographic analysis. The resulting NF and QRF reconstructions are shown in Fig. 4 and Fig. 5.

The quantitative comparison is summarized in Table 2. Each NF metric is averaged over five independent training realisations using the same training and testing samples but different random shuffling and seeds. The quoted uncertainty is the standard deviation across these realisations. For the QRF, the seed-to-seed variations are negligible compared with the NF run-to-run uncertainties and are therefore not reported.

As shown in Table 2, Fig. 4, and Fig. 5, the NF and QRF show different performance in recovering the mean redshift and the overall shape of the reconstructed $N(z)$. Considering the central values, the NF yields smaller $|\Delta\sigma|$ and JS divergence, and larger overlap, in

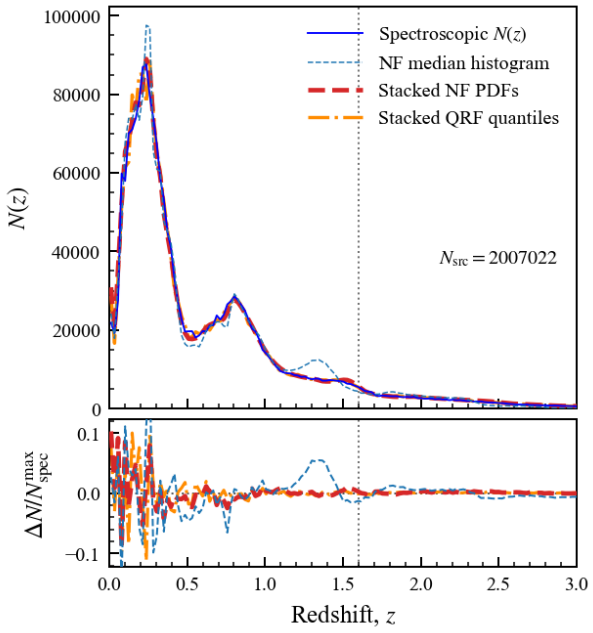


Figure 3. Comparison of the reconstructed redshift distributions, $N(z)$, for the testing sample displayed over $0 < z < 3$ for clarity. The upper panel shows the spectroscopic reference distribution alongside the reconstructions obtained by stacking the normalizing-flow (NF) and quantile random forest (QRF) PDFs, as well as the distribution of NF posterior-median redshift estimates. The lower panel shows the corresponding residuals relative to the spectroscopic distribution, normalised by its maximum value, $\Delta N/N_{\rm spec}^{\rm max}$, where $\Delta N = N_{\rm pred} - N_{\rm spec}$. The vertical dotted line at $z = 1.6$ approximately separates the galaxy-dominated and quasar-dominated redshift ranges.

all eight tomographic bins. The Wasserstein distance is also smaller for the NF in seven of the eight bins, with the only exception being $1.0 < z_{\rm median} < 1.6$. In contrast, the mean-redshift bias shows a less systematic trend: QRF gives a smaller $|\Delta z_{\rm mean}|$ in four bins, whereas the NF performs better in $0 < z_{\rm median} < 0.2$, $0.8 < z_{\rm median} < 1.0$, $1.6 < z_{\rm median} < 2.5$ and $z_{\rm median} > 2.5$.

The largest difference between the NF and QRF is observed in the highest-redshift bin ($z_{\rm median} > 2.5$). The NF achieves a smaller absolute mean-redshift bias ($|\Delta z_{\rm mean}| = 0.060$) than the QRF (0.318), together with a lower Wasserstein distance ($W_1 = 0.067$ compared with 0.318) and JS divergence (0.011 compared with 0.0375). The overlap is also higher for the NF (0.905) than for the QRF (0.787). The high-redshift tomographic samples exhibit a sharp peak near $z_{\rm spec} = 0$, dominated by stellar contaminants, as confirmed by their spectroscopic classifications. These contaminants are assigned high photometric redshift point estimates and consequently enter the high-z tomographic samples, highlighting the limitations of point estimates in capturing redshift ambiguities. In contrast, the NF PDFs retain non-negligible probability at low redshifts, allowing PDF stacking to partially recover this component.

Differences are also evident at lower redshifts. For $0.4 < z_{\rm median} < 0.6$, the NF achieves a higher overlap (0.966 versus 0.926) and a smaller Wasserstein distance (4.7×10^{-3} versus 9.7×10^{-3}) than the QRF. Overall, the NF provides a more accurate reconstruction of the tomographic $N(z)$, particularly at high redshifts, where photometric degeneracies are more significant and the available spectroscopic samples are smaller.

It is also evident from Figs. 4 and 5 that the distributions constructed from the point estimates, $N(z_{\rm median})$, do not reproduce the

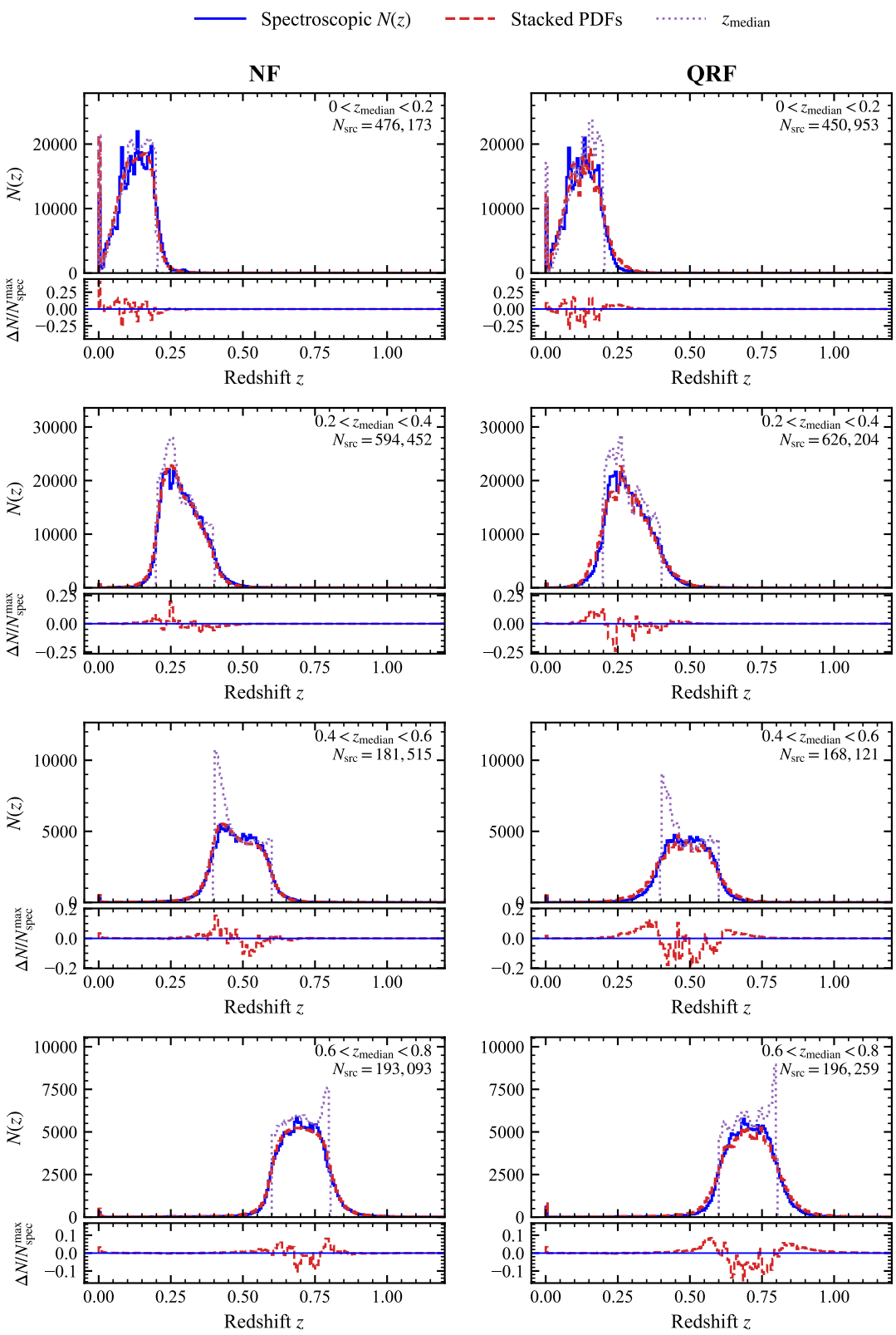


Figure 4. Tomographic redshift distributions, $N(z)$, reconstructed by stacking the predicted redshift PDFs from the NF (left) and QRF (right). The samples are divided into eight tomographic bins according to the median of the corresponding predicted redshift distribution, $z_{\rm median}$: $0.0 < z_{\rm median} < 0.2$ (top), $0.2 < z_{\rm median} < 0.4$ (second row), $0.4 < z_{\rm median} < 0.6$ (third row), and $0.6 < z_{\rm median} < 0.8$ (bottom). For all four tomographic bins, $N(z)$ is evaluated over $0 < z < 1.2$ using 200 uniform redshift bins. The dotted histograms show the distributions of the point estimates, $z_{\rm median}$, the solid blue lines show the spectroscopic reference distributions, $N_{\rm spec}(z)$, and the dashed red lines show the reconstructed distributions obtained by stacking the predicted PDFs, $N_{\rm pred}(z)$. The lower sub-panels show the normalised residuals, $\Delta N/N_{\rm spec}^{\rm max}$, where $\Delta N = N_{\rm pred} - N_{\rm spec}$.

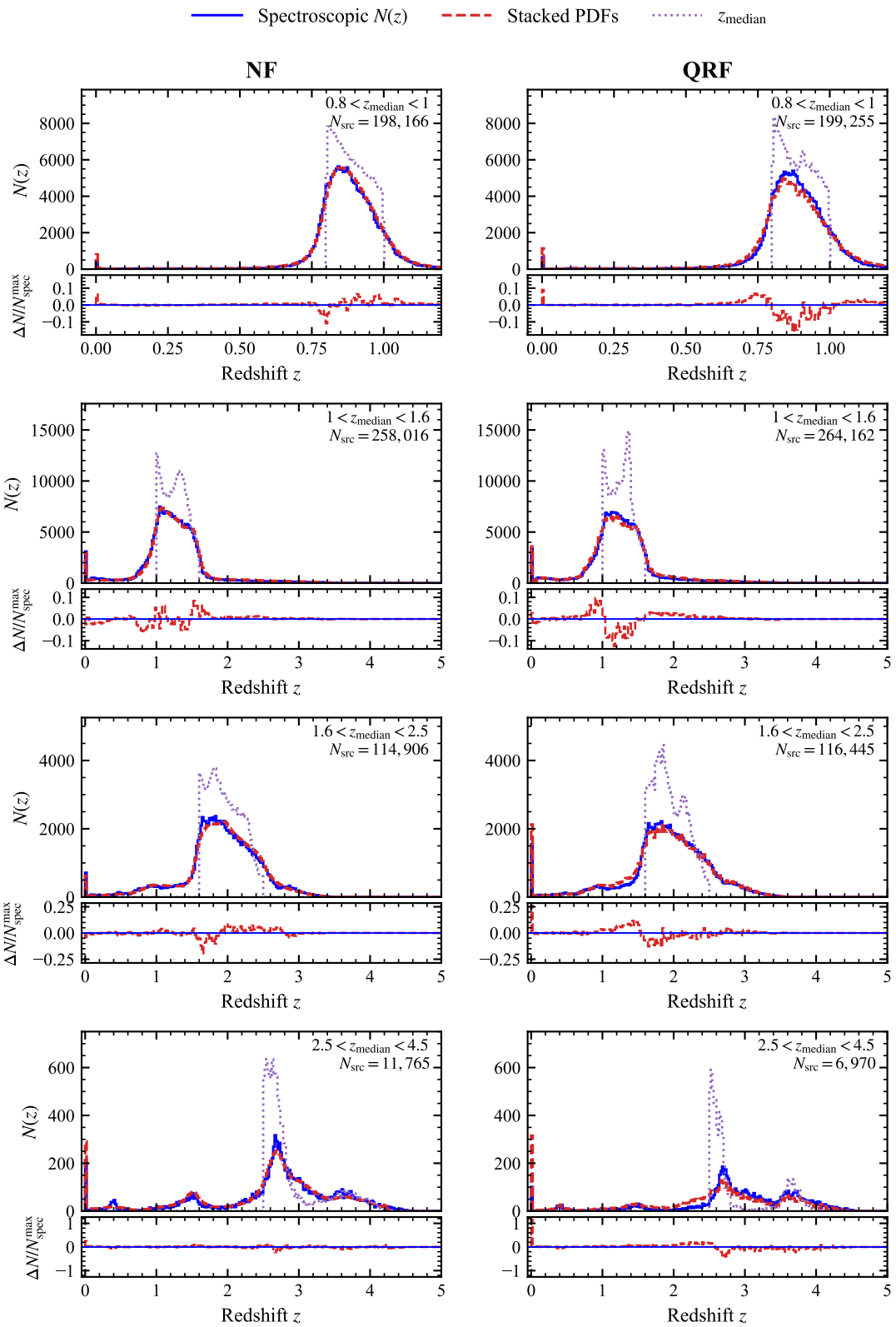


Figure 5. Same as Fig. 4 for the four bins defined by $z_{\rm median}$: $0.8 < z_{\rm median} < 1.0$, $1.0 < z_{\rm median} < 1.6$, $1.6 < z_{\rm median} < 2.5$, and $z_{\rm median} > 2.5$ from top to bottom. The first row uses $0 < z < 1.2$ with 200 uniform bins, while the remaining rows use $0 < z < 5.0$ with 250 bins. Dotted histograms show the $z_{\rm median}$ distributions, solid blue lines the spectroscopic $N_{\rm spec}(z)$, and dashed red lines the stacked-PDF reconstruction, $N_{\rm pred}(z)$. Lower panels show $\Delta N / N_{\rm spec}^{\rm max}$, where $\Delta N = N_{\rm pred} - N_{\rm spec}$.

Table 2. Comparison of the tomographic and full-sample $N(z)$ reconstructions obtained with the NF and QRF models. Sources are assigned to tomographic bins using the posterior-median point estimate, $z_{\rm median}$, predicted independently by each method. Consequently, the NF and QRF samples within the same nominal redshift interval may differ slightly (except the final tomographic bin which sample number are $\sim 12,000$ for NF and $\sim 7,000$ for QRF), and each reconstructed distribution is compared with the spectroscopic $N(z)$ of its corresponding selected sample. For the NF, quoted uncertainties are the sample standard deviations across five independent training realisations using identical training and test samples but different random seeds and data shuffling. The QRF seed-to-seed variations are below the displayed numerical precision and are therefore not shown here.

Tomographic selection	Method	$\lvert\Delta z_{\rm mean}\rvert$ ($\times 10^{-3}$)	$\lvert\Delta\sigma\rvert$ ($\times 10^{-3}$)	W_1 ($\times 10^{-3}$)	JS divergence ($\times 10^{-3}$)	Overlap(%)
$0 < z_{\rm median} < 0.2$	NF	2.6 ± 1.1	1.2 ± 1.9	3.0 ± 1.1	4.0 ± 0.5	93.1 ± 0.4
	QRF	4.2	7.9	5.1	7.4	90.7
$0.2 < z_{\rm median} < 0.4$	NF	3.2 ± 1.7	0.6 ± 0.4	3.5 ± 1.5	1.3 ± 0.2	96.7 ± 0.2
	QRF	3.0	4.2	4.1	3.4	94.6
$0.4 < z_{\rm median} < 0.6$	NF	3.6 ± 1.8	4.4 ± 1.1	4.7 ± 1.3	1.2 ± 0.2	96.6 ± 0.3
	QRF	3.0	12.7	9.7	5.6	92.6
$0.6 < z_{\rm median} < 0.8$	NF	3.3 ± 2.1	6.6 ± 1.2	4.3 ± 1.5	1.1 ± 0.2	96.9 ± 0.3
	QRF	2.1	11.3	8.2	3.8	94.0
$0.8 < z_{\rm median} < 1.0$	NF	1.6 ± 0.5	12.5 ± 0.9	6.5 ± 0.8	1.3 ± 0.2	97.0 ± 0.5
	QRF	4.3	14.7	9.8	2.6	94.7
$1.0 < z_{\rm median} < 1.6$	NF	21 ± 8	5.3 ± 7.3	22 ± 7	1.2 ± 0.3	96.8 ± 0.2
	QRF	11.0	27.7	21.1	1.7	95.6
$1.6 < z_{\rm median} < 2.5$	NF	18 ± 14	6.5 ± 3.3	21 ± 11	1.7 ± 0.3	96.1 ± 0.5
	QRF	20.7	26.1	26.8	2.3	95.1
$z_{\rm median} > 2.5$	NF	60 ± 31	30 ± 20	67 ± 23	11 ± 2	90.5 ± 0.8
	QRF	318	135	318	37.5	78.7
Full sample	NF	2.8 ± 3.3	9.7 ± 5.6	7.8 ± 1.9	0.69 ± 0.08	97.5 ± 0.2
	QRF	0.6	3.6	4.5	0.73	97.2

full shape of the spectroscopic $N(z_{\rm spec})$, particularly at high redshift. A single point estimate cannot represent secondary redshift solutions, extended tails, or contamination associated with photometric degeneracies, all of which can remain encoded in the full conditional PDFs $p(z \mid \mathbf{x})$.

For the full sample case, the differences between models are smaller. Without tomographic selection, QRF gives smaller $|\Delta z_{\rm mean}|$, $|\Delta\sigma|$, and W_1, whereas the NF gives a slightly smaller JS divergence and a slightly larger overlap. Nevertheless, both methods recover the full-sample $N(z)$ well, without obvious large-scale bias or leakage. This also shows that good recovery of the global $N(z)$ does not necessarily imply equally accurate reconstruction after tomographic selection.

In summary, these results show that accurate reconstruction of tomographic $N(z)$ benefits from retaining the full photo-z distribution, $p(z \mid \mathbf{x})$, rather than relying only on point estimates. This is particularly important for cosmological tomography, where sources scattered across photo-z bin boundaries produce extended true-redshift tails that cannot be captured by $z_{\rm median}$ alone. However, no calibration has yet been applied to the conditional photo-z PDFs considered here. Systematic miscalibration of individual PDFs may propagate through PDF stacking and affect the reconstructed $N(z)$. We therefore assess the reliability of the predicted PDFs and investigate a practical calibration procedure in Section 4.3.

4.3 PDF calibration and its impact on N(z) reconstruction

In this section, we use the Probability Integral Transform (PIT) (Dawid 1984; Wittman et al. 2016), to assess the photo-z PDFs predicted by the NF model. Motivated by the deviations from uniformity observed in the PIT distributions, we then apply a post-hoc cumulative distribution function (CDF) calibration following the approaches of Kuleshov et al. (2018); Dey et al. (2021). Finally, we investigate whether the improved PDF estimation propagates to a more accurate reconstruction of the full and tomographic redshift distributions, $N(z)$.

For each galaxy with spectroscopic redshift $z_{{\rm spec},i}$, the PIT value is defined as:

$$\mathrm{PIT}_i = \int_0^{z_{{\rm spec},i}} P(z \mid \mathbf{x}_i)\, \mathrm{d}z. \tag{15}$$

If the PDFs are fully consistent with the distribution of the underlying spectroscopic sample, the PIT values should be uniformly distributed over $[0, 1]$. A U-shaped distribution indicates over-confident PDFs, whereas an inverted-U shape indicates under-confidence; asymmetry and skewness generally reflect the existence of systematic bias. The PIT is particularly useful in this work because conventional photo-z point-estimate metrics, such as the bias, $\sigma_{\rm NMAD}$, and outlier fraction, compress each predicted distribution into a single value and cannot determine whether the full PDF is correctly estimated. Similarly, an accurate stacked $N(z)$ does not necessarily imply that the individual PDFs provide reliable probabilistic predictions. The PIT therefore provides a complementary test on the quality of the continuous NF PDFs estimation.

Fig. 6 shows the PIT distributions in six intervals of the posterior-median redshift, $z_{\rm median}$ with redshift bins [0.0, 0.2, 0.4, 0.8, 1.6, 2.5, 4.5]. The deviation from the uniform expectation varies clearly with redshift, with the asymmetric trend appearing in the redshift ranges of $0.2 < z_{\rm median} < 0.4$ and $0.4 < z_{\rm median} < 0.8$, and the strongly irregular fluctuations in highest redshift range $z_{\rm median} > 2.5$. The variations in PIT shape across redshift intervals motivate us to make the redshift-dependent post-hoc calibration introduced below.

To correct the PIT deviations shown in Fig. 6, we first tested global temperature scaling (Guo et al. 2017),

$$p_T(z \mid \mathbf{x}) \propto p(z \mid \mathbf{x})^{1/T}. \tag{16}$$

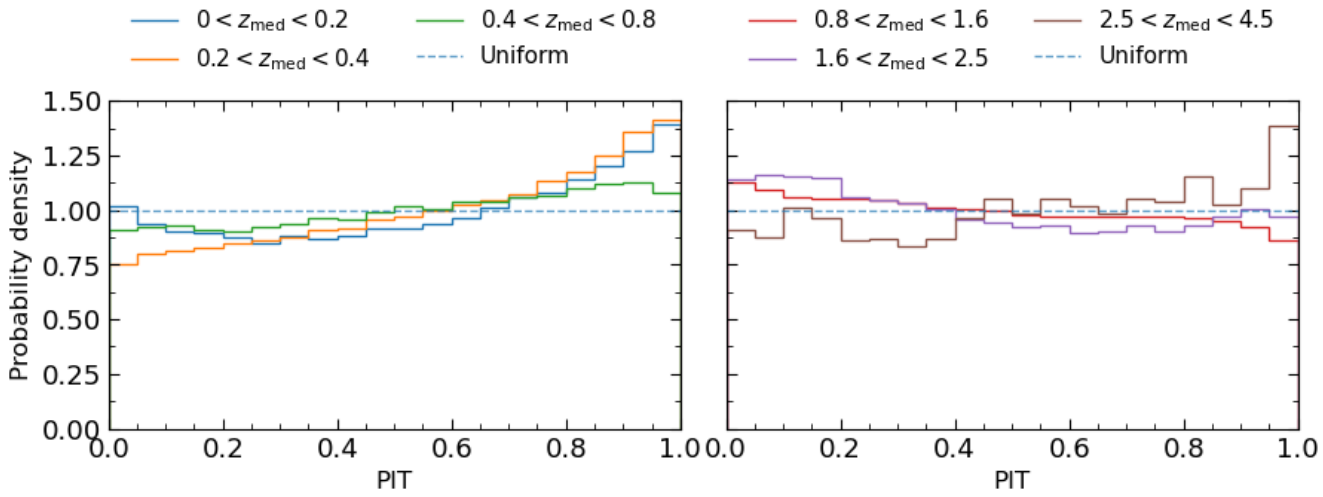


Figure 6. PIT distributions (20 bins) of the NF photo-z PDFs in six tomographic bins defined by the NF-median redshifts, [0.0, 0.2, 0.4, 0.8, 1.6, 2.5, 4.5]. The dashed horizontal line indicates the uniform distribution. The strongest deviation from uniformity occurs at low redshift, with a deficit at low PIT values and an excess towards PIT = 1. The deviation becomes progressively weaker with increasing redshift, while the higher-redshift bins show more irregular variations across the PIT range. This redshift dependence motivates the use of separate calibration mappings in each tomographic redshift bin.

Table 3. Comparison of the tomographic and full $N(z)$ reconstruction before and after post-hoc calibration. Sources are divided into eight tomographic bins according to the posterior-median estimate, $z_{\rm median}$. The quoted values are the mean and sample standard deviations across five independent NF realisations. Absolute values are used for $\Delta z_{\rm mean}$ and $\Delta\sigma$.

Tomographic selection	Calibration	$\lvert\Delta z_{\rm mean}\rvert(\times10^{-3})$	$\lvert\Delta\sigma\rvert(\times10^{-3})$	$W_1(\times10^{-3})$	JS divergence$(\times10^{-3})$	Overlap(%)
$0 < z_{\rm median} < 0.2$	Before	2.6 ± 1.1	1.2 ± 1.9	3.0 ± 1.1	4.0 ± 0.5	93.1 ± 0.4
	After	2.5 ± 1.8	7.3 ± 2.6	4.1 ± 1.6	4.4 ± 0.8	93.0 ± 0.7
$0.2 < z_{\rm median} < 0.4$	Before	3.2 ± 1.7	0.6 ± 0.4	3.5 ± 1.5	1.3 ± 0.2	96.7 ± 0.2
	After	1.6 ± 1.0	1.0 ± 0.5	2.3 ± 0.8	1.1 ± 0.2	97.0 ± 0.2
$0.4 < z_{\rm median} < 0.6$	Before	3.6 ± 1.8	4.4 ± 1.1	4.7 ± 1.3	1.2 ± 0.2	96.6 ± 0.3
	After	1.8 ± 1.5	4.0 ± 1.6	3.5 ± 1.3	1.1 ± 0.2	96.8 ± 0.3
$0.6 < z_{\rm median} < 0.8$	Before	3.3 ± 2.1	6.6 ± 1.2	4.3 ± 1.5	1.1 ± 0.2	96.9 ± 0.3
	After	1.8 ± 0.9	4.6 ± 0.9	3.1 ± 0.3	0.90 ± 0.08	97.2 ± 0.1
$0.8 < z_{\rm median} < 1.0$	Before	1.6 ± 0.5	12.5 ± 0.9	6.5 ± 0.8	1.3 ± 0.2	97.0 ± 0.5
	After	1.9 ± 1.8	8.6 ± 1.8	6.5 ± 0.8	1.2 ± 0.2	96.8 ± 0.5
$1.0 < z_{\rm median} < 1.6$	Before	21 ± 8	5.3 ± 7.3	22 ± 7	1.2 ± 0.3	96.8 ± 0.2
	After	26 ± 8	8.4 ± 4.8	28 ± 6	1.7 ± 0.3	96.4 ± 0.2
$1.6 < z_{\rm median} < 2.5$	Before	18 ± 14	6.5 ± 3.3	21 ± 11	1.7 ± 0.3	96.1 ± 0.5
	After	11 ± 6	18 ± 6	17 ± 2	1.6 ± 0.2	96.2 ± 0.2
$z_{\rm median} > 2.5$	Before	60 ± 31	30 ± 20	67 ± 23	11 ± 2	90.5 ± 0.8
	After	16 ± 11	28 ± 26	35 ± 7	11 ± 2	90.7 ± 1.0
Full Samples	Before	2.8 ± 3.3	9.7 ± 5.6	7.8 ± 1.9	0.69 ± 0.08	97.5 ± 0.2
	After	4.6 ± 3.6	8.1 ± 3.6	7.6 ± 1.1	0.64 ± 0.07	97.5 ± 0.1

However, a single temperature parameter T could not adequately correct the observed PIT deviations, indicating that global broadening or narrowing of the PDFs is insufficient. We therefore adopt a more flexible empirical CDF calibration method, following Kuleshov et al. (2018); Dey et al. (2021), using the independent validation sample described in Section 2.3. The core idea of this method is to construct a mapping bringing the estimated PIT distribution to the uniform distribution. This mapping is then applied to rescale the estimated PDFs.

The calibration is performed separately in six redshift bins defined by the predicted median redshift, $z_{\rm median}$, with boundaries $[0.0, 0.2, 0.4, 0.8, 1.6, 2.5, 4.5]$. For each galaxy in the validation sample, we first compute the PIT value

$$u_i = F_i(z_{{\rm spec},i}), \tag{17}$$

where $F_i(z)$ is the CDF predicted by the NF model. For each redshift bin b, we estimate the empirical CDF of the PIT values, denoted by $G_b(u)$.

To construct the calibration mapping, we define 101 equally spaced

quantile levels,

$$q_k = \frac{k}{100}, \qquad k = 0, \ldots, 100, \tag{18}$$

and calculate the corresponding empirical PIT quantiles, $u_{b,k}$, from the validation sample in each redshift bin. The boundary values are explicitly set to $u_{b,0} = 0$ and $u_{b,100} = 1$. The calibration function $G_b(u)$ is then approximated by linear interpolation between adjacent knots $(u_{b,k}, q_k)$, providing a continuous mapping from the original PIT values to calibrated quantile levels. The calibrated predictive CDF for galaxy i is defined as

$$F_{i,\mathrm{cal}}(z) = G_b\left[F_i(z)\right], \tag{19}$$

where b denotes the redshift bin assigned to the galaxy based on its original predicted median redshift.

In practice, the original predictive distribution of each galaxy is represented by 200 Monte Carlo redshift samples generated by the NF model. The calibration is implemented through quantile remapping. For each uniformly spaced quantile level q_j of the calibrated distribution, the corresponding quantile level of the original distribution is obtained through the inverse calibration mapping, $G_b^{-1}(q_j)$. The calibrated redshift values are then calculated as

$$z_{i,j}^{\mathrm{cal}} = \widehat{F}_i^{-1}\left[G_b^{-1}(q_j)\right], \qquad q_j = \frac{j - 1/2}{200}, \quad j = 1, \ldots, 200, \tag{20}$$

where $\widehat{F}_i^{-1}$ denotes the empirical quantile function estimated from the original Monte Carlo samples. This procedure modifies the shapes of the predictive redshift distributions while preserving the number of samples used for subsequent analyses. The calibration mappings are derived exclusively from the validation sample and subsequently applied to the independent testing sample.

We note that deriving the calibration mappings directly from the testing sample can generally lead to greater improvements in the tomographic $N(z)$ reconstruction metrics. However, such improvements are not necessarily consistent across all tomographic bins and metrics. Calibrating and evaluating on the same testing sample introduces data leakage and may lead to overly optimistic estimates of calibration performance. Such an approach is also not representative of realistic cosmological analysis pipelines, where spectroscopic redshifts are generally unavailable for the target photometric sample. We therefore adopt validation-based calibration to provide a more statistically reliable assessment of the method's generalisation to independent data.

The calibration mappings are constructed exclusively from the spectroscopic validation sample of 121,594 sources (Section 2) and applied to the spatially independent testing sample. This ensures an out-of-sample evaluation, with testing z_{spec} values used only to assess the calibrated PIT distributions and $N(z)$ reconstruction. For each NF realisation, a separate mapping is derived from its validation-set PIT distribution and applied to the corresponding testing PDFs.

Fig. 7 shows the PIT distributions before and after calibration. As expected, the validation PIT becomes approximately uniform by construction, while the testing PIT also shows reduced deviations from uniformity, indicating that the calibration transfers to the independent sample. We further quantify the calibration using the empirical CDF, $G(u)$, and its residual, $\Delta G(u) = G(u) - u$, shown in Fig. 8. For perfectly calibrated PDFs, $G(u) = u$, deviations from the identity relation directly indicate cumulative miscalibration.

We then use the calibrated NF PDFs to reconstruct the redshift distributions $N(z)$. Table 3 compares the reconstruction performance before and after calibration. At intermediate redshifts ($0.2 < z_{\mathrm{median}} < 0.8$), the absolute mean-redshift bias decreases

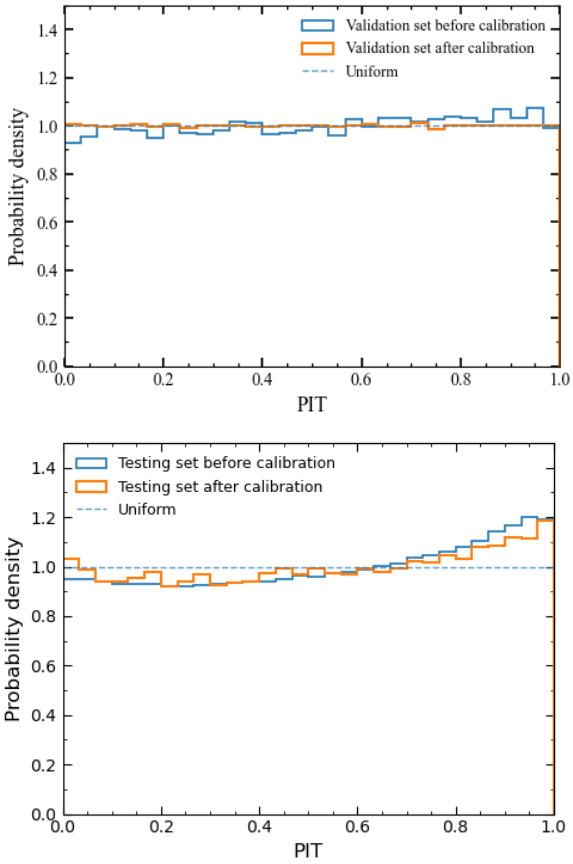


Figure 7. PIT distributions (30 bins) for the independent validation sample (upper panel) and testing sample (lower panel), shown before and after applying the calibration mapping derived from the validation set. The dashed horizontal line marks the ideal uniform distribution. After calibration, the validation-sample PIT is close to uniform, while the testing-sample PIT shows only modest improvement, with residual excesses remaining.

to $(1.6\text{–}1.8)\times10^{-3}$, alongside reductions in the Wasserstein distance and JS divergence. The largest improvement occurs at $z_{\mathrm{median}} > 2.5$, where the bias decreases from 0.060 to 0.016 and the Wasserstein distance from 0.067 to 0.035. However, calibration increases the distribution-width discrepancies in some bins and slightly worsens the full-sample mean-redshift bias, while the overlap remains broadly unchanged.

Fig. 9 compares the absolute mean-redshift bias with the Euclid weak-lensing requirement, $|\Delta z_{\mathrm{mean}}| < 0.002(1+z)$ (Euclid Collaboration 2020, 2021), where z is taken as the mean spectroscopic redshift of each tomographic bin. After calibration, the bias in the first bin decreases slightly from 0.0026 to 0.0025, approaching but remaining above the Euclid benchmark. The second to fifth bins ($0.2 < z_{\mathrm{median}} < 1.0$) satisfy the benchmark after calibration. At higher redshifts, the bias increases from 0.021 to 0.026 for $1.0 < z_{\mathrm{median}} < 1.6$, but decreases from 0.018 to 0.011 for $1.6 < z_{\mathrm{median}} < 2.5$ and from 0.060 to 0.016 for $z_{\mathrm{median}} > 2.5$. Nevertheless, all three higher-redshift bins remain above the Euclid benchmark.

We emphasise that satisfying this benchmark within the DESI-selected spectroscopic sample does not demonstrate that the same accuracy can be achieved for Euclid weak-lensing source galaxies, which have different selection functions and photometric properties.

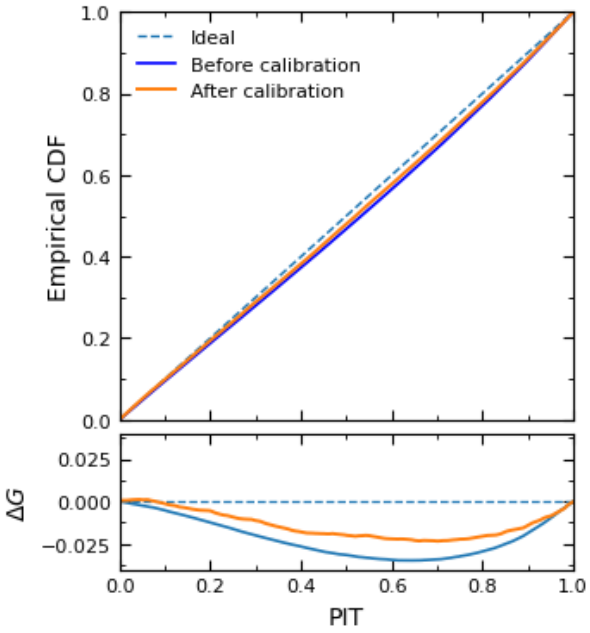


Figure 8. Empirical Cumulative Distribution Functions (CDF) of the PIT values for the independent testing sample before and after calibration. The dashed diagonal shows the ideal relation expected for uniformly distributed PIT values. The lower panel shows the residual, $\Delta G(u) = G(u) - u$, relative to the ideal CDF. After calibration, the empirical CDF moves closer to the uniform expectation and the magnitude of the residual is reduced over most of the PIT range, indicating improved out-of-sample probabilistic calibration.

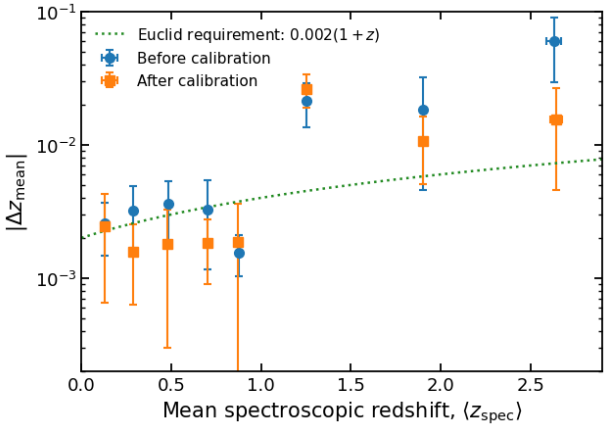


Figure 9. Absolute mean-redshift bias, $|\Delta z_{\rm mean}| = |\langle z_{\rm pred}\rangle - \langle z_{\rm spec}\rangle|$, as a function of the mean spectroscopic redshift for eight tomographic bins before (blue circles) and after (orange squares) post-hoc calibration. The points represent the mean absolute biases over independent NF realisations, with error bars indicating the sample standard deviations. The green dotted line indicates the Euclid mean-redshift accuracy benchmark, $0.002(1 + z)$. Tomographic bins are defined using the posterior-median redshift, $z_{\rm median}$, and galaxy membership is recomputed after calibration, accounting for the small differences in the mean spectroscopic redshifts before and after calibration. The lower error bar for the fifth tomographic bin extends below the plotted lower limit and is therefore partially truncated on the logarithmic y-axis.

Further validation on representative target populations is therefore required to assess the applicability of the framework to Euclid.

Generally speaking, redshift-dependent post-hoc calibration improves the probabilistic consistency of NF photo-z PDFs and leads to modest improvements in the reconstructed $N(z)$. The validation-derived mapping brings both the PIT distribution and its empirical CDF closer to the uniform expectation when applied to the independent test sample, although residual miscalibration remains. For the full sample, the Wasserstein distance and JS divergence decrease slightly, while the mean-redshift bias increases. The tomographic results show more substantial improvements, particularly at low and intermediate redshifts, where the mean-redshift biases in the second to fifth bins ($0.2 < z_{\rm median} < 1.0$) fall within the Euclid accuracy benchmark after calibration. At higher redshifts, the improvements are less consistent, although substantial reductions in the mean-redshift bias are observed in selected bins. The incomplete transfer of calibration may partly reflect differences between the validation and test samples, which occupy different sky regions. Nevertheless, the improvements achieved without using spectroscopic redshifts from the test sample highlight the potential of PIT-based recalibration for improving probabilistic photo-z estimates and tomographic redshift distributions in different catalogues.

5 DISCUSSION

5.1 PDF calibration

Accurate calibration of predicted redshift PDFs is essential for reliable reconstruction of the ensemble redshift distribution, $N(z)$. As shown in Section 4.3, empirical CDF calibration (Kuleshov et al. 2018; Dey et al. 2021) produces an approximately uniform PIT distribution for the validation sample, while modest deviations remain when the mapping is applied to the independent testing sample. Calibration also improves $|\Delta z_{\rm mean}|$ in several tomographic bins. However, the validation and testing samples exhibit similar photometric features and redshift distributions (Fig. A2), leaving the effectiveness of this approach for more distinct galaxy populations uncertain.

Compared with simpler approaches such as temperature scaling (Guo et al. 2017), empirical CDF recalibration with 101 interpolation knots provides greater flexibility in correcting complex PIT deviations. However, the resulting mapping is determined by the PIT distribution of the calibration sample and is therefore not necessarily universal. Even an accurately determined calibration mapping may not remain valid for a target population with different photometric properties or redshift distributions. This limitation is particularly relevant to practical photo-z applications, where representative spectroscopic samples are often unavailable for the target population.

Therefore, the key challenge is not simply to achieve a uniform PIT distribution on the validation sample, but to develop calibration methods that remain reliable across different galaxy populations. Future work could explore alternative post-hoc calibration approaches, such as isotonic regression (Niculescu-Mizil & Caruana 2005), and methods that account for photometric features or individual PDF properties, such as non-Gaussianity and multimodality, while evaluating their transferability using independent datasets. Importantly, achieving PIT uniformity does not necessarily guarantee accurate reconstruction of tomographic $N(z)$, further motivating calibration methods that are both transferable across galaxy populations and reliable for distribution-level cosmological analyses (Rau et al. 2015; Euclid Collaboration 2020, 2021).

5.2 Characteristics of NF-based photo-z estimation

In our tomographic analysis, we find that the NF results generally outperform the QRF ones, whereas the advantage is less pronounced for the full sample. These findings, together with the point estimate results, are consistent with the interpretation that NF photo-z estimation excels when the data are noisy and sparse (Ren et al. 2026).

Our analysis also highlights several practical characteristics of NF-based photo-z inference. First, repeated training runs show noticeable run-to-run variations even when the same training and testing samples are used (see Section 4.2). This training-induced variability introduces an additional source of uncertainty beyond the systematic effects associated with the data. Further investigation of convergence, optimization, model architecture, and initialization is therefore required to determine whether this variability can be substantially reduced.

At the same time, the continuous probabilistic representation provided by NFs offers useful flexibility for PDF-based analyses. For diagnostics such as PIT, the NF provides an explicit continuous predictive distribution, while individual PDFs can be sampled at arbitrarily fine numerical resolution without retraining the model. For the smaller high-redshift sample, the reduced computational cost allows us to draw 1,000 or more Monte Carlo samples per source. Hence, the continuous formulation of the NF is convenient for probabilistic diagnostics and individual-PDF studies.

5.3 Limitations and future work

Although the training and testing samples occupy non-overlapping sky regions, both are constructed from the DESI spectroscopic sample using the same photometric and quality selections from DESI Legacy DR10. As the primary goal of this study is to validate the probabilistic photo-z estimation and $N(z)$ reconstruction methodology, this controlled setting allows us to assess the performance of the NF model without introducing additional differences in sample selection.

However, this experimental design does not fully investigate its generalisability to broader photometric populations with different distributions of colours, magnitudes, or galaxy types. In particular, the tomographic $N(z)$ accuracy achieved in this work should not be interpreted as the expected performance for Euclid weak-lensing source galaxies. Recent studies have highlighted the importance of spectroscopic incompleteness and sample representativeness for redshift calibration (Lange et al. 2026; Blanco et al. 2026).

Nevertheless, the NF framework itself is not specific to DESI but offers considerable flexibility in modelling conditional redshift distributions. It can be adapted to different photometric surveys, including Euclid, LSST and CSST by incorporating their respective photometric features and retraining or recalibrating the model using suitable reference data. The method may also be particularly promising for lens galaxy samples, whose photometric properties can be more similar to those of DESI spectroscopic targets. Future applications will require careful assessment of selection effects and differences between the spectroscopic training and target photometric populations.

6 CONCLUSIONS

In this work, we investigate conditional Normalizing Flows (NFs) for probabilistic photometric-redshift estimation and redshift-distribution reconstruction using DESI DR1 spectroscopic data and DESI Legacy Imaging Surveys photometry. The main analysis uses an independent test sample of 2,028,371 sources spanning around $0 < z_{\rm spec} < 4.5$, with the NF evaluated over five independent training realisations and compared with the Random Forest (RF) and Quantile Random Forest (QRF) models. For point estimation, the NF achieves lower $\sigma_{\rm NMAD}$ and outlier fractions compare to RF and QRF across the full redshift range, although its bias is slightly larger than that of QRF and exhibits greater run-to-run variation. At higher redshifts ($z_{\rm spec} > 1.6$), where the performance of all three models generally deteriorates due to the limited samples' number at high redshift (see Fig. 2), the NF demonstrates a clearer advantage in bias, scatter, and outlier fraction, consistent with Ren et al. (2026). However, point-estimation metrics alone do not fully capture the advantages of the NF, which provides continuous conditional redshift distributions that retain information about their widths, asymmetries, and extended tails.

This advantage becomes more apparent in tomographic $N(z)$ reconstruction. Across the first seven tomographic bins, the NF generally recovers the spectroscopic redshift distributions more accurately than QRF in terms of distribution width, Wasserstein distance, JS divergence, and probability overlap, although QRF achieves smaller mean-redshift biases in some bins. In the highest-redshift bin ($z_{\rm median} > 2.5$), the NF shows a particularly substantial improvement over QRF across the reconstruction metrics, including its ability to recover the extended low-redshift tail. For the full-sample $N(z)$, both probabilistic methods perform well, indicating that the advantages of the NF are more pronounced in tomographic reconstruction than in global distribution-level statistics.

We further evaluate the probabilistic calibration of the NF PDFs using the probability integral transform (PIT) and apply a redshift-dependent empirical CDF recalibration derived from an independent validation sample. When transferred to the spatially independent test sample, the calibrated PIT distributions move closer to the uniform expectation, although residual miscalibration remains. The recalibration also leads to improvements in tomographic $N(z)$ reconstruction in several redshift bins, particularly in the mean-redshift accuracy, while largely preserving the overall distribution shapes. After calibration, the mean absolute redshift biases in the second through fifth tomographic bins at $0.2 < z_{\rm median} < 1$ fall below the Euclid accuracy benchmark of $0.002(1 + z)$. At higher redshifts, the results are less consistent, although substantial reductions in the mean-redshift bias are achieved for $1.6 < z_{\rm median} < 2.5$ and $z_{\rm median} > 2.5$. These findings demonstrate the flexibility of continuous NF PDFs for post-hoc calibration, while also showing that improved probabilistic consistency does not necessarily translate into better $N(z)$ reconstruction across all metrics.

Collectively, our results demonstrate that the advantages of Normalizing Flows for photometric-redshift inference extend beyond point-estimation accuracy to their flexible representation of the full conditional redshift probability distribution. Retaining this probabilistic information can improve tomographic $N(z)$ reconstruction, particularly when broad or multimodal redshift distributions lead to substantial leakage across photo-z bin boundaries. Nevertheless, the observed run-to-run variability, the uncertain generalisation to different photometric populations, and the limitations of post-hoc PDF calibration remain important challenges. Future work should explore more transferable calibration strategies, evaluate NFs on different target populations and against a broader range of probabilistic photo-z methods, and incorporate the reconstructed redshift distributions into galaxy-clustering and weak-lensing likelihood analyses to quantify their impact on cosmological inference.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Marc Manera (IFAE) for his valuable suggestions on the reconstruction of redshift distributions, $N(z)$, and for his helpful comments on the Introduction and Willian d'Assignies for his valuable comments on the manuscript. ZC and KCC acknowledge the support by the National Natural Science Foundation of China under the grant number 12273121 and 12533002, and the science research grants from the China Manned Space Project with NO.CMS-CSST-2021-B01. AR acknowledges support from the Swiss National Science Foundation (SNF) "Cosmology with 3D Maps of the Universe" research grant 200020_207379.

This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations is managed by the Lawrence Berkeley National Laboratory. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: www.desi.lbl.gov/collaborating-institutions. The DESI collaboration is honored to be permitted to conduct scientific research on I'oligam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or any of the listed funding agencies. The DESI Legacy Imaging Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS), the Beijing-Arizona Sky Survey (BASS), and the Mayall z-band Legacy Survey (MzLS). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. Pipeline processing and analyses of the data were supported by NOIRLab and the Lawrence Berkeley National Laboratory (LBNL). Legacy Surveys also uses data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), a project of the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. Legacy Surveys was supported by: the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility; the U.S. National Science Foundation, Division of Astronomical Sciences; the National Astronomical Observatories of China, the Chinese Academy of Sciences and the Chinese National Natural Science Foundation. LBNL is managed by the Regents of the University of California under contract to the U.S. Department of Energy. The complete acknowledgments can be found at https://www.legacysurvey.org/acknowledgment/.

DATA AVAILABILITY

The spectroscopic data used in this work are publicly available from the DESI Data Release 1 (DR1) at https://data.desi.lbl.gov/public/dr1/. The photometric data are publicly available from the DESI Legacy Imaging Surveys Data Release 10 (DR10) at https://www.legacysurvey.org/dr10/.

The training, validation, and testing catalogues constructed by cross-matching these datasets and applying the selection criteria described in Section 2 are available on Zenodo at https://doi.org/10.5281/zenodo.23642619.

The source code used for model training, photometric-redshift estimation, and redshift-distribution reconstruction, together with the five trained NF models, is publicly available at https://github.com/chenzhuo0623-pixel/Normalizing-flow-photometric-redshift-PDFs-with-DESI-DR1.

REFERENCES

Almosallam I. A., Jarvis M. J., Roberts S. J., 2016, MNRAS, 462, 726
Blanco D., Leauthaud A., Lange J. U., Wright A. H., Hildebrandt H., Heydenreich S., et al., 2026, The Open Journal of Astrophysics, 9
Bonnett C., 2015, Monthly Notices of the Royal Astronomical Society, 449, 1043
Bordoloi R., Lilly S. J., Amara A., 2010, Monthly Notices of the Royal Astronomical Society, 406, 881
Breiman L., 2001, Machine Learning, 45, 5
Buchs R., et al., 2019, MNRAS, 489, 820
CSST Collaboration et al., 2026, Science China Physics, Mechanics, and Astronomy, 69, 239501
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Carrasco Kind M., Brunner R. J., 2014, MNRAS, 438, 3409
Cawthon R., et al., 2022, MNRAS, 513, 5517
Chaussidon E., et al., 2022, Target Selection and Validation of DESI Quasars, doi:10.3847/1538-4357/acb3c2, https://arxiv.org/abs/2208.08511v2
Choppin de Janvry J., Gontcho A Gontcho S., Seljak U., Baleato Lizancos A., Chaussidon E., et al., 2026, Journal of Cosmology and Astroparticle Physics, 2026, 073
Collister A. A., Lahav O., 2004, PASP, 116, 345
Crenshaw J. F., Kalmbach J. B., Gagliano A., Yan Z., Connolly A. J., Malz A. I., Schmidt S. J., The LSST Dark Energy Science Collaboration 2024, AJ, 168, 80
DESI Collaboration 2022, Astronomical Journal, 164, 207
DESI Collaboration 2026, AJ, 171, 285
D'Isanto A., Polsterer K. L., 2018, Astronomy & Astrophysics, 609, A111
Dawid A. P., 1984, Journal of the Royal Statistical Society A, 147, 278
De Vicente J., Sánchez E., Sevilla-Noarbe I., 2016, Mon. Not. Roy. Astron. Soc., 459, 3078
Dey A., et al., 2019, Astronomical Journal, 157, 168
Dey B., Newman J. A., Andrews B. H., Izbicki R., Lee A. B., Zhao D., Rau M. M., Malz A. I., 2021, arXiv e-prints, p. arXiv:2110.15209
Dinh L., Krueger D., Bengio Y., 2015, in International Conference on Learning Representations (ICLR), Workshop Track. (arXiv:1410.8516), https://arxiv.org/abs/1410.8516
Durkan C., Bekasov A., Murray I., Papamakarios G., 2019, NeurIPS
Euclid Collaboration 2020, Astronomy & Astrophysics, 642, A191
Euclid Collaboration 2021, A&A, 647, A117

Flaugher B., Diehl H. T., Honscheid K., et al., 2015, The Astronomical Journal, 150, 150
Gatti M., et al., 2018, MNRAS, 477, 1664
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823
Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, in Proceedings of the 34th International Conference on Machine Learning (ICML). pp 1321–1330, http://proceedings.mlr.press/v70/guo17a.html
Hahn C., et al., 2023, The Astronomical Journal, 165, 253
Hildebrandt H., et al., 2021, A&A, 647, A124
Hoyle B., 2016, Astronomy and Computing, 16, 34
Ivezić Ž., et al., 2019, The Astrophysical Journal, 873, 111
Kuleshov V., Fenner N., Ermon S., 2018, in Proceedings of the 35th International Conference on Machine Learning. PMLR, pp 2796–2804
Kullback S., Leibler R. A., 1951, Annals of Mathematical Statistics, 22, 79
Lange J. U., Blanco D., Leauthaud A., Wright A. H., Fisher A., Ratajczak J., et al., 2026, The Astrophysical Journal, 998, 263
Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193
Li R., Napolitano N. R., Roy N., Tortora C., La Barbera F., Sonnenfeld A., Qiu C., Liu S., 2022, ApJ, 929, 152
Lin J., 1991, IEEE Transactions on Information Theory, 37, 145
Lu J., et al., 2024, MNRAS, 527, 12140
Masters D., et al., 2015, ApJ, 813, 53
McQuinn M., White M., 2013, MNRAS, 433, 2857
Meinshausen N., 2006, Journal of Machine Learning Research, 7, 983
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, arXiv e-prints, p. arXiv:1303.4722
Newman J. A., 2008, ApJ, 684, 88
Newman J. A., Gruen D., 2022, Annual Review of Astronomy and Astrophysics, 60, 363
Niculescu-Mizil A., Caruana R., 2005, in Proceedings of the 22nd International Conference on Machine Learning (ICML). pp 625–632, https://dl.acm.org/doi/10.1145/1102351.1102430
Peng H., Xu H., Zhang L., Chen Z., Yu Y., 2022, MNRAS, 516, 6210
Raichoor A., et al., 2023, The Astronomical Journal, 165, 126
Rau M., Seitz S., Brimioulle F., Frank E., 2015, Monthly Notices of the Royal Astronomical Society, 452, 3710
Rau M. M., et al., 2023, MNRAS, 524, 5109
Ren Y., Chan K. C., Zhang L., Li Y., et al., 2026, ApJ, 997, 65
Rezende D., Mohamed S., 2015, in International conference on machine learning. pp 1530–1538
Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502
Salvato M., et al., 2019, Nature Astronomy, 3, 212
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, The Astrophysical Journal, 500, 525
Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, MNRAS, 431, 3307
Schneider M., Knox L., Zhan H., Connolly A., 2006, ApJ, 651, 14
Schuldt S., Suyu S. H., Cañameras R., Taubenberger S., Meinhardt T., Leal-Taixé L., Hsieh B. C., 2021, A&A, 651, A55
Simonyan K., Vedaldi A., Zisserman A., 2014, in Workshop at International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1312.6034
Sun Z., Speagle J. S., Huang S., Ting Y.-S., Cai Z., 2023, arXiv e-prints, p. arXiv:2310.20125
Tabak E. G., Turner C. V., 2013, Commun. Pure Appl. Math., 66, 145
Tian D.-C., Wen Z.-L., Xia J.-Q., 2026, arXiv e-prints, p. arXiv:2602.01548
Villani C., 2009, Optimal Transport: Old and New. Springer
Wittman D., et al., 2016, MNRAS, 457, 4005
Wright E. L., Eisenhardt P. R. M., Mainzer A. K., et al., 2010, The Astronomical Journal, 140, 1868
Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., 2020, A&A, 637, A100
Xu H., et al., 2023, MNRAS, 520, 161
Zhang P., Pen U.-L., Bernstein G., 2010, MNRAS, 405, 359
Zhang L., Yu Y., Zhang P., 2017, ApJ, 848, 44
Zheng W., Chan K. C., Xu H., Zhang L., Song R., 2024, A&A, 692, A186
Zhou R., et al., 2021, MNRAS, 501, 3309
Zhou X., et al., 2022, MNRAS, 512, 4593
Zhou R., et al., 2023, The Astronomical Journal, 165, 58

APPENDIX A: SUPPLEMENTARY FIGURES

We present the $z_{\rm spec}$–$z_{\rm phot}$ scatter relation for the low redshift $0 < z_{\rm spec} < 1.6$ and high redshift $z_{\rm spec} > 1.6$ for three models in Fig. A1. While the low redshift bin results are visually very similar among the models, the NF results for the high redshift bin are slightly more symmetrical along the diagonal relative to the others.

Fig. A2 compares the distributions of the input photometric features between training, validation patch and testing patch. The distributions are normalised independently, and the Jensen–Shannon divergence, $D_{\rm JS}$, is used to quantify their differences. The two patches exhibit broadly similar marginal distributions of magnitudes and colours, with relatively larger differences in the PSF depths, particularly in the g band.

APPENDIX B: PERCENTILE-BASED TOMOGRAPHIC $N(Z)$ TEST

In Section 4.2, we find that the NF model substantially outperforms the QRF model in reconstructing the tomographic redshift distribution for the high-redshift selection $z_{\rm median} > 2.5$. However, the two models select substantially different numbers of sources, with approximately 12,000 sources for NF and 7,000 for QRF. To investigate whether this difference in sample size affects the comparison, we perform an additional tomographic test using equal-sized high-redshift samples.

Specifically, we independently select the top 0.5 per cent of sources ranked by their predicted median redshifts, $z_{\rm median}$, for each model. The resulting lower redshift thresholds are $z_{\rm median} = 2.55$ for NF and $z_{\rm median} = 2.40$ for QRF. This percentile-based selection ensures an identical sample size of 10,142 sources for both models, although the selected samples do not necessarily contain the same objects. For each model, the reconstructed $N(z)$ is compared with the spectroscopic redshift distribution of its corresponding selected sample.

The resulting tomographic redshift distributions are presented in Fig. B1. The stacked NF PDFs reproduce the spectroscopic $N(z)$ more accurately than the QRF PDFs, particularly around the main high-redshift peak. Quantitatively, NF achieves a distribution overlap of approximately 90%, an absolute mean-redshift bias of $|\Delta z_{\rm mean}| = 0.11$, a Wasserstein distance of 0.11, and a Jensen–Shannon (JS) divergence of 0.013. In comparison, QRF yields an overlap of approximately 82%, $|\Delta z_{\rm mean}| = 0.24$, a Wasserstein distance of 0.25, and a JS divergence of 0.024. These results are consistent with the findings of the original high-redshift tomographic analysis, demonstrating that the superior $N(z)$ reconstruction achieved by NF persists when the sample sizes are matched, although differences in the selected source populations may also contribute to the observed performance.

APPENDIX C: SENSITIVITY OF PHOTOMETRIC FEATURES IN NF MODEL

To investigate the sensitivity of the trained NF model to individual input features, we perform an input-gradient saliency analysis (Simonyan et al. 2014). For each source i, we define the input feature vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$, where x_{ij} denotes the j-th photometric

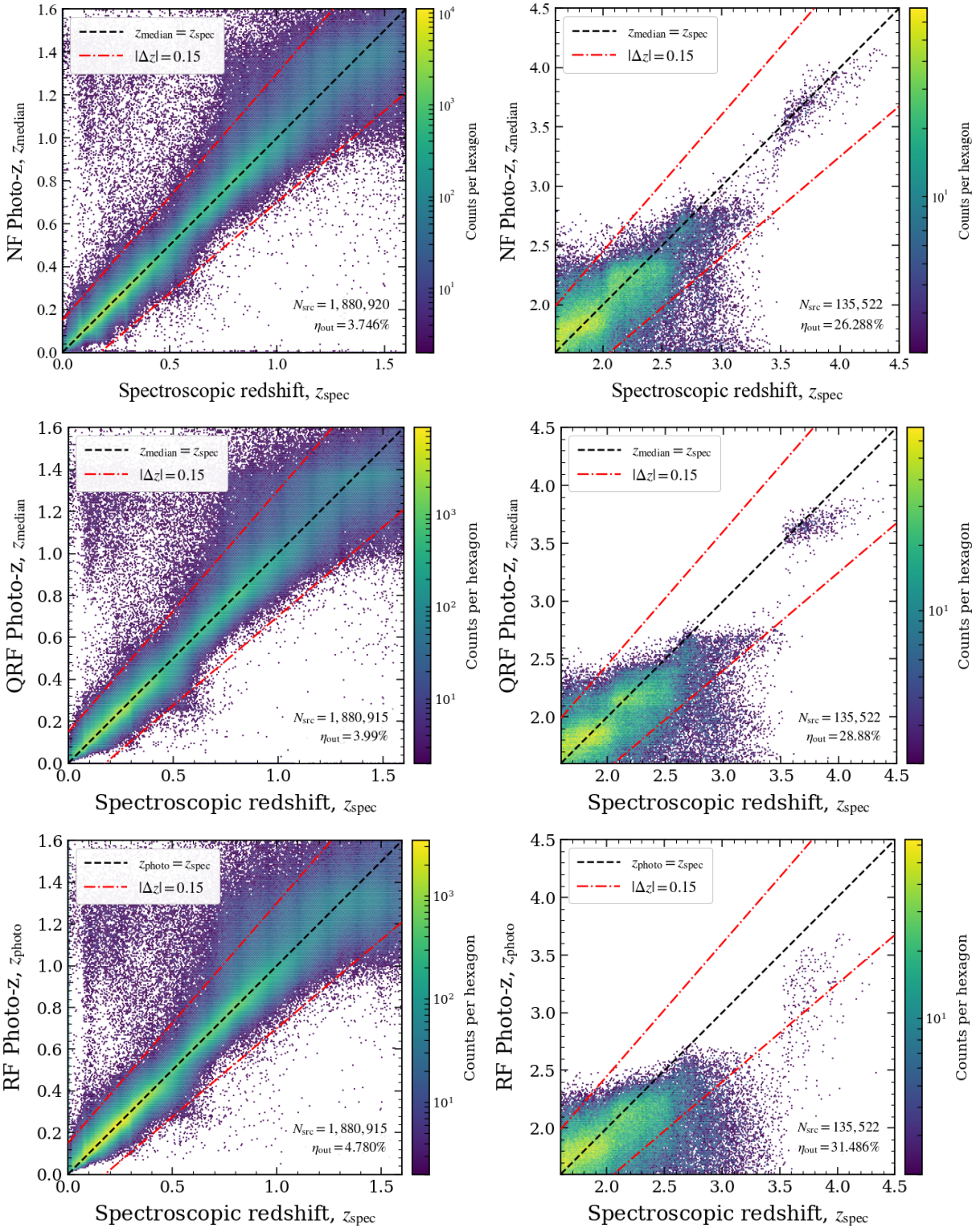


Figure A1. $z_{\rm spec} - z_{\rm median}$ scatter plots for 1,880,915 sources with $0.0 < z_{\rm spec} < 1.6$ and for 135,522 sources with $z_{\rm spec} > 1.6$, showing the Normalizing Flow (upper panel), Quantile Random Forest (middle panel), and Random Forest (lower panel). The NF redshift estimate is obtained from the median of each posterior PDF, while the QRF point estimate corresponds to the median derived from 99 uniformly spaced quantile levels between 0.01 and 0.99. The dot dash orange lines indicate the outlier threshold, $|\Delta z| = 0.15$, while the dashed black line shows the one-to-one relation, $z_{\rm spec} = z_{\rm photo}$. The colour scale indicates the number density of targets. In redshift range $z_{\rm spec} > 1.6$, the NF predictions are distributed more symmetrically around the one-to-one relation ($z_{\rm phot} = z_{\rm spec}$), while the QRF and RF predictions exhibit more pronounced asymmetric deviations.

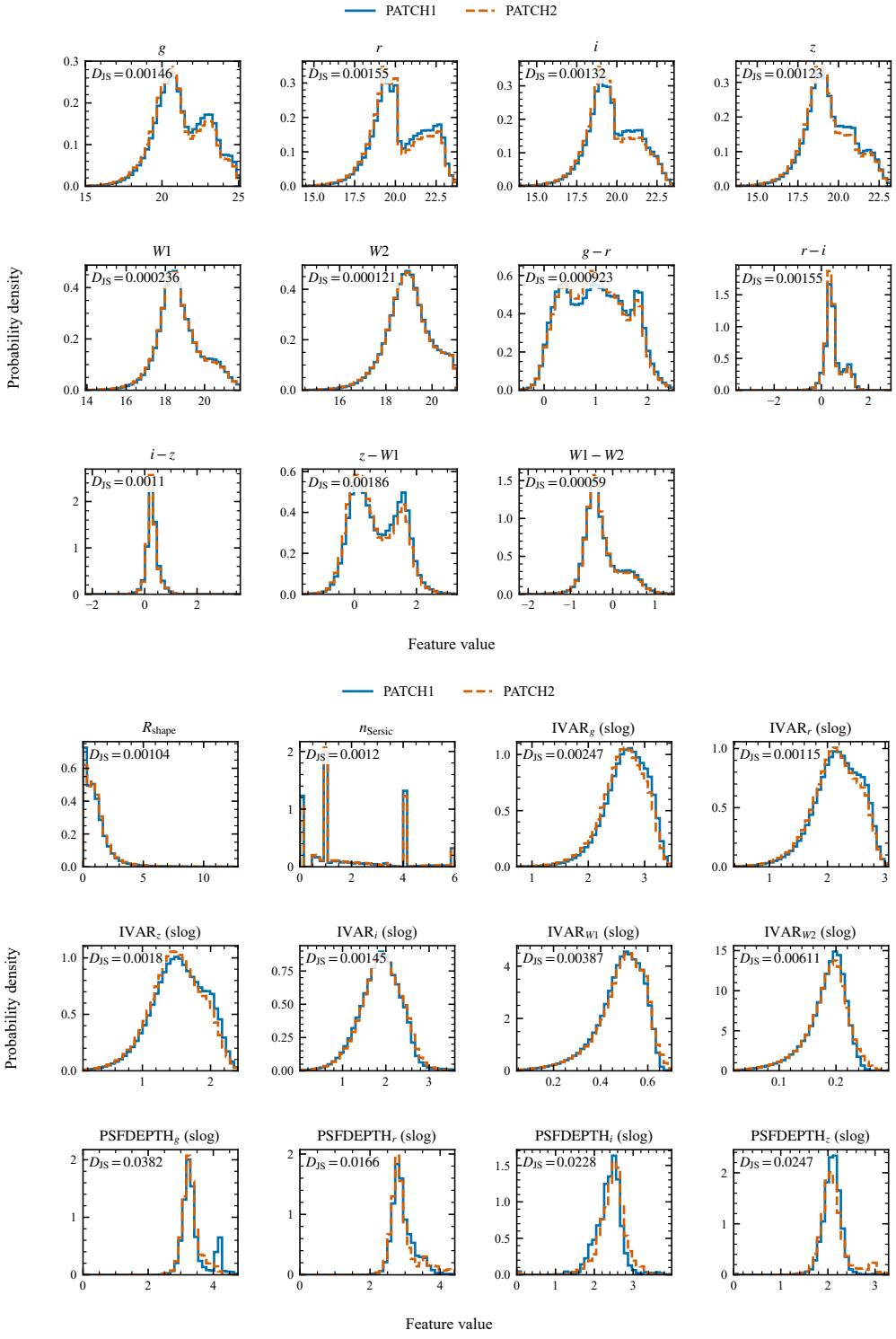


Figure A2. Comparison of the 23 photometric features distributions between training, validation footprint (1,215,914 sources) and testing footprint (2,028,371 sources), including magnitudes (g, r, i, z, $W1$, and $W2$), five colour indices ($g-r$, $r-i$, $i-z$, $z-W1$, and $W1-W2$), morphological parameters (SHAPE_R and SERSIC), flux inverse variances in six bands, and PSF depths in four optical bands. The blue solid and orange dashed lines represent the training, validation patch and testing patch, respectively. The histograms are normalised by the total number of sources in each patch and the corresponding bin widths. For visualization, the displayed range of each feature is determined by the minimum 0.05th and maximum 99.9th percentiles of the two samples. A signed logarithmic transformation, $\mathrm{sgn}(x)\log_{10}(1+|x|)$, is applied to the flux inverse variances and PSF depths to compress their ranges. The Jensen–Shannon divergence, D_{JS}, indicated in each panel, is calculated using the full samples with 100 common histogram bins, including boundary bins that retain probability outside the displayed ranges. The two patches exhibit broadly similar distributions, with relatively larger differences in the PSF depths.

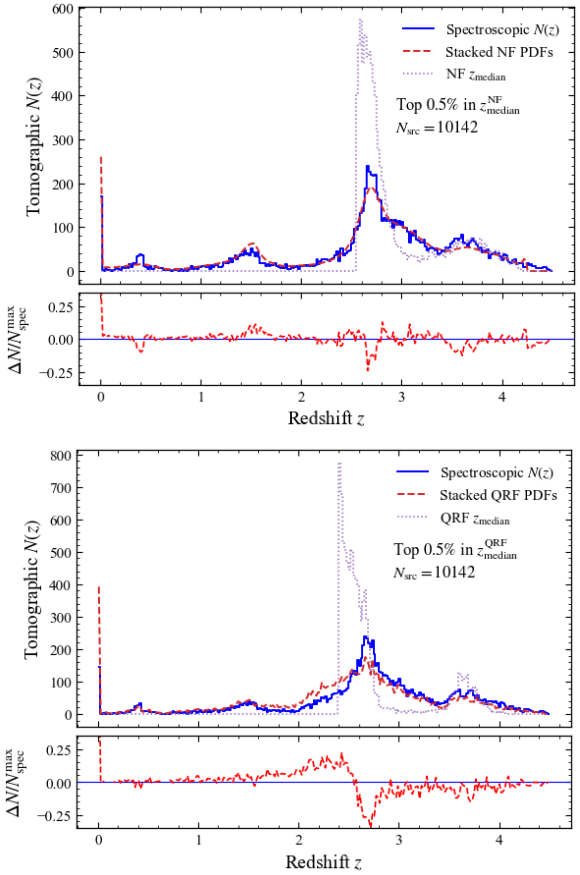


Figure B1. Equal-sample-size robustness test for high-redshift tomographic $N(z)$ reconstruction using the top 0.5 percent of sources ranked by their predicted median redshifts, $z_{\rm median}$, independently for the NF (upper panel) and QRF (lower panel). The solid blue lines show the spectroscopic reference distributions, the dashed red lines show the reconstructed distributions obtained by stacking the corresponding predicted PDFs, and the dotted purple lines show the distributions of the posterior-median estimates. The lower sub-panels show the residuals, $\Delta N/N^{\rm max}_{\rm spec}$, relative to the corresponding spectroscopic distributions.

feature of galaxy i. The gradient-based saliency is then calculated as:

$$S_{ij} = \left| \frac{\partial \log p_\theta(z_{{\rm spec},i} \mid \mathbf{x}_i)}{\partial x_{ij}} \right|. \tag{C1}$$

The gradients are evaluated using the standardised input features, with the trained model parameters held fixed. A larger saliency indicates greater local sensitivity of the predicted log-probability density to variations in the corresponding input feature.

For each feature, we compute the median absolute saliency across all evaluated sources, $\widetilde{S}_j = {\rm median}_i(S_{ij})$, and normalise it by the largest median saliency among the input features:

$$S^{\rm norm}_j = \frac{\widetilde{S}_j}{\max_k \widetilde{S}_k}. \tag{C2}$$

The resulting normalised saliencies are presented in Fig. C1, with error bars indicating the 16th–84th percentile ranges of the individual-source saliencies, normalised by the same factor.

The NF exhibits the strongest sensitivity to colour features, particularly $r-i$, followed by $g-r$ and $i-z$. Individual magnitudes also show appreciable sensitivity, suggesting that the model responds to both colour and magnitude information. The morphological parameters exhibit moderate saliency, whereas the flux inverse variances and PSF depths generally show lower values. The relatively broad 16th–84th percentile ranges of the colour features, especially $r-i$, indicate substantial source-to-source variations in their local influence on the predicted log-probability density.

Table C1. Random Forest point-estimation performance for all sources (~ 2.0 million targets in $0.0 < z_{\rm spec} < 4.5$) after removing individual photometric bands from the input feature set. For each ablation, the corresponding magnitude and all colours containing that band are removed, and the colour set is reconstructed from the remaining adjacent bands (e.g. removing r replaces $g-r$ and $r-i$ with $g-i$). The full RF model is shown for reference.

Removed band	bias($\times 10^{-4}$)	$\sigma_{\rm NMAD}$ (10^{-2})	Outlier fraction (%)
None	4.65	2.92	6.58
g	-4.77	3.56	7.46
r	-0.33	3.51	7.47
i	7.79	3.24	6.91
z	6.88	3.26	7.06
$W1$	6.84	3.09	7.41
$W2$	5.82	3.07	7.10

However, a relatively low gradient saliency does not necessarily imply that a feature has a negligible contribution to the overall photo-z performance. As a complementary analysis, we investigate the impact of individual photometric bands on the point-estimation performance of the RF model. For each band-ablation test, we remove the corresponding magnitude and compute the colour features using the remaining bands. For example, removing the r band replaces $g-r$ and $r-i$ with $g-i$. The resulting performance metrics are summarised in Table C1.

Removing either the g or r band leads to the largest degradation in both $\sigma_{\rm NMAD}$ and the outlier fraction, highlighting the importance of optical photometric information for redshift estimation. Removing the i or z band produces a more moderate degradation. Interestingly, excluding the $W1$ band results in a relatively small increase in $\sigma_{\rm NMAD}$ but a substantial increase in the outlier fraction, comparable to the effects of removing the g or r band. In contrast, removing $W2$ produces smaller changes in both metrics. These results suggest that features with relatively low gradient saliency may still contribute significantly to the overall predictive performance, particularly in reducing catastrophic outliers. Although the NF saliency and RF ablation analyses probe different aspects of feature sensitivity and are performed using different models, they provide complementary insights into the role of photometric information in photo-z estimation.

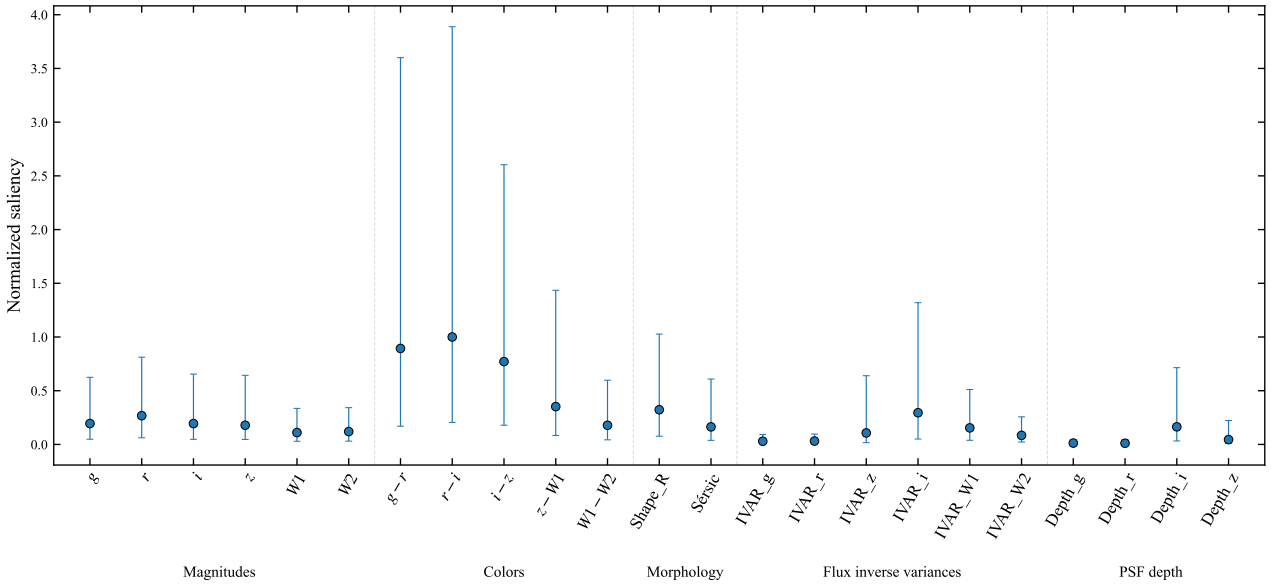


Figure C1. Feature saliency of the trained NF model for the input photometric parameters using 1.1 million training sources. Points show the median absolute input-gradient saliency of each feature, normalised by the maximum median value, while the error bars indicate the 16th–84th percentile range across individual sources. The features are grouped into magnitudes, colours, morphology, flux inverse variances, and PSF depths, with vertical dashed lines separating the groups. A larger saliency indicates greater local sensitivity of the predicted redshift PDF to the corresponding input feature.

This paper has been typeset from a TEX/LATEX file prepared by the author.