Weight Pruning and Uncertainty in Radio Galaxy Classification

Devina Mohan
Department of Physics & Astronomy
University of Manchester, UK
devina.mohan@postgrad.manchester.ac.uk

Anna Scaife[*]
Department of Physics & Astronomy
University of Manchester, UK
anna.scaife@manchester.ac.uk

Abstract

In this work we use variational inference to quantify the degree of epistemic uncertainty in model predictions of radio galaxy classification and show that the level of model posterior variance for individual test samples is correlated with human uncertainty when labelling radio galaxies. We explore the model performance and uncertainty calibration for a variety of different weight priors and suggest that a sparse prior produces more well-calibrated uncertainty estimates. Using the posterior distributions for individual weights, we show that signal-to-noise ratio (SNR) ranking allows pruning of the fully-connected layers to the level of 30% without significant loss of performance, and that this pruning increases the predictive uncertainty in the model. Finally we show that, like other work in this field, we experience a cold posterior effect. We examine whether adapting the cost function in our model to accommodate model misspecification can compensate for this effect, but find that it does not make a significant difference. We also examine the effect of principled data augmentation and find that it improves upon the baseline but does not compensate for the observed effect fully. We interpret this as the cold posterior effect being due to the overly effective curation of our training sample leading to likelihood misspecification, and raise this as a potential issue for Bayesian deep learning approaches to radio galaxy classification in future.

1 Introduction

A new generation of radio astronomy facilities around the world such as the Low-Frequency Array [LOFAR; 1], the Murchison Widefield Array [MWA; 2], the MeerKAT telescope [3], and the Australian SKA Pathfinder (ASKAP) telescope [4] are generating increasingly larger and larger data rates. In order to extract scientific impact from these facilities on reasonable timescales, a natural solution has been to automate the data processing as far as possible and this has lead to the increased adoption of machine learning methodologies.

In particular for new sky surveys, automated classification algorithms are being developed to replace the *by eye* approaches that were possible historically. In radio astronomy specifically, studies looking at morphological classification using convolutional neural networks (CNNs) and deep learning have become increasingly common, in particular with respect to the classification of radio galaxies.

The Fanaroff-Riley (FR) classification of radio galaxies was introduced over four decades ago [5], and has since been widely adopted and applied to many catalogues since then. The morphological divide seen in this classification scheme has historically been explained primarily as a consequence of differing jet dynamics. Fanaroff-Riley type I (FRI) radio galaxies have jets that are disrupted at shorter distances from the central super-massive black hole host and are therefore centrally brightened,

[*] The Alan Turing Institute, 96 Euston Rd, London, UK a.scaife@turing.ac.uk

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

whilst Fanaroff-Riley type II (FRII) radio galaxies have jets that remain relativistic to large distances, resulting in bright termination shocks. These observed structural differences may be due to the intrinsic power in the jets, but will also be influenced by local environmental densities [6, 7].

Intrinsic and environmental effects are difficult to disentangle using radio luminosity alone as systematic differences in particle content, environmental effects and radiative losses make radio luminosity an unreliable proxy for jet power [8]. Hence the use of morphology for inferring the environmental impact on radio galaxy populations is therefore important for gaining a better physical understanding of the FR dichotomy, and of the full morphological diversity of the population.

From a deep-learning perspective, the ground work for morphological classification in this field was done by [9]. This was followed by other works involving the use of deep learning in source classification [e.g. 10, 11, 12]. More recently, [13] showed that an attention-gated CNN could perform classification of radio galaxies with equivalent performance to other applications in the literature, but using $\sim$50% fewer learnable parameters, [14] showed that using group-equivariant convolutional layers that preserved the rotational and reflectional isometries of the Euclidean group resulted in improved overall model performance and stability of model confidence for radio galaxies at different orientations, and [15] generated synthetic populations of radio galaxies using structured variational inference.

With the exception of [14], there has been little work done on understanding the degree of confidence with which CNN models predict the class of individual radio galaxies; however, for radio astronomy, where modern astrophysical analysis is driven by population analyses, quantifying the confidence with which each object is assigned to a particular classification is crucial for understanding the propagation of uncertainties within that analysis.

This work In this work we use variational inference (VI) to quantify the degree of epistemic uncertainty in model predictions of radio galaxy classification. This differs from the approach of [14] who used dropout as a Bayesian approximation [16] to estimate model confidence as a function of image orientation, which is just one specific aspect of model performance and not directly comparable to this work. We compare the variance of our posterior predictions to qualifications present in our test data that indicate the level of human confidence in assigning a classification label and show that model uncertainty is correlated with human uncertainty.

An added advantage of Bayesian CNNs is their ability to identify the significance of individual weights within the model. In this work we show that weight signal-to-noise ratio (SNR) allows pruning of our fully-connected layers to the level of 30% without significant loss of model performance. We also find that pruning based on a Fisher analysis is able to thin our model more effectively, to the level of 60%, but that both pruning methods increase the uncertainty calibration error of the model in different ways.

Finally we show that, like other work in this field, we experience a *cold posterior* effect [see e.g. 17]. As suggested by [18], we examine whether a PAC Bayes approach adapting the cost function in our model to accommodate model misspecification can compensate for this effect, but find that it does not make a significant difference. Other works in this area have suggested that unprincipled data augmentation could be a contributing factor to the cold posterior effect [e.g. 19, 20]. We examine the effect of principled data augmentation and find that the cold posterior effect observed in our work *reduces* slightly with data augmentation. Consequently we suggest that the cold posterior effect in this case is likely to be due to the manner in which the training sample is curated, rather than model misspecification.

2 Uncertainty in radio galaxy classification

For this work we use the MiraBest radio galaxy dataset[2] [21], which is based on the catalogue of [22], who used a parent galaxy sample taken from [23] that cross-matched the Sloan Digital Sky Survey [SDSS; 24] data release 7 [DR7; 25] with the Northern VLA Sky Survey [NVSS; 26] and the Faint Images of the Radio Sky at Twenty centimetres [FIRST; 27]. The parameters of the dataset itself and the image preprocessing are described in [14]. For this work we extract the objects labelled as

[2] The data used in this work is provided under a Creative Commons license at `https://zenodo.org/record/4288837`

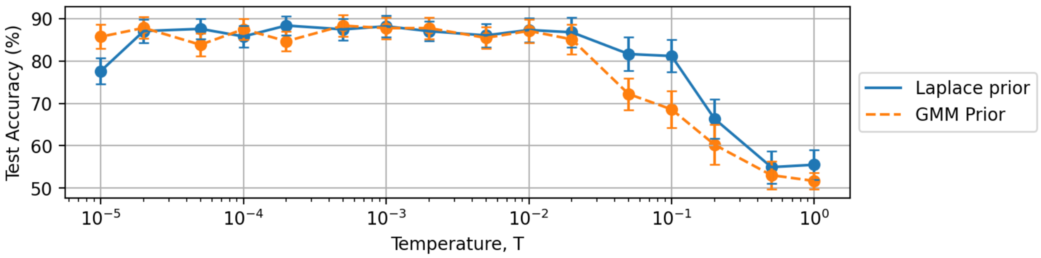

Figure 1: The "cold posterior" effect: for the MiraBest classification problem presented here we can improve the generalization performance significantly by cooling the posterior with a temperature $T \ll 1$, deviating from the Bayes posterior. Data are shown for the BBB models with no data augmentation and the original ELBO cost function trained with a Laplace prior (solid blue line), and trained with a GMM prior (orange dashed line).

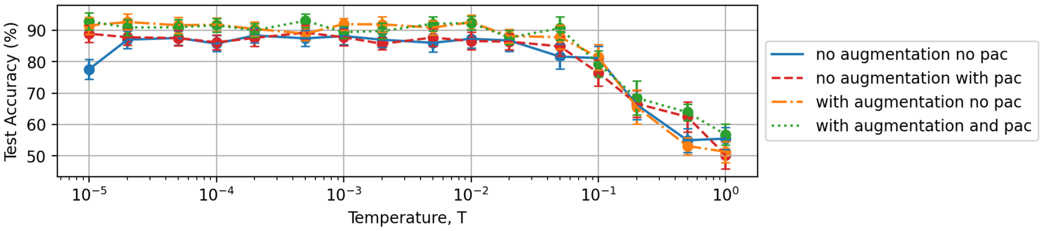

Figure 2: The "cold posterior" effect: Data are shown for the BBB model trained with a Laplace prior with no data augmentation and the original ELBO cost function (solid blue line), the BBB model with no data augmentation and the Masegosa posterior cost function (red dashed line), the BBB model with data augmentation and the original ELBO function (orange dot-dash line), and the BBB model with data augmentation and the Masegosa posterior cost function (green dotted line).

FRI and FRII galaxies. This creates a binary dataset and we do not employ sub-classifications. Each object within the MiraBest dataset was assigned a confidence qualification: *Confident* or *Uncertain* by its original human classifiers. Training is performed on the *Confident* subset.

We use an expanded LeNet-5 architecture incorporating two additional convolutional layers with 26 and 32 channels, respectively, and implement the variational inference approach described in [28] using the Adam optimiser with an initial learning rate of 5.10^{-5}. Hyper-parameters were tuned using a grid search. We use a Gaussian variational approximation to the true posterior and we consider four different prior distributions over the model parameters: a simple Gaussian, a two component Gaussian mixture model (GMM) prior, a Laplace prior and a Laplace Mixture Model (LMM) prior. We find that model performance is optimised using a Laplace prior. Using a GMM prior allows us to achieve comparable accuracy to the model trained with a Laplace prior but it has worse uncertainty calibration error and experiences a more significant cold posterior effect, see Figure 1. In the following analysis we report results for the Laplace prior.

We train the model in batches of 50 images and implement an early stopping criterion based on validation accuracy. We note that while other works have used deeper models for classification of radio galaxies, no significant difference in performance is seen for binary FR classification. We also note that, unlike previous works, we do not use any data augmentation in our training. Models were trained on an Intel i5 CPU for 500 epochs. Each training run took approximately 1 hour.

The cold posterior effect It has been observed in the literature that in order to obtain good predictive performance from Bayesian neural networks, it is necessary to down-weight or *lower the temperature* of the Bayesian posterior. Several explanations have been proposed to explain this effect, including model or prior misspecification [17], and data augmentation or curation issues [29]. We also observe this effect in the classification of radio galaxies in this work, see Figure 2.

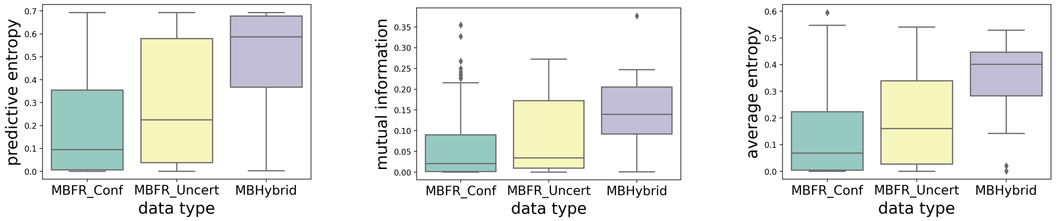

Figure 3: Distribution of predictive uncertainty (entropy; left), epistemic uncertainty (mutual information; centre) and aleatoric uncertainty (average entropy; right) for the *Confident* (MBFR_Conf) and *Uncertain* (MBFR_Uncert) test samples. Data are also shown for the *MiraBest* Hybrid test sample.

Assuming that all models are likely to be misspecified at some level, [18] suggested that the use of more generalised posteriors that treat the true Bayes posterior as a solution of a loose PAC-Bayes generalization bound on the predictive cross-entropy might provide a more desirable approximation target than the Bayes posterior itself. We retrain our models using the PAC-Bayes cost function defined in [18], but find that this does not make a significant difference to the observed cold posterior effect.

If we include data augmentation using random rotations from 0 to 360 degrees, the cold posterior effect observed in our work *reduces* slightly, see Figure 2. We suggest that this is because we have augmented the MiraBest dataset using principled methods that correspond to an informed prior for how radio galaxies are oriented, as radio galaxy class is assumed to be equivariant to orientation and chirality [see e.g. 30, 14]. Since the Masegosa posterior is a more complete test of model misspecification than the data augmentation used here is of likelihood misspecification, we suggest that a key element for exploring this problem in future may be the availability of radio astronomy training sets that do not only present average or consensus target labels, but instead include all individual labels from human classifiers. For the following results we use a posterior temperature of $T = 10^{-2}$.

Model Confidence As well as the MiraBest *Confident* test set, we use 49 samples from the MiraBest *Uncertain* subset to test the trained model's ability to correctly represent epistemic uncertainty. These samples can be considered to be drawn from the same data generating distribution as the MiraBest *Confident* samples, but have a lower degree of belief in their classification. Using Monte Carlo samples obtained from the trained predictive posterior distribution, we obtain 200 Softmax probabilities for each test sample. Following [16], we find that the predictive uncertainty (epistemic + aleatoric), as quantified by the entropy, and the model uncertainty (epistemic), as quantified by the mutual information of these samples, show that on average test data samples labelled as *Uncertain* have higher epistemic uncertainty than those labelled as *Confident*, see Figure 3. To quantify aleatoric uncertainty for in-distribution data samples for classical NNs [31] demonstrate that the entropy of a single pass can be used. Here we extend that definition to our Bayesian NN and take the average entropy for a single input using MC samples to capture the aleatoric uncertainty associated with each data point.

We also examine the uncertainty metrics for the MiraBest *Hybrid* subsample, which contains galaxies that have characteristics of both classes. We find that all uncertainty measures for this sample are even higher than for *Uncertain* FRI and FRII type objects, see Figure 3. We also note that Confidently classified *Hybrid* samples have higher predictive uncertainty than the Uncertain *Hybrid* samples. This could be because the Uncertain samples are more like the FRI/FRII galaxies that the model has seen during training, i.e. their classification as a *Hybrid* was considered Uncertain by a human classifier because the morphology was biased towards one of the standard FRI or FRII classifications. In which case their predictive uncertainty might be expected to be lower since the model was trained to predict those morphologies.

Uncertainty calibration We quantify the calibration of our posterior uncertainties using the class-wise expected Uncertainty Calibration Error (cUCE; [32]). We find that predictive entropy and average entropy for the unpruned model are better calibrated than the mutual information, see Table 1.

Table 1: Percentage classwise Uncertainty Calibration Error (cUCE) on MiraBest Confident test set for our BBB-CNN model trained with a Laplace prior. Results are shown for the unpruned model and the model pruned to its threshold limit for SNR and FIM-based pruning. The percentage cUCE is shown separately for the predictive entropy (PE), mutual information (MI) and average entropy (AE) as calculated on the MiraBest Confident test set.

		% cUCE		
Prior	**Pruning**	**PE**	**MI**	**AE**
Laplace	Unpruned	9.69	16.37	10.84
	SNR	14.35	16.82	13.93
	FIM	13.43	15.29	11.25

3 Weight pruning based on posterior variance

For a typical variational posterior such as the Gaussian distribution, the number of parameters in a Bayesian model will double compared its standard counterpart because both the mean and standard deviation values need to be learned. Consequently, to reduce the computational cost and memory overhead during deployment, network pruning approaches which remove uninformative parameters are often used. Several authors have also considered pruning to improve the generalisation performance of the network [e.g. 33].

Following [28], we calculate the signal-to-noise ratio (SNR) of individual weights as $|\mu|/\sigma$, where the trained posterior distribution on a weight is $\mathcal{N}(\mu, \sigma^2)$. By ranking all of the model weights in order of SNR, we are able to prune a given percentage of the lowest SNR-valued weights. We then calculate the error on the test set using the average of 100 forward passes.

The choice of prior directly influences the shape of the SNR distribution and the proportion of weights that can be pruned without affecting model performance on the test set. For the Laplace prior, we find that up to 30% of the weights in the fully-connected layers can be pruned without a significant drop in performance. However, following the work of [34], we also explore a pruning approach that uses the Fisher Information Matrix (FIM) of the weights. As also observed by [34], pruning the weights based on the Fisher information alone does not allow for a large number of parameters to be pruned effectively because many values in the FIM diagonal are close to zero; however, combining Fisher-based pruning and magnitude-based pruning allows for a larger number of weights to be pruned - up to 60% in this case.

We find that both the SNR and Fisher pruned models have higher predictive uncertainty than the unpruned model. This is also reflected in their distributions of aleatoric uncertainty. The epistemic uncertainty narrows slightly for both the pruning methods. Pruning also affects the uncertainty calibration error, see Table 1. We find that while cUCE of mutual information does not change significantly, the predictive entropy and average entropy values have a higher calibration error, the effect being worse for SNR pruning.

4 Conclusions

In this work we find that whilst Bayesian neural networks using variational inference are useful for returning the degree of model confidence in an individual classification, and that on average this seems to agree with more general degrees of confidence assigned by human classifiers, there is still work that needs to be done on the development of these models. Most specifically, the effect of overly curated training data samples requires additional investigation and we suggest that a key element for exploring this problem in future may be the availability of radio astronomy training sets that do not only present average or consensus target labels, but instead include all individual labels from human classifiers.

Code for this work is available at: `https://github.com/devinamhn/RadioGalaxies-BBB`.

Acknowledgments and Disclosure of Funding

AMS gratefully acknowledges support from an Alan Turing Institute AI Fellowship EP/V030302/1.

References

[1] M. P. van Haarlem, M. W. Wise, A. W. Gunst, et al. LOFAR: The LOw-Frequency ARray. *A&A*, 556:A2, August 2013.

[2] A. P. Beardsley, M. Johnston-Hollitt, C. M. Trott, et al. Science with the Murchison Widefield Array: Phase i results and Phase II opportunities. *Publications of the Astronomical Society of Australia*, 2019.

[3] Matt J. Jarvis, A. R. Taylor, I. Agudo, et al. The MeerKAT international GHz tiered extragalactic exploration (MIGHTEE) survey. In *Proceedings of Science*, 2016.

[4] S. Johnston, R. Taylor, M. Bailes, et al. Science with ASKAP : The Australian square-kilometre-array pathfinder. *Experimental Astronomy*, 2008.

[5] B. L. Fanaroff and J. M. Riley. The morphology of extragalactic radio sources of high and low luminosity. *MNRAS*, 167:31P–36P, May 1974.

[6] Geoffrey V. Bicknell. Relativistic Jets and the Fanaroff-Riley Classification of Radio Galaxies. *ApJS*, 101:29, November 1995.

[7] Christian R. Kaiser and Philip N. Best. Luminosity function, sizes and FR dichotomy of radio-loud AGN. *MNRAS*, 381(4):1548–1560, November 2007.

[8] J. H. Croston, J. Ineson, and M. J. Hardcastle. Particle content, radio-galaxy morphology, and jet power: all radio-loud AGN are not equal. *MNRAS*, 476(2):1614–1623, May 2018.

[9] A. K. Aniyan and K. Thorat. Classifying Radio Galaxies with the Convolutional Neural Network. *The Astrophysical Journal Supplement Series*, 2017.

[10] V. Lukic, M. Brüggen, J. K. Banfield, et al. Radio Galaxy Zoo: compact and extended radio source classification with deep learning. *MNRAS*, 476(1):246–260, May 2018.

[11] J. K. Banfield, O. I. Wong, K. W. Willett, et al. Radio Galaxy Zoo: host galaxies and radio morphologies derived from visual inspection. *MNRAS*, 453(3):2326–2340, November 2015.

[12] Chen Wu, Oiwei Ivy Wong, Lawrence Rudnick, et al. Radio Galaxy Zoo: Claran – a deep learning classifier for radio morphologies. *Monthly Notices of the Royal Astronomical Society*, 482(1):1211–1230, 10 2018.

[13] Micah Bowles, Anna M. M. Scaife, Fiona Porter, et al. Attention-gating for improved radio galaxy classification. *MNRAS*, 501(3):4579–4595, March 2021.

[14] Anna M. M. Scaife and Fiona Porter. Fanaroff-Riley classification of radio galaxies using group-equivariant convolutional neural networks. *MNRAS*, 503(2):2369–2379, May 2021.

[15] David J. Bastien, Anna M. M. Scaife, Hongming Tang, et al. Structured variational inference for simulating populations of radio galaxies. *MNRAS*, 503(3):3351–3370, May 2021.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[17] Florian Wenzel, Kevin Roth, Bastiaan Veeling, et al. How good is the Bayes posterior in deep neural networks really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020.

[18] Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. In H. Larochelle, M. Ranzato, R. Hadsell, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5479–5491. Curran Associates, Inc., 2020.

[19] Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, et al. Data augmentation in bayesian neural networks and the cold posterior effect. *arXiv preprint arXiv:2106.05586*, 2021.

[20] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021.

[21] Fiona A. M. Porter. Mirabest batched dataset 10.5281/zenodo.4288837, Nov 2020.

[22] H. Miraghaei and P. N. Best. The nuclear properties and extended morphologies of powerful radio galaxies: the roles of host galaxy and environment. *Monthly Notices of the Royal Astronomical Society*, 466(4):4346–4363, 01 2017.

[23] P. N. Best and T. M. Heckman. On the fundamental dichotomy in the local radio-AGN population: accretion, evolution and host galaxy properties. *MNRAS*, 421(2):1569–1582, April 2012.

[24] Donald G. York, J. Adelman, Jr. Anderson, John E., et al. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120(3):1579–1587, September 2000.

[25] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, Marcel A. Agüeros, et al. The Seventh Data Release of the Sloan Digital Sky Survey. *ApJS*, 182(2):543–558, June 2009.

[26] J. J. Condon, W. D. Cotton, E. W. Greisen, et al. The NRAO VLA Sky Survey. *AJ*, 115(5):1693–1716, May 1998.

[27] Robert H. Becker, Richard L. White, and David J. Helfand. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *ApJ*, 450:559, September 1995.

[28] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *arXiv e-prints*, page arXiv:1505.05424, May 2015.

[29] Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2021.

[30] Kushatha Ntwaetsile and James E. Geach. Rapid sorting of radio galaxy morphology using Haralick features. *MNRAS*, 502(3):3417–3425, April 2021.

[31] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, et al. Deterministic Neural Networks with Inductive Biases Capture Epistemic and Aleatoric Uncertainty. *arXiv e-prints*, page arXiv:2102.11582, February 2021.

[32] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Calibration of model uncertainty for dropout variational inference. *CoRR*, abs/2006.11584, 2020.

[33] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.

[34] Ming Tu, Visar Berisha, Yu Cao, and Jae-sun Seo. Reducing the model order of deep neural networks using information theory. In *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 93–98. IEEE, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] Data and impacts are not societal.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No] No theoretical results
   (b) Did you include complete proofs of all theoretical results? [No] No theoretical results
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Github link for code and Zenodo link for data re included.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figures 1, 2 & 3
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 2
   (b) Did you mention the license of the assets? [Yes] See Section 2
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] No new assets
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] No personal data used; only public astronomy data
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] No personal data used; only public astronomy data
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] No human data used; only public astronomy data
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] No human data used; only public astronomy data
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] No human data used; only public astronomy data