The Astrophysical Journal, 954:201 (15pp), 2023 September 10 https://doi.org/10.3847/1538-4357/aceafa

Photo-zSNthesis: Converting Type Ia Supernova Lightcurves to Redshift Estimates via Deep Learning

Helen Qu and Masao Sako
Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA; helenqu@sas.upenn.edu
Received 2023 April 28; revised 2023 July 20; accepted 2023 July 21; published 2023 September 7

Abstract

Upcoming photometric surveys will discover tens of thousands of Type Ia supernovae (SNe Ia), vastly outpacing the capacity of our spectroscopic resources. In order to maximize the scientific return of these observations in the absence of spectroscopic information, we must accurately extract key parameters, such as SN redshifts, with photometric information alone. We present Photo-zSNthesis, a convolutional neural network-based method for predicting full redshift probability distributions from multi-band supernova lightcurves, tested on both simulated Sloan Digital Sky Survey (SDSS) and Vera C. Rubin Legacy Survey of Space and Time data as well as observed SDSS SNe. We show major improvements over predictions from existing methods on both simulations and real observations as well as minimal redshift-dependent bias, which is a challenge due to selection effects, e.g., Malmquist bias. Specifically, we show a 61× improvement in prediction bias $\langle \Delta z \rangle$ on PLAsTiCC simulations and 5× improvement on real SDSS data compared to results from a widely used photometric redshift estimator, LCFIT +Z. The PDFs produced by this method are well constrained and will maximize the cosmological constraining power of photometric SNe Ia samples.

Unified Astronomy Thesaurus concepts: Type Ia supernovae (1728); Observational cosmology (1146); Convolutional neural networks (1938); Neural networks (1933); Astrostatistics (1882)

1. Introduction

The study of Type Ia supernovae (SNe Ia) has proven to be a crucial tool in modern cosmology, providing insight into the expansion rate of the universe and the properties of dark energy (Riess et al. 1998; Perlmutter et al. 1999). Measuring the cosmological redshift of each SN, a proxy quantity for recessional velocity, is essential for accurate estimation of the distance–redshift relation and resulting cosmological analyses. However, traditional methods of measuring redshifts are time-consuming and resource-intensive, primarily relying on spectroscopic observations of the SNe themselves or their host galaxies. Using host galaxy redshifts can also lead to cosmological biases if the host is incorrectly identified (Qu et al. 2023).

Most SNe Ia cosmological analyses so far have relied on SN spectra for SN type confirmation as well as redshift information, but only with samples of up to ∼1500 SNe (e.g., Abbott et al. 2019; Brout et al. 2022). Spectroscopic follow-up of all SNe Ia candidates or their host galaxies will become infeasible with future sky surveys such as the Legacy Survey of Space and Time (LSST) at the Vera C. Rubin Observatory, which will discover tens of thousands of SNe Ia over the course of their observational lifetimes (Abell et al. 2009). Recent improvements in photometric SN classification (e.g., Möller & de Boissière 2020; Qu et al. 2021) have drastically reduced the likelihood of non-Ia contamination in photometric SNe Ia samples and enabled cosmological analyses with photometrically classified samples (Vincenzi et al. 2023). However, spectroscopic redshifts were still available for the host galaxies of these photometrically confirmed SNe Ia and were used as the SN redshifts. SN redshift estimates independent of host galaxy redshift can also serve as an independent cross-check for host galaxy association, ensuring accurate studies of host galaxy correlations and corrections for the mass step (e.g., Rigault et al. 2020). Accurate photometric redshift estimates for SNe that are independent of host galaxy spectroscopic redshifts are thus the final building block required to enable SN Ia cosmology for the LSST era.

Cosmological inference frameworks that account for the inflated uncertainties from photometric redshifts are currently being developed. Mitra et al. (2023) and Dai et al. (2018) show promising results with simulated LSST samples and SN photometric redshifts fitted using host galaxy redshift priors. In particular, Dai et al. (2018) recover a fitted Ω_m value consistent with the input cosmology when using SN photo-zs fitted with a host galaxy photo-z prior. Mitra et al. (2023) show a 2% effect on fitted cosmological parameters of an assumed systematic uncertainty due to the use of SN photo-zs of 0.01. Using observed data from the Dark Energy Survey, Chen et al. (2022) performed a cosmological analysis using a subset of ∼100 SNe Ia hosted by galaxies in the redMaGiC catalog, which have both photometric and spectroscopic redshifts. The difference in best-fit cosmological parameters between using spectroscopic and photometric redshifts was found to be minimal, $\Delta w \sim 0.005$. redMaGiC galaxies were chosen for this analysis due to their particularly well-constrained photometric redshift estimates, which are not representative of the full population of SN host galaxies. However, even with a sound cosmological inference framework, galaxy photometric redshifts are often inaccurate or plagued with large uncertainties, and requiring host galaxy information to produce SN photometric redshift estimates may introduce additional biases.

These issues with traditional redshift determination along with the development of a cosmological framework for photometric redshifts have led to a growing interest in alternative techniques for predicting redshifts for SNe Ia. In

this work, we introduce a novel machine-learning algorithm to predict full redshift probability distributions for SNe Ia based solely on lightcurve data. Our estimator harnesses the constraining power on the redshift of the SN lightcurves and can additionally provide an independent cross-check on host galaxy matches, minimizing mismatch rates. We present a detailed analysis of our model, including its accuracy and limitations, and discuss the potential implications of our findings for future cosmological studies.

1.1. Photo-z Estimation

Most of the existing literature on photometric redshift estimation is on galaxy photo-zs. These approaches generally use either (1) a training set of galaxy photometric observables, such as colors and magnitudes, to determine a mapping to spectroscopic redshifts (e.g., Brunner et al. 1997), or (2) template fitting, in which observed properties are compared with redshifted template spectra to determine the best-fit redshift value (e.g., Benitez 2000). There have also been successful machine-learning-based galaxy photo-z models developed, e.g., D'Isanto & Polsterer (2018) and Pasquet et al. (2018) using convolutional neural networks on galaxy images, and Buchs et al. (2019) using a self-organizing map to relate color–magnitude space and redshifts.

Photometric redshift estimation for SNe uses many of the same techniques. Kim & Miquel (2007) used the SALT2 lightcurve model (Guy et al. 2007) to determine photometric redshifts and distances as well as uncertainties using the Fisher information matrix. Palanque-Delabrouille et al. (2010) and Kessler et al. (2010, LCFIT+Z) extended the SALT2 lightcurve fit to include redshift as a fitted parameter, and incorporating the host galaxy redshift as a prior. These approaches find the best-fit redshift by comparing photometric observables to those expected from the SALT2 model via χ^2 minimization. An analytic approach was presented in Wang et al. (2015), which assumed a functional form for the redshifts as a function of multi-band photometric fluxes and fit for free parameters using a training set of SNe Ia with spectroscopic redshifts. de Oliveira et al. (2022) applied machine-learning techniques to this problem, performing regression using features obtained from principal component analysis. However, many of these results suffer from redshift-dependent bias, in which photo-zs for high-redshift SNe are systematically underestimated (see, e.g., Figure 6 of Kessler et al. 2010, Figure 7 of de Oliveira et al. 2022). Most of these studies also estimate a redshift value and uncertainty rather than the full probability density function (PDF).

1.2. Overview

In this work, we present Photo-zSNthesis, a convolutional neural network model that uses multi-band photometry to predict a full redshift PDF. Our approach uses raw photometric data and does not require any manual feature engineering to determine the most predictive features. We evaluated Photo-zSNthesis using LSST and SDSS simulated SNe Ia as well as the SDSS photometric SN Ia sample, and show that our redshift predictions have low scatter and suffer from minimal redshift-dependent bias on all tested SN samples. Our results are a promising step toward precision photometric SN Ia cosmology.

We introduce the simulated and observed data samples used for evaluation in Section 2, as well as the preprocessing step used to transform multi-band lightcurves into convolutional neural network inputs. In Section 3, we describe the model architecture and training strategy. We present results, comparisons with LCFIT+Z, and further experiments in Section 4, and conclude in Section 5.

2. Data

2.1. Data Sources

We present results on simulated SNe Ia from two surveys: LSST (Ivezić et al. 2019) and the SDSS-II SN survey (Frieman et al. 2008). We also demonstrate that our model generalizes well to an observed photometric SN Ia data set by showing photo-z predictions on SDSS SNe classified as type Ia by SuperNNova (Möller & de Boissière 2020). We use the SuperNova ANAlysis software (SNANA; Kessler et al. 2009) to produce all simulated SN lightcurves used in this work.

2.1.1. Simulated LSST SNe Ia (PLAsTiCC)

LSST is a ground-based dark energy survey program that will discover millions of SNe over the 10 yr survey duration. The 8.4 m Simonyi Survey optical telescope at the Rubin Observatory uses a state-of-the-art 3200 megapixel camera with a 9.6 deg^2 field of view that will provide deeper and wider views of the universe with unprecedented quality. LSST will observe nearly half the night sky each week to a depth of 24th magnitude in *ugrizY* photometric bands spanning wavelengths from ultraviolet to near-infrared.

We simulate LSST-like observations of SNe Ia in *ugrizY* photometric bands following the model, rates, and LSST observing conditions developed for the PLAsTiCC data set (Kessler et al. 2019; Allam, et al. 2020) for $0.05 < z < 1.2$. We use the SALT2 lightcurve model (Guy et al. 2007) with training parameters derived from the Joint Lightcurve Analysis (Betoule et al. 2014) extended into the ultraviolet and near-infrared by Hounsell et al. (2018) following the procedure described in Pierel et al. (2018). While PLAsTiCC included two LSST observing strategies, the Deep Drilling Fields (DDF) as well as the Wide-Fast-Deep (WFD), we simulate only the DDF subsample. We co-add all observations within the same night, following PLAsTiCC, resulting in observations that are $\sim$2.5$\times$ more frequent and $\sim$1.5 mag deeper than the WFD sample. Though we focused on the DDF sample for this work, evaluating the performance of Photo-zSNthesis on WFD simulations is an important direction for future studies.

As PLAsTiCC is the name of the data set we emulated while LSST is the survey we simulate, we will use both interchangeably to denote "simulated LSST SN lightcurves following the PLAsTiCC data set."

2.1.2. Simulated SDSS SNe Ia

The SDSS-II Supernova Survey identified and measured lightcurves for intermediate-redshift ($0.05 < z < 0.4$) SNe Ia using repeated five-band (*ugriz*) imaging of Stripe 82, a stripe 2.°5 wide centered on the celestial equator in the Southern Galactic Cap. The primary instrument for this survey is the SDSS CCD camera mounted on a dedicated 2.5 m telescope at Apache Point Observatory, New Mexico. Over the three observing seasons between 2005 and 2007, SDSS discovered 10,258 transient and variable objects, with 536 spectroscopically confirmed SNe Ia and an additional 907

Table 1
Number of SDSS Lightcurves Remaining After Each Selection Cut

Cut	Number of SNe
Full sample	10,258
Lightcurve selection (Section 2.1.4)	2044
Likely SNe Ia ($P_{\rm Ia} \geqslant 0.5$)	1037
SALT fit cuts (Section 2.2)	555
Successful LCFIT+Z fit	489

Note. We evaluate Photo-zSNthesis on the remaining 489 SNe.

photometrically classified SNe Ia candidates (Sako et al. 2018). Frieman et al. (2008) provide an in-depth review of the SDSS-II SN survey.

We simulate SDSS SNe Ia in *ugriz* photometric bands using the SALT3 model described by Kenworthy et al. (2021) due to its improved wavelength range coverage in the near-infrared, allowing z-band observations to be simulated for SNe at low redshifts. We additionally extend this model to 500 Å to simulate u-band observations at high redshifts. We use simulated observing conditions from Kessler et al. (2013) and host galaxy spectroscopic detection efficiency from Kessler et al. (2009). We simulate a *photometric* SDSS SNe Ia data set using the host spectroscopic detection efficiency rather than an SN spectroscopic detection efficiency to demonstrate the performance of Photo-zSNthesis on the practical use case of a photometric sample.

2.1.3. Observed SDSS SNe Ia

In addition to simulated SNe, we test Photo-zSNthesis on SDSS lightcurves from Sako et al. (2018) classified by SuperNNova (Möller & de Boissière 2020) as likely SNe Ia, defined as SNe with SN Ia probability $P_{\rm Ia} \geqslant 0.5$. *True* redshifts $z_{\rm true}$ for the sample are from spectra of the SNe themselves or their host galaxies. Details on the training data and training procedure for the SuperNNova model used here can be found in B. P. Popovic et al. (2023, in preparation) 489 lightcurves remain after the selection cuts described in Sections 2.1.4 and 2.2. The number of SDSS lightcurves remaining after each set of cuts is shown in Table 1.

2.1.4. Lightcurve Selection

We apply basic selection cuts to the simulated LSST sample. We require

1. At least five observations of each SN
2. Signal-to-noise ratio (S/N) $>$ 3 for at least one observation each in two of the *griz* filters.

We apply selection cuts, following Popovic et al. (2020), to both the simulated and observed SDSS data to remove poor quality lightcurves. We define a rest-frame age, $T_{\rm rest} = (t - t_{\rm peak})/(1+z)$, where t is the observation date, $t_{\rm peak}$ is the estimated epoch of SN peak brightness from SNANA, and z is the redshift of the event. We require

1. $0 < T_{\rm rest} < 10$
2. S/N > 5 for at least one observation each in two of the *griz* filters.

We additionally require SNe in the SDSS data to have a spectroscopic redshift from either the SN spectrum or the SN host galaxy.

2.2. Data Preprocessing

All SNe in all data sets are fit with a χ^2-minimization program included in SNANA to determine several rest-frame parameters under the assumption that the event is an SN Ia: the time of SN peak brightness $t_{\rm peak}$, a stretch-like parameter x_1, a color parameter c, and the lightcurve normalization parameter x_0, as well as their uncertainties and covariances (i.e., σ_{x_1}, etc.). These parameters are used to calculate the distance modulus μ, allowing the SNe to be placed on the Hubble diagram. SDSS data sets are fit with the SALT3 model while the LSST data set is fit with the same extended SALT2 model used for the simulations.

We additionally select only SNe in our data sets that are well-described by the SALT model. These criteria were chosen following those used in past cosmology analyses, e.g., Betoule et al. (2014):

1. $|c| < 0.3$
2. $|x_1| < 3$
3. $\sigma_{x_1} \leqslant 1$
4. $\sigma_{t_{\rm peak}} \leqslant 2$.

Finally, we choose SNe within the redshift range $0.01 \leqslant z \leqslant 0.4$ for SDSS and $0.01 \leqslant z \leqslant 1.2$ for LSST that have a successful LCFIT+Z (Kessler et al. 2010) fit, in order to draw direct comparisons between the performance of Photo-zSNthesis and LCFIT+Z. A more detailed description of LCFIT+Z can be found in Section 4.1.2.

146,069 simulated SDSS SNe Ia and 81,734 simulated LSST SNe Ia remain after all cuts, as well as 489 observed SDSS SNe photometrically classified as SNe Ia. The number of observed SDSS lightcurves remaining after each cut is described in Table 1. The redshift distributions of the PLAsTiCC training and test data sets after selection cuts are shown in Figure 1, while the redshift distributions of the SDSS simulated and observed data sets are shown in Figure 2. We note that, while most machine-learning classifiers perform better when trained on a data set with balanced classes (i.e., flat redshift distribution), we found that this was not the case for Photo-zSNthesis. While our simulations are able to generate any artificial redshift distribution, an unphysical one such as a flat distribution could lead to strange artifacts in the data set, confusing the classifier.

2.2.1. Lightcurve Preprocessing

To preprocess each lightcurve, we first take observations within the range $t_{\rm peak} - 30 \leqslant t \leqslant t_{\rm peak} + 150$, where $t_{\rm peak}$ is the epoch of SN peak brightness estimated by SNANA. This is to ensure the lightcurves contain just the SN and no extraneous information.

We use two-dimensional Gaussian process (GP) regression to model each SN lightcurve in time (t) and wavelength (λ) space, then use the predictions of the fitted GP on a fixed (λ, t) grid as the input to our neural network model. Modeling lightcurves of astronomical transients using 2D Gaussian processes was originally introduced in Boone (2019); and Qu et al. (2021, Q21) used the GP models to create *images* from lightcurve data, which is the technique used for this work.

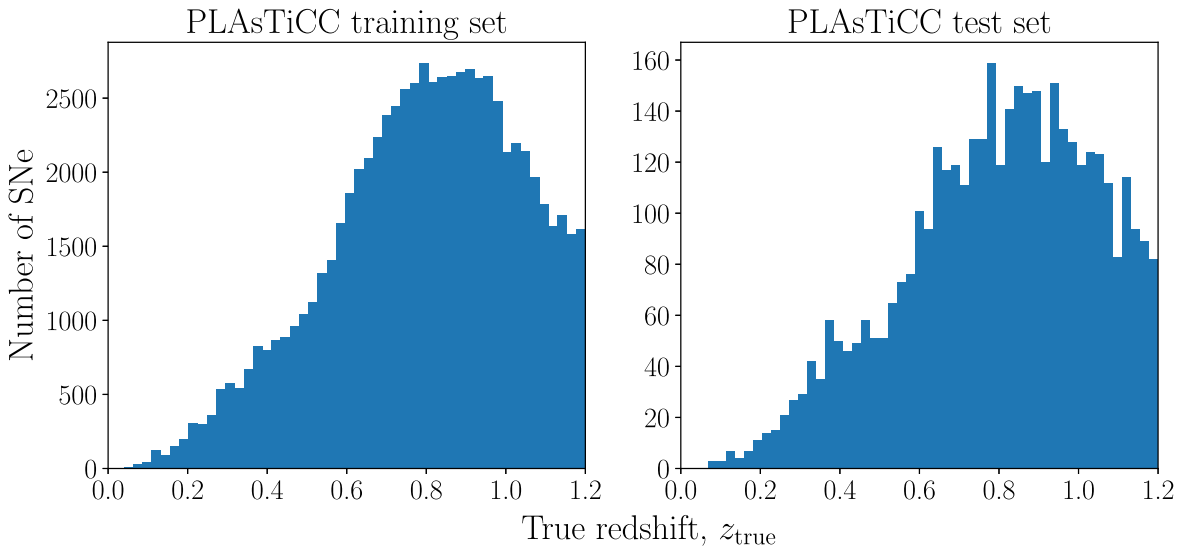

Figure 1. True redshift distribution of the 73,620 simulated SNe Ia in the LSST (PLAsTiCC) training data set (left) and 4057 simulated SNe Ia in the test data set (right).

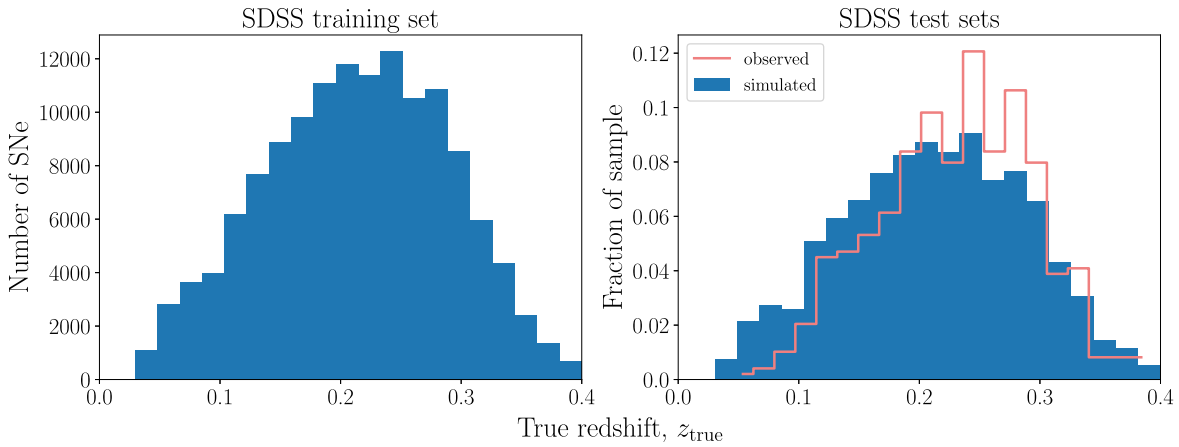

Figure 2. True redshift distribution of the 135,521 simulated SNe Ia in the SDSS training data set (left) and the 5274 simulated SNe Ia and 489 observed photometrically confirmed SNe Ia in the SDSS test sets (right).

Following Q21, we use the Matérn kernel ($\nu = 3/2$) with a fixed 6000 Å length scale in wavelength space and fit for the time length scale as well as the amplitude using maximum likelihood estimation. This GP model is fit separately to each lightcurve and used to produce a smooth 2D representation of the lightcurve by predicting flux values at each point in a (λ, t) grid. We choose 32 equally spaced points in the range $3000\ \text{Å} \leqslant \lambda \leqslant 10100\ \text{Å}$ ($\delta\lambda = 221.875$ Å) and 180 points in the range $t_{\rm peak} - 30 \leqslant t \leqslant t_{\rm peak} + 150$ ($\delta t = 1$ day). This produces a 32×180 matrix of predicted flux values. We also produce a matrix of prediction uncertainties at each λ_i, t_j of equal size.

We stack the flux and uncertainty matrices depth-wise to produce a $32 \times 180 \times 2$ tensor and divide element-wise by the maximum flux value to constrain all entries to [0,1].

We show low, medium, and high redshift examples of PLAsTiCC lightcurves, the fitted GP models and their uncertainties, and the resulting flux matrices in Figure 3. Redshifting increases the wavelength of light, which we see in the figure as the yellow (high flux) region moving down (toward $\lambda = 10100$ Å) with increasing redshift. We also expect to observe a longer duration for higher redshift transients, which is evident in the increase in the width of the yellow region. This data format is particularly well suited to the redshift prediction task, as we are able to visibly see the expected physical results of redshifting in the flux matrices.

2.2.2. Redshift Preprocessing

The redshift range for each survey is discretized into n_z discrete and non-overlapping bins ($0.01 \leqslant z \leqslant 0.4$, $n_z = 50$ for SDSS and $0.01 \leqslant z \leqslant 1.2$, $n_z = 150$ for LSST). We chose these n_z values to preserve the approximate width of each redshift bin across surveys ($\delta z \sim 0.0078$). The bin corresponding to the true redshift of each SN is one-hot encoded and passed into the model as the training label.

3. Model

Photo-zSNthesis is a convolutional neural network model (e.g., LeCun et al. 1989; Simonyan & Zisserman 2014; Zeiler & Fergus 2014; Krizhevsky et al. 2017) that takes in GP-interpolated lightcurves as well as the GP prediction uncertainties, prepared as described in Section 2.2, and predicts a probability distribution over fine-grained redshift bins. This

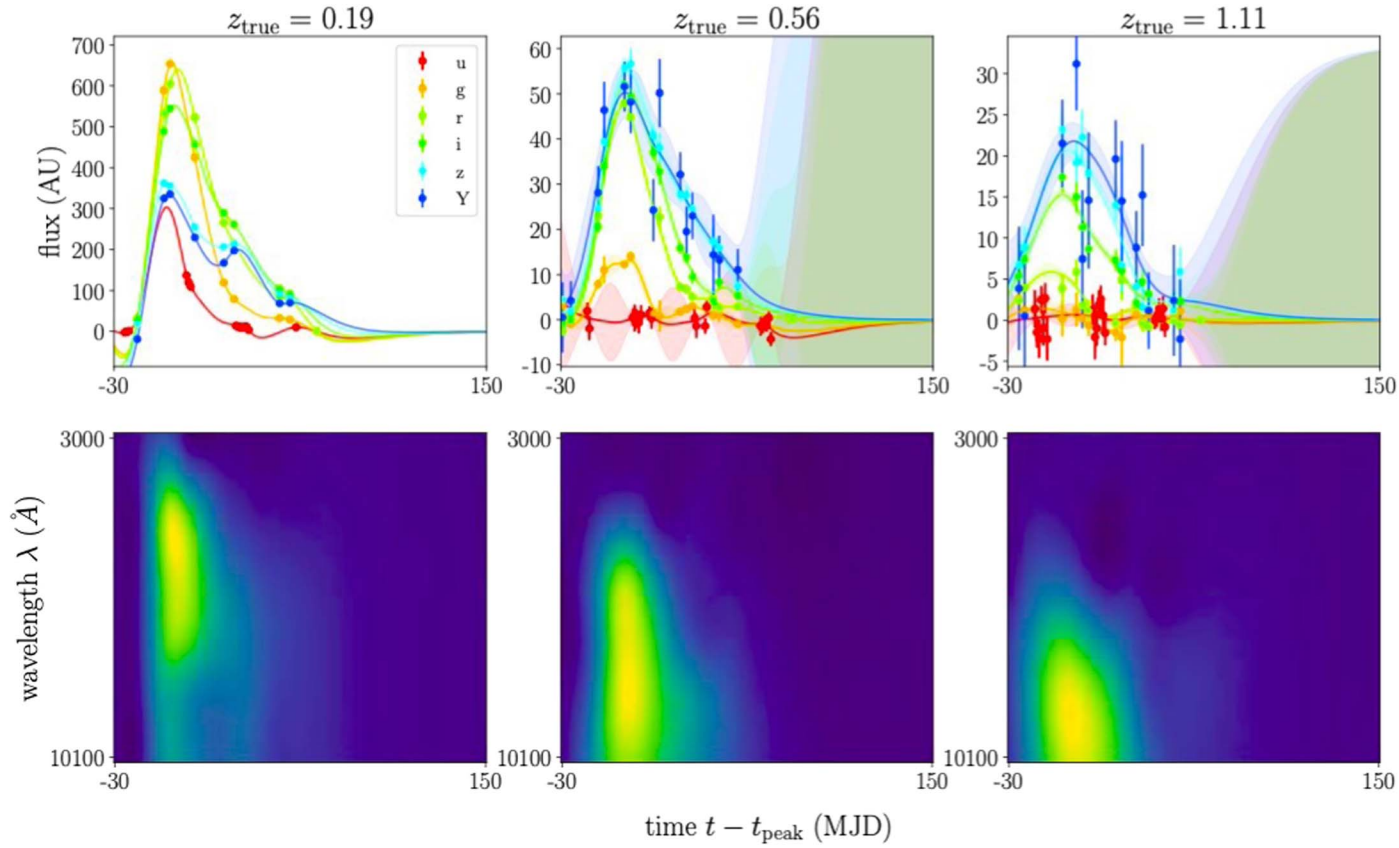

Figure 3. Examples of a low (left), medium (middle), and high (right) redshift simulated PLAsTiCC SN. (Top panel) The multi-band lightcurves are shown in colored points, with Gaussian process predictions and uncertainties for this lightcurve shown in colored lines and shaded regions, respectively. (Bottom panel) The flux matrix created from the lightcurve in the top panel (see Section 2.2.1 for details). Large flux values are colored in yellow while lower flux values are in dark blue. The physical effects of redshifting, including longer wavelengths and longer transient durations, are observed in the flux matrices in the downward (toward higher wavelengths) shift of the yellow high flux region and the increased width of the yellow region with increasing redshift.

approach allows us to produce a discretized full PDF over redshift space for each individual SN without any assumptions on the underlying distribution, and has been used in a variety of contexts including image generation (van den Oord et al. 2016) and prediction of precipitation probabilities (Sønderby et al. 2020). Treating this as a categorical classification problem using the cross-entropy loss function has also been shown to accurately approximate Bayesian posterior probabilities (Richard & Lippmann 1991).

3.1. Convolutional Neural Networks

The convolutional neural network (CNN) is a class of artificial neural networks with properties particularly suited to object and image recognition. It requires fewer trainable parameters than the standard feedforward network due to the convolution operation, learning a single weight matrix for small neighborhoods of the input image. This property is not only parameter efficient but also imparts CNNs with translation equivariance, i.e., the same feature shifted by n pixels will produce the same response shifted by n pixels. These convolutional layers are paired with pooling layers, which downsample the input to allow for the next set of convolutional layers to learn hierarchically more complex features. CNNs are prized in the machine-learning community for being simple yet performant on image recognition benchmark data sets such as ImageNet (Russakovsky et al. 2015).

3.2. Residual Learning

We implement residual learning (He et al. 2016) due to its state-of-the-art performance on the ImageNet benchmark at the time of publication as well as its widespread adoption over vanilla CNNs. Residual learning was presented as a solution to the degradation problem in CNNs, in which performance degraded past a certain threshold of network depth. Since the additional layers could simply act as identity mappings and not affect the network performance, He et al. (2016) decreased the difficulty by not only feeding inputs sequentially through a series of layers, but also adding the inputs back into the outputs of those layers. This allows the layers to learn zero mapping rather than identity mapping. Since these layers are tasked with learning the *residual* with respect to the input, this is known as *residual learning* and the stack of layers is known as a *residual block*. The residual connections in the Photo-zSNthesis architecture are shown as lines curving around each residual block in Figure 4.

3.3. Architecture

Our model (Figure 4) takes as input a $32 \times 180 \times 2$ tensor containing the GP-interpolated lightcurve data on a 32×180 grid of wavelength and time values, and the GP uncertainties at each of those points. Table 2 shows the input and output dimensions of example layers in the model architecture. We also provide the model with the one-hot encoded vector specifying the correct redshift bin as the training label.

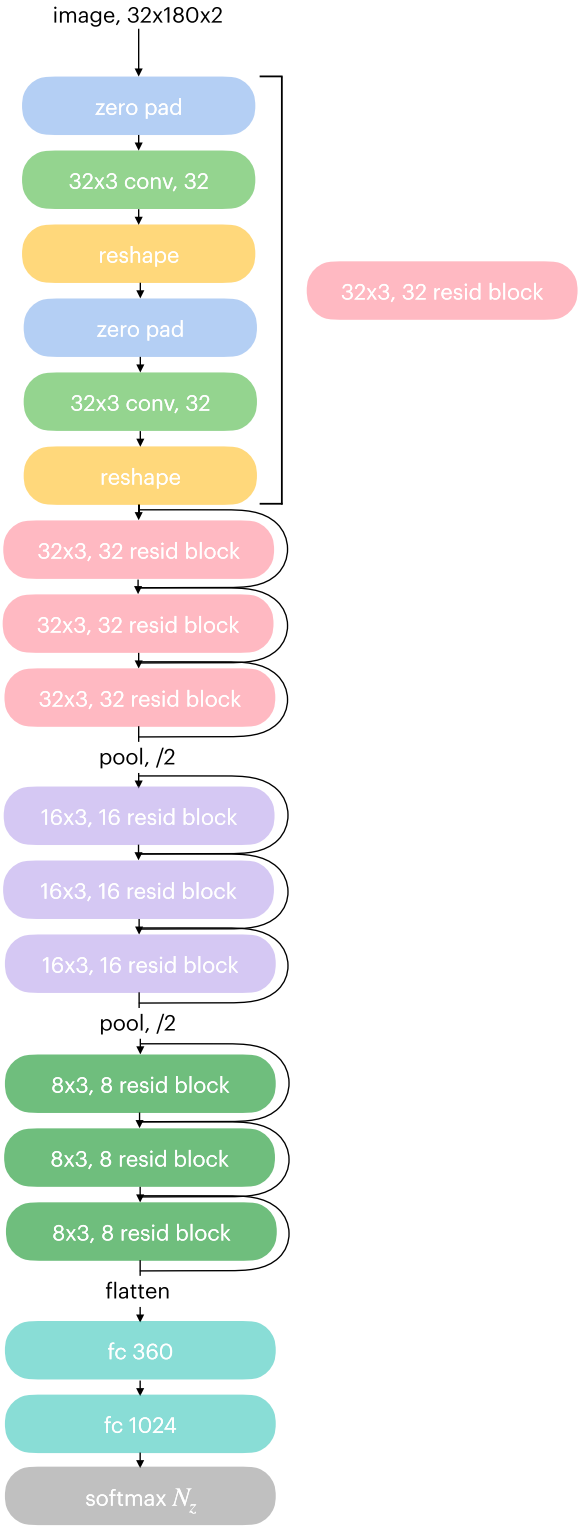

Figure 4. The model architecture developed for this work, described in the text in Section 3.3. A layer shown in the figure as "$h \times d$ conv, $n_{\rm filters}$" is a 2D convolutional layer with a kernel size of $h \times d$ and $n_{\rm filters}$ filters. Similarly, "$h \times d$, $n_{\rm filters}$ resid block" defines a residual block containing two $h \times d$ conv, n_f convolutional layers. "fc $n_{\rm nodes}$" denotes a fully connected layer with $n_{\rm nodes}$ nodes. The pooling layers apply max pooling, downsampling both the height and width by 2. The output softmax layer has N_z nodes, $N_z = 50$ for SDSS and $N_z = 150$ for LSST.

The input is zero padded on both sides to $32 \times 182 \times 2$ to ensure that feature maps output by the convolutional layers retain the original shape, then passed through a convolutional layer with 32 filters and kernel size 32×3. Since the convolutional kernel determines the receptive field of each unit in the layer, we choose a convolutional kernel that spans the wavelength space to allow each unit to learn from all wavelengths simultaneously while preserving the linearity of time. The output of this convolutional layer is $1 \times 180 \times 32$, which we then reshape back to $32 \times 180 \times 1$. This first feature map is now passed through a series of residual blocks.

Table 2
Description of Example Layers in the Model Architecture

Layer	Input Shape	Output Shape
Zero padding	$32 \times 180 \times 2$	$32 \times 182 \times 2$
32×3 convolutional	$32 \times 182 \times 2$	$180 \times 32 \times 1$
Reshape	$180 \times 32 \times 1$	$32 \times 180 \times 1$
Max pooling	$32 \times 180 \times 1$	$16 \times 90 \times 1$

Each residual block contains two convolutional blocks, each consisting of a ReLU nonlinearity, batch normalization, and the zero-padding, convolutional layer, and reshaping layer identical to the ones described above. The input to each residual block is then added to the output as described in Section 3.2. After each series of three residual blocks, the output is passed through a 2×2 max-pooling layer, downsampling the height and width of the output by a factor of 2.

Finally, the output is flattened and passed through a fully connected layer with 1024 hidden nodes, which connects to the final softmax layer. The nodes in this layer correspond to redshift bins; thus, it has 50 nodes for processing SDSS data and 150 for LSST data. The array of output probabilities is interpreted as the probability density over redshifts for our input SN.

3.4. Calibration

We also performed temperature scaling (Guo et al. 2017) to ensure that the probabilities output by Photo-zSNthesis are properly calibrated. In this process, we learn a single *temperature* parameter used to scale the output probabilities.

Before scaling, the output probabilities p_i for input SN i are derived from the softmax function

$$\boldsymbol{p}_i^{(k)} = \sigma_{\rm SM}(\boldsymbol{q}_i)^{(k)} = \frac{\exp(\boldsymbol{q}_i^{(k)})}{\sum_{j=1}^{K} \exp(\boldsymbol{q}_i^{(j)})}, \quad (1)$$

where $\boldsymbol{q}_i$ is the vector of network logits corresponding to SN i, i.e., the output of the final hidden layer of the network. The temperature parameter T is learned by minimizing the cross-entropy loss between the one-hot encoded labels and the scaled probabilities, $\boldsymbol{p}_i'$,

$$\boldsymbol{p}_i' = \sigma_{\rm SM}(\boldsymbol{q}_i/T). \quad (2)$$

Calibration is typically evaluated using reliability diagrams (DeGroot & Fienberg 1983; Niculescu-Mizil & Caruana 2005) and the expected calibration error statistic (ECE; Naeini et al. 2015). Reliability diagrams show the prediction accuracy as a function of *confidence*, which is defined as the probability associated with the predicted class: $\max(\boldsymbol{p}_i)$. A reliability diagram for a perfectly calibrated classifier will show the identity function, and any deviation from a perfect diagonal is a sign of miscalibration. The reliability diagram before and after temperature scaling for the PLAsTiCC model is shown in

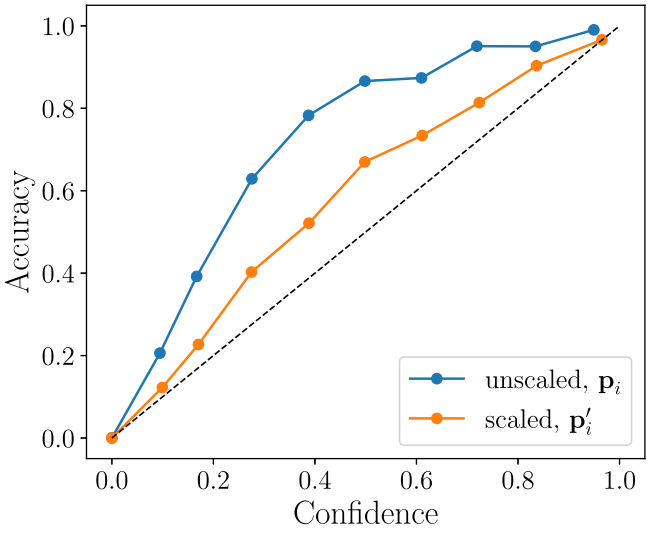

Figure 5. Reliability diagram showing PLAsTiCC model calibration before and after temperature scaling.

Figure 5. The scaled probabilities are much closer to the diagonal, representing a significant improvement in calibration.

ECE is a weighted average of the difference between the accuracy and the confidence in bins of confidence values. ECE is more precisely defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|, \tag{3}$$

where B_m is the mth confidence bin and N is the total number of samples in the data set. Prior to temperature scaling, the ECE for the PLAsTiCC model probabilities was 0.24, which improved to 0.08 after scaling.

3.5. Implementation Details

We split both the SDSS and LSST data sets into 90% training, 5% validation, and 5% test data sets and trained both models with batch sizes of 2048 for 750 epochs. We minimize the cross-entropy loss function with the Adam optimizer (Kingma & Ba 2014) with an initial learning rate of 1e-3 that is halved after 25 epochs of no improvement in the validation loss. The total number of trainable parameters in the SDSS model is 451,654 and 554,154 for the LSST model due to the difference in the number of redshift bins and the resulting difference in output layer size. To prevent overfitting, we use a weight decay of 1e-3 as well as dropout layers (Hinton et al. 2012). Calibration was performed with the validation set of both data sets and with an initialization of $T = 1$. We minimize a cross-entropy loss function using Adam and find $T_{\mathrm{SDSS}} = 0.82$ and $T_{\mathrm{LSST}} = 0.64$.

4. Results and Discussion

Examples of lightcurves and their corresponding predicted redshift PDFs are shown in Figure 6. We chose spectroscopically confirmed SDSS SNe Ia lightcurves, and selected a high accuracy ($\Delta z < 0.005$, $\Delta z \equiv \frac{z_{\mathrm{pred}} - z_{\mathrm{true}}}{1 + z_{\mathrm{true}}}$), medium accuracy ($0.01 \leqslant \Delta z \leqslant 0.05$), and an outlier ($\Delta z > 0.05$) example.

Table 3
Evaluation Metrics Computed for the PLAsTiCC Test Data Set for Both the Mean(PDF) and Max(PDF) Point Estimates for Our Model as Well as the LCFIT+Z

Metric	This Work		LCFIT+Z
	Mean	Max	
Bias $\langle \Delta z \rangle$	0.00095	**0.00075**	0.058
σ_{MAD}	0.0081	**0.0025**	0.0450
Outlier rate η	**3.87%**	4.24%	32.3%

Note. The best result for each metric is shown in bold.

4.1. Evaluation Metrics and Basis for Comparison

4.1.1. Point Estimates and Metrics

Although full photometric redshift PDFs are preferred for further statistical analyses (i.e., cosmological analyses), point estimates can also be computed from each PDF. We require these point estimates to evaluate our model performance against the true redshift values. We compare two possible methods to condense a PDF into a point estimate:

$$\mathrm{mean(PDF)} = \frac{1}{2} \sum_i p(Z_i) \frac{\lceil Z_i \rceil^2 - \lfloor Z_i \rfloor^2}{\lceil Z_i \rceil - \lfloor Z_i \rfloor} \tag{4}$$

where Z represents the vector of redshift bins, $\lceil Z_i \rceil$ the right edge of bin i, $\lfloor Z_i \rfloor$ the left edge of bin i, and $p(Z_i)$ the output probability assigned by the model to bin i; and

$$\mathrm{max(PDF)} = \tilde{Z}_{\mathrm{argmax}_i(p(Z_i))}, \tag{5}$$

where $\tilde{Z}$ represents the array of midpoints of the redshift bins Z. Taking the weighted mean is a common summary statistic for PDFs, while taking the bin with maximum probability as the predicted output is typical for classification tasks. As shown in Tables 3 and 4 and Figures 7 and 9, the two methods give similar results with mean(PDF) performing slightly better on the real SDSS data set, and thus we use the mean point estimate for Figures 8, 10, 11, 12, and 13.

We compute the following metrics for both of our point estimates, following, e.g., Pasquet et al. (2018):

1. The residuals $\Delta z \equiv \frac{z_{\mathrm{pred}} - z_{\mathrm{true}}}{1 + z_{\mathrm{true}}}$,
2. The bias $\langle \Delta z \rangle$,
3. The mean absolute deviation $\sigma_{\mathrm{MAD}} = 1.4826 \times \mathrm{median}(|\Delta z - \mathrm{median}(\Delta z)|)$,
4. The fraction of outliers η with $|\Delta z| > 0.05$.

We evaluate Photo-zSNthesis using these metrics on the PLAsTiCC data set as well as the simulated and observed SDSS data sets, and show the results in Tables 3 and 4.

4.1.2. Comparison with LCFIT+Z

LCFIT+Z (Kessler et al. 2010) is an extension of the lightcurve fitting code described in Section 2.1 that treats redshift as an additional free parameter. It determines the best-fit lightcurve parameters by minimizing χ^2 values of the observed and model fluxes, which are computed as a function of the free parameters (four lightcurve parameters t_{peak}, stretch x_1, color c, flux normalization x_0; and redshift z_{phot}). LCFIT+Z is actively used for recent and ongoing experiments performing full cosmological analyses with photometric redshifts (e.g.,

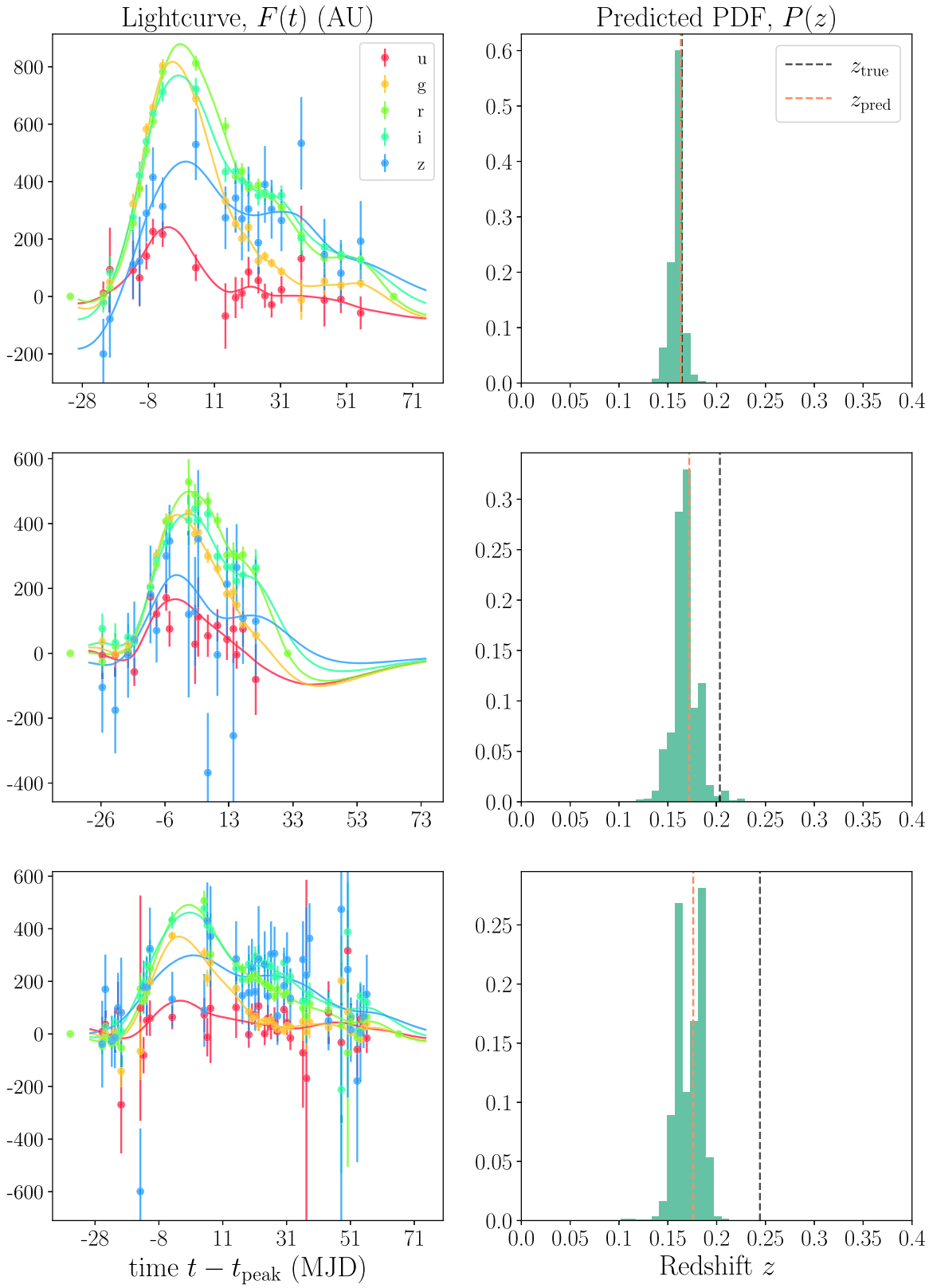

Figure 6. Examples of a high accuracy ($\Delta z < 0.005$, top), medium accuracy ($0.01 \leqslant \Delta z \leqslant 0.05$, middle), and outlier ($\Delta z > 0.05$, bottom) lightcurve and GP fit from the SDSS spectroscopic sample and their predicted PDFs. The true (spectroscopic) redshift is shown in the black dotted line, while the predicted redshift (mean(PDF)) is shown in orange. The medium and poor accuracy lightcurves have larger errors and scatter than the high accuracy lightcurve, and the low accuracy lightcurve appears to have much fewer points.

Dai et al. 2018; Mitra et al. 2023). LCFIT+Z produces point estimates with uncertainties as opposed to a full PDF, so we perform comparisons using our point estimates.

We compare our results with LCFIT+Z as opposed to other SN photometric redshift estimators (e.g., Wang et al. 2015; de Oliveira et al. 2022) due to its demonstrated performance,

Table 4
Evaluation Metrics Computed for the SDSS Simulated and Real Test Data sets for both the mean(PDF) and max(PDF) Point Estimates for Our Model as Well as the LCFIT+Z

Metric	SDSS Simulated			SDSS Real		
	This Work (Mean)	This Work (Max)	LCFIT+Z	This Work (Mean)	This Work (Max)	LCFIT+Z
Bias $\langle \Delta z \rangle$	**7.8e-5**	0.00073	0.023	**0.0050**	0.0052	0.027
$\sigma_{\rm MAD}$	0.011	**0.010**	0.028	**0.018**	0.020	0.047
Outlier rate η	**0.85%**	1.16%	19.9%	**5.14%**	5.78%	26.1%

Note. The best result for each metric and data set is shown in bold.

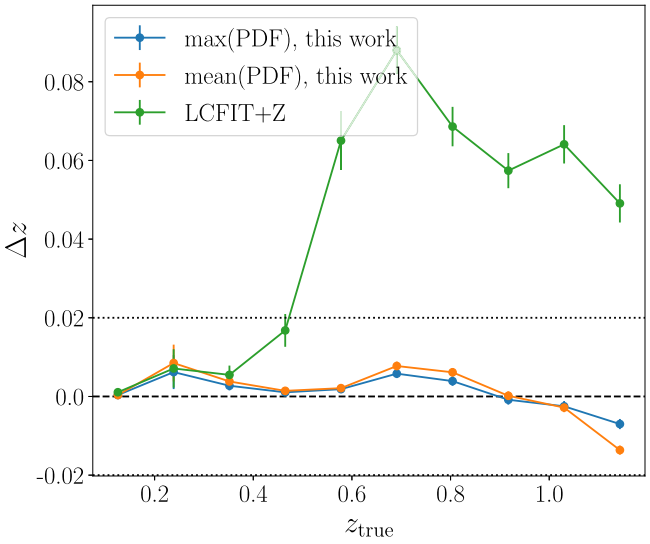

Figure 7. Mean binned residuals Δz as a function of true redshift $z_{\rm true}$ for the PLAsTiCC simulated SNe Ia test set. Predictions from our model have much lower biases as well as scatter compared to predictions from LCFIT+Z. The max(PDF) and mean(PDF) point estimates for our model also agree quite well, resulting in similar Δz values. Dotted lines are plotted at $\Delta z = \pm 0.02$ for reference.

widespread use, and integration into SNANA. We run LCFIT +Z on the same data sets used to evaluate our model, enabling a direct comparison. Note that most documented uses of LCFIT +Z place a prior on the redshift fit using the redshift of the SN host galaxy, resulting in much more constrained fits. However, since (1) host spectroscopic redshifts are unavailable in future survey environments, and (2) to draw direct comparisons with Photo-zSNthesis, which does not require any host galaxy information, we omit the host galaxy redshift prior when running LCFIT+Z. In order to compare how these methods perform using all available information in the photometric survey era, we test Photo-zSNthesis and LCFIT+Z with host galaxy *photometric* redshift priors in Section 4.4.2.

4.2. LSST (PLAsTiCC) Results

We evaluate our model and our baseline for comparison, LCFIT+Z, on a test set of 4057 simulated PLAsTiCC-like lightcurves with true redshift distribution shown in the right panel of Figure 1. The values of the evaluation metrics for results from this work as well as LCFIT+Z are shown in Table 3. The two methods for obtaining point estimates from Photo-zSNthesis PDFs, mean and max, give similar results, though the mean point estimates have a degraded $\sigma_{\rm MAD}$ value compared to the max point estimates but a smaller outlier rate. However, the differences between mean and max point estimates are orders of magnitude smaller than the improvement we see relative to LCFIT+Z. The mean gives a $\sim$3.2$\times$ larger result than max on $\sigma_{\rm MAD}$, compared with an $\sim$180$\times$ larger result from LCFIT+Z.

We show the mean binned residuals, Δz, as a function of true redshift for our model and LCFIT+Z in Figure 7. We see that while the residuals of our model and LCFIT+Z match quite well up to $z_{\rm true} \sim 0.4$, LCFIT+Z has a tendency to dramatically overestimate at higher redshifts. We also see that the spread within each residual bin in LCFIT+Z results is much larger, as shown in the size of the error bars on each point. In contrast, our model shows a relatively flat Δz over the full redshift range with minimal redshift-dependent bias, which has not been achieved by other SN photometric redshift estimators.

Figure 8 more clearly shows the spread of predicted redshifts, where $z_{\rm pred}$ is calculated as the mean(PDF) for our model. The predictions produced by our model (left panel) lie much closer to the $z_{\rm pred} = z_{\rm true}$ line with approximately equal amounts of scatter on either side, reinforcing the minimal redshift-dependent bias shown in Figure 7. The dark red area along the $z_{\rm pred} = z_{\rm true}$ line also indicates that most of the sample is localized there. The LCFIT+Z results (right panel) have a much larger spread and little localization, as evidenced by the lack of red in the plot.

4.3. SDSS Results

We evaluate our model and LCFIT+Z on a test set of 5274 simulated SDSS lightcurves and 489 real observed SDSS lightcurves with true redshift distributions shown in the right panel of Figure 2. Table 4 shows the evaluation metrics calculated with the mean and max point estimates on the simulated and observed SDSS data sets, as well as the LCFIT +Z point estimates. With the SDSS data sets, the similarity between the mean and max point estimates is more pronounced. We also note that though our model's performance degrades slightly between the simulated test set and the observed test set, all metrics still strongly favor our model as the better performer compared to LCFIT+Z. However, the $\sigma_{\rm MAD}$ values differ less between our model and LCFIT+Z compared to the PLAsTiCC test data set, only offering a 2$\times$ improvement as opposed to a 55–180$\times$ improvement for the PLAsTiCC test data set. This could be due to the narrower redshift range of SDSS, as the LCFIT+Z performance on the PLAsTiCC data set degrades significantly after $z_{\rm true} \sim 0.4$.

The mean binned residuals for our model and LCFIT+Z evaluated on the simulated and real SDSS data sets are shown in Figure 9. The residuals from both models are much more constrained for SDSS than PLAsTiCC, as expected from the less significant differences in evaluation metrics. LCFIT+Z

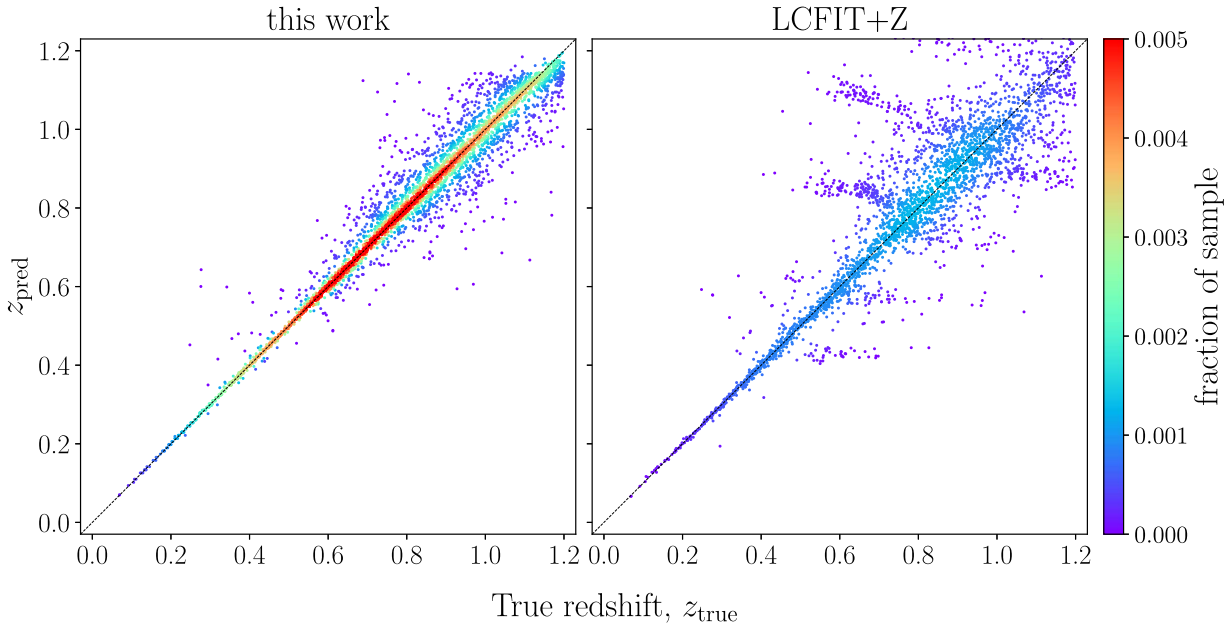

Figure 8. Predicted vs. true redshifts for the PLAsTiCC simulated SNe Ia sample colored by the fraction of each sample represented by each point. (Left) Predictions from Photo-zSNthesis described in this work, (right) predictions from LCFIT+Z.

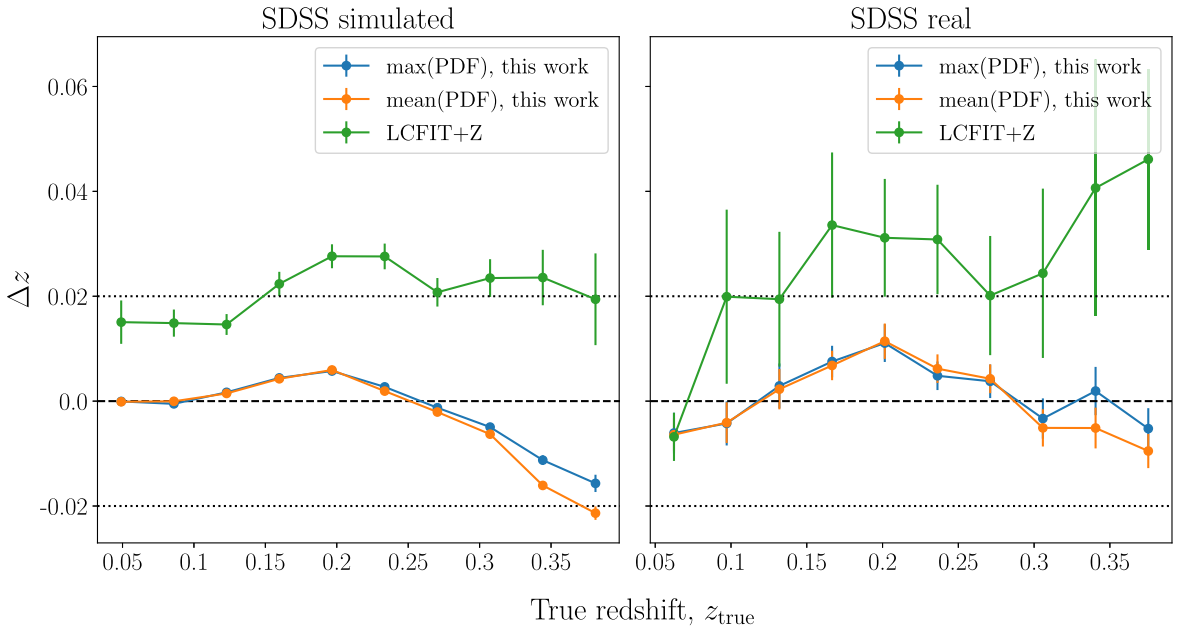

Figure 9. Mean binned residuals Δz as a function of true redshift $z_{\rm true}$ for the SDSS simulated and real SNe Ia samples. Predictions from this work have much lower biases as well as scatter compared to predictions from LCFIT+Z. We also show that two common methods of condensing redshift PDFs into point estimates, max (PDF) and mean(PDF), resulting in similar errors for our model. Dotted lines are plotted at $\Delta z = \pm 0.02$ for reference.

still exhibits larger mean Δz values for both the simulated and real data sets as well as a larger spread in each bin. Both models generalize relatively well from simulations to real data, with our mean residuals staying within $|\Delta z| < 0.02$ and LCFIT within $|\Delta z| < 0.05$; however, an overall increase in $|\Delta z|$ values is noticeable between simulated and real data.

We show the comparison between predicted and true redshifts in Figure 10 for the same data sets used in Figure 9. Our model (top row) clearly produces more constrained predictions, as they lie much closer to the $z_{\rm true} = z_{\rm pred}$ line with a high density of points (shown in red) along the line. In contrast with the PLAsTiCC results, in which LCFIT+Z performed relatively well at lower redshifts, the performance over the full SDSS redshift range is poor.

4.3.1. Bias Correction

Simulations not only allow deep learning methods such as Photo-zSNthesis to train on large data sets that would be infeasible with real data alone, but also give valuable estimates of the biases produced from those models that can be used to correct the results on real data. Here, we test this bias correction method on the results of Photo-zSNthesis on the SDSS simulated and real data. We compute the average Δz values for the simulated SDSS sample in 10 bins of $z_{\rm true}$ values and bin

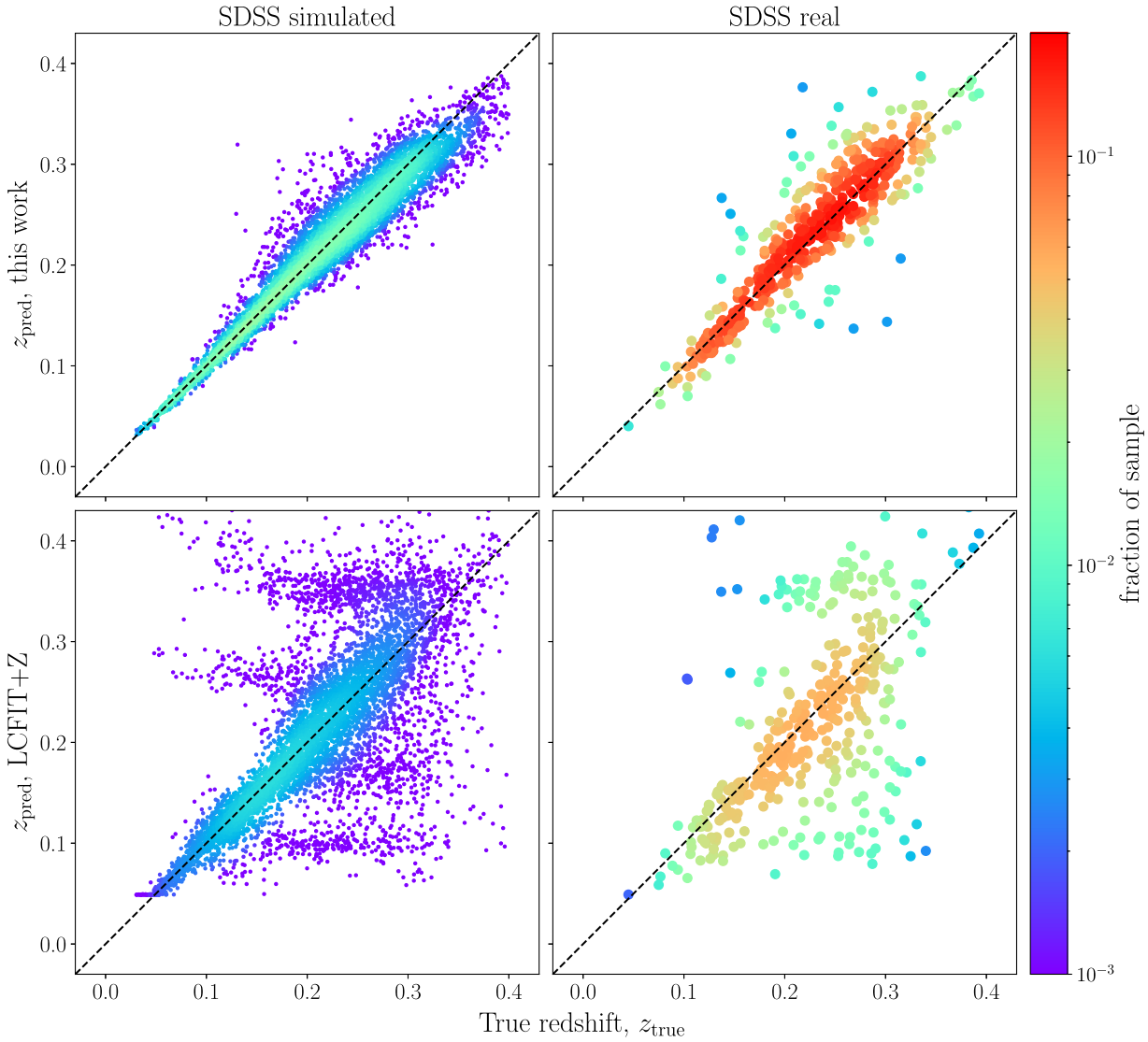

Figure 10. Predicted vs. true redshifts for the SDSS simulated and real SNe Ia samples colored by the fraction of each sample represented by each point. (Top row) Predictions from Photo-zSNthesis described in this work, (bottom row) predictions from LCFIT+Z.

the real SDSS sample in the same way. We then subtract the simulated Δz value associated with the bin of each real $z_{\rm pred}$ estimate to produce the corrected curve in Figure 11. Specifically, the corrected value of the ith SN in the real SDSS data set belonging to bin m is computed as

$$z_{{\rm corrected},i} = z_i - \langle \Delta z_{{\rm sim},m} \rangle, \tag{6}$$

where $z_{{\rm corrected},i}$ is the corrected prediction, z_i is the uncorrected prediction, and $\langle \Delta z_{{\rm sim},m} \rangle$ is the average Δz from the simulated SNe in bin m. The corrected Δz values are computed as defined in Section 4.1.1,

$$\Delta z_{\rm corrected} = \frac{z_{\rm corrected} - z_{\rm true}}{1 + z_{\rm true}}. \tag{7}$$

The corrected curve exhibits smaller biases than the uncorrected curve, showing that this bias correction method is valid in cases where the simulation is sufficiently representative of the real data.

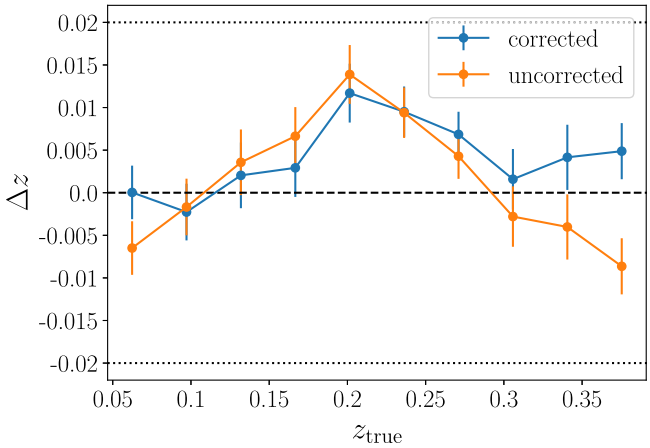

Figure 11. Predicted redshift error, Δz, as a function of true redshift, $z_{\rm true}$, for the SDSS real SNe Ia sample with and without bias correction computed from the simulated results. The point estimates used in this figure are computed using the mean(PDF) method. Dotted lines are plotted at $\Delta z = \pm 0.02$ for reference.

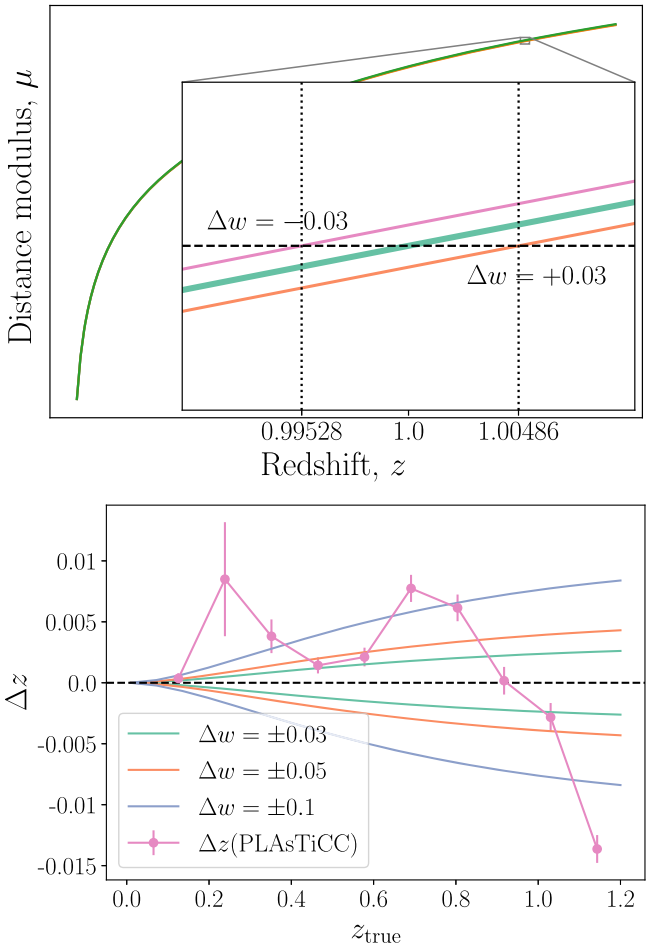

Figure 12. An illustration of our estimated cosmological biases arising from photo-z errors. (Top) A Hubble diagram zoomed-in on the neighborhood of $z_{\rm model} = 1$ showing a fiducial cosmology (green: $w = -1$, $\Omega_m = 0.3$) and two biased cosmologies (pink: $\Delta w = -0.03$, orange: $\Delta w = +0.03$). The dashed black line shows the distance modulus value at $z_{\rm model} = 1$ for the fiducial cosmology, $\mu_{\rm model}(z_{\rm model} = 1)$. The two dotted lines show the redshifts that correspond to the value of $\mu_{\rm model}(z_{\rm model} = 1)$ for the two biased cosmologies, i.e., $\mu_{\rm model}(z_{\rm model} = 1) = \mu_{\Delta w=-0.03}(z = 0.99528) = \mu_{\Delta w=+0.03}(z = 1.00486)$. The values of these redshifts are labeled on the redshift axis. We approximate the redshift error required to create a bias of, e.g., $\Delta w = +0.03$, as the difference between the biased and fiducial redshift values, $dz_{\rm bias} = 1 - 1.00486 = -0.00486$. (bottom) $\Delta z = dz/(1+z)$ values for various choices of Δw, compared to the mean binned residuals Δz produced by the mean point estimates of Photo-zSNthesis PDFs (pink, reproduced from Figure 7).

4.4. Further Experiments

4.4.1. Using Photo-zSNthesis Photo-zs for Cosmology

Though producing a full cosmological analysis using photo-zs predicted by Photo-zSNthesis is outside the scope of this work, we provide some intuition for the quality of our estimates in the context of cosmology. The redshift error $dz_{\rm bias}$ that results in a w bias of Δw at a particular redshift, $z_{\rm model}$, can be approximated by the redshift difference

$$dz_{\rm bias} = z_{\rm model} - z_{\Delta w}, \tag{8}$$

where $z_{\Delta w}$ is the redshift associated with the distance modulus $\mu_{\rm model}(z_{\rm model})$ for the biased cosmology, i.e.,

$$\mu_{\rm model}(z_{\rm model}) = \mu_{\Delta w}(z_{\Delta w}), \tag{9}$$

where $\mu_{\rm model}$, $\mu_{\Delta w}$ are the distance moduli associated with the model and biased cosmologies, respectively.

A concrete example of this is illustrated in the top panel of Figure 12, which shows a zoomed-in portion of the Hubble diagram in the neighborhood of our chosen $z_{\rm model} = 1$ for our fiducial cosmology ($w = -1$, $\Omega_m = 0.3$) as well as examples of biases on w ($\Delta w \pm 0.03$) plotted on either side. Our choice of $z_{\rm model} = 1$ is motivated by the importance of high-redshift SNe on the constraining power on w, so we compare it with our PLAsTiCC results. We choose to focus on the $\Delta w = +0.03$ cosmology as our redshifts are slightly underestimated at $z = 1$. In this example, $z_{\Delta w\,=\,+0.03} = 1.00486$ at $z_{\rm model} = 1$ is shown as the dotted vertical line on the right. We approximate $dz_{\rm bias}$ for $\Delta w = +0.03$ at $z_{\rm model} = 1$ to be

$$dz_{\rm bias} = z_{\rm model} - z_{\Delta w=+0.03} = 1 - 1.00486 = -0.00486. \tag{10}$$

This corresponds to half the distance between the dotted vertical lines in the figure. The mean redshift error for our results on the PLAsTiCC data set at $z_{\rm true} = 1$ are

$$\Delta z = \frac{z_{\rm pred} - z_{\rm true}}{1 + z_{\rm true}} = -0.002 \tag{11}$$

translating to an expected redshift difference of

$$z_{\rm pred} - z_{\rm true} = (1 + z_{\rm true}) \cdot -0.002 = -0.004. \tag{12}$$

This is below the expected $dz_{\rm bias}$ associated with a w shift of $\Delta w = +0.03$.

The bottom panel of Figure 12 shows Δz values for different choices of cosmological biases (here we choose to vary w) across the full LSST redshift range, along with the Δz values produced by Photo-zSNthesis on the PLAsTiCC data set. Note that the y-axis of this plot is $\Delta z = dz/(1+z)$, i.e., $dz_{\rm bias}$ normalized by $(1 + z_{\rm model})$, for ease of comparison with Photo-zSNthesis mean residual Δz values. This comparison over all redshifts shows that Photo-zSNthesis redshifts lack the precision to constrain cosmology to $\Delta w = \pm 0.03$, since the pink line representing Photo-zSNthesis redshift errors mostly does not lie in the area within the $\Delta w = \pm 0.03$ teal lines. However, this high-level analysis is an overestimate of the expected impact on cosmology, and a thorough cosmological analysis with Photo-zSNthesis photo-z's will be the subject of future work.

4.4.2. Comparison with LCFIT+Z with Host Galaxy Redshift Prior

In Sections 4.2 and 4.3, we showed results for Photo-zSNthesis and LCFIT+Z with no host galaxy information included. Photo-zSNthesis was formulated not to require a host redshift prior in order to prevent biases due to incorrect host redshifts. However, in a context in which host mismatches are rare and host redshifts are reliable, we want to use all available information to produce the most constrained SN photo-z estimates.

We test both Photo-zSNthesis and LCFIT+Z with a prior on $z_{\rm pred}$, the predicted redshift, given by the host galaxy photometric redshift estimate. We choose to use photometric redshifts as opposed to spectroscopic to more closely emulate a future scenario in which most host galaxies will not have spectroscopic information available.

We model this prior, $P(Z_{\rm host})$, as a Gaussian centered on the photo-z of the host galaxy, $z_{\rm host}$, and use the estimated

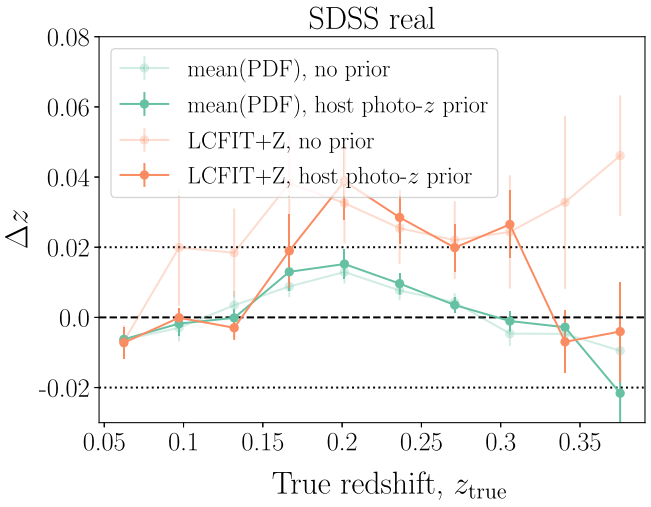

Figure 13. Mean binned residuals, $\Delta z \equiv \frac{z_{\rm pred} - z_{\rm true}}{1 + z_{\rm true}}$, as a function of true redshift, $z_{\rm true}$, for the SDSS observed photometric SNe Ia sample. Errors from both Photo-zSNthesis and LCFIT+Z are shown with (darker) and without (lighter) a prior on $z_{\rm pred}$ from the host galaxy photometric redshift. The lighter (no prior) curves are identical to those in Figure 9, with the max(PDF) curves omitted for clarity.

uncertainty, $\sigma_{z_{\rm host}}$, as the standard deviation:

$$Z_{\rm host} \sim \mathcal{N}(z_{\rm host}, \sigma_{z_{\rm host}}). \quad (13)$$

We treat Photo-zSNthesis PDFs, $P(Z_{{\rm Photo}-z{\rm SN}})$, as Bayesian posteriors and apply the host prior using Bayes' theorem:

$$P(Z_{\rm pred}) = \frac{P(Z_{{\rm Photo}-z{\rm SN}}|Z_{\rm host})P(Z_{\rm host})}{P(Z_{{\rm Photo}-z{\rm SN}})} = \frac{P(Z_{{\rm Photo}-z{\rm SN}})P(Z_{\rm host})}{\sum_i P(Z_{{\rm Photo}-z{\rm SN},i})P(Z_{{\rm host},i})} \quad (14)$$

simplified by the fact that $P(Z_{\rm host})$ and $P(Z_{{\rm Photo}-z{\rm SN}})$ are statistically independent.

LCFIT+Z uses a Markov chain Monte Carlo process to sample from the posterior distribution over a five-dimensional parameter space: four SALT fit parameters color c, stretch-luminosity parameter x_1, time of peak brightness t_0, flux normalization parameter x_0; and redshift $z_{\rm phot}$. The host prior constrains the search space, resulting in a higher likelihood of convergence to a global minimum.

In Figure 13, we show Δz values for the SDSS photometric SNe Ia data set for Photo-zSNthesis and LCFIT+Z with and without a host photo-z prior. The Δz results with no host prior are identical to those in the right panel of Figure 9. The host prior noticeably improves LCFIT+Z results on $z_{\rm true} < 0.2$ and $z_{\rm true} > 0.3$, as well as the scatter across the full redshift range, shown in the smaller error bars. The Photo-zSNthesis results with and without host prior are very similar; however, as the host prior is often less constraining than the Photo-zSNthesis PDFs themselves. We show the bias, $\sigma_{\rm MAD}$, and outlier rate for both models with and without the host prior in Table 5. These results show that Photo-zSNthesis outperforms LCFIT+Z even in the presence of a host galaxy redshift prior.

Table 5
Evaluation Metrics Computed for the SDSS Observed Photometric SNe Ia Sample Both with and without a Host Galaxy Photo-z Prior

Metric	This Work		LCFIT+Z	
	No Prior	With Prior	No Prior	With Prior
Bias $\langle \Delta z \rangle$	**0.0050**	0.0061	0.027	0.021
$\sigma_{\rm MAD}$	0.018	**0.017**	0.057	0.025
Outlier rate η	5.14%	**4.93%**	26.1%	17.6%

Note. The best results for each metric and data set are shown in bold.

5. Conclusions

In this work, we presented Photo-zSNthesis, a convolutional neural network model that predicts full redshift PDFs from multi-band photometric SNe Ia lightcurves. We evaluated its performance on simulated SDSS and LSST lightcurves, as well as a photometrically confirmed SDSS SN Ia sample. We compared our results against LCFIT+Z, the most frequently used photometric redshift estimation method for SNe, and showed superior performance across all evaluation metrics. Our model also exhibits minimal redshift-dependent bias, which has plagued redshift estimators in the past, and generalizes well between simulated and observed data.

Though Photo-zSNthesis does not require host galaxy information to produce accurate SN photo-z estimates, host galaxy redshifts can be incorporated as a prior to further improve our predictions. We tested Photo-zSNthesis and LCFIT+Z with a host galaxy photometric redshift prior to demonstrate the performance of these models in the LSST era when host galaxy spectroscopic redshifts will not be widely available. We show that while the host galaxy prior improves LCFIT+Z estimates, Photo-zSNthesis still produces more accurate photo-z point estimates as well as constrained PDFs.

We envision Photo-zSNthesis to be useful for many tasks in supernova science, including precise volumetric rate calculations, discovery of incorrect host galaxy matches via redshift discrepancies, and photometric SN Ia cosmology. We briefly explored the cosmological constraints that can be expected from Photo-zSNthesis photo-zs and concluded that the bias on w may be on the order of $\Delta w \sim 0.1$.

In future work, we intend to develop a framework for incorporating redshift PDFs and their associated uncertainties into the Hubble diagram and produce more accurately modeled cosmological constraints from Photo-zSNthesis redshift PDFs. We also plan to explore the use of Photo-zSNthesis redshift predictions as a method of identifying host confusion and potential mismatches for LSST. Though Photo-zSNthesis was developed for SNe Ia with cosmological applications in mind, the data processing and model architecture can be generalized for use with any astronomical time-domain events.

Accurate photometric redshift estimation for SNe Ia will become vital in the imminent era of photometric SN Ia cosmology. We believe that the approach and model presented here will allow us to maximize the constraining power of these new data sets.

Acknowledgments

H.Q. and M.S. were supported by DOE grant DE-FOA-0002424 and NSF grant AST-2108094. The authors would like to thank Sang Michael Xie for insightful discussions on the network architecture and training process, Brodie Popovic for providing SDSS simulations, Rebecca Chen for helpful discussions and LCFIT+Z setup, Rick Kessler for assisting

with LCFIT+Z, SNANA, and LSST simulations, and Carles Sánchez and Jaemyoung Lee for helpful discussions.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

Appendix A
Survey-agnostic Model Performance

Due to the 2D Gaussian process regression in both wavelength and time dimensions that is performed in order to create the input images (details in Section 2.2.1), data from different surveys with different photometric bands can be processed into the same image format. This creates the potential for a survey-agnostic model that can be trained on data from a single survey and applied to data from others. To test this hypothesis, we evaluated our model trained on PLAsTiCC data, described in Section 4.2, on 248 spectroscopically confirmed SNe Ia from the first 3 yr of the Dark Energy Survey supernova program (DES3YR; Brout et al. 2019). We show these results in Figure 14.

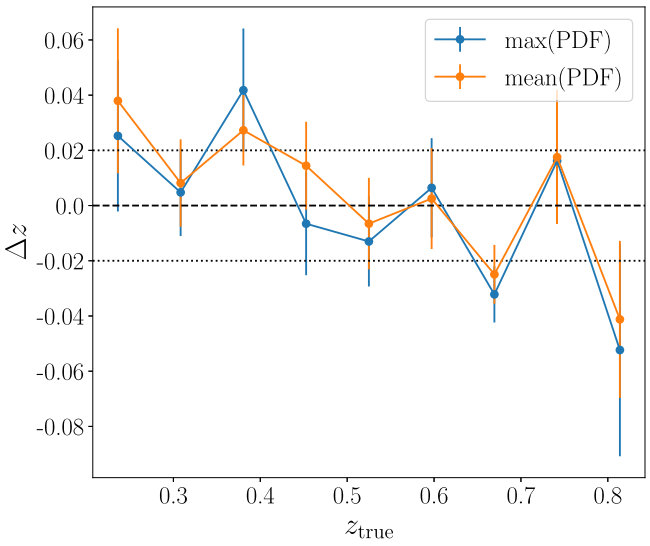

Figure 14. Mean binned residuals, $\Delta z \equiv \frac{z_{\rm pred} - z_{\rm true}}{1 + z_{\rm true}}$, as a function of true redshift, $z_{\rm true}$, for the DES3YR SNe Ia sample produced by a model trained on the PLAsTiCC data set. The max(PDF) and mean(PDF) methods of obtaining point estimates from Photo-zSNthesis PDFs are described in Section 4.1.1.

The scatter and redshift-dependent bias are certainly more prominent in this result than the others presented in this work (Figures 7 and 9), but is nevertheless an impressive result considering that this 2D Gaussian process regression method uniquely makes cross-survey results possible. We also highlight that these observed SNe Ia cover a much larger redshift range than the SDSS observed sample, demonstrating the applicability of Photo-zSNthesis on high-redshift, deep sky surveys such as LSST. While training a fresh model with a full suite of Dark Energy Survey SNe Ia simulations would certainly produce improved results, we posit that improvements could be attained relatively inexpensively by *fine-tuning* an existing trained model using a small volume of SNe Ia simulated in the target survey. We leave these interesting extensions to future work.

Appendix B
Redshift Prediction Outliers of the Real SDSS Data Set

We investigate the 29 Photo-zSNthesis prediction outliers, defined as objects with $\Delta z > 0.05$, in the observed SDSS data set. We find that the population of outliers is noticeably redder than their non-outlier counterparts at the same predicted redshift (Figure 15), which could explain the prediction failure.

We also attempted to use the predicted PDF shapes to remove outliers, assuming that poor predictions may be correlated with wider PDFs, representing the model's lack of confidence in less accurate predictions. Although we found a slight correlation, it is not strong enough for this metric to be useful: 48% of outliers have *wide* PDFs (defined as $\sigma > 0.025$ empirically) as well as 18% of non-outliers.

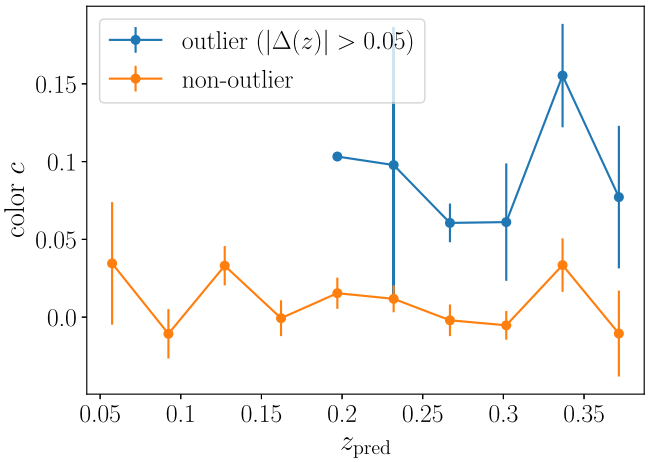

Figure 15. Mean binned SALT2 color, c, as a function of predicted redshift, $z_{\rm pred}$, for the Photo-zSNthesis prediction outliers and non-outliers in the SDSS observed photometric SN Ia sample.

ORCID iDs

Helen Qu https://orcid.org/0000-0003-1899-9791
Masao Sako https://orcid.org/0000-0003-2764-7093

References

Abbott, T. M. C., Allam, S., Andersen, P., et al. 2019, ApJL, 872, L30
Abell, P. A., Allison, J., (LSST Science Collaboration), et al. 2009, arXiv:0912.0201
Allam, A. T., Bahmanyar, A., (The PLAsTiCC Team), et al. 2018, arXiv:1810.00001
Benitez, N. 2000, ApJ, 536, 571
Betoule, M., Kessler, R., Guy, J., et al. 2014, A&A, 568, A22
Boone, K. 2019, AJ, 158, 257
Brout, D., Sako, M., Scolnic, D., et al. 2019, ApJ, 874, 106
Brout, D., Scolnic, D., Popovic, B., et al. 2022, ApJ, 938, 110
Brunner, R. J., Connolly, A. J., Szalay, A. S., & Bershady, M. A. 1997, ApJL, 482, L21
Buchs, R., Davis, C., Gruen, D., et al. 2019, MNRAS, 489, 820
Chen, R., Scolnic, D., Rozo, E., et al. 2022, ApJ, 938, 62
Dai, M., Kuhlmann, S., Wang, Y., & Kovacs, E. 2018, MNRAS, 477, 4142
de Oliveira, F. M. F., Vargas dos Santos, M., & Reis, R. R. R. 2022, MNRAS, 518, 2385
DeGroot, M. H., & Fienberg, S. E. 1983, J. R. Stat. Soc. Ser. D (The Statistician), 32, 12
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
Frieman, J. A., Bassett, B., Becker, A., et al. 2008, AJ, 135, 338
Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. 2017, in Proc. 34th Int. Conf. on Machine Learning, 70, 1321, https://proceedings.mlr.press/v70/guo17a.html
Guy, J., Astier, P., Baumont, S., et al. 2007, A&A, 466, 11
He, K., Zhang, X., Ren, S., & Sun, J. 2016, in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 770
Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv:1207.0580
Hounsell, R., Scolnic, D., Foley, R. J., et al. 2018, ApJ, 867, 23
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Kenworthy, W. D., Jones, D. O., Dai, M., et al. 2021, ApJ, 923, 265
Kessler, R., Bernstein, J. P., Cinabro, D., et al. 2009, PASP, 121, 1028
Kessler, R., Cinabro, D., Bassett, B., et al. 2010, ApJ, 717, 40
Kessler, R., Guy, J., Marriner, J., et al. 2013, ApJ, 764, 48
Kessler, R., Narayan, G., Avelino, A., et al. 2019, PASP, 131, 094501
Kim, A. G., & Miquel, R. 2007, APh, 28, 448
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2017, Commun. ACM, 60, 84
LeCun, Y., Boser, B., Denker, J., et al. 1989, in Advances in Neural Information Processing Systems, ed. D. Touretzky, Vol. 2 (San Mateo, CA: Morgan-Kaufmann)
Mitra, A., Kessler, R., More, S., Hlozek, R. & (LSST Dark Energy Science Collaboration) 2023, ApJ, 944, 212
Möller, A., & de Boissière, T. 2020, MNRAS, 491, 4277
Naeini, M. P., Cooper, G. F., & Hauskrecht, M. 2015, in Proc. Twenty-Ninth AAAI Conf. Artificial Intelligence, AAAI'15
Niculescu-Mizil, A., & Caruana, R. 2005, in Proc. 22nd Int. Conf. Machine Learning, ICML '05 (New York: ACM), 625
Palanque-Delabrouille, N., Ruhlmann-Kleider, V., Pascal, S., et al. 2010, A&A, 514, A63
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2018, A&A, 621, A26
Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565
Pierel, J. D. R., Rodney, S., Avelino, A., et al. 2018, PASP, 130, 114504
Popovic, B., Scolnic, D., & Kessler, R. 2020, ApJ, 890, 172
Qu, H., Sako, M., Möller, A., & Doux, C. 2021, AJ, 162, 67
Qu, H., Sako, M., Vincenzi, M., et al. 2023, arXiv:2307.13696
Richard, M. D., & Lippmann, R. P. 1991, Neural Comput., 3, 461
Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009
Rigault, M., Brinnel, V., Aldering, G., et al. 2020, A&A, 644, A176
Russakovsky, O., Deng, J., Su, H., et al. 2015, IJCV, 115, 211
Sako, M., Bassett, B., Becker, A. C., et al. 2018, PASP, 130, 064002
Simonyan, K., & Zisserman, A. 2014, arXiv:1409.1556
Sønderby, C. K., Espeholt, L., Heek, J., et al. 2020, arXiv:2003.12140
van den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. 2016, in Proc. 33rd Int. Conf. on Machine Learning, 48, 1747, https://proceedings.mlr.press/v48/oord16.html
Vincenzi, M., Sullivan, M., Möller, A., et al. 2023, MNRAS, 518, 1106
Wang, Y., Gjergo, E., & Kuhlmann, S. 2015, MNRAS, 451, 1955
Zeiler, M. D., & Fergus, R. 2014, in European Conf. on Computer Vision, ed. D. Fleet et al. (Berlin: Springer), 818