Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **491,** 1554–1574 (2020)
Advance Access publication 2019 October 7

doi:10.1093/mnras/stz2816

Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning

Mike Walmsley [1]★ Lewis Smith,[2] Chris Lintott [1] Yarin Gal,[2] Steven Bamford,[3] Hugh Dickinson,[4,5] Lucy Fortson,[4,5] Sandor Kruk,[6] Karen Masters [7,8] Claudia Scarlata,[4,5] Brooke Simmons,[9,10] Rebecca Smethurst [1] and Darryl Wright[4,5]

[1]*Oxford Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[2]*Oxford Computer Science, University of Oxford, 15 Parks Rd, Oxford OX1 3QD, UK*
[3]*School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, UK*
[4]*School of Physics and Astronomy, University of Minnesota, 116 Church St SE, Minneapolis, MN 55455, USA*
[5]*Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, MN 55455, USA*
[6]*European Space Agency, ESTEC, Keplerlaan 1, NL-2201 AZ Noordwijk, the Netherlands*
[7]*Haverford College, Department of Physics and Astronomy, 370 Lancaster Avenue, Haverford, PA 19041, USA*
[8]*Institute for Cosmology and Gravitation, University of Portsmouth, Dennis Sciama Building, Burnaby Road, Portsmouth PO1 3FX, UK*
[9]*Physics Department, Lancaster University, Lancaster LA1 4YB, UK*
[10]*Center for Astrophysics and Space Sciences (CASS), Department of Physics, University of California, San Diego, CA 92093, USA*

Accepted 2019 September 27. Received 2019 September 27; in original form 2019 May 14

ABSTRACT

We use Bayesian convolutional neural networks and a novel generative model of Galaxy Zoo volunteer responses to infer posteriors for the visual morphology of galaxies. Bayesian CNN can learn from galaxy images with uncertain labels and then, for previously unlabelled galaxies, predict the probability of each possible label. Our posteriors are well-calibrated (e.g. for predicting bars, we achieve coverage errors of 11.8 per cent within a vote fraction deviation of 0.2) and hence are reliable for practical use. Further, using our posteriors, we apply the active learning strategy BALD to request volunteer responses for the subset of galaxies which, if labelled, would be most informative for training our network. We show that training our Bayesian CNNs using active learning requires up to 35–60 per cent fewer labelled galaxies, depending on the morphological feature being classified. By combining human and machine intelligence, Galaxy zoo will be able to classify surveys of any conceivable scale on a time-scale of weeks, providing massive and detailed morphology catalogues to support research into galaxy evolution.

Key words: methods: data analysis – methods: statistical – galaxies: evolution – galaxies: statistics – galaxies: structure.

1 INTRODUCTION

Galaxy Zoo was created because Sloan Digital Sky Survey (SDSS) scale surveys could not be visually classified by professional astronomers (Lintott et al. 2008). In turn, Galaxy Zoo is being gradually outpaced by the increasing scale of modern surveys like DES (Flaugher 2005), PanSTARRS (Kaiser et al. 2010), the Kilo-Degree Survey (de Jong et al. 2015), and Hyper Suprime-Cam (Aihara et al. 2018).

Each of these surveys can image galaxies as fast or faster than those galaxies are being classified by volunteers. For example, DECaLS (Dey et al. 2018) contains (as of Data Release 5) approximately 350 000 galaxies suitable for detailed morphological classification (applying $r < 17$ and `petroR90_r`[1] > 3 arcsec, the cuts used for Galaxy Zoo 2 in Willett et al. 2013). Collecting 40 independent volunteer classifications for each galaxy, as for Galaxy Zoo 2 (Willett et al. 2013), would take approximately 5 yr at the current classification rate. The Galaxy Zoo science team must therefore both judiciously select which surveys to classify and, for the selected surveys, reduce the number of independent classifications per galaxy. The speed at which we can accurately classify galaxies severely limits the scale, detail, and quality of our morphology catalogues, diminishing the scientific value of such surveys.

★ E-mail: mike.walmsley@physics.ox.ac.uk

[1]`petroR90_r` is the Petrosian radius which contains 90 per cent of the *r*-band flux

The next generation of surveys will make this speed limitation even more stark. *Euclid*,[2] LSST[3], and *WFIRST* [4] are expected to resolve the morphology of unprecedented numbers of galaxies. This could be revolutionary for our understanding of galaxy evolution, but only if such galaxies can be classified. The future of morphology research therefore inevitably relies on automated classification methods. Supervised approaches (given human-labelled galaxies, predict labels for new galaxies) using convolutional neural networks (CNNs) are increasingly common and effective (Cheng et al. 2019). CNNs outperform previous non-parametric approaches (Dieleman, Willett & Dambre 2015; Huertas-Company et al. 2015), and can be rapidly adapted to new surveys (Domínguez Sánchez et al. 2019) and to related tasks such as light profile fitting (Tuccillo et al. 2018). Unsupervised approaches (cluster examples without any human labels) also show promise (Hocking et al. 2015).

However, despite major progress in raw performance, the increasing complexity of classification methods poses a problem for scientific inquiry. In particular, CNNs are 'black box' algorithms which are difficult to introspect and do not typically provide estimates of uncertainty. In this work, we combine a novel generative model of volunteer responses with Monte Carlo dropout (Gal, Islam & Ghahramani 2017a) to create Bayesian CNNs that predict *posteriors* for the morphology of each galaxy. Posteriors are crucial for drawing statistical conclusions that account for uncertainty, and so including posteriors significantly increases the scientific value of morphology catalogues. Our Bayesian CNNs can predict posteriors for surveys of any conceivable scale.

Limited volunteer classification speed remains a hurdle; we need to collect enough responses to train our Bayesian networks. How do we train Bayesian networks to perform well while minimizing the number of new responses required? Recent work suggests that transfer learning (Lu et al. 2015) may be effective. In transfer learning, models are first trained to solve similar tasks where training data are plentiful and then 'fine-tuned' with new data to solve the task at hand. Results using transfer learning to classify new surveys, or to answer new morphological questions, suggest that models can be fine-tuned using only thousands (Ackermann et al. 2018; Khan et al. 2019) or even hundreds (Domínguez Sánchez et al. 2019) of newly labelled galaxies, with only moderate performance losses compared to the original task.

Each of these authors randomly selects which new galaxies to label. However, this may not be optimal. Each galaxy, if labelled, provides information to our model; they are *informative*. Our hypothesis is that all galaxies are informative, but some galaxies are more informative than others. We use our galaxy morphology posteriors to apply an active learning strategy (Houlsby et al. 2011): *intelligently selecting the most informative galaxies for labelling by volunteers*. By prioritizing the galaxies that our strategy suggests would, if labelled, be most informative to the model, we can create or fine-tune models with even fewer newly labelled data.

In the first half of this work (Section 2), we present Bayesian CNNs that predict posteriors for the morphology of each galaxy. In the second (Section 3), we simulate using our posteriors to select the most informative galaxies for labelling by volunteers.

[2] 15 000 deg^2 at 0.30 arcsec half-light radius PSF from 2022, Laureijs et al. (2011)

[3] 18 000 deg^2 to 0.39 arcsec half-light radius PSF from 2023, LSST Science Collaboration (2009)

[4] 2 000 deg^2 at 0.12 arcsec half-light radius PSF from approximately 2025, Spergel et al. (2013)

2 POSTERIORS FOR GALAXY MORPHOLOGY

A vast number of automated methods have been used as proxies for 'traditional' visual morphological classification. Non-parametric methods such as CAS (Conselice 2003) and Gini (Lotz, Primack & Madau 2004) have been commonly used, both directly and to provide features which can be used by increasingly sophisticated machine learning strategies (Scarlata et al. 2007; Banerji et al. 2010; Huertas-Company et al. 2011; Freeman et al. 2013; Peth et al. 2016). Most of these methods provide imperfect proxies for expert classification (Lintott et al. 2008). The key advantage of CNNs is that they learn to approximate human classifications directly from data, without the need to hand-design functions aimed at identifying relevant features (LeCun, Bengio & Hinton 2015). CNNs work by applying a series of spatially invariant transformations to represent the input image at increasing levels of abstraction, and then interpreting the final abstraction level as a prediction. These transformations are initially random, and are 'learned' by iteratively minimizing the difference between predictions and known labels. We refer the reader to LeCun et al. (2015) for a brief introduction to CNNs and to Dieleman et al. (2015), Lanusse et al. (2018), Kim & Brunner (2017), and Hezaveh, Levasseur & Marshall (2017) for astrophysical applications.

Early work with CNNs immediately surpassed non-parametric methods in approximating human classifications (Dieleman et al. 2015; Huertas-Company et al. 2015). Recent work extends CNNs across different surveys (Domínguez Sánchez et al. 2019; Khan et al. 2019) or increasingly specific tasks (Domínguez Sánchez et al. 2018; Huertas-Company et al. 2018; Tuccillo et al. 2018; Walmsley et al. 2018). However, these previous CNNs do not account for uncertainty in training labels, limiting their ability to learn from all available data (one common approach is to train only on 'clean' subsets). Previous CNNs are also not designed to make probabilistic predictions (though they have been interpreted as such), limiting the reliability of conclusions drawn using such methods (see Appendix A).

Here, we present Bayesian CNNs for morphology classification. Bayesian CNNs provide two key improvements over previous work:

(i) We account for varying (i.e. heteroskedastic) uncertainty in volunteer responses
(ii) We predict full posteriors over the morphology of each galaxy

We first introduce a novel framework for thinking about Galaxy Zoo classifications in probabilistic terms, where volunteer responses are drawn from a binomial distribution according to an unobserved (latent) parameter: the 'typical' response probability (Section 2.1). We use this framework to construct CNNs that make probabilistic predictions of Galaxy Zoo classifications (Section 2.2). These CNNs predict a typical response probability for each galaxy by maximizing the likelihood of the observed responses. By maximizing the likelihood, they learn effectively from heteroskedastic labels; the likelihood reflects the fact that more volunteer responses are more indicative of the 'typical' response than fewer responses. To account for the uncertainty in the CNN weights, we use Monte Carlo dropout (Gal et al. 2017a) to marginalize over possible CNNs (Section 2.3). Our final predictions (Section 2.7) are posteriors of how a typical volunteer would have responded, had they been asked about each galaxy. These can then be used to classify surveys of any conceivable scale (e.g. LSST, *Euclid*), helping researchers make reliable inferences about galaxy evolution using millions of labelled galaxy images.

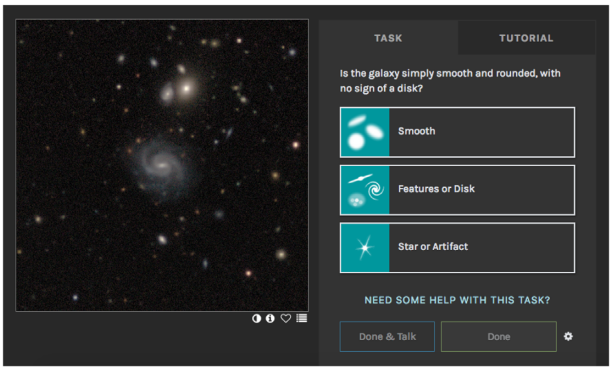


Figure 1. The Galaxy Zoo web interface as shown to volunteers. This screenshot shows the first question in the decision tree: is the galaxy smooth or featured?.

2.1 Probabilistic framework for Galaxy Zoo

Galaxy Zoo asks members of the public to volunteer as 'citizen scientists' and label galaxy images by answering a series of questions. Fig. 1 illustrates the web interface.

We aim to make a probabilistic prediction for the response of a typical volunteer. To do this, we need to model how each volunteer response is generated. Formally, each Galaxy Zoo decision tree question asks N_i volunteers to view galaxy image x_i and select the most appropriate answer A_j from the available answers $\{A\}$. This reduces to a binary choice; where there are more than two available answers ($|\{A\}| > 2$), we can consider each volunteer response as either A_j (positive response) or not A_j (negative response). We can therefore apply our model to questions with any number of answers.

Let k_{ij} be the number of volunteers (out of N_i) observed to answer A_j for image x_i. We assume that there is a true fraction ρ_{ij} of the population (i.e. all possible volunteers) who would give the answer A_j for image x_i. We assume that volunteers are drawn uniformly from this population, so that if we ask N_i volunteers about image x_i, we expect that the distribution over the number of positive answers k_{ij} to be binomial:

$$k_{ij} \sim \mathrm{Bin}(\rho_{ij}, N_i) \tag{1}$$

$$p(k_{ij}|x_{ij}, N_i) = \binom{N_i}{k_i} \rho_{ij}^{k_{ij}} (1-\rho_{ij})^{N_i - k_{ij}} \tag{2}$$

This will be our model for how each volunteer response k_{ij} was generated. Note that ρ_{ij} is a latent variable: we only observe the responses k_{ij}, never ρ_{ij} itself.

2.2 Probabilistic prediction with CNNs

Having established a novel generative model for our data, we now aim to infer the likelihood of observing a particular k for each galaxy x (for brevity, we omit subscripts).

Let us consider the scalar output from our neural network $f^w(x)$ as a (deterministic) prediction for ρ, and hence a probabilistic prediction for k:

$$p(k|x, w) = \mathrm{Bin}(k|f^w(x), N) \tag{3}$$

For each labelled galaxy, we have observed k positive responses. We would like to find the network weights w such that $p(k|x, N)$ is maximized (i.e. to make a maximum likelihood estimate given the observations):

$$\max_w [p(k|x, w)] = \max_w [\mathrm{Bin}(k|f^w(x), N)] \tag{4}$$

$$= \max_w [\log \binom{N}{k} + k \log f^w(x) + (N-k) \log(1 - f^w(x))]. \tag{5}$$

The combinatorial term is fixed and hence our objective function to minimize is

$$\mathcal{L} = k \log f^w(x) + (N-k) \log(1 - f^w(x)). \tag{6}$$

We can create a probabilistic model for k by optimizing our network to make maximum likelihood estimates $\hat{\rho} = f^w(x)$ for the latent parameter ρ from which k is drawn.

In short, each network w predicts the response probability ρ that a random volunteer will select a given answer for a given image.

2.3 From probabilistic to Bayesian CNN

So far, our model is probabilistic (i.e. the output is the parameter of a probabilistic model) but not Bayesian. If we asked N volunteers, we would predict k answers with a posterior of $p(k|w) = \mathrm{Bin}(k|f^w(x), N)$ (where $f^w(x)$ is our network prediction of ρ for galaxy x). However, this treats the model, w, as fixed and known. Instead, the Bayesian approach treats the model itself as a random variable.

Intuitively, there are many possible models that could be trained from the same training data $\mathcal{D}$. To predict the posterior of k given $\mathcal{D}$, we should marginalize over these possible models:

$$p(k|x, \mathcal{D}) = \int p(k|x, w) p(w|\mathcal{D}) dw. \tag{7}$$

We need to know how likely we were to train a particular model w given the data available, $p(w|\mathcal{D})$. Unfortunately, we do not know how likely each model is. We only observe the single model we actually trained.

Instead, consider dropout (Srivastava et al. 2014). Dropout is a regularization method that temporarily removes random neurons according to a Bernoulli distribution, where the probability of removal ('dropout rate') is a hyperparameter to be chosen. Dropout may be interpreted as taking the trained model and permuting it into a different one (Srivastava et al. 2014). Gal (2016) introduced the approach of approximating the distributions of models one might have trained, but didn't, with the distribution of networks from applying dropout:

$$p(w|\mathcal{D}) \approx q^* \tag{8}$$

removing neurons according to dropout distribution q^*. This is the Monte Carlo Dropout approximation (hereafter MC Dropout). See Appendix B for a more formal overview.

Choosing the dropout rate affects the approximation; greater dropout rates lead the model to estimate higher uncertainties (on average). Following convention, we arbitrarily choose a dropout rate of 0.5. We discuss the implications of using an arbitrary dropout rate, and opportunities for improvement, in Section 4.

Applying MC dropout to marginalize over models (equation 7):

$$p(k|x, \mathcal{D}) = \int p(k|x, w) q^* \mathrm{d}w. \tag{9}$$

In practice, following Gal (2016), we sample from q^* with T forward passes using dropout *at test time* (i.e. Monte Carlo integration):

$$\int p(k|x, w) q^* \mathrm{d}w \approx \frac{1}{T} \sum_t p(k|x, w_t). \tag{10}$$

Using MC Dropout, we can improve our posteriors by (approximately) marginalizing over the possible models we might have trained.

To demonstrate our probabilistic model and the use of MC dropout, we train models to predict volunteer responses to the 'Smooth or Featured' and 'Bar' questions on Galaxy Zoo 2 (Section 2.5).

2.4 Data - Galaxy Zoo 2

Galaxy Zoo 2 (GZ2) classified all 304 122 galaxies from the SDSS DR7 Main Galaxy Sample (Strauss et al. 2002; Abazajian et al. 2009) with $r < 17$ and `petroR90_r`[5] > 3 arcsec. Classifying 304 122 galaxies required ∼60 million volunteer responses collected over 14 months.

GZ2 is the largest homogenous galaxy sample with reliable measurements of detailed morphology, and hence an ideal data source for this work. GZ2 has been extensively used as a benchmark to compare machine learning methods for classifying galaxy morphology. The original GZ2 data release (Willett et al. 2013) included comparisons with (pre-CNN) machine learning methods by Baillard et al. (2011) and Huertas-Company et al. (2011). GZ2 subsequently provided the data for seminal work on CNN morphology classification (Dieleman et al. 2015) and continues to be used for validating new approaches (Domínguez Sánchez et al. 2018; Khan et al. 2019).

We use the 'GZ2 Full Sample' catalogue (hereafter 'GZ2 catalogue'), available from data.galaxyzoo.org. To avoid the possibility of duplicated galaxies or varying depth imaging, we exclude the 'stripe82' subset.

The GZ2 catalogue provides aggregate volunteer responses at each of the three post-processing stages: raw vote counts (and derived vote fractions), consensus vote fractions, and redshift-debiased vote fractions. The raw vote counts are simply the number of users who selected each answer. The consensus vote fractions are calculated by iteratively re-weighting each user based on their overall agreement with other users. The debiased fractions estimate how the galaxy would have been classified if viewed at $z = 0.03$ (Hart et al. 2016). Unlike recent work (Domínguez Sánchez et al. 2018; Khan et al. 2019), we use the raw vote counts. The redshift-debiased fractions estimate the *true* morphology of a galaxy, not what the image actually *shows*. To predict what volunteers would say about an image, we should only consider what the volunteers see. We believe that debiasing is better applied after predicting responses, not before. We caution the reader that our performance metrics are therefore not directly comparable to those of Domínguez Sánchez et al. (2018) and Khan et al. (2019), who use the debiased fractions as ground truth.

2.5 Application

2.5.1 *Tasks*

To test our probabilistic CNNs, we aim to predict volunteer responses for the 'Smooth or Featured' and 'Bar' questions.

The 'Smooth or Featured' question asks volunteers 'Is the galaxy simply smooth and rounded, with no sign of a disc?' with (common[6]) answers 'Smooth' and 'Featured or Disc'. As 'Smooth or Featured' is the first decision tree question, this question is always asked, and therefore every galaxy has ∼ 40 'Smooth or Featured' responses.[7] With N fixed to ∼40 responses, the loss function (equation 6) depends only on k (for a given model w).

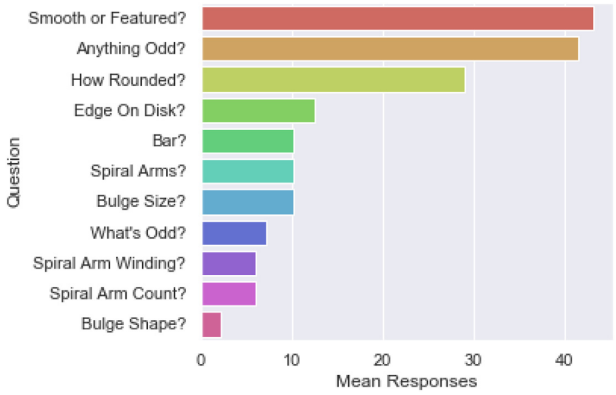


Figure 2. Mean responses (N) by GZ2 question. Being the first question, 'Smooth or Featured' has an unusually high (∼40) number of responses. Most questions (6 of 11), including 'Bar', are only asked for 'Featured' galaxies, and hence have only ∼10 responses. Training CNNs while accounting for the label uncertainty caused by low N responses is a key goal of this work.

The 'Bar' question asks volunteers 'Is there a sign of a bar feature through the centre of the galaxy?' with answers 'Bar (Yes)' and 'No Bar'. Because 'Bar' is only asked if volunteers respond 'Featured' and 'Not Edge-On' to previous questions, each galaxy can have anywhere from 0 to 40 total responses – typically around 10 (Fig. 2). This scenario is common; only two questions are always asked, and most questions have $N < <40$ total responses (Fig. 2). Building probabilistic CNNs that learn better by appreciating the varying count uncertainty in volunteer responses is a key advantage of our design. We achieve this by maximizing the likelihood of the observed responses given our predicted 'typical' response and N (Section 2.2).

2.5.2 *Architecture*

Our CNN architecture is shown in Fig. 3. This architecture is inspired by VGG16 (Simonyan & Zisserman 2015), but scaled down to be shallower and narrower in order to fit our computational budget. We use a softmax final layer to ensure the predicted typical vote fraction ρ lies between 0 and 1, as required by our binomial loss function (equation 6).

We are primarily concerned with accounting for label uncertainty and predicting posteriors, rather than maximizing performance metrics. That said, our architecture is competitive with, or outperforms, previous work (Section 2.7.1). Our overall performance can likely be significantly improved with more recent architectures (He et al. 2015; Szegedy et al. 2015; Huang et al. 2017) or a larger computational budget.

[5] `petroR90_r` is the Petrosian radius which contains 90 per cent of the r-band flux

[6] 'Smooth or Featured' includes a third 'Artefact' answer. However, artefacts are sufficiently rare (0.08 per cent of galaxies have 'Artefact' as the majority response) that predicting 'Smooth' or 'Not Smooth' is sufficient to separate smooth and featured galaxies in practice

[7] Technical limitations during GZ2 caused 26 530 galaxies to have $N < 36$. We exclude these galaxies for simplicity.

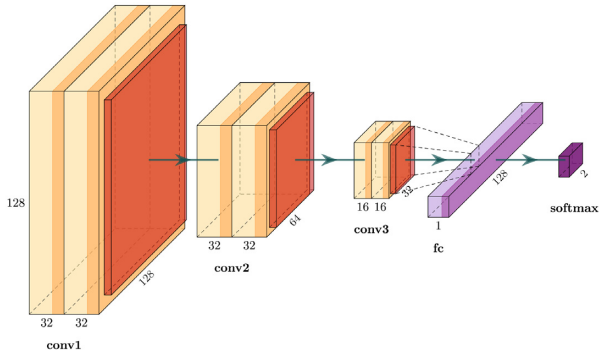


Figure 3. The CNN architecture used throughout. The input image, after applying augmentations (Section 2.5.3), is of dimension 128x128x1. The first pair of convolutional layers are each of dimension $128 \times 128 \times 32$ with 3×3 kernels. We then max-pool down to a second pair of convolutional layers of dimension $64 \times 64 \times 32$ with 3×3 kernels, then again to a final pair of dimension $32 \times 32 \times 16$ with 3×3 kernels. We finish with a 128 neuron linear dense layer and a 2 neuron softmax dense layer.

2.5.3 Augmentations

To generate our training and test images, we resize the original $424 \times 424 \times 3$ pixel GZ2 `png` images shown to volunteers into $256 \times 256 \times 3$ `uint8`[8] matrices and save these matrices in TFRecords (to facilitate rapid loading). When serving training images to our model, each image has the following transformations applied:

(i) Average over channels to create a greyscale image
(ii) Random horizontal and/or vertical flips
(iii) Rotation through an angle randomly selected from 0° to 90° (using nearest-neighbour interpolation to fill pixels)
(iv) Adjusting the image contrast to a contrast uniformly selected from 98 per cent to 102 per cent of the original contrast
(v) Cropping either randomly ('Smooth or Featured') or centrally ('Bar') according to a zoom level uniformly selected from 1.1 × to 1.3 × ('Smooth or Featured') or 1.7 × to 1.9 × ('Bar')
(vi) Resizing to a target size of $128 \times 128(\times 1)$

We train on greyscale images because colour is often predictive of galaxy type (E and S0 are predominantly redder, while S are bluer; Roberts & Haynes 1994) and we wish to ensure that our classifier does not learn to make biased predictions from this correlation. For example, a galaxy should be classified as smooth because it appears smooth, and not because it is red and therefore more likely to be smooth. Otherwise, we bias any later research investigating correlations between morphology and colour.

Random flips, rotations, contrast adjustment, and zooms (via crops) help the CNN learn that predictions should be invariant to these transformations – our predictions should not change because the image is flipped, for example. We choose a higher zoom level for 'Bar' because the original image radius for GZ2 was designed to show the full galaxy and any immediate neighbours (Willett et al. 2013) yet bars are generally found in the centre of galaxies (Kruk et al. 2017). We know that the 'Bar' classification should be invariant to all but the central region of the image, and therefore choose to sacrifice the outer regions in favour of increased resolution in the centre. Cropping and resizing are performed last to minimize resolution loss due to aliasing. Images are resized to match our computational budget.

We also apply these augmentations at test time. This allows us to marginalize over any unlearned invariance using MC Dropout, as part of marginalizing over networks (Section 2.3). Each permuted network makes predictions on a uniquely augmented image. The aggregated posterior (over many forward passes T) is therefore independent of e.g. orientation, enforcing our domain knowledge.

[8]Unsigned 8-bit integer i.e. 0–255 inclusive. After rescaling, this is sufficient to express the dynamic range of the images (as judged by visual inspection) while significantly reducing memory requirements versus the original 32-bit float flux measurements.

2.6 Experimental setup

For each question, we randomly select 2500 galaxies as a test subset and train on the remaining galaxies (following the selection criteria described in Section 2.4). Unlike Domínguez Sánchez et al. (2018) and Khan et al. (2019), we do not select a 'clean' sample of galaxies with extreme vote fractions on which to train. Instead, we take full advantage of the responses collected for every galaxy by carefully accounting for the vote uncertainty in galaxies with fewer responses (equation 6).

For 'Smooth or Featured', we use a final training sample of 176 328 galaxies. For 'Bar', we train and test only on galaxies with $N_{\text{bar}} \geq 10$ (56 048 galaxies). Without applying this cut, we find that models fail to learn; performance fails to improve from random initialization. This may be because galaxies with $N_{\text{bar}} < 10$ must have $k_{\text{featured}} < 10$ and so are almost all smooth and unbarred, leading to increasingly unbalanced typical vote fractions ρ.

Training was performed on an Amazon Web Services (AWS) `p2.xlarge` EC2 instance with an NVIDIA K80 GPU. Training each model from random initialization takes approximately 8 h.

Using the trained models, we make predictions $\hat{\rho}$ for the typical vote fraction ρ of each galaxy in the test subsets. We then evaluate performance by comparing $p(k|\hat{\rho}, N)$, our posterior for k positive responses from N volunteers, with the observed k from the N Galaxy Zoo volunteers asked.

2.7 Results

We find that our probabilistic CNNs produces posteriors which are reliable and informative.

For each question, we first compare a random selection of posteriors from either 1 or 30 MC Dropout forward passes (i.e. 1 or 30 MC-dropout-approximated 'networks'). Figs 4 and 5 show our posteriors for 'Smooth or Featured' and 'Bar', respectively.

Without MC Dropout, our posteriors are binomial. The spread of each posterior reflects two effects. First, the spread reflects the extremity of $\hat{\rho}$ that previous authors have expressed as 'volunteer agreement' or 'confidence' (Dieleman et al. 2015; Domínguez Sánchez et al. 2018). $\text{Bin}(k|\hat{\rho}, N)$ is narrower where $\hat{\rho}$ is close to 0 or 1. Secondly, the spread reflects N, the number of volunteers asked. For 'Smooth or Featured', where N is approximately fixed, this second effect is minor. For 'Bar', where N varies significantly between 10 and ∼40, the posteriors are more spread (less precise) where fewer volunteers have been asked.

With MC Dropout, our posteriors are a superposition of Binomials from each forward pass, each centred on a different $\hat{\rho}_t$. In consequence, the MC Dropout posteriors are more uncertain. This matches our intuition – by marginalizing over the different weights and augmentations we might have used, we expect our predictions to broaden.

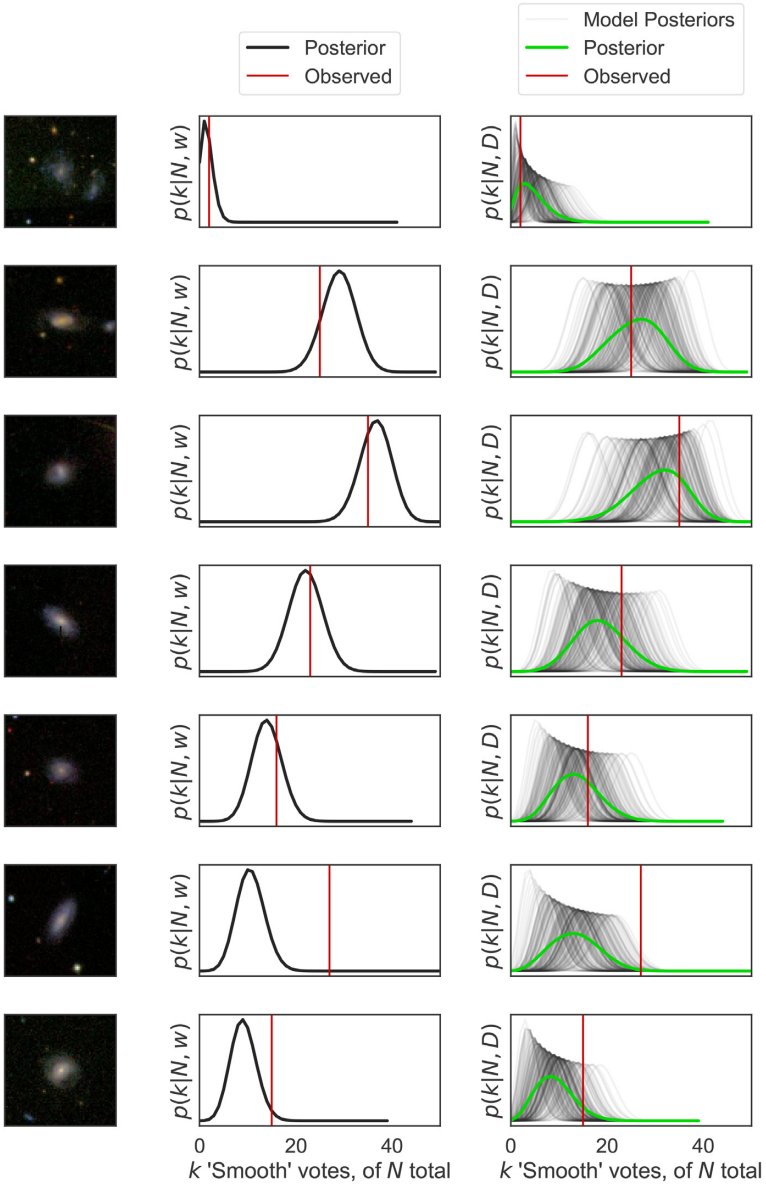


Figure 4. Posteriors for k of N volunteers answering 'Smooth' to the question 'Smooth or Featured?'). Each row is a randomly selected galaxy. Overplotted in red is the actual k measured from $N \sim 40$ volunteers. The left-hand column shows the galaxy in question, as presented to the network (following the augmentations described in Section 2.5.3). The central column shows the posterior predicted by a single network (black), while the right-hand column shows the posterior marginalized (averaged) over 30 MC-dropout-approximated 'networks' (green) as well as from each 'network' (grey). While the posterior from a single network is fixed to a binomial form, the marginalized posteriors from many 'networks' can take any form. The posterior from a single network is generally more confident (narrower); we later show that a single network is overconfident, and many 'networks' are better calibrated.

Given that each single network is relatively confident and the MC-dropout-marginalized model is relatively uncertain, which should be used? We prefer posteriors which are well-calibrated i.e. which reflect the true uncertainty in our predictions.

To quantify calibration, we introduce a novel method; we compare the predicted and observed vote fractions $\frac{k}{N}$ within increasing ranges of acceptable error. We outline this procedure below.

Choose some maximum acceptable error ϵ in predicting each vote fraction $v = \frac{k}{N}$. Over all galaxies, sum the total probability (from our predicted posteriors) that $v_i = \hat{v}_i \pm \epsilon$ for each galaxy i. We call this the expected count: how many galaxies the posterior suggests should have v within ϵ of the model prediction $\hat{v}$. For example, our 'Bar' model expects 2320 of 2500 galaxies in the 'Bar' test set to have an observed v within ± 0.20 of $\hat{v}$.

$$C_{\text{expected}} = \sum_{i}^{N_{\text{galaxies}}} \sum_{j>\hat{k}-N\epsilon}^{j<\hat{k}+N\epsilon} p(j|\hat{\rho}_i, N_i). \tag{11}$$

Next, over all galaxies, count how often v_i is within that maximum error $v_i = \hat{v}_i \pm \epsilon$. We call this the 'actual' count: how many galaxies are actually observed to have v_i within ϵ of the model prediction $\hat{v}_i$. For example, we observe 2075 of 2500 galaxies in the 'Bar' test set to have v_i within ± 0.20 of $\hat{v}$.

$$C_{\text{actual}} = \sum_{i}^{N_{\text{galaxies}}} \sum_{j>\hat{k}_i-N\epsilon}^{j<\hat{k}_i+N\epsilon} \delta(k_i - j). \tag{12}$$

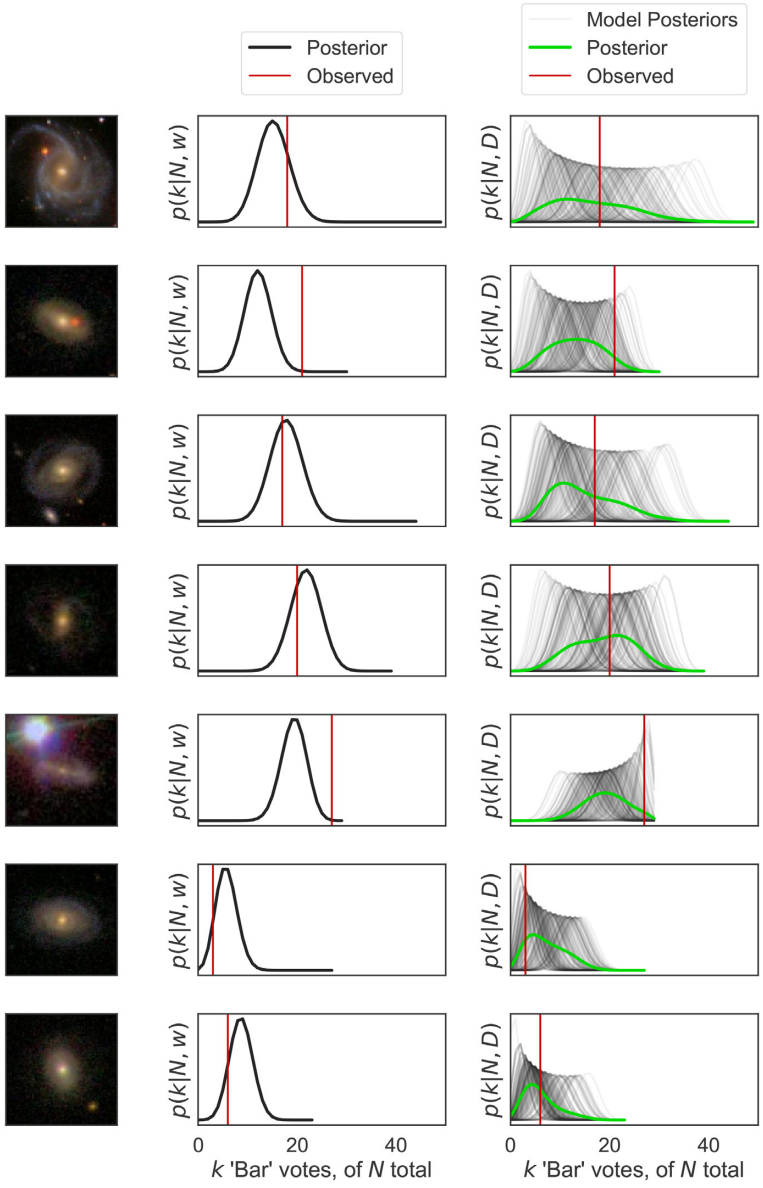


Figure 5. As for Fig. 4, but showing posteriors for k of N volunteers answering 'Bar (Yes)' to the question 'Bar?'. Unlike 'Smooth or Featured', N varies significantly between galaxies, and hence so does the spread (uncertainty in k) and absolute width (highest possible k) of the posterior.

For a perfectly calibrated posterior, the actual and expected counts would be identical: the model would be correct (within some given maximum error) as often as it expects to be correct. For an overconfident posterior, the expected count will be higher, and for an underconfident posterior, the actual count will be higher.

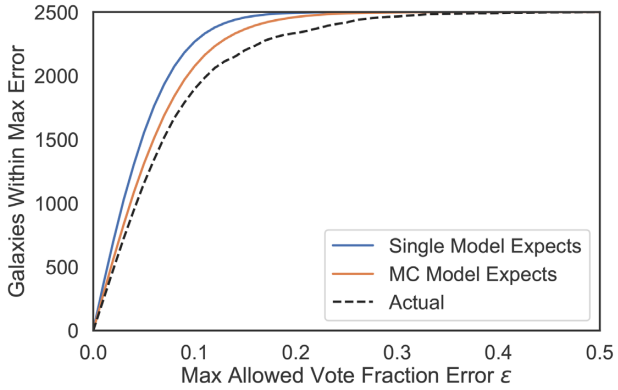
We find that our predicted posteriors of volunteer votes are fairly well-calibrated; our model is correct approximately as often as it *expects* to be correct. Fig. 6 compares the expected and actual counts for our model, choosing ϵ between 0 and 0.5. Tables 1 and 2 show calibration results for our 'Smooth' and 'Bar' models, with and without MC Dropout, evaluated on their respective test sets. Coverage error is calculated as:

$$\text{Coverage error} = \frac{C_{\text{expected}} - C_{\text{actual}}}{C_{\text{actual}}}. \quad (13)$$

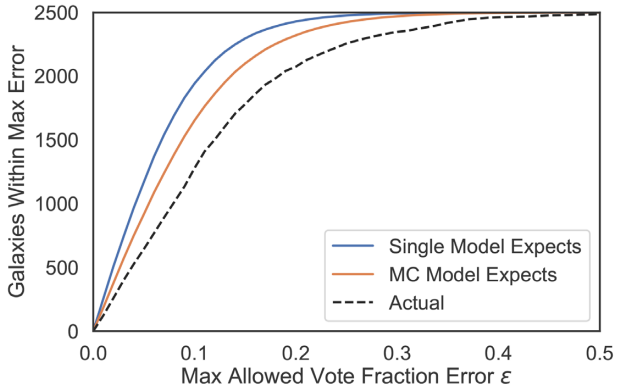
For both questions, the single network (without using MC Dropout) is visibly overconfident. The MC-dropout-marginalized network shows a significant improvement in calibration over the single network. We interpret this as evidence for the importance of marginalizing over both networks and augmentations in accurately estimating uncertainty (Section 2.3).

When making precise predictions, the MC-dropout-marginalized network remains somewhat overconfident. However, as the acceptable error ϵ is allowed to increase, the network is increasingly well-calibrated. For example, the predicted probability that $v \pm 0.02$ (i.e. $\epsilon = 0.02$) k of N volunteers respond 'Bar' is overestimated by ∼45 per cent. In contrast, the predicted probability that $k \pm 0.2$ (i.e. $\epsilon = 0.2$) of N volunteers respond 'Bar' is ∼ 10 per cent of the true probability. We discuss future approaches to further improve calibration in Section 4.

A key method for galaxy evolution research is to compare the distribution of some morphology parameter across different samples (e.g. are spirals more common in dense environments, Wang et al. 2018; do bars fuel AGN, Galloway et al. 2015; do mergers inhibit

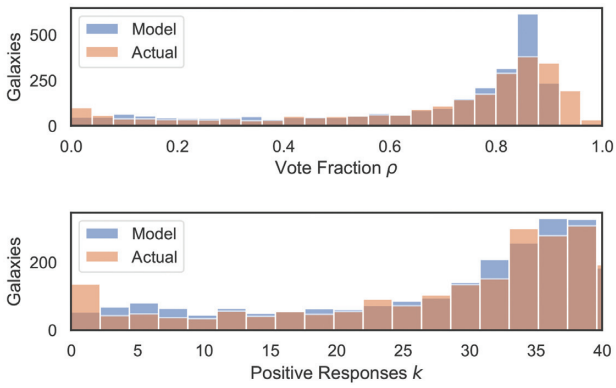


(a) Calibration for 'Smooth or Featured'

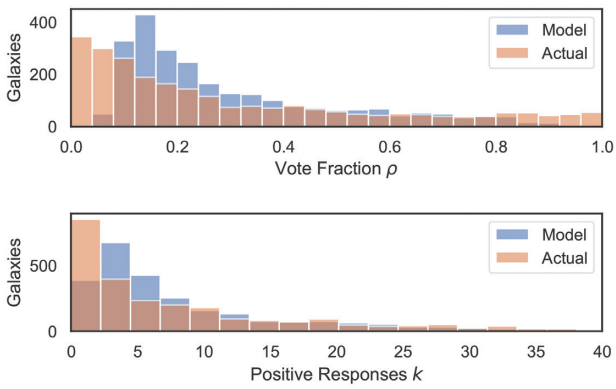


(b) Calibration for 'Bar'

Figure 6. Calibration of CNN-predicted posteriors, showing the expected versus actual count of galaxies within each acceptable maximum vote fraction error range (ϵ). Our probabilistic model is fairly well-calibrated (similar expected and actual counts), with a significant improvement from applying MC Dropout.



(a) Distribution of k and ρ for 'Smooth or Featured'



(b) Distribution of k and ρ for 'Bar'

Figure 7. Comparison between the distribution of predicted or observed ρ and k over all galaxies, for each question. Upper: comparison for 'Smooth or Featured'. Lower: comparison for 'Bar'. The observed ρ is approximated as $\rho_{\text{proxy}} = \frac{k}{N}$ The distributions of predicted ρ and k closely match the observed distributions, indicating our models are globally unbiased. The only significant deviation is near extreme ρ and k, which our models are 'reluctant' to predict.

LERGs, Gordon et al. 2019, etc.) We would therefore like the distribution of predicted $\hat{\rho}$ and $\hat{k}$, over all galaxies, to approximate the observed distribution of ρ[9] and k. In short, we would like our predictions to be *globally unbiased*. Fig. 7 compares our predicted and actual distributions of ρ and k. We find that our predicted distributions for ρ and k match well with the observed distributions for most values of ρ and k. Our model appears somewhat reticent to predict extreme ρ (and therefore extreme k) for both questions. This may be a consequence of the difficulty in predicting the behaviour of single volunteers. We discuss this further in Section 4.

Reliable research conclusions also require that model performance should not depend strongly on non-morphological galaxy parameters (mass, colour, etc). For example, if a researcher would like to investigate correlations between galaxy mass and bars, it is important that our model is equally able to recognize bars in high-mass and low-mass galaxies. To check if our model is sensitive to non-morphological parameters, we use an Explainable Boosting Machine (EBM) model (Lou, Caruana & Gehrke 2012; Caruana et al. 2015). EBM aims to predict a target variable based on tabular features by separating the impact of those features into single (or, optionally, pairwise) effects on the target variable. They are a specific[10] implementation of Generalized Additive Models (GAM; Hastie & Tibshirani 1990). GAM are of the form:

$$g(y) = f_1(x) + ... + f_n(x_n), \tag{14}$$

Table 1. Calibration results for predicting the probability that $v \pm \epsilon$ fraction of volunteers respond 'Smooth', with and without applying MC Dropout.

Max error ϵ	Coverage error without MC (percent)	Coverage error with MC (percent)
0.02	49.6	16.5
0.05	38.5	13.4
0.10	26.1	9.4
0.20	7.9	5.4

[9]The 'observed' ρ is approximated as $\rho_{\text{proxy}} = \frac{k}{N}$, which has a similar distribution to the true (latent, unobserved) ρ over a large sample.

[10]https://github.com/microsoft/interpret

Table 2. Calibration results for predicting the probability that $v \pm \epsilon$ fraction of volunteers respond 'Bar', with and without applying MC Dropout.

Max error ϵ	Coverage error without MC (percent)	Coverage error with MC (percent)
0.02	92.2	45.5
0.05	85.5	42.4
0.10	57.8	29.2
0.20	22.6	11.8

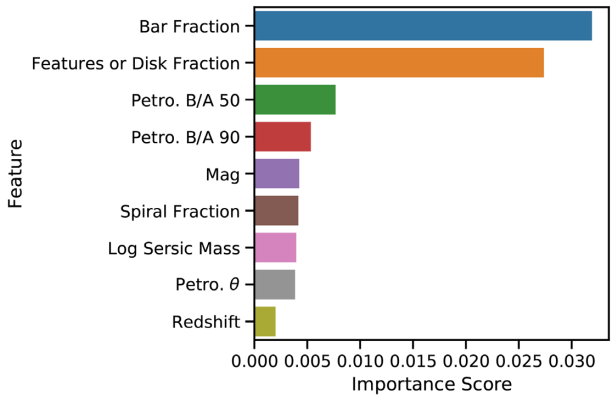
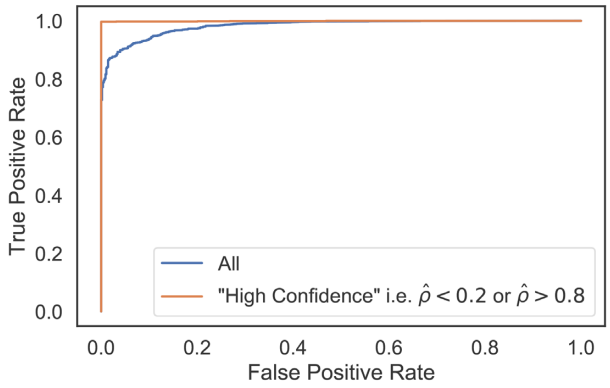


Figure 8. Relative importance of morphological (Features or Disc, Bar, Spiral) and non-morphological (Petro B/A, Mag, etc.) features for BCNN performance. Morphology fractions are the (human-reported) $\frac{k}{N}$ values from Galaxy Zoo 2. Petro B/A 50 and Petro B/A 90 measure the axial ratios at 50 per cent and 90 per cent of the half-light radius. Mag is the estimated B magnitude. Sersic mass is the approximate stellar mass, estimated from the single-component Sersic fit flux. Petro θ is the (r-band) Petrosian radius. Redshift is measured spectroscopically. The effect of each component is additive and independent; for example, the measured effect of spiral features does not include the effect of being featured in general. BCNN performance varies much less from the effect of non-morphological features than from morphological features.

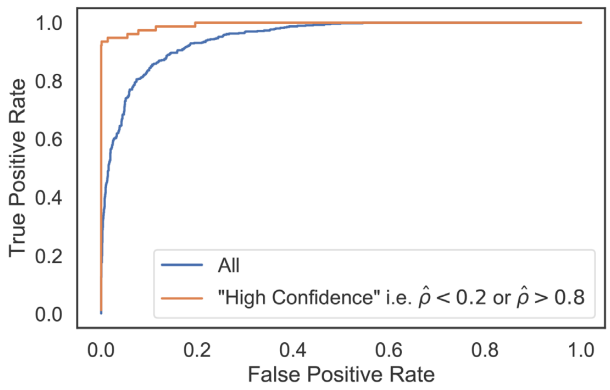
where g is identity for regression problems and f_i is any learnable function. For EBM, each f_i is learned using gradient boosting with bagging of shallow regression trees. They aim to answer the question 'What is the effect on the target variable of *this particular feature alone*?' We train an EBM to predict the surprise[11] of our 'Bar' model when making test set predictions (Section 2.6), using the human-reported morphologies and key non-morphological parameters reported in the NASA Sloan Atlas (v1.01; Albareti et al. 2017).

The interested reader can find our full investigation at http://www.walmsley.dev/2019/bias, recorded as a Jupyter Notebook. Fig. 8 shows the key result; the relative importance of each feature on BCNN model surprise. We find that performance variation with respect to non-morphological parameters is much smaller than variation with respect to morphology. Our network performs better on smooth galaxies and unbarred galaxies (plausibly because there are more training examples of such galaxies to learn from). Inclination is the non-morphological parameter with the strongest effect on performance, and this effect is approximately 3.5–4× weaker than the effect of either smoothness or barredness above. We are therefore confident that our model introduces no new major biases with respect to key non-morphological parameters.

[11]Recall that we quantify surprise as the likelihood of our prediction given the observed votes $\frac{k}{N}$ (equation 3).



(a) ROC curve for the 'Smooth or Featured' question.



(b) ROC curve for the 'Bar' question.

Figure 9. ROC curves for the 'Smooth or Featured' (above) and 'Bar' (below) questions, as predicted by our probabilistic model. To generate scalar class predictions on which to threshold, we reduce our posteriors to mean vote fractions. For comparison to DS + 18, we also include ROC curves of the subsample they describe as 'high confidence' – galaxies where the class probability (for us, $\hat{\rho}$) is extreme (1420 galaxies for 'Smooth', 1174 for 'Bar').

2.7.1 *Comparison to previous work*

The key goals of this paper are to introduce probabilistic predictions for votes and (in the following section) to apply this to perform active learning. However, by reducing our probabilistic predictions to point estimates, we can also provide conventional predictions and performance metrics.

Previous work has focused on deterministic predictions of either the votes (Dieleman et al. 2015) or the majority response (Domínguez Sánchez et al. 2018; Khan et al. 2019). While differences in sample selection and training data prevent a precise comparison, our model performs well at both tasks.

When reducing our posteriors to the most likely vote count $\hat{k}$, we achieve a root-mean-square error of 0.10 (approximately ±3 votes) for 'Smooth or Featured' and 0.15 for 'Bar'. We can also reduce the same posteriors to the most likely majority responses. Below, we present our results in the style of the ROC curves in Domínguez Sánchez et al. (2018) (hereafter DS + 18; Fig. 9) and the confusion matrices in Khan et al. (2019) (hereafter K + 18; Fig. 10) using our reduced posteriors. We find that our model likely outperforms Domínguez Sánchez et al. (2018) and is likely comparable with Khan et al. (2019).

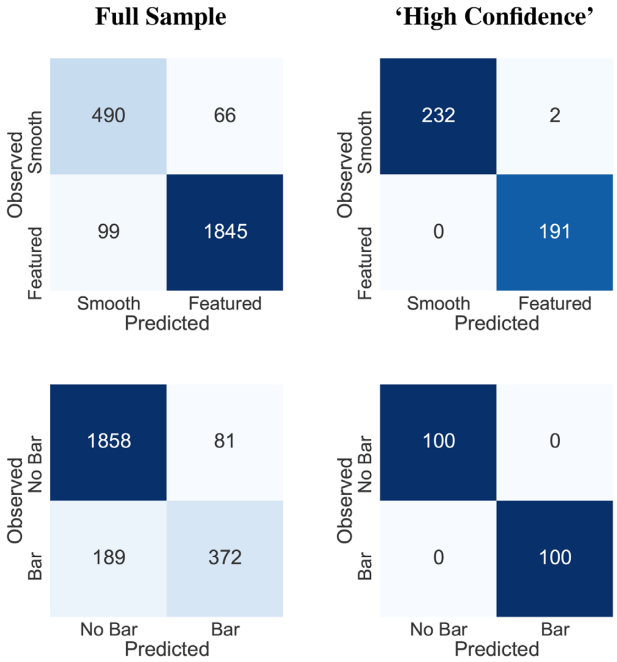


Figure 10. Confusion matrices for 'Smooth or Featured' (upper row) and 'Bar' (lower row) questions. For comparison to K + 18, we also include confusion matrices for the most confident predictions (right-hand column) Following K + 18, we include the most confident ~ 7.7 per cent of spirals and ~ 9.3 per cent of ellipticals (upper right). Of the two galaxies where humans select 'Smooth' ($\frac{k}{N} > 0.5$) and the model selects 'Featured' ($\hat{\rho} < 0.5$), one is an ongoing smooth/featured major merger and one is smooth with an imaging artefact. Generalizing (K + 18 do not consider bars), we also show the most confident ~ 8 per cent of barred and unbarred galaxies. We achieve perfect classification for 'Bar'.

Overall, these conventional metrics demonstrate that our models are sufficiently accurate for practical use in galaxy evolution research even when reduced to point estimates.

3 ACTIVE LEARNING

In the first half of this paper, we presented Bayesian CNNs that predict posteriors for the morphology of each galaxy. In the second, we show how we can use these posteriors to select the most informative galaxies for labelling by volunteers, helping humans and algorithms work together to do better science than either alone.

CNNs, and other deep learning methods, rely on vast training sets of labelled examples (He et al. 2015; Russakovsky et al. 2015; Simonyan & Zisserman 2015; Szegedy et al. 2015; Huang et al. 2017). As we argued in Section 1, we urgently need methods to reduce this demand for labelled data in order to fully exploit current and next-generation surveys.

Previous approaches in morphology classification have largely used fixed data sets of labelled galaxies acquired prior to model training. This is true both for authors applying direct training (Huertas-Company et al. 2015, 2018; Domínguez Sánchez et al. 2018; Fischer, Dom & Bernardi 2018; Walmsley et al. 2018) and those applying transfer learning (Ackermann et al. 2018; Pérez-Carrasco et al. 2019; Domínguez Sánchez et al. 2019). Instead, we ask: to train the best model, which galaxies should volunteers label?

Selecting the most informative data to label is known as active learning. Active learning is useful when acquiring labels is difficult (expensive, time-consuming, requiring experts, private, etc). This scenario is common for many, if not most, real-world problems. Terrestrial examples include detecting cardiac arrhythmia (Rahhal et al. 2016), sentiment analysis of online reviews (Zhou, Chen & Wang 2013), and Earth observation (Tuia et al. 2011; Liu, Zhang & Eom 2017). Astrophysical examples include stellar spectral analysis (Solorio et al. 2005), variable star classification (Richards et al. 2012), telescope design and time allocation (Xia, Protopapas & Doshi-Velez 2016), redshift estimation (Hoyle et al. 2016), and spectroscopic follow-up of supernovae (Ishida et al. 2019).

3.1 Active learning approach for Galaxy Zoo

Given that only a small subset of galaxies can be labelled by humans, we should intelligently select which galaxies to label. The aim is to make CNNs which are just as accurate without having to label as many galaxies.

Our approach is as follows. First, we train our CNN on a small randomly chosen initial training set. Then, we repeat the following active learning loop:

(i) Measure the CNN prediction uncertainty on all currently unlabelled galaxies (excluding a fixed test set)
(ii) Apply an acquisition function (Section 3.2) to select the most uncertain galaxies for labelling
(iii) Upload these galaxies to Galaxy Zoo and collect volunteer classifications (in this work, simulated with historical classifications)
(iv) Re-train the CNN and repeat

Other astrophysics research has combined crowdsourcing with machine learning models. Wright et al. (2017) classified supernovae in PanSTARRS (Kaiser et al. 2010) by aggregating crowdsourced classifications with the predictions of expert-trained CNN and show that the combined human–machine ensemble outperforms either alone. However, this approach is not directly feasible for Galaxy Zoo, where scale prevents us from recording crowdsourced classifications for every image.

A previous effort to consider optimizing task assignment was made by Beck et al. (2018), who developed a 'decision engine' to allocate galaxies for classification by either human or machine (via a random forest). Their system assigns each galaxy to the categories 'Smooth' or 'Featured'[12] , using SWAP (Marshall et al. 2016) to decide how may responses to collect. This is in contrast to the system presented here which only requests responses for informative galaxies, but (for simplicity) requests the same number of responses for each informative galaxy. Another important difference is that Beck et al. (2018) train their model exclusively on galaxies which can be confidently assigned to a class, while the use of uncertainty in our model allows learning to occur from every classified galaxy.

This work is the first time active learning has been used for morphological classification, and the first time in astrophysics that active learning has been combined with CNNs or crowdsourcing.

In the following Sections (3.2, 3.3, 3.4), we derive an acquisition function that selects the most informative galaxies for labelling by volunteers. We do this by combining the general acquisition strategy BALD (MacKay 1992; Houlsby et al. 2011) with our

[12]The actual categories used were 'Featured' or 'Not Featured' (Smooth + Artefact), but they argue that Artefact is sufficiently rare to not affect the results.

probabilistic model and Monte Carlo Dropout (Gal 2016). We then use historical data to simulate applying active learning strategy to Galaxy Zoo (Section 3.5) and compare the performance of models trained on galaxies selected using the mutual information versus galaxies selected randomly (Section 3.6).

3.2 BALD and mutual information

Bayesian Active Learning by Disagreement, BALD (MacKay 1992; Houlsby et al. 2011), is a general information-theoretic acquisition strategy. BALD selects subjects to label by maximizing the mutual information between the model parameters θ and the probabilistic label prediction y. We begin deriving our acquisition function by describing BALD and the mutual information.

We have observed data $\mathcal{D} = (x_i, y_i)_{i=1}^{n}$. Here, x_i is the ith subject and y_i is the label of interest. We assume there are (unknown) parameters θ that model the relationship between input subjects x and output labels y, $p(y|x, \theta)$. We would like to infer the posterior of θ, $p(\theta|\mathcal{D})$. Once we know $p(y|x, \theta)$, we can make predictions on new galaxy images.

The mutual information measures how much information some random variable A carries about another random variable B, defined as:

$$\mathbb{I}[A, B] = H[p(A)] - E_{p(B)}H[p(A|B)], \tag{15}$$

where H is the entropy operator and $E_{p(B)}H[p(A|B)]$ is the expected entropy of $p(A|B)$, marginalized over $p(B)$ (Murphy 2012)

We would like to know how much information each label y provides about the model parameters θ. We can then pick subjects x to maximize the mutual information $\mathbb{I}[y, \theta]$, helping us to learn θ efficiently. Substituting A and B for x and y:

$$\mathbb{I}[y, \theta] = H[p(y|x, \mathcal{D})] - \mathbb{E}_{p(\theta|\mathcal{D})}[H[p(y|x, \theta)]]. \tag{16}$$

The first term is the entropy of our prediction for x given the training data, implicitly marginalizing over the possible model parameters θ. We refer to this as the predictive entropy. The predictive entropy reflects our overall uncertainty in y given the training data available.

The second term is the expected entropy of our prediction made with a given θ, sampling over each θ we might have inferred from $\mathcal{D}$. The expected entropy reflects the typical uncertainty of each particular model on x. Expected entropy has a lower bound set by the inherent difficulty in predicting y from x, regardless of the available labelled data.

Confident disagreement between possible models leads to high mutual information. For high mutual information, we should be highly uncertain about y after marginalizing over all the models we might infer (high $H[p(y|x, \mathcal{D})]$), but have each particular model be confident (low expected $H[p(y|x, \theta)]$). If we are uncertain overall, but each particular model is certain, then the models must confidently disagree.

Throughout this work, when we refer to galaxies as informative, we mean specifically that they have a high mutual information; they are *informative for the model*. These are not necessarily the galaxies which are the most *informative for science*; any overlap will depend upon the research question at hand. The scientific benefit of our approach is that we improve our morphological predictions for all galaxies using minimal newly labelled examples.

3.3 Estimating mutual information

Rewriting the mutual information explicitly, replacing y with our labels k and θ with the network weights w:

$$\mathbb{I}[k, w] = \mathbb{H}[\int p(k|x, w)p(w|\mathcal{D})\mathrm{d}w] - \int p(w|\mathcal{D})\mathbb{H}[p(k|x, w)]\mathrm{d}w. \tag{17}$$

Gal et al. (2017a) showed that we can use equation (8) to replace $p(w|\mathcal{D})$ in the mutual information (equation 17):

$$\mathbb{I}[k, w] = \mathbb{H}[\int p(k|x, w)q^*\mathrm{d}w] - \int q^*\mathbb{H}[p(k|x, w)]\mathrm{d}w \tag{18}$$

and again sample from q^* with T forward passes using dropout at test time (i.e. Monte Carlo integration):

$$\mathbb{I}[k, w] = \mathbb{H}[\frac{1}{T}\sum_t p(k|x, w)] - \frac{1}{T}\sum_t \mathbb{H}[p(k|x, w)]. \tag{19}$$

Next, we need a probabilistic prediction for k, $p(k|x, w)$. Here, we diverge from previous work.

Recall that we trained our network to make probabilistic predictions for k by estimating the latent parameter ρ from which k is Binomially drawn (equation 3). Substituting the probabilistic predictions of equation (3) into the mutual information:

$$\mathbb{I}[k, w] = \mathbb{H}[\frac{1}{T}\sum_t \mathrm{Bin}(k|f^w(x), N)] - \frac{1}{T}\sum_t \mathbb{H}[\mathrm{Bin}(k|f^w(x), N)] \tag{20}$$

Or concisely:

$$\mathbb{I}[k, w] = \mathbb{H}[\langle \mathrm{Bin}(k|f^w(x), N)\rangle] - \langle \mathbb{H}[\mathrm{Bin}(k|f^w(x), N)]\rangle. \tag{21}$$

A novel complication is that we do not know N, the total number of responses, prior to labelling. In GZ2, each subject is shown to a fixed number of volunteers, but (due to the decision tree) N for each question will depend on responses to the previous question. Further, technical limitations mean that even for the first question ('Smooth or Featured'), N can vary (Fig. 2). We (implicitly, for clarity) approximate N with the expected $\langle N\rangle$ for that question. In effect, we calculate our acquisition function with N set to the value that, *were we to ask volunteers to label this galaxy, we would expect N responses*.

To summarize, equation (21) asks: how much additional information would be gained about network parameters that we use to predict ρ and k, were we to ask $\langle N\rangle$ people about subject x?

3.4 Entropy evaluation

Having approximated $p(w|\mathcal{D})$ with dropout and calculated $p(k|x, w)$ with our probabilistic model, all that remains is to calculate the entropies $\mathbb{H}$ of each term.

k is discrete and hence we can directly calculate the entropy over each possible state:

$$\mathbb{H}[\mathrm{Bin}(k|f^w(x), N)] = -\sum_{k=0}^{N} \mathrm{Bin}(k|f^w(x), N)\log[\mathrm{Bin}(k|f^w(x), N)]. \tag{22}$$

For $\mathbb{H}[\langle \mathrm{Bin}(k|f^w(x), N)\rangle]$, we can also enumerate over each possible k, where the probability of each k is the mean of the

posterior predictions (sampled with dropout) for that k:

$$\mathbb{H}[\langle \mathrm{Bin(k|f^w(x), N)} \rangle] = -\sum_{k=0}^{N} \langle \mathrm{Bin}(k|f^w(x), N) \rangle \log[\langle \mathrm{Bin(k|f^w(x), N)} \rangle] \quad (23)$$

and hence our final expression for the mutual information is:

$$\mathbb{I}[k, w] = -\sum_{k=0}^{N} \langle \mathrm{Bin}(k|f^w(x), N) \rangle \log[\langle \mathrm{Bin}(k|f^w(x), N) \rangle] + \sum_{k=0}^{N} \mathrm{Bin}(k|f^w(x), N) \log[\mathrm{Bin}(k|f^w(x), N)]. \quad (24)$$

3.5 Application

To evaluate our active learning approach, we simulate applying active learning during GZ2. We compare the performance of our models when trained on galaxies selected using the mutual information versus galaxies selected randomly. For simplicity, each simulation trains a model to predict either 'Smooth or Featured' responses or 'Bar' responses.

For the 'Smooth or Featured' simulation, we begin with a small initial training set of 256 random galaxies. We train a model and predict $p(k|\rho, N)$ (where N is the expected number of volunteers to answer the question, calculated as the mean total number of responses for that question over all previous galaxies; see Fig. 2). We then use our BALD acquisition function (equation 21) to identify the 128 most informative galaxies to label. To simulate uploading the informative galaxies to GZ and receiving classifications, we retrieve previously collected GZ2 classifications. Finally, we add the newly labelled informative galaxies to our training set. We refer to each execution of this process (training our model, selecting new galaxies to label, and adding them to the training set) as an *iteration*. We repeat for 20 iterations, recording the performance of our model throughout.

We selected 256 initial galaxies and 128 further galaxies per iteration, to match the training data size over which our 'Smooth or Featured' model performance varies. Our relatively shallow model reaches peak performance on around 3000 random galaxies; more galaxies do not significantly improve performance.

For the 'Bar' simulation, we observe that performance saturates after more galaxies (approximately 6000) and so we double the scale; we start with 512 galaxies and acquire 256 further galaxies per iteration. This matches previous results (and intuition) that 'Smooth or Featured' is an easier question to answer than 'Bar'. Identifying bars, particularly weak bars, is challenging for both humans (Masters et al. 2012; Kruk et al. 2018), and machines (including CNNs; Domínguez Sánchez et al. 2018).

To measure the effect of our active learning strategy, we also train a baseline classifier by providing batches of randomly selected galaxies. We aim to compare two acquisition strategies for deciding which galaxies to label: selecting galaxies with maximal mutual information (active learning via BALD and MC Dropout) or selecting randomly (baseline). We evaluate performance on a fixed test set of 2500 random galaxies. We repeat each simulation four times to reduce the risk of spurious results from random variations in performance.

3.6 Results

For both 'Smooth' and 'Bar' simulations, our probabilistic models achieve equal performance on fewer galaxies using active learning versus random galaxy selection. We show model performance by iteration for the 'Smooth' (Fig. 11) and 'Bar' (Fig. 12) simulations. We display three metrics: training loss (model surprise on previously seen images, measured by equation 6), evaluation loss (model surprise on unseen images), and root-mean-square error (RMSE). We measure the RMSE between our maximum-likelihood-estimates $\hat{\rho}$ and $\rho_{\mathrm{proxy}} = \frac{k}{N}$ as ρ itself is never observed and hence cannot be used for evaluation. Due to the high variance in metrics between batches, we smooth our metrics via LOWESS (Cleveland 1979) and average across four simulation runs.

For 'Smooth', we achieve equal RMSE scores with, at best, ∼60 per cent fewer newly labelled galaxies (RMSE of 0.117 with 256 versus 640 new galaxies; Fig. 11). Similarly for 'Bar', we achieve equal RMSE scores with, at best, ∼35 per cent fewer newly labelled galaxies (RMSE of 0.17 with 1280 versus 2048 new

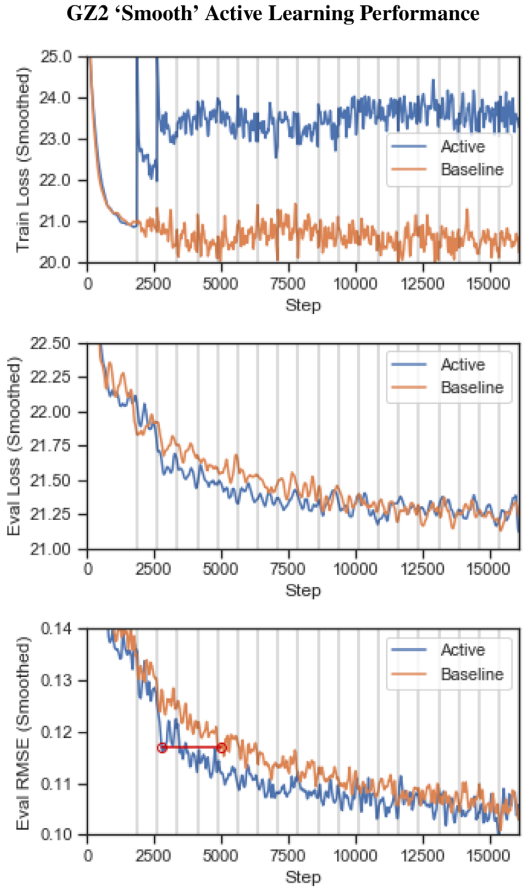


Figure 11. Training loss (upper), evaluation loss (middle), and RMSE (lower) of model performance on 'Smooth or Featured' during active learning simulations, by iteration (set of new galaxies). The vertical bars denote new iterations, where new galaxies are acquired and added to the training set. Prior to 2000 training iterations, both the random selection (baseline) models and active learning models train on only the initial random training set of 256 galaxies, and hence show similar performance. Around 2000 to 3500 iterations, after acquiring 128–256 additional galaxies, the active learning model shows a clear improvement in evaluation performance over the baseline model. We annotate in red where each model achieves the maximal relative RMSE improvement, highlighting the reduction in newly labelled galaxies required (vertical bars = 128 new galaxies). Note that active learning leads to a dramatically higher training loss, indicating that more challenging galaxies are being identified as informative and added to the training set.

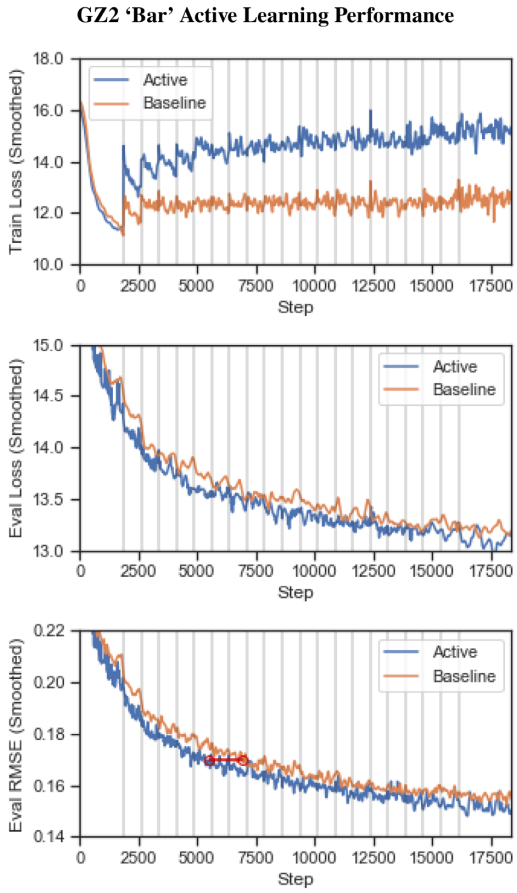


Figure 12. As with 11, but for the 'Bar' active learning simulations. Again, active learning leads to a clear improvement in evaluation performance and a dramatically higher training loss (indicating challenging galaxies are being selected). We annotate in red where each model achieves the maximal relative RMSE improvement, highlighting the reduction in newly labelled galaxies required (vertical bars $= 256$ new galaxies).

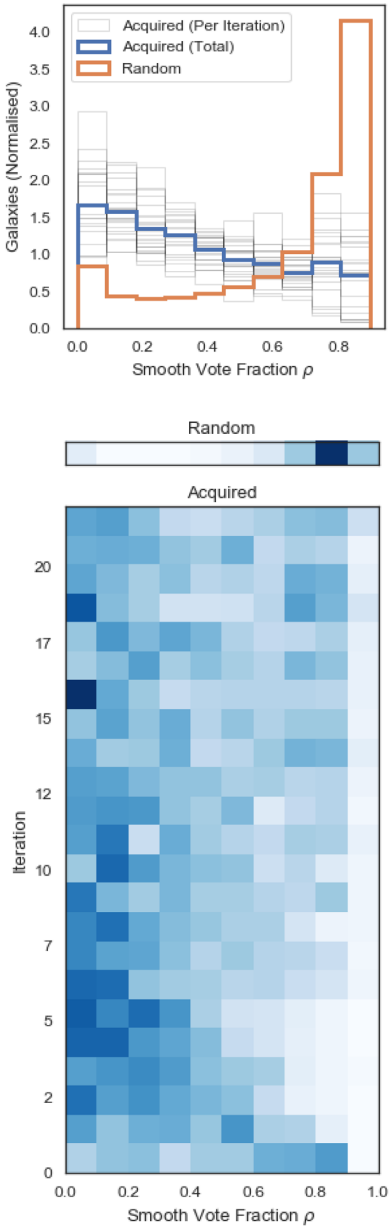


Figure 13. Distribution of observed 'Smooth' vote fraction p in galaxies acquired during Galaxy Zoo 'Smooth or Featured' active learning simulation. Above: Distribution of acquired p over all iterations, compared against random selection. While randomly selected galaxies are highly smooth, our acquisition function selects galaxies from across the p range, with a moderate preference towards featured. Below: Distribution of p by iteration, compared against random selection (upper inset). Our acquisition function strongly prefers featured galaxies in early ($n < \sim 7$) iterations, and then selects a more balanced sample. This likely compensates for the initial training sample being highly smooth.

galaxies; Fig. 12). Active learning outperforms random selection in every run.

Given sufficient (∼3000 for 'Smooth', ∼6000 for 'Bar') galaxies, our models eventually converge to similar performance levels – regardless of galaxy selection. We speculate that this is because our relatively shallow model architecture places an upper limit on performance. In general, model complexity should be large enough to exploit the information in the training set yet small enough to avoid fitting to spurious patterns. Model complexity increases with the number of free parameters, and decreases with regularization (Friedman, Hastie & Tibshirani 2001). Our model is both shallow and well-regularized (recall that dropout was originally used as a regularization technique, Section 2.3). A more complex (deeper) model may be able to perform better by learning from additional galaxies.

3.6.1 Selected galaxies

Which galaxies do the models identify as informative? To investigate, we randomly select one 'Smooth or Featured' and one 'Bar' simulation.

For the 'Smooth or Featured' simulation, Fig. 13 shows the observed 'Smooth' vote fraction distribution, per iteration (set of new galaxies) and in total (summed over all new galaxies). Highly smooth galaxies are common in the general GZ2 catalogue. Random selection therefore leads to a training sample skewed towards highly smooth galaxies. In contrast, our acquisition function is far more likely to select galaxies which are featured, leading to a more balanced sample. This is especially true for the first few iterations; we speculate that this counteracts the skew towards smooth in the randomly selected initial training sample. By the final training sample, featured galaxies become moderately more common than

smooth (mean $\frac{k_{\text{smooth}}}{N} = 0.38$). This suggests that featured galaxies are (on average) more informative for the model – over and above correcting for the skewed initial training sample. We speculate that featured galaxies may be more visually diverse, leading to a greater challenge in fitting volunteer responses, more disagreement between dropout-approximated-models, and ultimately higher mutual information.

For the 'Bar' simulation, Fig. 14 shows the 'Bar' vote fraction distribution, per iteration and in total, as well as the total redshift distribution. Again, our acquisition function selects a more balanced sample by prioritizing (rarer) barred galaxies. This selection remains approximately constant (within statistical noise) as more galaxies are acquired. With respect to redshift, our acquisition function prefers to select galaxies at lower redshifts. Based on inspection of the selected images (Fig. 15), we suggest that these galaxies are more informative to our model because such galaxies are better resolved (i.e. less ambiguous) and more likely to be barred.

We present the most and least informative galaxies from the (fixed and never labelled) test subset for 'Smooth' (Fig. 16) and Bar (Fig. 15), as identified by our novel acquisition function and the final models from each simulation.

4 DISCUSSION

Learning from fewer examples is an expected benefit of both probabilistic predictions and active learning. Our models approach peak performance on remarkably few examples: 2816 galaxies for 'Smooth' and 5632 for 'Bar'. With our system, volunteers could complete Galaxy Zoo 2 in weeks[13] rather than years if the peak performance of our models would be sufficient for their research. Further, reaching peak performance on relatively few examples indicates that an expanded model with additional free parameters is likely to perform better (Murphy 2012).

For this work, we rely on GZ2 data where N (the number of responses to a galaxy) is unknown before making a (historical) classification request. Therefore, when deriving our acquisition function, we approximated N as $\langle N \rangle$ (the expected number of responses). However, during live application of our system, we can control the Galaxy Zoo classification logic to collect exactly N responses per image, for any desired N. This would allow our model to request (for example) one more classification for *this* galaxy, and three more for *that* galaxy, before retraining. Precise classification requests from our model will enable us to ask volunteers exactly the right questions, helping them make an even greater contribution to scientific research.

We also hope that this human–machine collaboration will provide a better experience for volunteers. Inspection of informative galaxies (Figs 13, 14) suggests that more informative galaxies are more diverse than less informative galaxies. We hope that volunteers will find these (now more frequent) informative galaxies interesting and engaging.

Our results motivate various improvements to the probabilistic morphology models we introduce. In Section 2.7, we showed that our models were approximately well-calibrated, particularly after applying MC Dropout. However, the calibration was imperfect; even after applying MC Dropout, our models remain slightly overconfident (Fig. 6). We suggest two reasons for this remaining

[13]For example, classifying ∼10 000 galaxies (sufficient to train our models to peak performance) at the mean GZ2 classification rate of ∼ 800 galaxies per day would take ∼ 13 d.

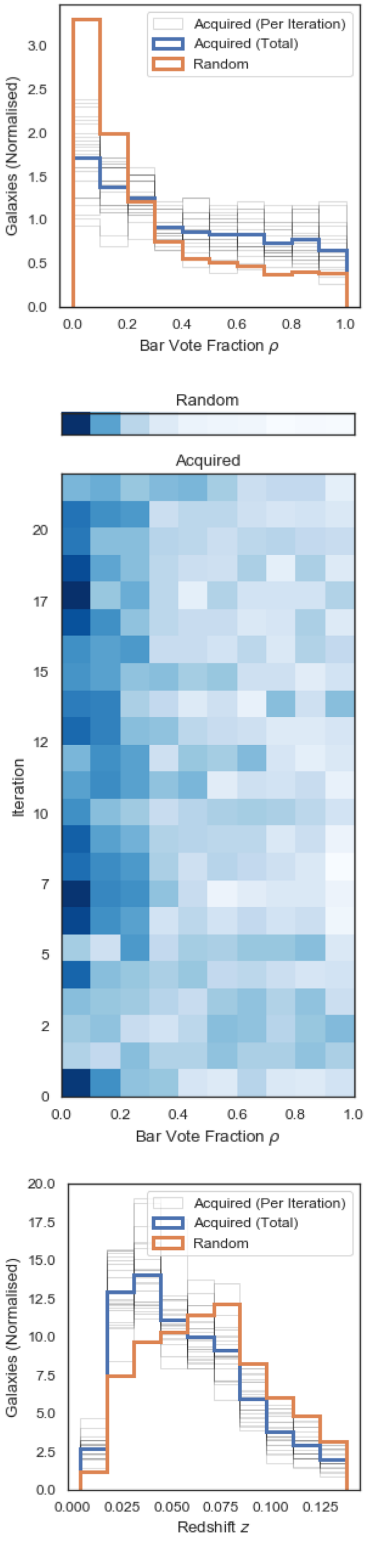
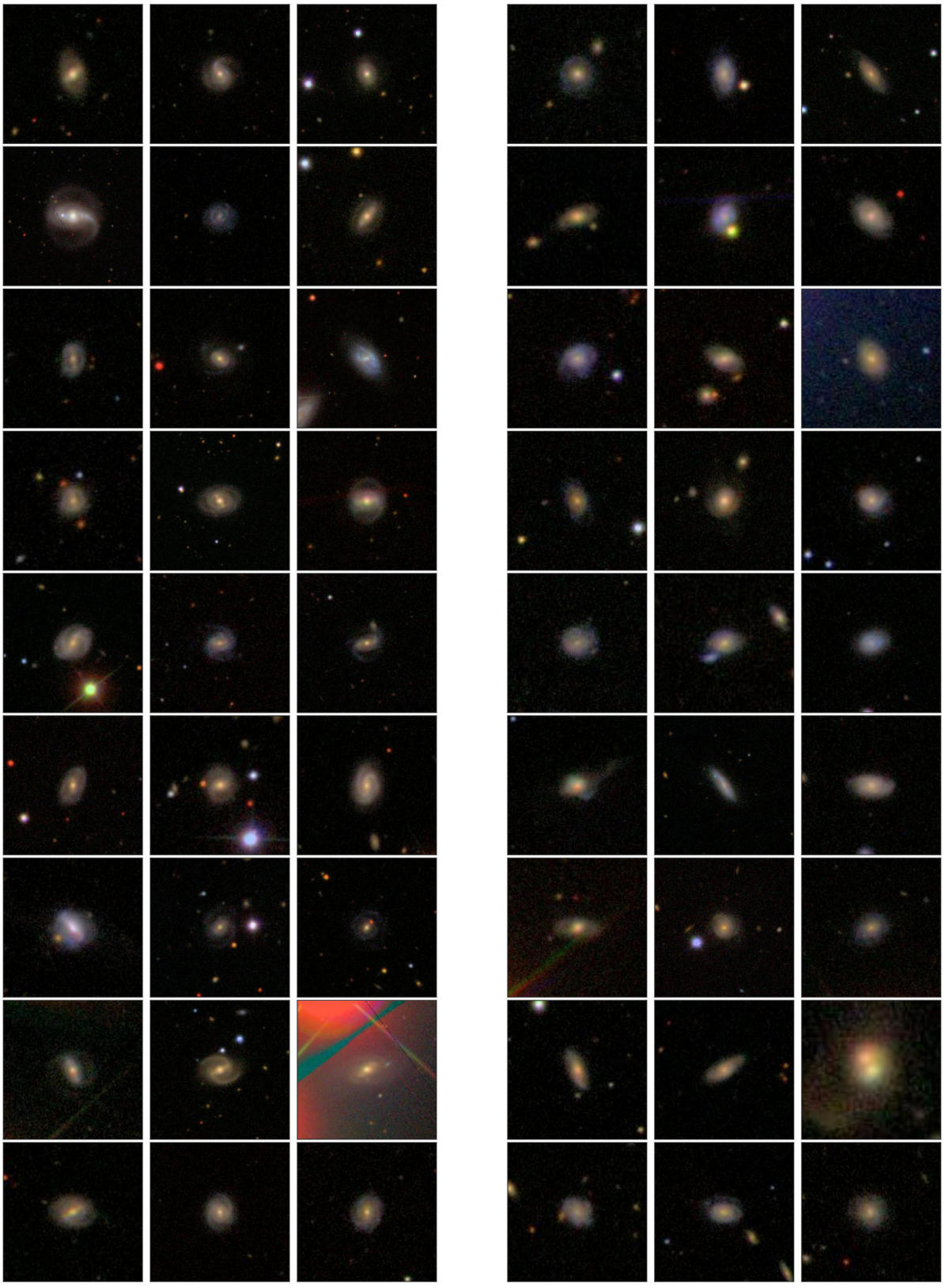


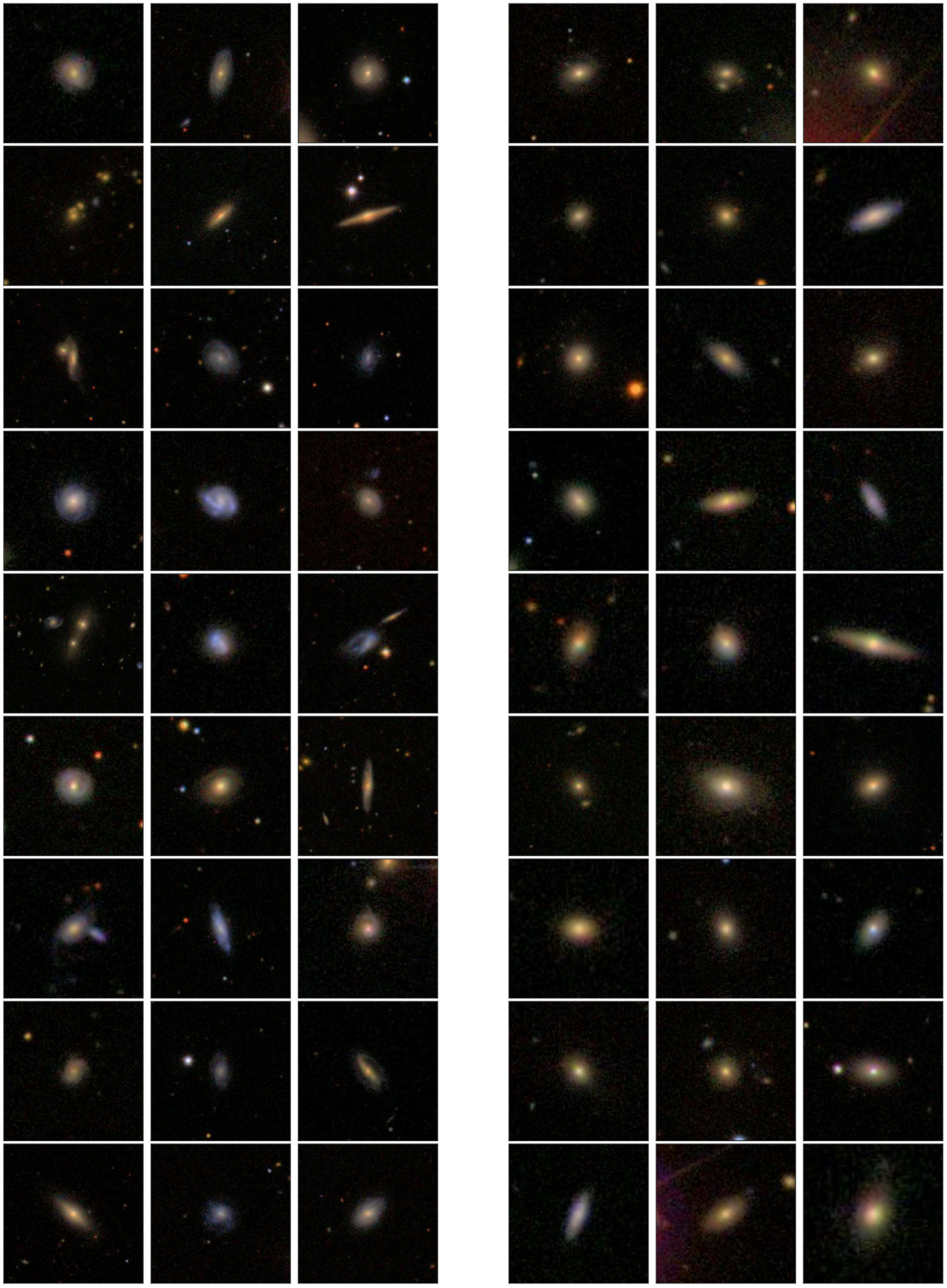
Figure 14. Upper: Distribution of observed 'Bar' vote fraction p in galaxies acquired during Galaxy Zoo 'Bar' active learning simulation. While randomly selected galaxies are highly non-barred, the 'Bar' model selects a more balanced sample. Middle: Distribution of 'Bar' p by iteration, compared against random selection (upper inset). Our acquisition function selects a similar *rho* distribution at each iteration. Lower: Redshift distribution of acquired galaxies over all iterations, compared against random selection. The 'Bar' model selects lower redshift galaxies, which are both more featured and better resolved (i.e. less visually ambiguous).



(a) Galaxies with maximum mutual information for 'Bar'

(b) Galaxies with minimum mutual information for 'Bar'

Figure 15. As with Fig. 16 above, but showing galaxies identified by the final model from a 'Bar' simulation.



(a) Galaxies with maximum mutual information for ‘Smooth or Featured’ (b) Galaxies with minimum mutual information for ‘Smooth or Featured’

Figure 16. Informative and uninformative galaxies from the (hidden) test subset, as identified by our novel acquisition function and the final model from a ‘Smooth or Featured’ simulation. When active learning is applied to Galaxy Zoo, volunteers will be more frequently presented with the most informative images (left-hand panel) than the least (right-hand panel).

overconfidence. First, within the MC Dropout approximation, the dropout rate is known to affect the calibration of the final model (Gal, Hron & Kendall 2017b). We choose our dropout rate arbitrarily (0.5); however, this rate may not sufficiently vary the model to approximate training many models. One solution is to 'tune' the dropout rate until the calibration is correct (Gal et al. 2017b). Secondly, the MC Dropout approximation is itself imperfect; removing random neurons with dropout is not identical to training many networks. As an alternative, one could simply train several models and ensemble the predictions (Lakshminarayanan, Pritzel & Blundell 2016). Both of these approaches are straightforward given a sufficient computational budget.

We also showed the distribution of model predictions over all galaxies generally agrees well with the distribution of predictions from volunteers (i.e. we are globally unbiased, Section 2.7). However, we noted that the models are 'reluctant' to predict extreme ρ (the typical response probability, Section 2.1). We suggest that this is a limitation of our generative model for volunteer responses. The binomial likelihood becomes narrow when p (here, ρ) is extreme, and hence the model is heavily penalized for incorrect extreme p estimates. If volunteer responses were precisely binomially distributed (i.e. N independent identically distributed trials per galaxy, each with a fixed p of a positive response), this heavy penalty would correctly reflect the significance of the error. However, our binomial model of volunteers is only approximate; one volunteer may give consistently different responses to another. In consequence, the true likelihood of non-extreme k responses given ρ is wider than the binomial likelihood from the 'typical' response probability ρ suggests, and the network is penalized 'unfairly'. The network therefore learns to avoid making risky extreme predictions.

If this suggestion is correct, the risk-averse prediction shift will be monotonic (i.e. extreme galaxies will have slightly different ρ but still be ranked in the same order) and hence researchers selecting galaxies near extreme ρ may simply choose a slightly higher or lower $\hat{\rho}$ threshold. To resolve this issue, one could apply a monotonic rescaling to the network predictions (as we do in Appendix A), introduce a more sophisticated model of volunteer behaviour (Marshall et al. 2016; Beck et al. 2018; Dickinson et al. 2019), or calibrate the loss to reflect the scientific utility of extreme predictions (Cobb, Roberts & Gal 2018). As predictions are globally unbiased for all non-extreme ρ, and extreme ρ predictions can be corrected post hoc (above), our network is ready for use.

Throughout this work, our goal has been to predict volunteer responses at scale. These responses are known to vary systematically with e.g. redshift (Willett et al. 2013; Hart et al. 2016) and colour (Cabrera-Vives, Miller & Schneider 2018), and hence require calibration prior to scientific analysis. Unlike Domínguez Sánchez et al. (2018) and Khan et al. (2019), who train on redshift-calibrated 'debiased' responses, we expect and intend to reproduce these systematics. We prefer to apply calibration methods to our predictions. A calibrated CNN-predicted catalogue will be presented as part of a future Galaxy Zoo data release.

Finally, we highlight that our approach is highly general. We hope that Bayesian CNNs and active learning can contribute to the wide range of astrophysical problems where CNNs are applicable (e.g. images, time-series), uncertainty is important, and the data is expensive to label, noisy, imbalanced, or includes rare objects of interest. In particular, imbalanced data sets (where some labels are far more common than others) are common throughout astrophysics. Topics include transient classification (Wright et al. 2017), fast radio burst searches (Zhang et al. 2018), and exoplanet detection (Osborn et al. 2019). Active learning is known to be effective at correcting such imbalances (Ishida et al. 2019). Our results suggest that this remains true when active learning is combined with CNNs (this work is the first astrophysics application of such a combination). Recall that smooth galaxies are far more common in GZ2 but featured galaxies are strongly preferentially selected by active learning – automatically, without our instruction – apparently to compensate for the imbalanced data (Fig. 13). If this observation proves to be general, we suggest that Bayesian CNNs and active learning can drive intelligent data collection to overcome research challenges throughout astrophysics.

5 CONCLUSION

Previous work on predicting visual galaxy morphology with deep learning has either taken no account of uncertainty or trained only on confidently labelled galaxies. Our Bayesian CNNs model exploit the uncertainty in Galaxy Zoo volunteer responses using a novel generative model of volunteers. This enables us to accurately answer detailed morphology questions using only sparse labels ($\sim$10 responses per galaxy). Our CNNs can also express uncertainty by predicting probability distribution parameters and using Monte Carlo Dropout (Gal et al. 2017a). This allows us to predict posteriors for the expected volunteer responses to each galaxy. These posteriors are reliable (i.e. well-calibrated), show minimal systematic bias, and match or outperform previous work when reduced to point estimates (for comparison). Using our posteriors, researchers will be able to draw statistically powerful conclusions about the relationships between morphology and AGN, mass assembly, quenching, and other topics.

Previous work has also treated labelled galaxies as a fixed data set from which to learn. Instead, we ask: which galaxies should we label to train the best model? We apply active learning (Houlsby et al. 2011) – our model iteratively requests new galaxies for human labelling and then retrains. To select the most informative galaxies for labelling, we derive a custom acquisition function for Galaxy Zoo based on BALD (MacKay 1992). This derivation is only possible using our posteriors. We find that active learning provides a clear improvement in performance over random selection of galaxies. The galaxies identified as informative are generally more featured (for the 'Smooth or Featured' question) and better resolved (for the 'Bar' question), matching our intuition.

As modern surveys continue to outpace traditional citizen science, probabilistic predictions and active learning become particularly crucial. The methods we introduce here will allow Galaxy Zoo to produce visual morphology measurements for surveys of any conceivable scale on a time-scale of weeks. We aim to launch our active learning strategy on Galaxy Zoo in 2019.

ACKNOWLEDGEMENTS

MW would like to thank H. Domínguez Sanchez and M. Huertas-Company for helpful discussions.

MW acknowledges funding from the Science and Technology Funding Council (STFC) Grant Code ST/R505006/1. We also acknowledge support from STFC under grant ST/N003179/1. LF, CS, HD, and DW acknowledge partial support from one or more of the US National Science Foundation grants IIS-1619177, OAC-1835530, and AST-1716602.

This research made use of the open-source PYTHON scientific computing ecosystem, including SCIPY (Jones et al. 2001), MATPLOTLIB (Hunter 2007), SCIKIT-LEARN (Pedregosa et al. 2011),

SCIKIT-IMAGE (van der Walt et al. 2014), and PANDAS (McKinney 2010).

This research made use of ASTROPY, a community-developed core PYTHON package for Astronomy (The Astropy Collaboration 2013, 2018).

This research made use of TENSORFLOW (Abadi et al. 2015).

All code is publicly available on GitHub at www.github.com/mwalmsley/galaxy-zoo-bayesian-cnn (Walmsley 2019).

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available at: https://www.tensorflow.org/

Abazajian K. N. et al., 2009, ApJS, 182, 543

Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415

Aihara H. et al., 2018, PASJ, 70, S8

Albareti F. D. et al., 2017, ApJS, 233, 25

Baillard A., Bertin E., Lapparent V. D., Fouqué P., Arnouts S., Mellier Y., Pelló R., Leborgne J., 2011, A&A, 532, A74

Banerji M. et al., 2010, MNRAS, 406, 342

Beck M. R. et al., 2018, MNRAS, 476, 5516

Cabrera-Vives G., Miller C. J., Schneider J., 2018, AJ, 156, 284

Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N., 2015, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Sydney, Australia, p. 1721

Cheng T.-Y., 2019, Optimising Automatic Morphology Classification of Galaxies with Machine Learning and Depp Learning using Dark Enery Survey Imaging, preprint (arXiv:1908.03610)

Cleveland W. S., 1979, J. Am. Stat. Assoc., 74, 829

Cobb A. D., Roberts S. J., Gal Y., 2018, Loss-Calibrated Approximate Inference in Bayesian Neural Networks, preprint (arXiv:1805.03901)

Conselice C. J., 2003, ApJS, 147, 1

de Jong J. T. A. et al., 2015, A&A, 582, A62

Dey A. et al., 2019, AJ, 157, 168

Dickinson H., Fortson L., Scarlata C., Beck M., Walmsley M., 2019, in Boquien M., Lusso E., Gruppioni C., Tissera P., eds, Proc. IAU Symp. 341, Challenges in Panchromatic Galaxy Modelling with Next Generation Facilities. Int. Astron. Un., Paris

Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441

Domínguez Sánchez H., Huertas-Company M., Bernardi M., Tuccillo D., Fischer J. L., 2018, MNRAS, 476, 3661 (DS + 18)

Domínguez Sánchez H. et al., 2019, MNRAS, 484, 93

Fischer J.-L., Dom H., Bernardi M., MNRAS, 2018, 483, 2057

Flaugher B., 2005, Int. J. Mod. Phys. A, 20, 3121

Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M., 2013, MNRAS, 434, 282

Friedman J., Hastie T., Tibshirani R., 2001, The Elements of Statistical Learning. Springer, New York

Gal Y., 2016, PhD thesis. University of Cambridge

Gal Y., Islam R., Ghahramani Z., 2017a, Proceedings of the 24th International Conference on Machine Learning, Sydney, Australia, 70, 1183

Gal Y., Hron J., Kendall A., 2017b, Advances in Neural Information Processing Systems 30 (NIPS). Curran Associates, Inc, p. 3581, Available at: http://papers.nips.cc/paper/6949-concrete-dropout.pdf

Galloway M. A. et al., 2015, MNRAS, 448, 3442

Gordon Y. A. et al., 2019, ApJ, 878, 88

Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, Int. Conf. Mach. Learn., 70, 1321

Hart R. E. et al., 2016, MNRAS, 461, 3663

Hastie T. J., Tibshirani R., 1990, Generalized Additive Models. 1 edn., Chapman and Hall, London

He K., Zhang X., Ren S., Sun J., 2016, in The IEEE Conference on Computer Vision and Pattern Recognition, p. 770

Hezaveh Y. D., Levasseur L. P., Marshall P. J., 2017, Nature, 548, 555

Hocking A., Geach J. E., Davey N., Sun Y., 2015, MNRAS, 473, 1108

Houlsby N., Huszár F., Ghahramani Z., Lengyel M., 2011, PhD thesis, Universty of Cambridge

Hoyle B., Paech K., Rau M. M., Seitz S., Weller J., 2016, MNRAS, 458, 4498

Huang G., Liu Z., van der Maaten L., Weinberger K. Q., 2017, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Long Beach, USA, p. 4700

Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Almeida J. S., 2011, A&A, 525, 1

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, 99

Ishida E. E. O. et al., 2019, MNRAS, 483, 2

Jones E., Oliphant T., Peterson P. et al., 2001, SciPy: Open source scientific tools for Python. Available at: http://www.scipy.org/

Kaiser N. et al., 2010, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Conf. Ser. Vol. 7733, Ground-based and Airborne Telescopes III. SPIE, Bellingham, 77330E

Khan A., Huerta E. A., Wang S., Gruendl R., Jennings E., Zheng H., 2019, Phys. Lett. B, 795, 248 (K + 18)

Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463

Kruk S. J. et al., 2017, MNRAS, 469, 3363

Kruk S. J. et al., 2018, MNRAS, 473, 4731

Lakshminarayanan B., Pritzel A., Blundell C., 2016, Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, preprint (arXiv:1612.01474)

Lanusse F., Ma Q., Li N., Collett T. E., Li C. L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, MNRAS, 473, 3895

Laureijs R., et al., 2011, Euclid Definition Study Report, Report number ESA/SRE(2011)12, preprint (arXiv:1110.3193)

LeCun Y. A., Bengio Y., Hinton G. E., 2015, Nature, 521, 436

Lintott C. J. et al., 2008, MNRAS, 389, 1179

Liu P., Zhang H., Eom K. B., 2017, IEEE J. Topics Appl. Earth Obs. Remote Sensing, 10, 712

Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163

Lou Y., Caruana R., Gehrke J., 2012, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, p. 150

LSST Science Collaboration, 2009, LSST Science Book, Version 2.0, preprint (arXiv:0912.0201)

Lu J., Behbood V., Hao P., Zuo H., Xue S., Zhang G., 2015, Knowledge-Based Systems, 80, 14

MacKay D. J. C., 1992, Neural Comput., 4, 590

Marshall P. J. et al., 2016, MNRAS, 455, 1171

Masters K. L. et al., 2012, MNRAS, 424, 2180

McKinney W., 2010, Data Structures for Statistical Computing in Python. Available at: http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Murphy K. P., 2012, Machine Learning: A Probabilisitic Perspective. MIT Press, Boston, MA

Nair P. B., Abraham R. G., 2010, ApJS, 186, 427

Osborn H. P. et al., 2019, Rapid Classification of TESS Planet Candidates with Convolutional Neural Networks, preprint (arXiv:1902.08544)

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Pérez-Carrasco M., Cabrera-Vives G., Martinez-Marin M., Cerulo P., Demarco R., Protopapas P., Godoy J., Huertas-Company M., 2019, PASP, 131, 108002

Peth M. A. et al., 2016, MNRAS, 458, 963

Rahhal M. A., Bazi Y., AlHichri H., Alajlan N., Melgani F., Yager R., 2016, Inform. Sci., 345, 340

Richards J. W. et al., 2012, ApJ, 744, 192

Roberts M. S., Haynes M. P., 1994, ARA&A, 32, 115

Russakovsky O. et al., 2015, Int. J. Comput. Vision, 115, 211

Scarlata C. et al., 2007, ApJS, 172, 406

Simonyan K., Zisserman A., 2015, in Bengio Y., LeCun Y., eds, 3rd International Conference on Learning Representations, San Diego, CA, USA, preprint (arXiv:1409.1556)

Solorio T., Fuentes O., Terlevich R., Terlevich E., 2005, MNRAS, 363, 543

Spergel D. et al., 2013, WFIRST-2.4: What Every A stronomer Should Know, preprint (arXiv:1305.5425)
Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, J. Mach. Learn. Res., 15, 1929
Strauss M. A. et al., 2002, AJ, 124, 1810
Szegedy C. et al., 2015, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Colombus, OH, USA, preprint (arXiv:1409.4842)
The Astropy Collaboration, 2013, A&A, 558, A33
The Astropy Collaboration, 2018, AJ, 156, 123
Tuccillo D., Huertas-Company M., Decencière E., Velasco-Forero S., Sánchez H. D., Dimauro P., 2018, MNRAS, 475, 894
Tuia D., Volpi M., Copa L., Kanevski M., Munoz-Mari J., 2011, IEEE J. Select. Topics Signal Process., 5, 606
van der Walt S., Schönberger J. L., Nunez-Iglesias J., Boulogne F., Warner J. D., Yager N., Gouillart E., Yu T., 2014, PeerJ, 2, e453
Walmsley M., 2019, Galaxy Zoo Bayesian CNN: Initial public release, doi:10.5281/ZENODO.2677874
Walmsley M., Ferguson A. M. N., Mann R. G., Lintott C. J., 2018, MNRAS, 483, 2968
Wang L. et al., A&A, 2018, 618, A1
Willett K. W. et al., 2013, MNRAS, 435, 2835
Wright D. E. et al., 2017, MNRAS, 472, 1315
Xia X., Protopapas P., Doshi-Velez F., 2016, in Venkatasubramanian S. C., Meira W., eds, Proceedings of the 2016 SIAM International Conference on Data Mining. SIAM, Philadelphia, p. 477
Zhang Y. G., Gajjar V., Foster G., Siemion A., Cordes J., Law C., Wang Y., 2018, ApJ, 866, 149
Zhou S., Chen Q., Wang X., 2013, Neurocomputing, 120, 536

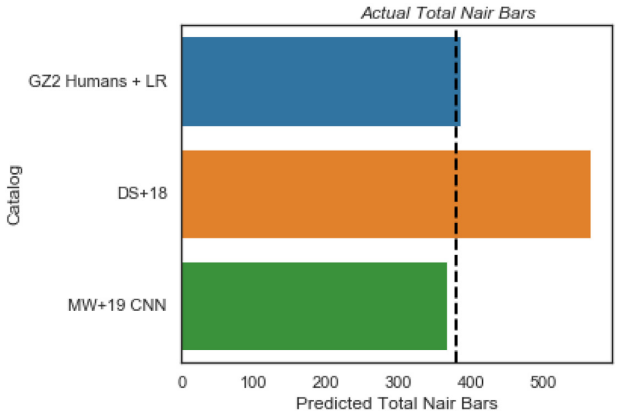


Figure A1. Predictions for the total number of galaxies labelled as 'Bar' by human expert N10 in test galaxy subset (correct answer: 379). DS + 18 overestimates the number of Nair Bars (559). We find that GZ2 vote fractions from volunteers can be used to make an improved estimate (396) with a rescaling correction calculated via logistic regression (GZ2 Humans + LR). Applying the same correction to the vote fractions predicted by the Bayesian CNN in this work (MW + 19) also produces an improved estimate (372). By accurately predicting the vote fractions, and then applying a correction to map from vote fractions to expert responses, we can predict what N10 would have said for the full SDSS sample.

APPENDIX A:

CNN predictions are not (in general) well-calibrated probabilities (Lakshminarayanan et al. 2016; Guo et al. 2017). Interpreting them as such may cause systematic errors in later analysis. To illustrate this problem, we show how the CNN probabilities published in DS + 18 (Domínguez Sánchez et al. 2018) significantly overestimate the prevalence of expert-classified barred galaxies. We chose DS + 18 as the most recent deep learning morphology catalogue made publicly available, and thank the authors for their openness. We do not believe this issue is unique to DS + 18. We highlight this issue not as a criticism of DS + 18 specifically, but to emphasize the advantages of using probabilistic methods.

DS + 18 trained a CNN to predict the probability that a galaxy is barred (DS + 18 Section 5.3). Barred galaxies were defined as those galaxies labelled as having any kind of bar (weak/intermediate/strong) in expert catalogue (Nair & Abraham 2010, N10). We refer to such galaxies as Nair Bars. We chose to investigate this particular DS + 18 model because it explicitly aims to reproduce the (expert) N10 classifications, allowing for direct comparison of the predicted probabilities against the true labels.

We first show that these CNN probabilities are not well-calibrated. We then demonstrate a simple technique to infer probabilities for Nair Bars from GZ2 vote fractions. Finally, we show that, as our Bayesian CNN estimates of GZ2 vote fractions are well-calibrated, these vote fractions can be used to estimate probabilities for Nair Bars. The practical application is to predict what Nair & Abraham (2010) would have recorded, had the expert authors visually classified every SDSS galaxy.

We select a random subset of 1211 galaxies classified by N10 (this subset is motivated below). How many barred galaxies are in this subset? The DS + 18 Nair Bar 'probabilities' p_i (for each galaxy i) predict $\sum_i p_i = 559$ Nair Bars. However, only 379 are actually Nair Bars (Fig. A1). This error is caused by the DS + 18 Nair Bar 'probabilities' being, on average, skewed towards predicting 'Bar', as shown by the calibration curve of the DS + 18 Nair Bar probabilities (Fig. A2).

How can we better predict the total number of Nair Bars? GZ2 collected volunteer responses for many galaxies classified by N10 (6051 of 14 034 match within 5 arcsec, after filtering for total 'Bar?' votes $N_{\rm bar} > 10$ as in Section 2.6). The fraction of volunteers who responded 'Bar' to the question 'Bar?' is predictive of Nair Bars, but is not a probability (Lintott et al. 2008). For example, volunteers are less able to recognize weak bars than experts (Masters et al. 2012), and hence the 'Bar' vote fraction only slightly increases for galaxies with weak Nair Bars versus galaxies without. We need to rescale the GZ2 vote fractions. To do this, we divide the N10 catalogue into 80 per cent train and 20 per cent test subsets and use the train subset to fit (via logistic regression) a rescaling function (Fig. A3) mapping GZ2 vote fractions to p(NairBar|GZ2Fraction). We then evaluate the calibration of these probabilities on the test subset, which is the subset of 1211 galaxies used above. We predict 396 Nair Bars, which compares well with the correct answer of 379 versus the DS + 18 answer of 559 (Fig. A1). This directly demonstrates that our rescaled GZ2 predictions are correctly calibrated over the full test subset. The calibration curve shows no systematic skew, unlike DS + 18 (Fig. A2).

Since the GZ2 vote fractions can be rescaled to Nair Bar probabilities, and the Bayesian CNN makes predictions of the GZ2 vote fractions, we can also rescale the Bayesian CNN predictions into Nair Bar probabilities using the same rescaling function. The rescaled Bayesian CNN GZ2 vote predictions correctly estimate the count of Nair Bars (372 bars predicted versus 379 observed bars; Fig. A1).

Finally, we note that if the research goal is simply to identify samples of e.g. Nair Bars, one can do so by interpreting each prediction as a score (i.e. an arbitrary scalar, as opposed to a probability). When interpreted as scores, the rescaled GZ2 votes –

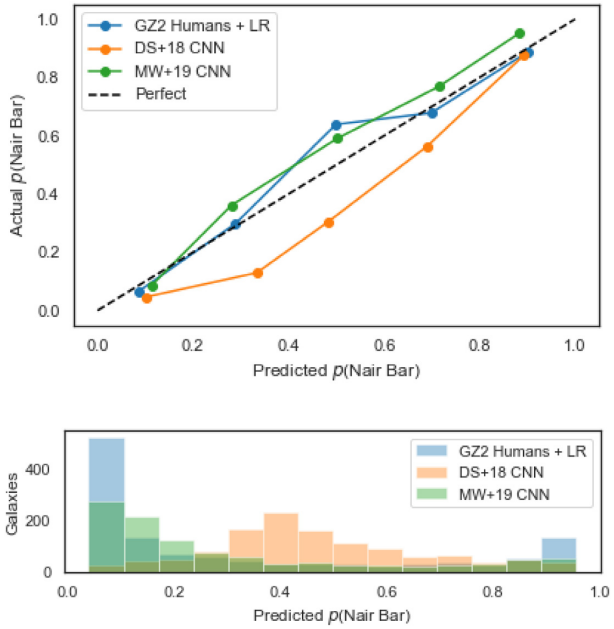


Figure A2. Above: Comparison of calibration curves for each predictive model. The calibration curve is calculated by binning the predicted probabilities and counting the fraction of Nair Bars in each bin. The fraction of Nair Bars in a given bin approximates the true (frequentist) probability of each binned galaxy being a Nair Bar. Points compare the predicted fraction of Nair Bars (x axis) with the actual fraction (y axis) for five equally spaced bins. For well-calibrated probabilities, the predicted and actual fractions are equal (black dashed line). Below: the distribution of Nair Bar predictions from each model. DS + 18 typically predicts $p \sim 0.4$ (below) and has a relatively poor calibration near $p \sim 0.4$ (above), leading to a significant overestimate of the total number of Nair Bars.

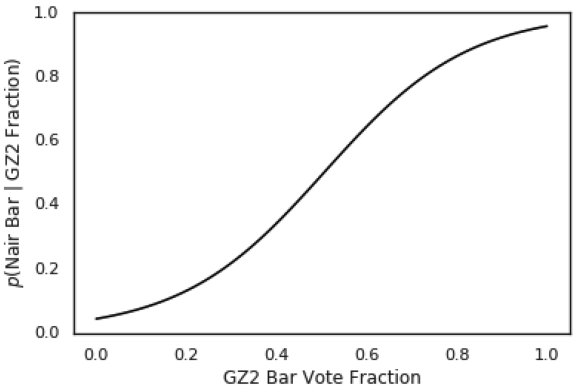


Figure A3. The rescaling function used to map GZ2 vote fractions to p(NairBar|GZ2Fraction), estimated via logistic regression. This rescaling function is also used (without modification) to map Bayesian CNN GZ2 vote fraction predictions to p(NairBar|BCNN-predictedGZ2Fraction).

both observed from volunteers and predicted by the Bayesian CNN – outperform DS + 18 in identifying Nair Bars at all thresholds (Fig. A4). This may be because our BCNN can learn to detect bars from the extensive GZ2 sample (56 048 galaxies with $N_{\rm bar} \geq 10$) before those predictions are rescaled to correspond to Nair Bars, rather than DS + 18's approach of training only on the much smaller set of galaxies (7000) directly labelled in Nair & Abraham (2010).

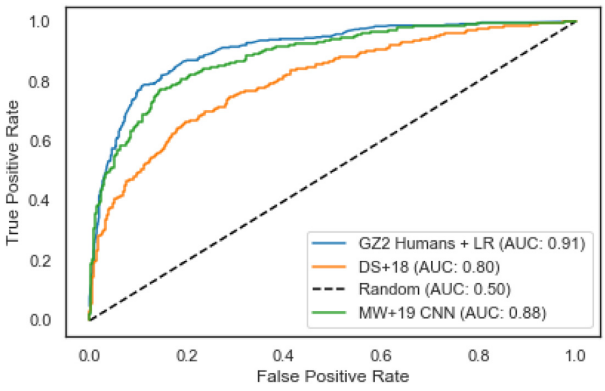


Figure A4. Comparison of ROC curves for predicting Nair Bars using each model.

Nair Bars are initially defined through repeated expert classification (as close to 'gold standard' ground truth as exists for imaging data) and hence accurate automated identification of Nair Bars is directly useful for morphology research.

APPENDIX B: THEORETICAL BACKGROUND ON VARIATIONAL INFERENCE

The general problem of Bayesian inference can be framed in terms of a probabilistic model where we have some observed random variables Z and some latent variables θ and we wish to infer $P(\theta|Z)$ after observing some data. Our probabilistic model $P(Z, \theta)$ allows us to use Bayes rule to do so; $P(\theta \mid Z) = \frac{P(\theta, Z)}{P(Z)} = \frac{P(Z|\theta)p(\theta)}{P(Z)}$. In the setting of discriminative learning, the observed variables are the inputs and outputs of our classification task X and Y, and we directly parametrize the distribution $P(y|x, \theta)$ in order to make predictions by marginalizing over the unknown weights, that is, the prediction for an unseen point x given training data X is

$$p(y \mid x, X, Y) = \int p(y \mid x, \theta) p(\theta \mid X, Y) \mathrm{d}\theta. \tag{B1}$$

While this is a simple framework, in practice the integrals required to normalize Bayes' rule and to take this marginal are often not analytically tractable, and we must resort to numerical approaches.

While there are many possible ways to perform approximate Bayesian inference, here we will focus on the framework of *variational inference*. The essential idea of variational inference is to approximate the posterior $P(\theta|Z)$ with a simpler distribution $q(\theta)$ which is 'as close as possible' to $P(\theta|Z)$, and then use q in place of the posterior. This can take the form of analytically finding the optimal q subject only to some factorization assumptions using the tools of the calculus of variations, but the case that is relevant to our treatment is when we fix q to be some family of distributions $q_\xi(\theta)$ parametrized by ξ and fit ξ, changing an integration problem to an optimization one.

The measure of 'as close as possible' used in variational inference in the Kullback–Leibler (KL) divergence, or the relative entropy, a measure of distance between two probability distributions defined as

$$D_{\rm KL}(p : q) = \int p(x)(\log p(x) - \log q(x)) \mathrm{d}x. \tag{B2}$$

The objective of variational inference is to choose the q such that $D_{\rm KL}(q(\theta): p(\theta|X))$ is minimized. Minimizing this objective can

be shown to be equivalent to maximizing the 'log Evidence Lower BOund', or ELBO,

$$L(q) = \mathbb{E}_{q(\theta)} - [\log p(Y \mid X, \theta)p(\theta) - \log q(\theta)]. \tag{B3}$$

The reason for the name is the relationship

$$\log P(X) = D_{\mathrm{KL}}(q(\theta) : p(\theta \mid X)) + L(q), \tag{B4}$$

which implies, since the KL divergence is strictly positive, that L provides a lower bound on the log of the evidence $P(X)$, the denominator in Bayes rule above. By optimizing the parameters of q *xi*, with respect to L, one can find the best approximation to the posterior in the family of parametrized distributions chosen in terms of the ELBO.

The key advantage of this formalism is that the ELBO only involves the tractable terms of the model, $P(X|\theta)$ and $P(\theta)$. The expectation is over the approximating distribution, but since we are able to choose q we can make a choice that is easy to sample from, and therefore it is straightforward to obtain a Monte Carlo approximation of L via sampling, which is sufficient to obtain stochastic gradients of L which can be used for optimization. The integral over the posterior on θ in the marginalization step can likewise be approximated via sampling from q if necessary.

For neural networks, a common approximating distribution is dropout (Srivastava et al. 2014). The dropout distribution over the weights of a single neural network layer is parametrized by a weight matrix M and a dropout probability p. Draws from this distribution are described by

$$W_{ij} = M_{ij} z_j, \tag{B5}$$

where $z_j \sim$ Bernoulli(p). Gal (2016) introduced approximating $p(w|\mathcal{D})$, with a dropout distributions over the weights of a network, and showed that in this case optimizing the standard likelihood based loss is equivalent to the variational objective that would be obtained for the dropout distribution, so we may interpret the dropout distribution over the weights of a trained model as an approximation to the posterior distribution $p(w \mid \mathcal{D})$.

We can use this approximating distribution as a proxy for the true posterior when we marginalize over models to make predictions;

$$\int p(k|x, w)p(w|\mathcal{D})\mathrm{d}w \approx \int p(k|x, w)q^*\mathrm{d}w. \tag{B6}$$

A more detailed mathematical exposition of dropout as variational inference can be found in Gal (2016).

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