Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **484,** 4683–4694 (2019)
Advance Access publication 2019 February 1

doi:10.1093/mnras/stz333

Using convolutional neural networks to predict galaxy metallicity from three-colour images

John F. Wu ★ and Steven Boada

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

Accepted 2019 January 29. Received 2019 January 29; in original form 2018 October 30

ABSTRACT

We train a deep residual convolutional neural network (CNN) to predict the gas-phase metallicity (Z) of galaxies derived from spectroscopic information ($Z \equiv 12 + \log(\mathrm{O/H})$) using only three-band *gri* images from the Sloan Digital Sky Survey. When trained and tested on 128×128-pixel images, the root mean squared error (RMSE) of $Z_{\mathrm{pred}} - Z_{\mathrm{true}}$ is only 0.085 dex, vastly outperforming a trained random forest algorithm on the same data set (RMSE = 0.130 dex). The amount of scatter in $Z_{\mathrm{pred}} - Z_{\mathrm{true}}$ decreases with increasing image resolution in an intuitive manner. We are able to use CNN-predicted Z_{pred} and independently measured stellar masses to recover a mass–metallicity relation with 0.10 dex scatter. Because our predicted MZR shows no more scatter than the empirical MZR, the difference between Z_{pred} and Z_{true} cannot be due to purely random error. This suggests that the CNN has learned a representation of the gas-phase metallicity, from the optical imaging, beyond what is accessible with oxygen spectral lines.

Key words: methods: data analysis – surveys – galaxies: evolution – galaxies: general.

1 INTRODUCTION

Large-area sky surveys, both ongoing and planned, are revolutionizing our understanding of galaxy evolution. The Dark Energy Survey (DES; The Dark Energy Survey Collaboration 2005) and upcoming Large Synoptic Survey Telescope (LSST; LSST Dark Energy Science Collaboration 2012) will scan vast swaths of the sky and create samples of galaxies of unprecedented size. Spectroscopic follow-up of these samples will be instrumental in order to understand their properties. Previously, the Sloan Digital Sky Survey (SDSS; York et al. 2000) and its spectroscopic campaign enabled characterization of the mass–metallicity relation (hereafter MZR; Tremonti et al. 2004) and the fundamental metallicity relation, (hereafter FMR; e.g. Mannucci et al. 2010). As future surveys are accompanied by larger data sets, individual spectroscopic follow-up observations will become increasingly impractical.

Fortunately, the large imaging data sets to be produced are ripe for application of machine learning (ML) methods. ML is already showing promise in studies of galaxy morphology (e.g. Dieleman et al. 2015; Huertas-Company et al. 2015; Beck et al. 2018; Dai & Tong 2018; Hocking et al. 2018), gravitational lensing (e.g. Hezaveh et al. 2017; Lanusse et al. 2018; Petrillo et al. 2017, 2019), galaxy clusters (e.g. Ntampaka et al. 2015, 2017), star–galaxy separation (e.g. Kim & Brunner 2017), creating mock galaxy catalogues (e.g. Xu et al. 2013), asteroid identification (e.g. Smirnov & Markov 2017), and photometric redshift estimation (e.g. Hoyle 2016; D'Isanto & Polsterer 2018; Pasquet et al. 2019), among many others. ML methods utilizing neural networks have grown to prominence in recent years. While neural networks are a relatively old technique (e.g. LeCun et al. 1989), their recent increase in popularity is driven by the widespread availability of affordable graphics processing units (GPUs) that can be used to do general purpose, highly parallel computing. Also, unlike more 'traditional' ML methods, neural networks excel at image classification and regression problems.

Inferring spectroscopic properties from the imaging taken as part of a large-area photometric survey is, at a basic level, an image regression problem. These problems are most readily solved by use of convolutions in multiple layers of the network (see e.g. Krizhevsky, Sutskever & Hinton 2012). Convolutional neural networks (CNNs, or convnets) efficiently learn spatial relations in images whose features are about the same sizes as the convolution filters (or kernels) that are to be learned through training. CNNs are considered *deep* when the number of convolutional layers is large. Visualizing their filters reveals that increased depth permits the network to learn more and more abstract features (e.g. from Gabor filters, to geometric shapes, to faces; Zeiler & Fergus 2014).

In this work, we propose to use supervised ML by training CNNs to analyse pseudo-three-colour images and predict the gas-phase metallicity. We use predicted metallicities to recover the empirical Tremonti et al. (2004) MZR. This paper is organized as follows: In Section 2, we describe the acquisition and cleaning of the SDSS data sample. In Section 3, we discuss selection of the network's hyperparameters and outline training of the network. We present the main results in Section 4. In Section 5, we interpret

★ E-mail: jfwu@physics.rutgers.edu

the CNN's performance and discuss our findings in the context of current literature. In Section 6, we characterize the MZR using the metallicity predicted by our CNN. We summarize our key results in Section 7.

Unless otherwise noted, throughout this paper, we use a concordance cosmological model ($\Omega_\Lambda = 0.7$, $\Omega_m = 0.3$, and $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$), assume a Kroupa initial mass function (Kroupa 2001), and use *AB* magnitudes (Oke 1974).

2 DATA

To create a large training sample, we select galaxies from the SDSS (York et al. 2000) DR7 MPA/JHU spectroscopic catalogue (Kauffmann et al. 2003; Brinchmann et al. 2004; Tremonti et al. 2004; Salim et al. 2007). The catalogue provides spectroscopically derived properties such as stellar mass ($M_\star$) and gas-phase metallicity (Z) estimates (Tremonti et al. 2004). We select objects with low reduced chi-squared of model fits (`rChi2` <2), and median Z estimates available (`oh_p50`). We supplement the data from the spectroscopic catalogue with photometry in each of the five SDSS photometric bands (u, g, r, i, z), along with associated errors from SDSS DR14 (Abolfathi et al. 2018).

We require that galaxies magnitudes are $10 < ugriz < 25$ mag, in order to avoid saturated and low-signal-to-noise detections. We enforce a colour cut, $0 < u - r < 6$, in order to avoid extremely blue or extremely red objects, and require objects to have spectroscopic redshifts greater than $z = 0.02$ with low errors ($z_{err} < 0.01$). The median redshift is 0.07 and the highest redshift object has $z = 0.38$. We also require that the r-band magnitude measured inside the Petrosian radius (`petroMag_r`; Petrosian 1976) be less than 18 mag, corresponding to the spectroscopic flux limit. With these conditions we construct an initial sample of 142 182 objects (there are four objects with duplicate SDSS DR14 identifiers). We set aside 25 000 objects for later testing, and use the rest for training and validation.

We create RGB image cut-outs of each galaxy with the SDSS cut-out service,[1] which converts *gri* bands to RGB channels according to the algorithm described in Lupton et al. (2004) (with modifications by the SDSS SkyServer team). Since images are not always available, we are left with 116 429 SDSS images with metallicity measurements, including 20 466/25 000 of the test subsample. We create 128 × 128-pixel JPG images with a pixel scale of 0.″396, which corresponds to 38 × 38 arcsec on the sky. We do not further preprocess, clean, or filter the images before using them as inputs to our CNN.

3 METHODOLOGY

Before the CNN can be asked to make predictions, it must be trained to learn the relationships between the input data (the images described above) and the desired output (metallicity). The CNN makes predictions using the input images, and the error (or loss) is determined based on the differences between true and predicted values. The CNN then updates its parameters, or weights, in a way that minimizes the loss function. We use the root mean squared error (RMSE) loss function:

$$\mathrm{RMSE} \equiv \sqrt{\langle |y_{\mathrm{true}} - y_{\mathrm{pred}}|^2 \rangle}, \quad (1)$$

where y_{true} is the 'true' and y_{pred} is the predicted value, and y represents the target quantity.

It is worth emphasizing that the Z_{true} is the metallicity estimated by model fits to strong emission lines in the SDSS spectra. Tremonti et al. (2004) determine a likelihood distribution of metallicities based on the model fits, and we define their 50th percentile metallicity estimates to be the *true* metallicity (Z_{true}) for the purpose of training our network. The typical systematic uncertainty in their metallicity model fits is about 0.03 dex.

We randomly split our training sample of ∼96 953 images into 80 per cent (76 711) training and 20 per cent (19 192) validation data sets, respectively. The test data set of 20 466 images is isolated for now, and is not accessible to the CNN until all training is completed. Images and Z_{true} answers are given to the CNN in 'batches' of 256 at a time, until the full training data set has been used for training. Each full round of training using all of the data is called an epoch, and we compute the loss using the validation data set at the end of each epoch. We use gradient descent for each batch to adjust weight parameters, and each weight's fractional contribution of loss is determined by the backpropagation algorithm (LeCun et al. 1989), during which finite partial derivatives are computed and propagated backwards through layers (i.e. using the chain rule for derivatives).

We use a 34-layer residual CNN architecture (He et al. 2015) initialized to weights pretrained on the ImageNet data set, which consists of 1.7 million images belonging to 1000 categories of objects found on Earth (e.g. cats, horses, cars, or books; Russakovsky et al. 2014). The CNN is trained for a total of 10 epochs. For more details about the CNN architecture, transfer learning, hyperparameter selection, data augmentation, and the training process, see the Appendix. In total, our training process requires 25–30 min on our GPU and uses under 2 GB of memory.

We evaluate predictions using the RMSE loss function, which approaches the standard deviation for Gaussian-distributed data. We also report the NMAD, or the normal median absolute deviation (e.g. Ilbert et al. 2009; Dahlen et al. 2013; Molino et al. 2017):

$$\mathrm{NMAD}(x) \approx 1.4826 \times \mathrm{median}\left(\left|x - \mathrm{median}(x)\right|\right), \quad (2)$$

where for a Gaussian-distributed x, the NMAD will also approximate the standard deviation, σ. NMAD has the distinct advantage in that it is insensitive to outliers and can be useful for measuring scatter. However, unlike the RMSE, which quantifies the typical scatter distributed about a centre of zero, NMAD only describes the scatter around the (potentially non-zero) median.

4 RESULTS

4.1 Example predictions

In Fig. 1, we show examples of 128 × 128 pixel *gri* SDSS images that are evaluated by the CNN. Rows (a) and (b) depict the galaxies with lowest predicted and lowest true metallicities, respectively. The CNN associates blue, edge-on disc galaxies with low metallicities, and is generally accurate in its predictions. In rows (c) and (d), we show the galaxies with highest predicted and highest true metallicities, respectively. Here we find that red galaxies containing prominent nuclei are predicted to be high in metallicity, and that their predictions generally match Z_{true}.

Galaxies predicted by our CNN to have high metallicities ($Z_{\mathrm{pred}} > 9.0$) tend to be characterized by high Z_{true}, and the equivalent is true for low-metallicity galaxies. Conversely, galaxies with the highest (lowest) *true* metallicities in the sample are also predicted

[1] http://skyserver.sdss.org/dr14/en/help/docs/api.aspx

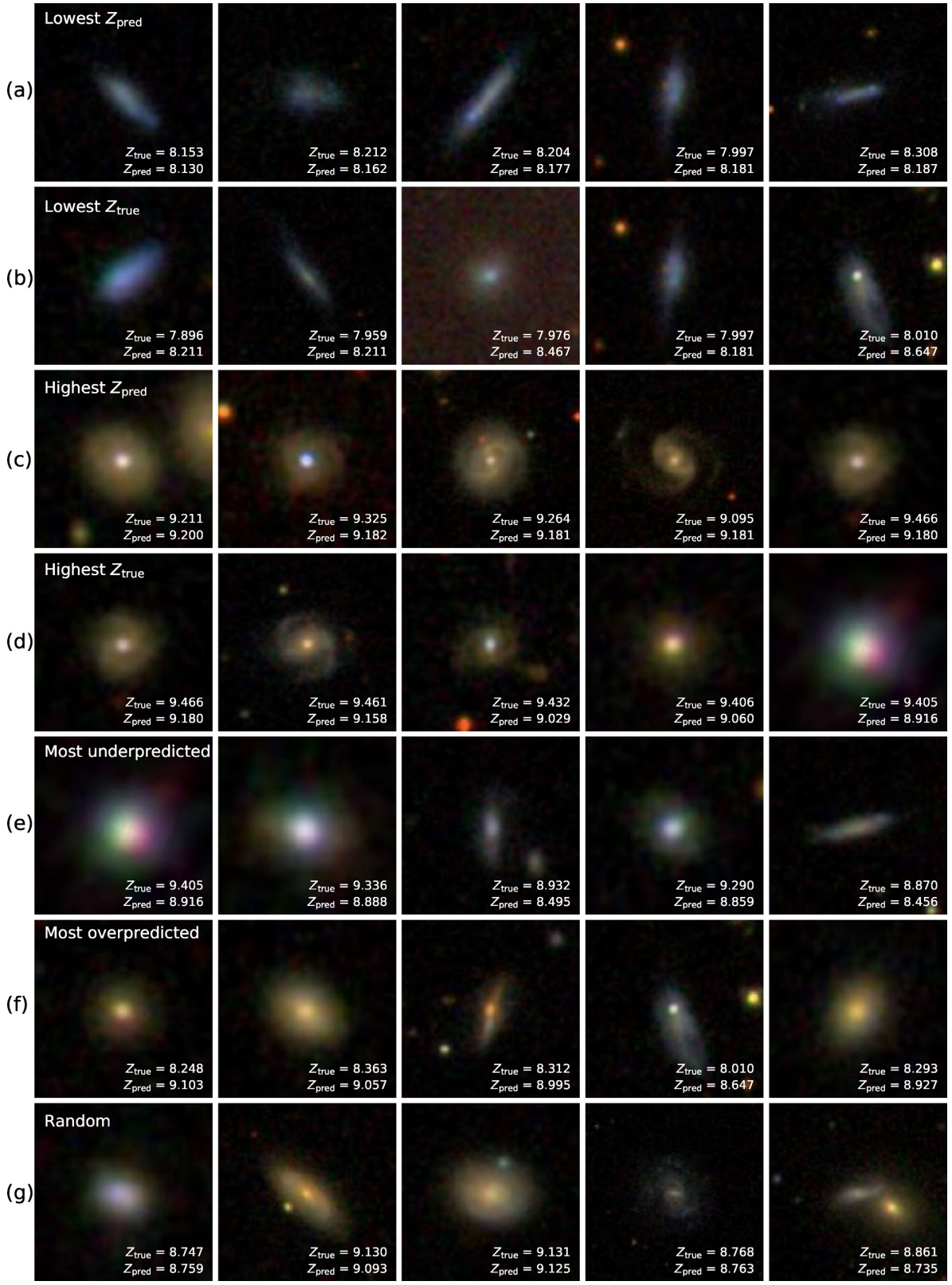

Figure 1. SDSS imaging with predicted and true metallicities from the test data set. Five examples are shown from each of the following categories: (a) lowest predicted metallicity, (b) lowest true metallicity, (c) highest predicted metallicity, (d) highest true metallicity, (e) most underpredicted metallicity, (f) most overpredicted metallicity, and (g) a set of randomly selected galaxies.

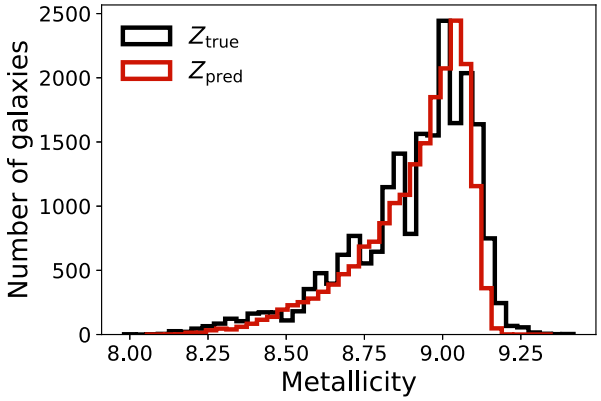

Figure 2. Distributions of the true (black) and predicted (red) galaxy metallicities. Note that the bin widths are different for the two distributions. See text for details.

to have high (low) metallicities. Note that inclined galaxies tend to be lower in metallicity whereas face-on galaxies appear to be higher in metallicity. Tremonti et al. (2004) explain this correlation by suggesting that the SDSS fibre aperture captures more column of a projected edge-on disc, allowing the metal-poor, gas-rich, and less-extincted outer regions to more easily be detected and depress the integrated $Z_{\rm true}$.

We will now consider examples of the most incorrectly predicted galaxies. In rows (e) and (f), we show instances in which the CNN predicted too low metallicity and too high metallicity, respectively. The two galaxies with the most negative residuals $\Delta Z \equiv Z_{\rm pred} - Z_{\rm true}$ (i.e. most underpredicted metallicities) suffer from artefacts that cause unphysical colour gradients, and/or are labelled as quasars on the basis of their SDSS spectra (for which we expect $Z_{\rm true}$ to be biased). It is not unsurprising that the CNN has made mistakes in some of these cases, since they go against astronomers' usual heuristics: blue, disc-dominated sources are generally thought of as lower in metallicity, and redder, more spheroidal objects tend to be higher in metallicity.

In the bottom row (g) of Fig. 1, we show five randomly selected galaxies. The random SDSS assortment consists of lenticular, spiral, and possibly even an interacting pair of galaxies. Residuals are low (below 0.15 dex), and we again find that the CNN predictions track with human visual intuition.

4.2 Comparing predicted and true metallicities

In Fig. 2, we show histograms of the true and predicted metallicities in black and red, respectively. The histogram bin sizes are chosen according to the Freedman & Diaconis (1981) rule for each distribution. The discreet striping of the Tremonti et al. (2004) and Brinchmann et al. (2004) metallicity estimator appears in the $Z_{\rm true}$ distribution but does not appear in our CNN predictions. This striping should increase the scatter in our distribution of residuals.

The range of $Z_{\rm pred}$ is more limited than the range of $Z_{\rm true}$, which can also be seen from Fig. 1 for extreme values of $Z_{\rm true}$. Too narrow a domain in $Z_{\rm pred}$ will lead to systematic errors, as the CNN will end up never predicting very high or very low metallicities. Although the two distributions are qualitatively consistent with each other at low metallicities (e.g. $Z < 8.5$), the fraction of galaxies with high $Z_{\rm true} > 9.1$ ($2573/20466 = 12.6$ per cent) is higher than the fraction with high $Z_{\rm pred} > 9.1$ ($1174/20466 = 5.7$ per cent).

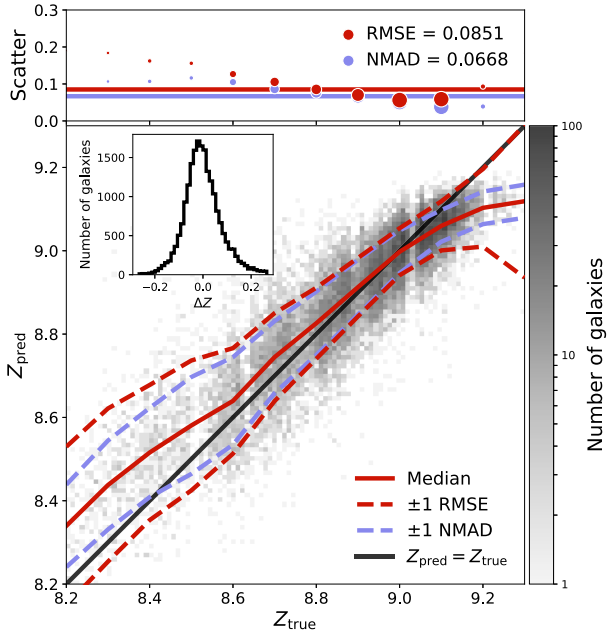

Figure 3. Bivariate distribution of true galaxy metallicity ($Z_{\rm true}$) and CNN prediction ($Z_{\rm pred}$) is shown in the main panel. Overlaid are the median predicted metallicity (solid red line), RMSE scatter (dashed red lines), and NMAD scatter (dashed violet lines), in bins of $Z_{\rm true}$. The solid black line shows the one-to-one relation. The distribution of residuals ($Z_{\rm pred} - Z_{\rm true}$) is shown in the inset plot. In the upper panel, we again show the binned scatter, where the size of each marker is proportional to the number of galaxies in that bin. Each horizontal line corresponds to the average scatter over the entire test data set (and the global value indicated in the upper panel legend).

We find that the mode of the binned predicted metallicity distribution is higher than that of $Z_{\rm true}$. This result may be a consequence of the CNN overcompensating for its systematic underprediction of metallicity for galaxies with $Z_{\rm true} > 9.1$. However, its effect on the entire distribution is small, and may be remedied simply by increasing the relative fraction of very high $Z_{\rm true}$ objects. We find overall good qualitative agreement between the $Z_{\rm pred}$ and $Z_{\rm true}$ distributions.

4.3 Scatter in $Z_{\rm pred}$ and $Z_{\rm true}$

In Fig. 3, we compare the distributions of $Z_{\rm true}$ and $Z_{\rm pred}$ using a two-dimensional histogram (shown in grey-scale in the main, larger panel). We also show the median predictions varying with binned $Z_{\rm true}$ (solid red line), in addition to the scatter in RMSE (dashed red) and NMAD (dashed violet), and also the one-to-one line (solid black). The running median agrees well with the one-to-one line, although at low metallicity we find that the CNN makes overpredictions.

A histogram of metallicity residuals is shown in the inset plot of the Fig. 3 main panel. The ΔZ distribution is characterized by an approximately normal distribution with a heavy tail at large positive residuals; this heavy tail is likely due to the systematic overprediction for low-$Z_{\rm true}$ galaxies. There is also an overabundance of large negative ΔZ corresponding to underpredictions for high $Z_{\rm true}$, although this effect is smaller. We do not find significant correlations between ΔZ and galaxy observables including spectroscopic redshift, any combination of photometric colour (including

u and z bands), emission line signal-to-noise ratios, observed *gri* magnitudes, or axis ratios.

We now turn our attention to the upper panel of Fig. 3, which shows how the scatter varies with spectroscopically derived metallicity. The RMSE scatter and outlier-insensitive NMAD are both shown. Marker sizes are proportional in area to the number of samples in each $Z_{\rm true}$ bin, and the horizontal lines are located at the average loss (RMSE or NMAD) for the full test data set.

Predictions appear to be both accurate and low in scatter for galaxies with $Z_{\rm true} \approx 9.0$, which is representative of a typical metallicity in the SDSS sample. Where the predictions are systematically incorrect, we find that the RMSE increases dramatically. However, the same is not true for the NMAD; at $Z_{\rm true} < 8.5$, it asymptotes to $\sim$0.10 dex, even though the running median is incorrect by approximately the same amount. This discrepancy is because the NMAD determines the scatter about the *median* and not $\Delta Z = 0$, and thus, this metric becomes somewhat unreliable when the binned samples do not have a median value close to zero. Fortunately, the global median of ΔZ is -0.006 dex, or less than 10 per cent of the RMSE, and thus the global NMAD = 0.067 dex is representative of the outlier-insensitive scatter for the entire test data set.

This effect partly explains why the global NMAD (0.067 dex) is higher than the weighted average of the binned NMAD ($\sim$0.05 dex). Also, each binned NMAD is computed using its local scatter, such that the outlier rejection criterion varies with $Z_{\rm true}$. To illustrate this effect with an example: $\Delta Z \approx 0.2$ dex would be treated as a $3\,\sigma$ outlier at $Z_{\rm true} = 9.0$, where the CNN is generally accurate, but the same residual would not be rejected as an outlier using NMAD for $Z_{\rm true} = 8.5$. Since the binned average NMAD depends on choice of bin size, we do not include those results in our analysis and only focus on the global NMAD. RMSE is a robust measure of both local and global scatter (although it becomes biased high by outliers).

5 INTERPRETING THE CNN

5.1 Uncertainty in the scatter

It is worth examining how reliable our estimate of RMSE = 0.085 dex is. Because we have a large data set, we can calculate uncertainties on the RMSE through multiple training/test realizations. One possible method is by dividing our full data set into cross-validation and nested cross-validation splits in order to see how the RMSE varies.

For the first method (five-fold cross-validation), we take the entire data set from Section 2 and split it into five 80 per cent/20 per cent training/test subsets, each of which is optimized independently. We then compute the mean and standard deviation of the five test samples' $Z_{\rm pred}$, and find that the RMSE = 0.0836 ± 0.0005. Because there are more training examples here than in our original training set, the mean RMSE is lower than what we have previously found in Section 4.

For the second method (nested cross-validation), we split the full data set into five 80 per cent/20 per cent training/validation test splits, and then further divide the training/validation data sets into 75 per cent/25 per cent cross-validation splits. We compute the mean and standard deviation of $Z_{\rm pred}$ for the ensemble of five-fold test splits. When we select the model with the best cross-validation score, the RMSE is 0.0823 ± 0.0009. The unweighted average of all training/validation models is RMSE = 0.0831 ± 0.0011.

There is an additional source of scatter due to noise in the SDSS images' pixels. This noise is not uniform across SDSS images, and the Lupton et al. (2004) intensity scaling makes estimating or resampling the noise distribution challenging. Therefore, we do not account for the contribution of image noise to our estimate of the uncertainties.

5.2 Impact of artificially increasing scatter

In order to simulate additional uncertainty that may arise from noisier measurements of spectral lines, we add normally distributed scatter to the $Z_{\rm true}$ values. We train our CNN as before (in Section 4), except that we add random $\sigma = \{0.03, 0.05, 0.10, 0.20\}$ dex of scatter to the target $Z_{\rm true}$. The smallest value, 0.03 dex, is the same as the systematic uncertainty in the Tremonti et al. (2004) measurements, and 0.20 dex represents the standard deviation for the entire $Z_{\rm true}$ distribution. We compare CNN-predicted $Z_{\rm pred}$ with the original $Z_{\rm true}$ (i.e. the underlying values without artificial scatter introduced) using the RMSE metric as before (equation 1). For all values of additional scatter, the CNN is able to estimate $Z_{\rm pred}$ to RMSE = $\{0.0851, 0.0851, 0.0869, 0.0882\}$ dex respectively. These results show that the CNN is robust to extra scatter added in an unbiased way.

As a second test, we include random scatter drawn from a normal distribution centred at zero and with standard deviation equal to the galaxy's redshift. This simple model can test how the CNN responds to redshift-dependent scatter in $Z_{\rm true}$. We note that this toy model disproportionately impacts higher metallicity galaxies because our sample of lower mass (and thus, lower metallicity) galaxies is less complete at higher redshifts. After training on the original images with these modified $Z_{\rm true}$ values, we find that the resulting $Z_{\rm pred}$ are not strongly affected: RMSE = 0.0862 dex.

These tests demonstrate that our CNN is robust to normally distributed scatter in $Z_{\rm true}$. Our results show initial promise for cases in which training data are more uncertain, such as at higher redshift. However, more testing is necessary to understand the effects of biased or correlated sources of uncertainty, and to account for the evolving relationships between metallicity and other observed properties (e.g. Zahid et al. 2013; Salim et al. 2015).

5.3 Resolution and colour effects

Because our methodology is so computationally light, we can run the same CNN training and test procedure on images scaled to different sizes in order to understand the effects of image resolution. Our initial results use SDSS 38 × 38 arcsec cut-outs resized to 128 × 128 pixels, and we now downsample the same images to 64 × 64, 32 × 32, , 2 × 2, and even 1 × 1 pixels via rebinning. All images retain their three channels, so the smallest 1 × 1 image is effectively the pixels in each of the *gri* bands averaged together with the background and possible neighbouring sources.

In Fig. 4, we show the effects of image resolution by measuring the global scatter in ΔZ using the RMSE and NMAD metrics (shown in red and violet circular markers, respectively). Also shown is the scatter in ΔZ if we always predict the mean value of $Z_{\rm true}$ over the data set (shown using a square marker). This constant prediction effectively delivers the worst possible scatter, and the Tremonti et al. (2004) systematic uncertainty in $Z_{\rm true}$ of $\sim$0.03 dex yields the best possible scatter. We find that both RMSE and NMAD decrease with increasing resolution, as expected if morphology or colour gradients are instrumental to predicting metallicity.

There appears to be little improvement in scatter going from 1 × 1 to 2 × 2 pixel images. 1 × 1 three-colour images contain similar information to three photometric data points (although because background and neighbouring pixels are averaged in, they are

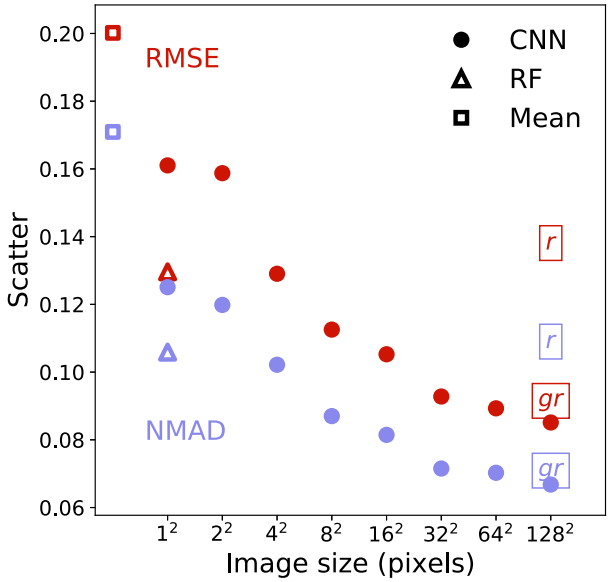

Figure 4. The effects of image resolution and colour on CNN performance. Red and violet circular markers indicate scatter in the residual distribution (ΔZ) measured using RMSE and NMAD, respectively, for *gri* imaging. (Each point is analogous to the horizontal lines shown in Fig. 3.) We also show predictions from a random forest algorithm as open triangle markers, and constant $\langle Z_{\rm true} \rangle$ predictions as open square markers. Large, labelled squares indicate the scatter for images consisting of only *r*-band imaging and a combination of *g* and *r* bands.

less information dense than photometry), which can be used to perform a crude spectral energy distribution (SED) fit. Therefore it is unsurprising that the 1 × 1 CNN predictions perform so much better than the baseline mean prediction. A 2 × 2 three-colour image contains four times as many pixels as a 1 × 1 image, but because the object is centred between all four pixels, information is still averaged among all available pixels. Therefore, the scatter does not improve appreciably going from 1 × 1 to 2 × 2 resolution.[2]

The scatter is a strong function of resolution as the images are resolved from 2 × 2 to about 32 × 32 pixels. With further increasing resolution, improvement is still evident, although the scaling with scatter is noticeably weaker. Because the angular size of each image cut-out stays the same, the pixel scale changes from 1.″184 pixel^{-1} for 32 × 32 images, to 0.″592 pixel^{-1} for 64 × 64 images, to 0.″296 pixel^{-1} for 128 × 128 images. The native SDSS pixel resolution is 0.″396 pixel^{-1}, such that the 64 × 64 and 128 × 128 resolutions result in the oversampling of each image. Thus, scatter is expected to plateau for images larger than 128 × 128. It is worth noting, however, that the CNN attempts to learn filters that depend on the size of the input image, so smaller images may result in the CNN training filters that are too low in resolution to be completely effective for prediction. Therefore, it is also not surprising that the CNN makes incremental gains for images with increasing resolution beyond 64 × 64 pixels.

We also train the CNN to predict metallicity using only the central 16 × 16-pixel regions of each SDSS *gri* image. We find that the network is able to predict metallicity to within RMSE = 0.0965 dex. Because this value is higher than the scatter found when using the full-sized images, we conclude that the CNN loses valuable information when only the central regions are considered, and that relevant information for predicting global metallicity can be found in the galaxies' outer regions that are not probed by the 3 arcsec SDSS spectroscopic fibres.

As a way of testing how the CNN responds to reduced colour information, we have also repeated our training and testing routines using *r*-band and *gr*-band 128 × 128 images. In order to make use of our pretrained network, we modify the original, three-colour JPG images to correspond to either one- or two-band SDSS images. For the single-band imaging, we duplicate the *r*-band data into the blue and red JPG channels. For the *gr*-band images, the blue and red channels correspond to the *g* and *r* filters, while the green channel is the mean of the two.

The large squares labelled '*r*' in Fig. 4 show that the network trained and tested on single-band images performs relatively poorly (RMSE = 0.1381 dex) compared to *gri* imaging even at low resolution. The addition of a second colour improves CNN performance significantly. When a second band is added to the training images (box labelled '*gr*' in Fig. 4), the RMSE improves to a level similar (0.0915 dex) to that of the original three-colour images (0.0851 dex). This enhancement may be due to extra information that the bluer *g* band provides about younger stellar populations. In both examples, the CNN is able to utilize spatial information about a galaxy to improve metallicity estimates.

5.4 Random forest predictions for metallicity

We also construct a random forest (RF) of decision trees in order to predict metallicity using the implementation from SCIKIT-LEARN (Pedregosa et al. 2012). Hyperparameters are selected according to the optimal RF trained by Acquaviva (2016). We use exactly the same data labels (i.e. galaxies) to train/validate or test the RF that we have used for training and testing the CNN, so that our measurements of scatter can be directly compared. However, we have used the *gri* three-band photometry data (given in magnitudes) to train and predict metallicity. Since each galaxy only has three pieces of photometric information, it can be compared to the 1 × 1 three-band 'images' processed by our CNN.

The RF predicts metallicity with RMSE = 0.130 dex, which is superior to our CNN trained and tested on 1 × 1 and 2 × 2 images. This result is unsurprising because the RF is supplied aperture-corrected photometry, whereas the CNN is provided 1 × 1 *gri* 'images' whose features have been averaged with their backgrounds. 2 × 2 images are only marginally more informative. When the resolution is further increased to 4 × 4 images, then the CNN can begin to learn rough morphological features and colour gradients, which is already enough to surpass the performance (measured by both RMSE and NMAD) of the RF. This result suggests that the CNN is able to learn a non-trivial representation of gas-phase metallicity based on three-band brightness distributions, even with extremely low-quality data.

5.5 Comparisons to previous work

CNNs have been used for a wide variety of classification tasks in extragalactic astronomy, including morphological classification (e.g. Dieleman et al. 2015; Huertas-Company et al. 2015; Simmons et al. 2017), distinguishing between compact and extended objects (Kim & Brunner 2017), selecting observational samples of

[2]There is extra information in the 2 × 2 pixel images in non-circularly symmetric cases. For an inclined disc, it is possible to roughly determine the orientation in the sky plane, but this information is not very useful. In the case of a major merger or interacting companion, the 2 × 2 images may be more powerful than 1 × 1 images.

rare objects based on simulations (Huertas-Company et al. 2018; Lanusse et al. 2018), and visualizing high-level morphological galaxy features (Dai & Tong 2018). These works seek to improve classification of objects into a discreet number of classes, i.e. visual morphologies. Our paper uses CNNs to tackle the different problem of regression, i.e. predict values from a continuous distribution.

Examples of regressing stellar properties in the astronomical ML literature (e.g. Bailer-Jones 2000; Fabbro et al. 2018) train on synthetic stellar spectra and test on real data. Their predicted measurements of stellar properties, e.g. stellar effective temperature, surface gravity, or elemental abundance, can be derived from the available training data set. Our work is novel because we predict metallicity, a spectroscopically determined galaxy property, using only three-colour images. Said another way, it is not necessarily the case that Z can be predicted from our training data. However, we find that galaxy shape supplements colour information in a way that is useful for predicting metallicity.

A study similar to this work is that of Acquaviva (2016), who uses a variety of ML methods including RFs, extremely random trees (ERTs), boosted decision trees (AdaBoost), and support vector machines (SVMs) in order to estimate galaxy metallicity. The Acquaviva (2016) data set consisted of a $z \sim 0.1$ sample (with $\sim$25 000 objects) and a $z \sim 0.2$ sample (with $\sim$3 000 objects), each of which has five-band SDSS photometry (*ugriz*) available as inputs. These samples are sparsely populated at low metallicities, and they contain smaller fractions of objects with $Z_{\rm true} < 8.5$ than our sample, but are otherwise similarly distributed in $Z_{\rm true}$ to ours. Our samples have different sizes because we require SDSS objects to have imaging available, whereas the Acquaviva (2016) criteria impose stronger spectroscopic redshift constraints.

We will first compare RF results, since this technique is common to both of our analyses, and they reveal important differences in our training data. Because outliers are defined differently in both works, we will use the RMSE metric to compare scatter between the two. Acquaviva (2016) obtain RMSE of 0.081 and 0.093 dex when using RFs on the five-band photometry for the $z \sim 0.1$ and 0.2 subsamples. Using exactly the same RF approach on a larger sample, while working with only *three* bands of photometric information, we find RMSE = 0.130 dex. Our scatter is larger than the value reported by Acquaviva (2016) by a factor of $\sim$1.5. This result may partly be explained by the fact that Acquaviva (2016) $Z_{\rm true}$ distribution is narrower than for our training data set, or the fact that our data set spans a broader range in galaxy redshift; however, some of this advantage is offset by our larger sample size. Ultimately, it appears that the extra *u* and *z* bands supply ML algorithms with valuable information for predicting metallicity.

Indeed, the *u* and *z* bands convey information about a galaxy's star formation rate (SFR) and stellar mass (see e.g. Hopkins et al. 2003). For this reason, it is possible that the RF trained on five-band photometry can estimate $Z_{\rm true}$ down to the limit of the FMR, which has very small scatter ($\sim$0.05 dex) at fixed $M_\star$ *and* SFR. The *g*, *r*, and *i* bands are less sensitive to the SFR, but can still provide some information about the stellar mass, and so our RF and CNN results are more linked to the MZR rather than the FMR.

Regardless of these limitations, our CNN is able to estimate metallicity with $\Delta Z = 0.085$ dex, which is comparable to the scatter in residuals using the best algorithms from Acquaviva (2016). There is evidence that the morphological information provided by using images rather than photometric data is helping the CNN perform so well: (1) the RMSE scatter decreases with increasing image resolution, and (2) it identifies edge-on galaxies as lower $Z_{\rm pred}$ and face-on galaxies as higher $Z_{\rm pred}$ (consistent with observational bias). Gradients in colour, or identification of mergers (e.g. Ackermann et al. 2018) may also be helpful for predicting metallicity.

6 THE MASS–METALLICITY RELATION

The MZR describes the tight correlation between galaxy stellar mass and nebular gas-phase metallicity. Scatter in this correlation is approximately $\sigma \approx 0.10$ dex in $Z_{\rm true}$ over the stellar mass range $8.5 < \log(\mathrm{M}_\star/\mathrm{M}_\odot) < 11.5$ (Tremonti et al. 2004), where σ is the standard deviation of the metallicity and is equivalent to the RMSE for a normal distribution. The MZR at $z = 0$ can be characterized empirically using a polynomial fit:

$$Z = -1.492 + 1.847 \log(M_\star/M_\odot) - 0.08026 \left[\log(M_\star/M_\odot)\right]^2 . \quad (3)$$

The physical interpretation of the MZR is that a galaxy's stellar mass strongly correlates with its chemical enrichment. Proposed explanations of this relationship's origin include metal loss through blowout (see e.g. Garnett 2002; Tremonti et al. 2004; Brooks et al. 2007; Davé, Finlator & Oppenheimer 2012), inflow of pristine gas (Dalcanton, Yoachim & Bernstein 2004), or a combination of the two (e.g. Lilly et al. 2013); however, see also Sánchez et al. (2013). Although the exact physical process responsible for the low (0.10 dex) scatter in the MZR is not known, its link to SFR via the FMR is clear, as star formation leads to both metal enrichment of the interstellar medium and stellar mass assembly.

The FMR connects the instantaneous ($\sim$10 Myr) SFR with the gas-phase metallicity ($\sim$1 Gyr time-scales; see e.g. Leitner & Kravtsov 2011) and $M_\star$ (i.e. the $\sim$13 Gyr integrated SFR). Our CNN is better suited for predicting $M_\star$ rather than SFR, using the *gri* bands, which can only weakly probe the blue light from young, massive stars. Therefore, we expect the scatter in CNN predictions to be limited by the MZR (with scatter $\sigma \sim 0.10$ dex) rather than the FMR ($\sigma \sim 0.05$ dex). It is possible that galaxy colour and morphology, in tandem with CNN-predicted stellar mass, can be used to roughly estimate the SFR, but in this paper we will focus on only the MZR.

6.1 Predicting stellar mass

Since galaxy stellar mass is known to strongly correlate with metallicity, and is easier to predict (than, e.g. SFR) from *gri* imaging, we consider the possibility that the CNN is simply predicting stellar mass ($M_{\star,\rm pred}$) accurately and then learning the simple polynomial transformation in order to estimate metallicity. We can simulate this scenario by training the CNN on $M_{\star,\rm true}$ and then converting the stellar mass predictions to metallicities using equation (3).

We re-run the CNN methodology to train and predict $M_\star$ using the 116 394 available images (out of the 142 145/142 186 original objects that have stellar mass measurements). These results are shown in the left-hand panel of Fig. 5. From the same subsample as before (minus three objects that do not have $M_\star$ estimates), we verify that $M_{\star,\rm true}$ median agrees with the median of $M_{\star,\rm true}$ for values between $9.0 \lesssim \log \mathrm{M}_\star/\mathrm{M}_\odot \lesssim 10.5$. The RMSE scatter in the $M_\star$ residuals is $\sim$0.22 dex, and the NMAD is $\sim$0.20 dex. The slope of the empirical MZR at $\log(\mathrm{M}_\star/\mathrm{M}_\odot) \sim 10$ is (0.4 dex in *Z*)/(1.0 dex in $M_\star$), implying that the CNN might be able to leverage the MZR and predict metallicity to $\sim$0.08 dex (plus any intrinsic scatter in the MZR, in quadrature).

We use equation (3) and $M_{\star,\rm pred}$ to predict metallicity, which we call $Z_{\rm MZR}$. In the right-hand panel of Fig. 5, we compare $Z_{\rm MZR}$

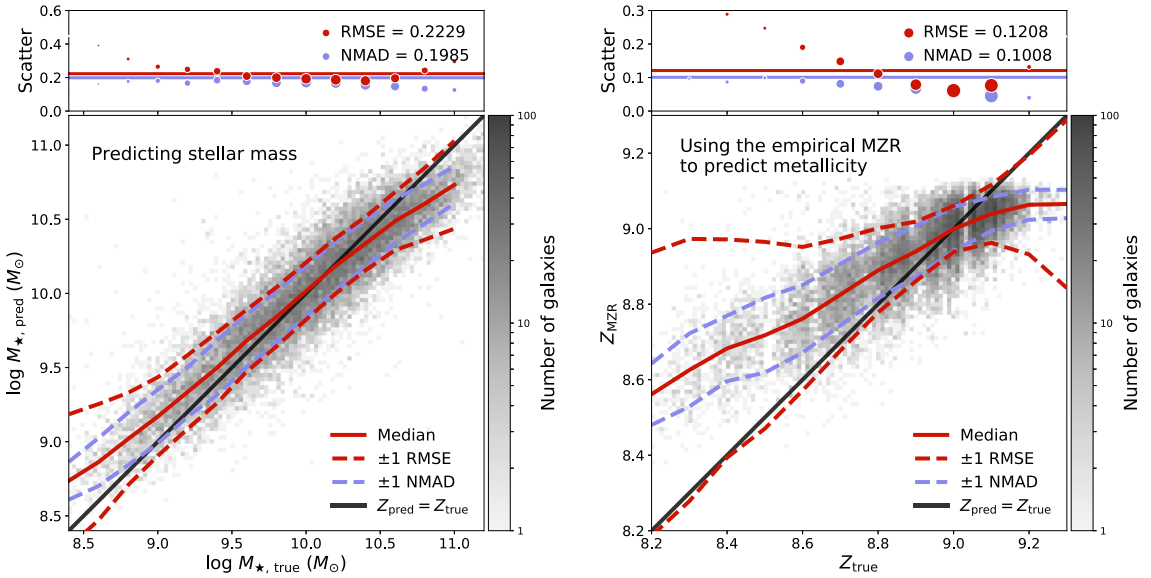

Figure 5. In the left-hand panel, we plot the CNN predicted galaxy stellar mass against true stellar mass. Colours and marker or line styles are the same as in Fig. 3. In the right-hand panel, we compare the predicted stellar mass converted to metallicity, assuming the Tremonti et al. (2004) MZR, against the true metallicity. These findings indicate that using the empirical MZR and CNN-predicted $M_{\star,\mathrm{pred}}$ yields poor results, unlike what we have observed in Fig. 3.

against Z_{true}. The scatter in residuals $Z_{\mathrm{MZR}} - Z_{\mathrm{true}}$ is 0.12 dex, which is significantly higher than the 0.085 dex scatter reported in Section 4. If the MZR alone were mediating the CNN's ability to estimate from *gri* imaging, then we would expect the scatter for Z_{pred} to be greater than for Z_{MZR}; instead we find that the opposite is true. This evidence suggests that the CNN has learned to determine metallicity in a more powerful way than by simply predicting $M_{\star,\mathrm{pred}}$ and then effectively applying a polynomial conversion.

6.2 An unexpectedly strong CNN-predicted mass–metallicity relation

The RMSE = 0.085 dex difference between the true and CNN-predicted metallicities can be interpreted in one of two ways: (1) the CNN is inaccurate, and Z_{pred} deviates randomly from Z_{true}, or (2) the CNN is labelling Z_{pred} according to some other hidden variable, and ΔZ residuals represent non-random shifts in predictions based on this variable. If the first scenario were true, we would expect the random residuals to increase the scatter of other known correlations such as the MZR when predicted by the CNN. If the second were true, we would expect the scatter of such correlations to remain unchanged or shrink. We can therefore contrast the MZR constructed from Z_{pred} and from Z_{true} in order to test these interpretations.

In the main panel of Fig. 6, we plot CNN-predicted metallicity versus true stellar mass. For comparison, we also overlay the Tremonti et al. (2004) MZR median relation and its $\pm 1\sigma$ scatter (which is ∼0.10 dex). Their empirical median relation (solid black) matches our predicted MZR median (solid red), and the lines marking observed scatter (dashed black) match Z_{pred} scatter as well (dashed red and violet). Over the range $9.5 \leq \log(M_{\star,\mathrm{true}}/\mathrm{M}_\odot) \leq 10.5$, the RMSE scatter in Z_{pred} appears to be even tighter than the observed $\pm 1\,\sigma$ (dashed black). The same is true for the NMAD, which is even lower over the same interval.

In the upper panel of Fig. 6, we present the scatter in both predicted and Tremonti et al. (2004) MZR binned by mass. We confirm that the CNN predicts an MZR that is at most equal

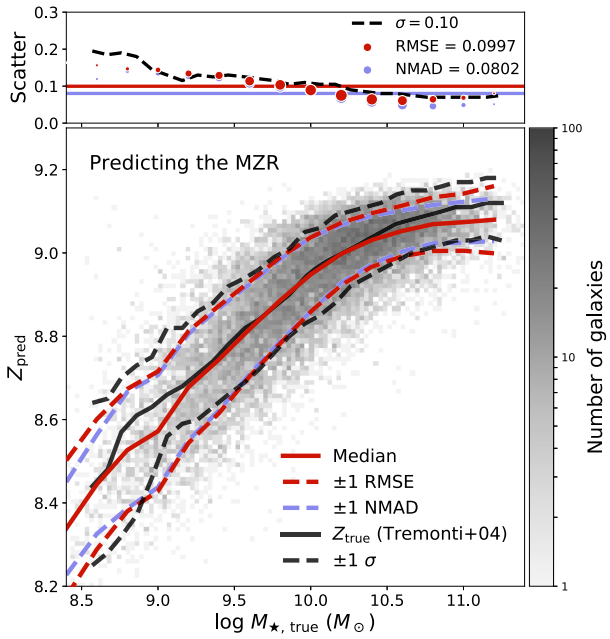

Figure 6. In the main panel, the predicted MZR comparing true $M_\star$ against CNN-predicted Z_{pred} is shown in grey-scale. The running median (solid red) and scatter (dashed red and violet) are shown in 0.2 dex mass bins. For comparison, we also show the Tremonti et al. (2004) observed median and scatter (solid and dashed black lines, respectively), which are binned by 0.1 dex in mass. In the upper panel, we show the scatter in the predicted and empirical MZR. The standard deviation of the scatter for the empirical MZR is shown as a dashed black line, while the red and violet circles respectively show RMSE and NMAD for the predicted MZR. Marker sizes are proportional to the number of galaxies in each stellar mass bin for the test data set. Global scatter in the CNN-predicted MZR appears to be comparable to, or even lower than, scatter in the empirical MZR.

in scatter than one constructed using the true metallicity. The stellar mass bins for which our CNN found tighter scatter than the empirical MZR are the same bins that happen to contain the greatest number of examples ($9.5 \leq \log(M_{\star,\rm true}/M_{\odot}) \leq 10.5$); thus, the strong performance of our network at those masses may be due to a wealth of training examples. If our data set were augmented to include additional low- and high-$M_{\star,\rm true}$ galaxies, then the scatter in the predicted MZR could be even lower overall.

The fact that a CNN trained on only *gri* imaging is able to predict metallicity accurately enough to reproduce the MZR in terms of median and scatter is not trivial. The error budget is very small: $\sigma = 0.10$ dex affords only, e.g. 0.05 dex of scatter when SFR is a controlled parameter plus a 0.03 dex systematic scatter in $Z_{\rm true}$ measurements, leaving only $\sim$0.08 dex remaining for CNN systematics, assuming that these errors are not correlated and are added in quadrature. This remaining error budget may barely be able to accommodate our result of RMSE(ΔZ) = 0.085. Interpreting the MZR scatter as the combination of intrinsic FMR scatter, $Z_{\rm true}$ systematics, and ΔZ systematics cannot be correct since it assumes that the CNN is recovering the FMR perfectly. As we have discussed previously, it is highly unlikely that the CNN is sensitive to the SFR, and therefore cannot probe the MZR at individual values of the SFR.

If we assume that the error budget for the MZR is not determined by the FMR, then the error 'floor' should be 0.10 dex. This is immediately exceeded, as we have found RMSE $\approx$ 0.10 dex for the predicted MZR without accounting for the fact that $Z_{\rm pred}$ and $Z_{\rm true}$ differ by RMSE = 0.085 dex. Consider the case in which all $Z_{\rm true}$ values are shifted randomly by a Gaussian noise distribution with $\sigma = 0.085$ dex. These shifted values should not be able to reconstruct a correlation without introducing additional scatter unless the shifts were not arbitrary to begin with.

We thus find more evidence that the CNN has learned something from the SDSS *gri* imaging that is different from, but at least as powerful as, the MZR. One possible explanation is that the CNN is measuring some version of metallicity that is more fundamentally linked to the stellar mass, rather than $Z_{\rm pred}$ as derived from oxygen spectral lines. Another possibility is that the MZR is a projection of a correlation between stellar mass, metallicity, and a third parameter, perhaps one that is morphological in nature. If this is the case, then the Tremonti et al. (2004) MZR represents a relationship that is randomly distributed in the yet unknown third parameter, while our CNN would be able to stratify the MZR according to this parameter (much as the FMR does with the SFR). We are unfortunately not able to identify any hidden parameter using the current CNN methodology, but we plan to explore this topic in a future work.

7 SUMMARY

We have trained a deep CNN to predict galaxy gas-phase metallicity using only 128 × 128-pixel, three-band (*gri*) JPG images obtained from SDSS. We characterize CNN performance by measuring scatter in the residuals between predicted ($Z_{\rm pred}$) and true ($Z_{\rm true}$) metallicities. Our conclusions are as follows:

(i) By training for a half-hour on a GPU, the CNN can predict metallicity well enough to achieve residuals characterized by RMSE = 0.085 dex (or outlier-insensitive NMAD = 0.067 dex). These findings may be promising for future large spectroscopy-limited surveys such as LSST.

(ii) We find that the residual scatter decreases in an expected way as resolution or number of channels is increased, suggesting that the CNN is leveraging both the spatial information about a galaxy's light distribution and the colour in order to predict metallicity.

(iii) The CNN outperforms a random forest trained on *gri* photometry if provided images larger than 4 × 4 pixels, and is as accurate as a random forest trained on *ugriz* photometry when given 128 × 128 pixel *gri* images.

(iv) We find that scatter in the MZR constructed using CNN-predicted metallicities is as tight as the empirical MZR (σ = 0.10 dex). Because predicted metallicities differ from the 'true' metallicities by RMSE = 0.085 dex, the only way that the predicted MZR can have such low scatter is if the CNN has learned a connection to metallicity that is more strongly linked to the galaxies' light distributions than their nebular line emission.

All of the code used in our analysis and for making the figures can be accessed at https://github.com/jwuphysics/galaxy-cnns.

ACKNOWLEDGEMENTS

SB is supported by NSF Astronomy and Astrophysics Research Program award number 1615657. The authors thank the anonymous referee for useful comments that have improved this paper, particularly in terms of its scientific content. The authors thank Andrew Baker, Eric Gawiser, and John Hughes for helpful comments and discussions, and also thank David Shih and Matthew Buckley for the use of their GPU cluster in the Rutgers University Experimental High Energy Physics department. JW thanks Jeremy Howard, Rachel Thomas, and the development team for creating the `fastai` on-line courses and deep learning library.[3] JW also thanks Florian Peter for valuable assistance with using the `fastai` library. This research made use of the IPYTHON package (Perez & Granger 2007) and MATPLOTLIB, a Python library for publication quality graphics (Hunter 2007).

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University

[3] https://github.com/fastai/fastai

of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

REFERENCES

Abolfathi B. et al., 2018, ApJS, 235, 42
Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415
Acquaviva V., 2016, MNRAS, 456, 1618
Bailer-Jones C. A. L., 2000, A&A, 357, 197
Beck M. R. et al., 2018, MNRAS, 476, 5516
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Brooks A. M., Governato F., Booth C. M., Willman B., Gardner J. P., Wadsley J., Stinson G., Quinn T., 2007, ApJ, 655, L17
D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111
Dahlen T. et al., 2013, ApJ, 775, 93
Dai J.-M., Tong J., 2018, preprint (arXiv:1807.05657)
Dalcanton J. J., Yoachim P., Bernstein R. A., 2004, ApJ, 608, 189
Davé R., Finlator K., Oppenheimer B. D., 2012, MNRAS, 421, 98
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, MNRAS, 475, 2978
Freedman D., Diaconis P., 1981, Z. Wahrscheinlichkeitstheor. Verwandte Geb., 57, 453
Garnett D. R., 2002, ApJ, 581, 1019
He K., Zhang X., Ren S., Sun J., 2015, preprint (arXiv:1512.03385)
Hezaveh Y. D., Levasseur L. P., Marshall P. J., 2017, Nature, 548, 555
Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R., 2012, preprint (arXiv:1207.0580)
Hocking A., Geach J. E., Sun Y., Davey N., 2018, MNRAS, 473, 1108
Hopkins A. M. et al., 2003, ApJ, 599, 971
Howard J., Gugger S., Bekman S., Ingham F., Monroe F., Shaw A., Thomas R., 2018, fastai
Hoyle B., 2016, Astron. Comput., 16, 34
Huertas-Company M. et al., 2015, ApJS, 221, 8
Huertas-Company M. et al., 2018, ApJ, 858, 114
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ilbert O. et al., 2009, ApJ, 690, 1236
Ioffe S., Szegedy C., 2015, preprint (arXiv:1502.03167)
Kauffmann G. et al., 2003, MNRAS, 341, 33
Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Krizhevsky A., Sutskever I., Hinton G. E., 2012, Pererira F.,, Burges C. J. C.,, Bottou L.,, Weinberger K. Q.,, Proc. 25th Int. Conf. on Neural Information Processing Systems - Volume 1, Lake Tahoe, Nevada, 60, 1097
Krogh A., Hertz J. A., 1992, in Hanson S. J., Cowan J. D., Giles C. L., eds, Advances in Neural Information Processing Systems 5, Margan Kaufmann Publ. Inc., San Francisco, CA, p. 950
Kroupa P., 2001, MNRAS, 322, 231
Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, MNRAS, 473, 3895
LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D., 1989, Neural Comput., 1, 541
Leitner S. N., Kravtsov A. V., 2011, ApJ, 734, 48
Lilly S. J., Carollo C. M., Pipino A., Renzini A., Peng Y., 2013, ApJ, 772, 119
Loshchilov I., Hutter F., 2016, preprint (arXiv:1608.03983)
Loshchilov I., Hutter F., 2017, preprint (arXiv:1711.05101)
LSST Dark Energy Science Collaboration, 2012, 133 preprint (arXiv:1211.0310)
Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133
Mannucci F., Cresci G., Maiolino R., Marconi A., Gnerucci A., 2010, MNRAS, 408, 2115
Molino A. et al., 2017, MNRAS, 470, 95
Nair V., Hinton G. E., 2010, in Fürnkranz J., , Joachims T.,eds, Proc. 27th Int. Conf. on Machine Learning. ICML'10. , Omnipress. USA, p. 807
Ntampaka M., Trac H., Sutherland D. J., Battaglia N., Póczos B., Schneider J., 2015, ApJ, 803, 50
Ntampaka M., Trac H., Cisewski J., Price L. C., 2017, ApJ, 835, 106
Oke J. B., 1974, ApJS, 27, 21
Pan S. J., Yang Q., 2010, IEEE Trans. Knowl. Data Eng., 22, 1345
Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26
Paszke A. et al., 2017, iNIPS-W.Available at: autodiff-workshop.github.io
Pedregosa F. et al., 2012, J. Mach. Learn. Res., 12, 2825
Perez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21
Petrillo C. E. et al., 2017, MNRAS, 472, 1129
Petrillo C. E. et al., 2019, MNRAS, 482, 807
Petrosian V., 1976, ApJ, 209, L1
Russakovsky O. et al., 2014, preprint (arXiv:1409.0575)
Salim S. et al., 2007, ApJS, 173, 267
Salim S., Lee J. C., Davé R., Dickinson M., 2015, ApJ, 808, 25
Sánchez S. F. et al., 2013, A&A, 554, A58
Scherer D., Müller A., Behnke S., 2010, Artificial Neural Networks–ICANN 2010. Springer, Thessaloniki, Greece, p. 92
Simmons B. D. et al., 2017, MNRAS, 464, 4420
Simonyan K., Zisserman A., 2014, preprint (arXiv:1409.1556)
Smirnov E. A., Markov A. B., 2017, MNRAS, 469, 2024
Smith L. N., 2015, preprint (arXiv:1506.01186)
The Dark Energy Survey Collaboration, 2005, 42 preprint (astro-ph/0510346)
Tremonti C. A. et al., 2004, ApJ, 613, 898
Xu X., Ho S., Trac H., Schneider J., Poczos B., Ntampaka M., 2013, ApJ, 772, 147
York D. G. et al., 2000, AJ, 120, 1579
Zahid H. J., Geller M. J., Kewley L. J., Hwang H. S., Fabricant D. G., Kurtz M. J., 2013, ApJ, 771, L19
Zeiler M. D., Fergus R., 2014, in Fleet D., Pajdla T., Schiele B., Tuytelaars T., eds, Computer Vision – ECCV. Springer, Cham, p. 818

APPENDIX A: CONVOLUTION NEURAL NETWORK DETAILS

A1 Residual neural network architecture

CNNs are divided into 'layers' that compute the convolutions of filters, or kernels, with each of the inputs. A Rectified Linear Unit (ReLU) activation function is applied to the convolved output (ReLUs have been shown to propagate information about the relative importances of different features, and are effective for training deep neural networks; Nair & Hinton 2010). In a residual CNN, multiple convolutional layers containing small (e.g. 3×3) filters are arranged sequentially, and a final 'shortcut connection' adds the first layer, unaltered, to the final output (before the final ReLU activation; see e.g. fig. 2 of He et al. 2015). Such combinations of convolutions, activations, and shortcuts are called residual building blocks.

We use a 34-layer residual CNNs with the architecture described by He et al. (2015), and implemented using PYTORCH (version 0.3.1; Paszke et al. 2017) provided by the FASTAI framework (version 0.7; Howard et al. 2018). A full description of the architecture's layers can be found in the online PYTORCH documentation,[4] but we also provide a brief overview below.

The resnet can be separated into three 'layer groups' that roughly correspond to the levels of abstraction able to be learned by the network. Once an image is fed into the network, e.g. a three-channel 128×128 SDSS image, it is effectively converted into

[4] https://pytorch.org/docs/stable/torchvision/models.html

activation maps that depend on how well the filters match the input image. These maps are further convolved with the next layer of filters, and this process continues until the last layer group is reached. The activation maps are periodically downsampled, max pooled, or average pooled, which effectively halve the map sizes in each spatial dimension (for more about pooling layers in CNNs, see Scherer, Müller & Behnke 2010). The first two layer groups comprise multiple residual building blocks, and the final layer group consists of two fully connected linear layers, with a ReLU activation after the first and no activation after the second. The last fully connected layer does not have an activation function because we are working on a regression problem, and so the weights trained in that layer should be tuned to predict metallicity in the desired range.

A2 Adaptive learning rates

Neural network performance tends to depend dramatically on choice of hyperparameters. After an image is fed forward and the residual (= prediction − true) value is computed, relative contributions of error are propagated backward through the network, starting from the final layer and ending at the first layer. Using gradient descent of the loss (in our case, the root mean squared error), the network layers' weights are adjusted according to their error contributions multiplied by the *learning rate*. The process of computing errors from known images and metallicities and updating weights is called *training*, and when all of the training data set has been used to adjust network weights, a training *epoch* is completed.

The learning rate can be thought of as the step size during each weight update. A high learning rate allows the network to improve quickly, but at some point the large step size may become too coarse for additional optimization; conversely, a low learning rate might allow the network to traverse every bump and wiggle in the error landscape, but might also take a very long time to reach convergence (or get stuck indefinitely in a local minimum). We first select a learning rate by using the method described by Smith (2015). Over a number of epochs, the learning rate is reduced (or *annealed*) as the network needs to make more fine-tuned updates in order to achieve better accuracy. We use a method called cosine annealing, during which the learning rate is annealed with the cosine function continuously over individual (or batches of) training examples. It has been shown that if the learning rate is annealed and then restarted after one or more epochs, the network is less likely to get caught in local minima and overall accuracy is improved. We refer to Loshchilov & Hutter (2016) for details about employing cyclical learning rates and gradient descent with restarts, which are implemented in our CNN.

A3 Optimization techniques and preventing overfitting

Losses are computed for small 'batches' of training examples at a time. Gradients that minimize each batch are expected to be noisier than gradients that are computed to optimize the entire training data set loss. This technique of *stochastic gradient descent* helps prevent the CNN from overfitting training data, which is a possibility given the huge number of parameters in a deep CNN. We also use weight decay, another commonly used regularization technique, which adds a decay term proportional to each layer weight during the update step of training (e.g. Krogh & Hertz 1992).

As the learning rate is annealed with increasing numbers of batches, the weight updates are also expected to diminish. The Adam optimizer adaptively smooths the gradient descent in a way that depends on previous gradients (Kingma & Ba 2014). Adam is analogous to rolling downhill with gravitational potential, momentum, and friction terms (whereas gradient descent would be analogous to movement dependent only on the potential at its given time-step). For caveats about combining weight decay and Adam, see Loshchilov & Hutter (2017), whose updated algorithm is implemented in FASTAI.

We implement batch normalization (BN), a technique developed to fix a problem that previously caused deep networks to train extremely slowly (Ioffe & Szegedy 2015). To briefly recap the issue: updates to the layer weights depend on the contribution of the backpropagated error, but when the number of layers is large (i.e. in a deep CNN), the contribution becomes vanishingly small. BN is simply the rescaling of each input to the non-linear activation so that it has mean of zero and standard deviation of unity (i.e. subtract the mean and divide by the standard deviation). A new choice of hyperparameter is the batch size, or the number of training examples from which the mean and variance are calculated); we choose 256 based on tests of performance in ten training epochs.

Dropout is a method of disabling a random subset of connections after linear layers in a network in order to improve the network's generalizability (Hinton et al. 2012). The ensemble of learned gradients is less prone to overfit the training data set because the network is forced to discard random (and potentially valuable) information. The resulting network is better able to, e.g. learn subtle differences in the data that would otherwise be ignored when more obvious features dominate the gradient descent process. We apply dropout layers only to the final fully connected layers in our deep CNN, and avoid dropout in the batch-normalized layers (as recommended by Ioffe & Szegedy 2015). We use dropout rates of 0.25 for the linear layer after the early group, and 0.50 at the later linear layer, both of which are FASTAI defaults.

A4 Training the network

We initialize the network using weights that have been pretrained on the 1.7 million example ImageNet data set (which contains 1000 classes of objects; Russakovsky et al. 2014). The network should more quickly optimize toward the global minimum loss through transfer of low-level features already learned in earlier layers of the network (known as transfer learning; see e.g. Pan & Yang 2010).

We train only the final layer group for the first two epochs, which can be accomplished by not updating weights in the first two layer groups. The learning rate is initially set to 0.1 and then annealed according to a cosine schedule over an epoch (and then restarted to 0.1 at the beginning of the following epoch). We then allow the updating of weights in all layer groups while setting the learning rates to 0.01, 0.03, and 0.1 for the first, second, and last layer groups, respectively. This approach allows the final group of fully connected layers to respond strongly to different types of training examples (e.g. galaxies that appear very different in *gri* imaging) while the earlier layers are trained very slowly in order to preserve their more general features. Using these layered learning rates, we train the full network using a cosine annealing schedule that spans one, one, two, and then four epochs (where the different learning rates are annealed by the same amount). Using this combination of learning rate schedules, we find that our network quickly achieves low training losses (RMSE ∼0.085 on validation data sets). Altogether, only ten epochs of training are needed, which takes under 30 min on a GPU. We find that further training does

yield some gains, but this improvement plateaus around RMSE ∼0.083 and takes many more hours.

A5 Data augmentation

Nearly all neural networks benefit from larger training samples because they help prevent overfitting. Beyond the local Universe, galaxies are seen at nearly random orientation; such invariance permits synthetic data to be generated from rotations and flips of the training images (see e.g. Simonyan & Zisserman 2014). Each image is fed into the network along with four augmented versions, thus increasing the total training sample by a factor of five.

This technique is called data augmentation, and is particularly helpful for the network to learn uncommon truth values (e.g. in our case, very metal-poor or metal-rich galaxies). Each augmented image is fed-forward through the network and gradient contributions are computed together as part of the same batch. A similar process is applied to the network during predictions, which is known as test-time augmentation (TTA), whereby synthetic images are generated according to the same rules applied to the training data set. The CNN predicts an ensemble average over the augmented images, which tends to further improve RMSE by a few per cent. We use the default hyperparameters in the FASTAI library.

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