Abstract

Investigating Galaxy Morphology in Large Surveys Using Machine Learning

Aritra Ghosh | অরিত্র ঘোষ

2023

asdjnfjins adsnfij snadijfn in iasndfinsjidnf ijjsndifjnsijdnf isjnadfn aoisjdfn iojansijdn

Investigating Galaxy Morphology in Large Surveys Using Machine Learning

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Aritra Ghosh | অরিত্র ঘোষ

Dissertation Director: Prof. C. M. Urry

December, 2023

Contents

Acknowledgements **xx**

1 Introduction **1**

2 Galaxy Morphology Network: A Convolutional Neural Network Used to Study Morphology and Quenching in $\sim 100,000$ SDSS and $\sim 20,000$ CANDELS Galaxies **2**

2.1 Introduction . . . 3

2.2 Data Sets Used . . . 6

2.3 Training our Convolutional Neural Network — GaMorNet . . . 8

2.3.1 Simulations . . . 9

2.3.2 The Network . . . 12

2.3.3 Initial Training . . . 15

2.3.4 Transfer Learning . . . 16

2.4 Results . . . 21

2.4.1 Morphology Results . . . 21

2.4.2 Color - Mass Results . . . 25

2.5 Summary and Discussion . . . 30

2.A Public Release of Code, Models, and Galaxy Morphological Classifications . . . 33

2.A.1 GaMorNet Source Code . . . 33

2.A.2 GaMorNet Trained Models . . . 33

2.A.3 Tables with predicted probabilities and classifications . . . 34

2.A.4 GalaxySim Source Code . . . 35

3 GaMPEN: A Machine Learning Framework for Estimating Bayesian Posteriors of Galaxy Morphological Parameters **36**

3.1 Introduction . . . 37

3.2 Simulated Galaxies . . . 39

3.3 GaMPEN Architecture . . . 43
3.3.1 The Spatial Transformer Network Module . . . 44
3.3.2 The Convolutional Neural Network Module . . . 48
3.4 Prediction of Posteriors . . . 49
3.4.1 Bayesian Implementation of GaMPEN and Epistemic Uncertainties . . . 50
3.4.2 Likelihood Calculation and Aleatoric Uncertainties . . . 51
3.4.3 Practical Implementation Details . . . 52
3.4.4 Combining Aleatoric and Epistemic Uncertainties . . . 54
3.5 Training GaMPEN . . . 54
3.6 Results . . . 58
3.6.1 Inspecting the Predicted Posteriors . . . 58
3.6.2 Evaluating the Accuracy of GaMPEN . . . 60
3.6.3 Inspecting the Predicted Uncertainties . . . 66
3.6.4 Qualitative Transformation of GaMPEN Predictions . . . 68
3.7 Discussion & Conclusions . . . 69
3.A Early Data Access . . . 76
3.B Extended Derivation for Bayesian Implementation of GaMPEN . . . 76
3.C Extended Derivation of the Loss Function . . . 77
3.D Additional Technical Details on GaMPEN . . . 78

4 Morphological Parameters and Associated Uncertainties for 8 Million Galaxies in the Hyper Suprime-Cam Wide Survey 81

4.1 Introduction . . . 82
4.2 Data . . . 85
4.2.1 Hyper Suprime-Cam Data . . . 85
4.2.2 Simulated Galaxies for Initial Training . . . 88
4.3 Brief Introduction to GaMPEN . . . 90
4.4 Training GaMPEN . . . 92
4.4.1 Data Transformations . . . 93
4.4.2 Initial Training of GaMPEN on simulated galaxies . . . 94
4.4.3 Transfer Learning using Real Data . . . 96
4.5 Galfitting Galaxies for Transfer Learning & Validation . . . 99
4.6 Evaluating GaMPEN's performance . . . 103

4.6.1 Inspecting the Predicted Posteriors . . . 103
4.6.2 Evaluating the STN performance . . . 105
4.6.3 Comparing GaMPEN predictions to GALFIT predictions . . . 106
4.6.4 Inspecting the Predicted Uncertainties . . . 112
4.6.5 Comparing GaMPEN's Uncertainty Estimates to Other Algorithms . . . 113
4.7 Comparing Our Predictions to Other Catalogs . . . 115
4.8 Conclusions & Discussion . . . 117
4.A Data Access . . . 122
4.B Additional Details About Data . . . 122
4.C Additional Details About GaMPEN . . . 125
4.D Additional Details About Trained GaMPEN Models . . . 125
4.E Identifying Issues with our Light Profile Fitting Pipeline . . . 126
4.F Additional Two-Dimensional Residual & Uncertainty Plots . . . 126
4.G Comparing GaMPEN and GALFIT's performance on smaller simulated galaxies . . . 127

5 Variation of Morphology With Large Scale Density 134

6 Conclusions 135

List of Figures

2.1 The above figure contains randomly chosen galaxies from both our data sets classified by GaMorNet as being disk-dominated (left column panels), bulge-dominated (middle column panels) or indeterminate (right column panels). Refer to § 2.4.1 for the definitions of these categories. The top two rows show SDSS cutouts, which are 33.07" × 33.07" (83 pixels × 83 pixels) and the bottom two rows show CANDELS cutouts, which are 4.98" ×4.98" (83 pixels × 83 pixels). During training, GaMorNet focuses on galaxies located at the center of the image and, thus, can process cutouts with other objects in the frame besides the central galaxy, as is evident from the images above. 4

2.2 Three stages in simulating an SDSS galaxy. Left (a): Light profile generated by GALFIT with a bulge-to-disk ratio of 0.24. Center (b): The left image convolved with the SDSS PSF. Right (c): SDSS noise added to the middle image. See § 2.3.1 for details of the PSF convolution and noise addition. 10

2.3 A schematic diagram showing a simple artificial neural network with a single hidden layer. 13

2.4 Schematic diagram of GaMorNet, a CNN optimized to identify whether galaxies are bulge-dominated or disk-dominated. Its architecture, which is based on AlexNet (Krizhevsky et al., 2012), consists of five convolutional layers and three fully connected layers. Between these layers are max-pooling, local response normalization, and dropout layers.t The numbers inside the circles refer to the layer number and corresponding details for each layer can be found by looking up the corresponding layer order number in Table 2.2. 15

2.5 Learning curves for the process of training GaMorNet on simulated galaxy images. The accuracy evaluated on the validation set (brown curves, left axes) and the value of the loss function after each epoch of training (blue curves, right axes) are shown for both GaMorNet-S and GaMorNet-C (left and right panels, respectively). GaMorNet-S achieves an accuracy of 93.55% after 1000 epochs of training, and GaMorNet-C achieves an accuracy of 88.33% after 400 epochs of training. For more details about the training process, see § 2.3.3. 17

2.6 The triangles show the input bulge-to-total ratio (L_B/L_T) versus fitted Sérsic index for the galaxies simulated by Simmons & Urry (2008; adapted from the lowest panel in their Figure 19). The plotted points are the median of each bin's distribution, and the error bars mark the central 68% of sources in the bin. The shaded regions correspond to our definitions of the three output classes used by GaMorNet-C. The histogram shows the distribution of the Sérsic index for all the galaxies in our CANDELS sample, most of which are disk-dominated (see § 2.2). Clearly, all galaxies with $n < 2$ are truly disk-dominated (i.e., have $L_B/L_T < 0.45$) but, because of the spread in Sérsic indices, some disk-dominated or intermediate galaxies may get misclassified as bulge-dominated. Although a higher n threshold (for, e.g., $n \sim 6$) would lead to a purer bulge-dominated sample, for reasons mentioned in § 2.3.4, it would make the transfer learning sample insufficiently small. Note that readers can choose different bin boundaries, doing their own transfer learning step on the simulation-trained network made available via § 2.A.2 19

2.7 The normalized distribution of correctly classified and misclassifed CANDELS galaxies in the test set as a function of the signal-to-noise ratio (S/N) and half-light radius (r_e). Both plots show that compared to the correctly classified galaxies, a higher fraction of the misclassified galaxies have a low S/N ratio and/or small r_e. 'Frequency density' refers to the number counts normalized to form a probability density. 24

2.8 Relation of confidence threshold to completeness and accuracy of classification, for the SDSS data set. Left (a): The fraction of indeterminate galaxies increases with increasing confidence threshold. Right (b): The accuracy of both disk-dominated (blue line, left axis) and bulge-dominated (orange line, right axis) classifications increases with increasing confidence threshold. We decided on a confidence threshold of 0.8 (or 80%) for GaMorNet-S (star in both plots) as the optimal compromise between accuracy and completeness. 25

2.9 Relation of confidence threshold to the accuracy (blue lines, left axes) and completeness (orange lines, right axes) of GaMorNet-C classification of the CANDELS data set. Stars denote the adopted confidence thresholds. Left (a): For the chosen disk confidence threshold of 0.36, provided the probability of being disk-dominated exceeds the probabilities of being bulge-dominated or indeterminate, the classification accuracy is better than 92% and the indeterminate fraction <40%. Right (b): For the chosen bulge confidence threshold of 0.55, we obtain an accuracy of >80% and indeterminate fraction <40%. 26

2.10 Color-mass diagrams for the galaxies in the SDSS test set, separated by morphology. Disk-dominated galaxies (panels (a) and (c)) are mostly blue until they reach high masses (and presumably high halo masses), at which point they evolve to the red. In contrast, bulge-dominated galaxies (panels (b) and (d)) are predominately red, and appear to evolve rapidly from a short-lived population of rare, blue ellipticals that likely formed from major mergers of disky star-forming galaxies. Panels (a) and (b) show individual data points, with color indicating the specific star formation rates (sSFR) for each galaxy in units of yr^{-1}. Contours show the linear density of galaxies in this plot, and the numbers refer to the levels of the contours. Panels (c) and (d) are the same data plotted in terms of galaxy density. The lines mark the position of the green valley. 27

2.11 Color-mass diagrams for the galaxies in the CANDELS test set, separated by morphology. Similar to Fig. 2.10, disk-dominated galaxies (panels (a) and (c)) show signs of secular evolution, while bulge-dominated galaxies (panels (b) and (d)) appear to evolve rapidly from a short-lived population of rare, blue ellipticals. Panels (a) and (b) show individual data points, with color indicating the specific star formation rates (sSFR) for each galaxy in units of yr^{-1}. Contours show the linear density of galaxies in this plot and the numbers refer to the levels of the contours. Panels (c) and (d) are the same data plotted in terms of galaxy density. The lines mark the position of the green valley. 28

2.12 The normalized distribution of the specific star formation rate (sSFR), separated by morphology, for the SDSS and CANDELS data sets as obtained from the MPA-JHU and 3D-HST catalogs, respectively. 'Frequency density' refers to the number counts normalized to form a probability density. 30

3.1 Two stages of simulating an HSC galaxy. (*Left*): A randomly chosen two-dimensional light profile generated by GalSim. (*Right*): The same image after PSF convolution and noise addition. The white pixels represent (small) negative values that arise from the process of noise addition. 41

3.2 Ten randomly selected galaxies from our simulated dataset. The simulation parameters are chosen such that the simulated galaxies represent a diverse range of light profiles and include most bright, local galaxies at $z \lesssim 0.25$. 42

3.3 A schematic diagram of the Galaxy Morphology Posterior Estimation Network. GaMPEN's architecture consists of a downstream CNN module preceded by an upstream STN module. The CNN module empowers GaMPEN to estimate posterior distributions of galaxy morphology parameters. The upstream STN module trains without any extra supervision and learns to apply appropriate cropping transformations to the input image before passing it on to the CNN (for more details about these modules, see §§ 3.3.1, 3.3.2). The numbers below each layer refer to the number of filters/neurons in each layer. The yellow boxes inside the convolutional layers show the kernel and the number beside it refers to the corresponding kernel size. Only one kernel is shown per set of convolutional layers; all other layers in the set have kernels of the same size. Conv2D and ReLU refer to Convolutional Layers and Rectified Linear Units, respectively (described in §.3.3.2). 44

3.4 Examples of the transformation applied by the STN to six randomly selected input galaxy images. The top row shows the input galaxy images, and the bottom row shows the corresponding output from the STN. The numbers in the top-left yellow boxes help correspond the output images to the input images. As can be seen, the STN learns to apply an optimal amount of cropping for each input galaxy. 45

3.5 (*Left*): Galaxies in the testing dataset with the lowest values of s (i.e., the most aggressive crops) (*Right*): Galaxies in the testing dataset with the highest values of s (i.e., the least aggressive crops). As can be seen, the STN correctly learns to apply the most aggressive crops to small galaxies; and the least aggressive crops to large galaxies. 47

3.6 Examples of the transformation applied by a trained STN to real HSC-Wide g-band galaxies. The STN helps the downstream CNN to focus on the galaxy of interest at the center of the cutout by cropping out most secondary galaxies present in the input frame. 48

3.7 Diagram outlining the training (*left*) and inference (*right*) phases of the GaMPEN workflow. Training consists of feeding 105,000 simulated images (with known parameter values) through the STN and CNN modules, minimizing the loss function (Eqn. 3.8) using Stochastic Gradient Descent. During this process, we re-scale the variables as described in the text, and return them to the original variable space during inference. After the STN+CNN are trained, the inference step consists of 1000 forward passes with dropout enabled for each galaxy image. We draw a sample from the predicted multivariate Gaussian distribution during each forward pass, and the collection of these samples gives us the predicted posterior distribution. 54

3.8 The calculated percentile coverage probabilities for different dropout rates. The top three rows show coverage probabilities for each output variable individually, while the bottom row shows the probabilities averaged over the three variables. The coverage probabilities are defined as the percentage of the total test examples where the true value of the parameter lies within a particular confidence interval of the predicted distribution. A dropout rate of 7×10^{-4} leads to coverage probabilities very close to their corresponding confidence levels. 56

3.9 Joint and marginalized probability distributions predicted by GaMPEN for a randomly chosen galaxy in our testing set. The red dotted lines show the true values of the parameters. 59

3.10 Examples of predicted posterior distributions for four randomly chosen simulated galaxies. The blue shaded histogram shows the predictions from GaMPEN and the blue solid lines show the associated probability distribution functions estimated by kernel density estimation. These are used to calculate the confidence intervals shown in the figure with pink, yellow, and green shading. The mode (red line) shows the most probable value of each morphological parameter. As expected, in most cases, the true value (purple line) lies within the 68.27% confidence interval. 60

3.11 The true values of the galaxy parameters plotted against the most probable values predicted by GaMPEN. The black dashed line marks the $y = x$ diagonal on which perfectly recovered parameters should lie. The color of each hexagon corresponds to the number of galaxies it contains, as indicated by the colorbar at right. 61

3.12 Histograms of residuals for all galaxies in the testing set. We define the residuals as the difference between the true value and the most probable value predicted by GaMPEN. The dashed vertical line represents $x = 0$, denoting cases with perfectly recovered parameter values. The mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution are listed in each panel. 62

3.13 Residuals of GaMPEN predicted parameter values plotted against the true values. The residual for each parameter is defined as the difference between the most probable predicted value and the true value, i.e., Mode($\hat{\boldsymbol{Y}}_{\boldsymbol{n}}$) $- \boldsymbol{Y}_{\boldsymbol{n}}$. The color of each hexagonal bin corresponds to the number of galaxies it contains, as shown by the colorbar on the right. The black dotted line ($y = 0$) represents perfectly recovered parameters. . 64

3.14 Residuals of the output parameters plotted against the predicted values. This figure allows us to assign quality labels to GaMPEN predictions (e.g., flagging parameters that are unreliable) based on the output values. See § 3.6.4 for details. 65

3.15 Histograms of L_B/L_T residuals shown separately for single component galaxies, all double component galaxies, and double component galaxies with $0.1 < L_B/L_T < 0.85$. The standard deviation (σ) for each distribution is also shown in the top left. The dashed vertical line represents $x = 0$, denoting cases with perfectly recovered L_B/L_T. The apparent hard cutoffs in the distributions of the single component, and the restricted range double-component galaxies arise from the fact that the y-scale is logarithmic. We have verified that when plotted on a linear scale, the apparent hard cutoffs disappear. 66

3.16 Fractional residuals for the effective radius and flux plotted against their corresponding true values. Note that since we are plotting the absolute values, the ideal situation of perfectly recovered parameters is at the bottom of each panel. The right two panels show that both the residuals increase for fainter galaxies, while the top-middle panel shows that the radius residuals increase for smaller galaxies. 67

3.17 Uncertainties predicted by GaMPEN for each parameter plotted against the true values. The σ for each parameter is defined as the width of the 68.27% confidence interval. Note that we plot fractional uncertainties for radius and flux in order to make the y-axis dimensionless for all three rows. 68

3.18 Uncertainties (widths of the 68.27% confidence intervals) predicted by GaMPEN for each parameter versus the corresponding residuals (predicted mode minus true value). Fractional uncertainties and residuals are plotted for radius and flux in order to make all the quantities dimensionless. The trend in all three cases is that GaMPEN-estimated uncertainties increase for cases where its predictions are less accurate. The coverage probabilities reported on the test set (Table 3.2) confirm that the predicted uncertainties are well-calibrated and correspond well to the quoted confidence intervals. 69

3.19 The left panels show the residuals for bulge-to-total light ratio and radius plotted against their predicted values. The black dashed regions show the parameter-space where we replace the quantitative predictions with qualitative flags. Each corresponding histogram on the right shows the distribution of residuals before and after the transformation of output values. 70

3.20 Confusion matrix between the labels we assign when GaMPEN predicts extreme bulge-to-total ratios, $L_B/L_T < 0.1$ or > 0.85, and their true L_B/L_T values. The number in each block shows how many galaxies correspond to that panel, resulting in an overall accuracy $> 99\%$. 71

4.1 The limiting absolute magnitudes probed by the Hyper Suprime-Cam (HSC) Wide Survey and Sloan Digital Sky Survey (SDSS) at different redshifts. 83

4.2 The filter used for each redshift bin is shown along with the wavelength range sampled by each filter. The blue line shows where rest-frame 450 nm emission falls for redshifts labeled on the x-axis. As this figure shows, the chosen filters allow us to consistently perform morphology determination in the rest-frame g-band. 85

4.3 Redshift (*top*) and magnitude (*bottom*) distributions for the $\sim$ 8 million galaxies used in this study. We used spectroscopic redshifts when available and high-quality photometric redshifts otherwise. The spectroscopic completeness of each sub-sample is shown in Table 4.1. 86

4.4 Four randomly chosen galaxy cutouts are shown here for each redshift bin, with the object of interest at the center of each cutout. Note that most of these cutouts have secondary objects in the frame, which can often cause ML algorithms to produce spurious classifications. GaMPEN uses a Spatial Transformer Network to crop most secondary objects out of the frame (see §3). 87

4.5 Two stages of simulating an HSC galaxy. (*Left*): A randomly chosen two-dimensional light profile generated by GalSim. (*Right*): The same image after PSF convolution and noise addition. The white pixels represent (small) negative values that arise from the process of noise addition. 90

4.6 Diagram outlining the training (*left*) and posterior inference (*right*) phases of the GaMPEN workflow. Training consists of feeding galaxies (with pre-determined parameter values) through the STN and CNN modules, minimizing the loss function using Stochastic Gradient Descent. During this process, we re-scale the variables described in the text, and return them to the original variable space during inference. After the STN+CNN networks are trained, the posterior inference step consists of 500 forward passes with dropout enabled for each galaxy image. We draw a sample from the predicted multivariate Gaussian distribution during each forward pass, and the collection of these samples gives us the predicted posterior distribution. 91

4.7 Diagram outlining the different stages of training GaMPEN. We first train GaMPEN using simulated light profiles described in §4.2.2. Thereafter, we fine-tune the simulation-trained framework using 0.5% of our real data sample, for which we predetermined the morphological parameters using light-profile fitting, as described in §4.5. Finally, we process all the $\sim$ 8 million galaxies in our dataset through the trained GaMPEN framework to obtain estimates of their morphological parameters and associated uncertainties. 93

4.8 Histograms of residuals for simulated galaxies (in the test set) across all three redshift bins. We define the residuals as the difference between the true value and the most probable value predicted by GaMPEN. The dashed vertical line represents $x = 0$, denoting cases with perfectly recovered parameter values. The mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution are listed in each panel. . . 95

4.9 Magnitude and redshift distributions for all galaxies in each redshift bin, plotted along with the galaxies selected for transfer learning (before and after applying quality cuts, as described in §4.5). Note that we plot density on the y-axis, not the number of samples. The total number of galaxies used for transfer learning is $\sim$ 0.5% of all the galaxies in our dataset. The relative density of some magnitude bins is higher than others (e.g., $18 < m \leq 20$ for low-z) because they span a smaller range while having roughly the same number of galaxies as the other bins. 96

4.10 The calculated percentile coverage probabilities for different dropout rates for the low-z bin. Note that the coverage probabilities have been averaged over the three output variables. The coverage probabilities are defined as the percentage of the total test examples where the value determined using light profile fitting lies within a particular confidence interval of the predicted distribution. A dropout rate of 4×10^{-4} leads to coverage probabilities very close to their corresponding confidence levels. A similar process of tuning in the mid-z and high-z bins leads to an optimal dropout rate of 2×10^{-4} in both of them. 98

4.11 Steps used in our light-profile fitting pipeline to determine morphological parameters, for two representative galaxies in each redshift bin. From left to right we show the input image, the mask generated by Source Extractor, the model generated by GALFIT, and the residuals. Note that since we do not explicitly model any Fourier bending modes or coordinate rotations, we expect features like spiral arms to show up in the residuals, as depicted in the second row. 100

4.12 Morphological parameters determined for HSC-imaged galaxies using our light-profile fitting pipeline, versus morphological parameters for the same galaxies determined by Simard et al. (2011) based on SDSS imaging. The black dashed diagonal corresponds to perfect agreement. The vertical line in the middle panel shows the median SDSS g-band seeing. 101

4.13 Examples of predicted posterior distributions for two randomly chosen galaxies from each redshift bin. The blue shaded histograms show the predictions from GaMPEN, and the solid blue lines show the associated probability distribution functions estimated by kernel density estimation. These are used to calculate the confidence intervals shown in the figure with pink, yellow, and green shading. The mode (solid red line) shows the most probable value of each morphological parameter. As expected, in most cases, the GALFIT-ed value (dashed black line) lies within the 68.27% confidence interval. 104

4.14 Examples of the transformation applied by the STN to two randomly selected galaxies from each redshift bin. The top row shows the input galaxy images, and the bottom row shows the corresponding output from the STN. The numbers in the top-left yellow boxes help correspond the output images to the input images. As can be seen, the STN learns to crop most secondary objects present in the input frame. 105

4.15 (*Left*): Galaxies in the low-z bin with the lowest values of s (i.e., the most aggressive crops) (*Right*): Galaxies in the low-z bin with the highest values of s (i.e., the least aggressive crops). The s parameter denotes the fraction of the input image that was retained in the STN output. As can be seen, the STN correctly learns to apply the most aggressive crops to small galaxies; and the least aggressive crops to large galaxies. 106

4.16 Percentile coverage probabilities achieved on the test set shown separately for each redshift bin. The leftmost set of bars in each panel shows the coverage probabilities when averaged over the three output parameters, and the right three sets of bars show the coverage probabilities for each parameter individually. The mean coverage probability never deviates by more than 4.5% from the claimed confidence interval, and when considered for each parameter separately, the coverage probability never deviates by more than 8.7%. This demonstrates that GaMPEN produces well-calibrated accurate uncertainties. 107

4.17 The most probable parameter values predicted by GaMPEN for all galaxies in the test set plotted against the values determined using GALFIT. Galaxies are plotted in hexagonal bins of roughly equal size, and the number of galaxies in each bin is represented according to the logarithmic colorbar to the right of each panel. The top, middle, and bottom rows show the results for the low-, mid-, and high-z bins, respectively. The dashed black $y = x$ line represents the line of equality. Across all three redshift bins, values predicted by GaMPEN closely mirror the values obtained using light-profile fitting. 108

4.18 Distributions of residuals for all galaxies in the test set; specifically, the differences between the values predicted by GaMPEN and those obtained via light-profile fitting. The top, middle, and bottom rows show the results for the low-, mid-, and high-z bins, respectively. The boxes in the top-left corner of each panel show the mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution. The σ of each distribution identifies the typical disagreement for each parameter (e.g., for the low-z bin, in 68.27% cases, the predicted magnitude is within ± 0.41 of the value determined by light profile fitting). The dashed red vertical line marks $x = 0$. 109

4.19 Residuals of the output parameters (difference between GaMPEN and GALFIT predictions) plotted against the values predicted by GaMPEN for all galaxies in the low-z test set. To make the y-axis dimensionless for all three parameters, we plot the fractional R_e residuals instead of absolute values. This figure allows us to assign quality labels to GaMPEN's predictions based on the output values (e.g., flagging regions of the parameter space with high levels of disagreement, as shown by the line-shaded region in the top-left panel). See § 4.6.3 for details. The equivalent figures for the mid- and high-z bins are shown in Appendix 4.F. 110

4.20 Uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the low-z test set. The σ for each parameter is defined as the width of the 68.27% confidence interval. Note that we plot fractional uncertainties for the radius in order to make the y-axis dimensionless for all three rows. The line-shaded region in the top-left panel shows the region where we recommend transforming quantitative L_B/L_T predictions to qualitative labels (see §4.6.3 for details). 112

4.21 Percentile coverage probabilities for the 68.27% confidence interval obtained by GaMPEN on our HSC sample compared to coverage probabilities obtained by various light-profile fitting algorithms on simulated Euclid data (from Bretonniere et al., 2022). The rightmost set of bars shows the values calculated on the entire dataset, while the other sets display values calculated on sub-samples of galaxies with specific magnitude ranges (AB mag, shown on the x-axis). Compared to light-profile fitting tools, the uncertainties predicted by GaMPEN are better calibrated by $\sim 15-25\%$ overall and by as much as $\sim 60\%$ for the brightest galaxies. 114

4.22 GaMPEN predictions plotted against values estimated by Simard et al. (2011) for a cross-matched sample of $\sim 20,000$ galaxies with $z < 0.2$ and $m < 19$. The density of points in each histogram is represented according to the logarithmic colorbar on the right. The black dots show the median y values in bins of equal width along the x-axis, with the error bars depicting the average 68.27% and 95.45% confidence intervals predicted by GaMPEN in that bin. The dash-dotted vertical line in the middle panel shows the median SDSS g-band seeing. 115

4.23 GaMPEN predictions plotted against values estimated by Kawinwanichakij et al. (2021) for a cross-matched sample of $\sim 200,000$ galaxies with $0.5 < z \leq 0.75$ and $m \leq 23$. Similar to Figure 4.22, we show the median y-values and associated error bars in bins of equal width along the x-axis. The dash-dotted vertical line in the middle panel shows the median HSC i-band seeing. 115

4.24 We exclude $\sim 20\%$ of our downloaded galaxies due to different flags being triggered. The distribution of various flags that contribute to galaxies being excluded from our sample is shown in this figure. As can be seen, the large majority of exclusions are due to cosmic-ray hits (and hence, interpolated pixels). 123

4.25 A schematic diagram of the Galaxy Morphology Posterior Estimation Network. GaMPEN's architecture consists of a downstream CNN module preceded by an upstream STN module. The CNN module empowers GaMPEN to estimate posterior distributions of galaxy morphology parameters. The upstream STN module trains without any extra supervision and learns to apply appropriate cropping transformations to the input image before passing it on to the CNN (for more details about these modules, see §4.3). The numbers below each layer refer to the number of filters/neurons in each layer. The yellow boxes inside the convolutional layers show the kernel and the number beside it refers to the corresponding kernel size. Only one kernel is shown per set of convolutional layers; all other layers in the set have kernels of the same size. Conv2D and ReLU refer to Convolutional Layers and Rectified Linear Units, respectively. 125

4.26 The different failure modes of the semi-automated light profile fitting code described in §4.5. From left to right, we show the input image, the mask generated by Source Extractor, the model generated by GALFIT, and the residuals. 128

4.27 Residuals of the output parameters (difference between GaMPEN and GALFIT predictions) plotted against the values predicted by GaMPEN for all galaxies in the mid-z test set. See § 4.6.3 for details. 129

4.28 Residuals of the output parameters (difference between GaMPEN and GALFIT predictions) plotted against the values predicted by GaMPEN for all galaxies in the high-z test set. See § 4.6.3 for details. 130

4.29 Uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the mid-z test set. The σ for each parameter is defined as the width of the 68.27% confidence interval. The line-shaded region in the top-left panel shows the region where we recommend transforming quantitative L_B/L_T predictions to qualitative labels. See §4.6.4 for details. 131

4.30 Uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the high-z test set. The σ for each parameter is defined as the width of the 68.27% confidence interval. The line-shaded region in the top-left panel shows the region where we recommend transforming quantitative L_B/L_T predictions to qualitative labels. See §4.6.4 for details. 132

4.31 Distribution of residuals for ~ 5000 simulated galaxies in each redshift bin with $R_e \leq 2''$. These galaxies were selected randomly from the simulation testing set. The top, middle, and bottom rows show the results for the low-, mid-, and high-z bins respectively. The boxes in the top-left corner of each panel show the median ($\tilde{\mu}$) and standard deviation (σ) of each residual distribution. The dashed black vertical line marks $x = 0$. 133

List of Tables

2.1 Parameter Ranges for Simulated Galaxies 8
2.2 Structure of GaMorNet 14
2.3 Transfer Learning Parameters 18
2.4 Classification Probabilities for 82,547 SDSS Galaxies 20
2.5 Classification Summary for 82,547 SDSS Galaxies 20
2.6 Classification Probabilities for 21,746 CANDELS galaxies 22
2.7 Classification Summary for 21,746 CANDELS Galaxies 22
2.8 Statistics of the Color-Mass Diagrams 29

3.1 Parameter Ranges of Simulated Galaxies 40
3.2 Coverage Probabilities on the Test Set 61
3.3 Structure of GaMPEN 78

4.1 Data Characteristics 86
4.2 Parameter Ranges of Simulated Galaxies 89
4.3 GALFIT Quality Cuts 101
4.4 Tuned Values of Various Hyper-parameters 126

Acknowledgements

Chapter 1

Introduction

This is the introoooooooo.

Chapter 2

Galaxy Morphology Network: A Convolutional Neural Network Used to Study Morphology and Quenching in $\sim 100,000$ SDSS and $\sim 20,000$ CANDELS Galaxies

Originally published by the American Astronomical Society in *The Astrophysical Journal*, Volume 895, Issue 2, pp. 112, DOI:10.3847/1538-4357/ab8a47

Aritra Ghosh, C. Megan Urry, Zhengdong Wang, Kevin Schawinski, Dennis Turp, and Meredith C. Powell

We examine morphology-separated color-mass diagrams to study the quenching of star formation in $\sim 100,000$ ($z \sim 0$) Sloan Digital Sky Survey (SDSS) and $\sim 20,000$ ($z \sim 1$) Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) galaxies. To classify galaxies morphologically, we developed Galaxy Morphology Network (GaMorNet), a convolutional neural network that classifies galaxies according to their bulge-to-total light ratio. GaMorNet does not need a large training set of real data and can be applied to data sets with a range of signal-to-noise ratios and spatial resolutions. GaMorNet's source code as well as the trained models are made public as part of this work (Link 1|Link 2). We first trained GaMorNet on simulations of galaxies with a bulge and a disk component and then transfer learned using $\sim 25\%$ of each data set to achieve misclassification rates of $\lesssim 5\%$. The misclassified sample of galaxies is dominated by small galaxies with low signal-to-noise ratios. Using the GaMorNet classifications, we find that bulge- and disk-dominated galaxies have distinct color-mass diagrams, in agreement with previous studies. For both SDSS and CANDELS galaxies, disk-dominated galaxies peak in the blue cloud, across a broad range of masses, consistent with the slow exhaustion of star-forming gas with no rapid quenching. A small population of red disks is found at high mass ($\sim 14\%$ of disks at $z \sim 0$ and 2% of disks

at $z \sim 1$). In contrast, bulge-dominated galaxies are mostly red, with much smaller numbers down toward the blue cloud, suggesting rapid quenching and fast evolution across the green valley. This inferred difference in quenching mechanism is in agreement with previous studies that used other morphology classification techniques on much smaller samples at $z \sim 0$ and $z \sim 1$.

2.1 Introduction

We know from large-scale surveys that both local and high-redshift galaxies show a bimodal distribution in the galaxy color-mass space (Baldry et al., 2006, 2004; Brammer et al., 2009; Strateva et al., 2001) with a "blue cloud," a "red sequence" and a "green valley." Galaxy color-mass diagrams are useful for studying galactic evolution, as the stellar mass of a galaxy indicates its growth over time, and the color tracks its rate of star formation. The standard interpretation of the bimodal color-mass distribution is that, because there are few galaxies in the green valley, star formation in blue cloud galaxies must be quenched rapidly, perhaps aided by emission from an active galactic nucleus (AGN; Bell et al., 2004; Faber et al., 2007). Direct evidence of this AGN feedback remains murky, however (Harrison, 2017).

Galaxy morphology adds a third interesting dimension to the color-mass space. Because elliptical galaxies typically form in major mergers, and galactic disks usually do not survive them, morphology can be used as a tracer of the recent merger history of a galaxy. The observed bimodality in the color-mass diagram (as well as interpretations therefrom) comes from superposing distinct populations with different morphological types, as first shown by Schawinski et al. (2014), who used Galaxy Zoo morphological classifications to study local ($z \sim 0$) galaxies. They suggested that there are two separate evolutionary tracks for galaxies: (1) major mergers forming ellipticals from disk-dominated galaxies, accompanied by AGN triggering and rapid quenching of star formation, and (2) slow, secular growth of disk-dominated galaxies, until they reach a critical halo mass, after which the remaining cold gas is slowly consumed and the stellar population gradually reddens. At $z \sim 0$, the latter population is an order of magnitude larger than the merger-created ellipticals.

Still, most star formation and the most pronounced galaxy evolution happen not locally but at $z \sim 1$ and above. Thus, it is important to investigate the galaxy color-mass diagram at $z \gtrsim 1$. Powell et al. (2017) studied galaxies from The Great Observatories Origins Deep Survey (GOODS)-N and GOODS-S at $z \sim 1$ and found that disks and spheroids have distinct color-mass distributions in rough agreement with the results at $z \sim 0$. From the distribution of X-ray-selected AGN hosts in this sample, they concluded that AGN feedback may quench star formation in galaxies that undergo

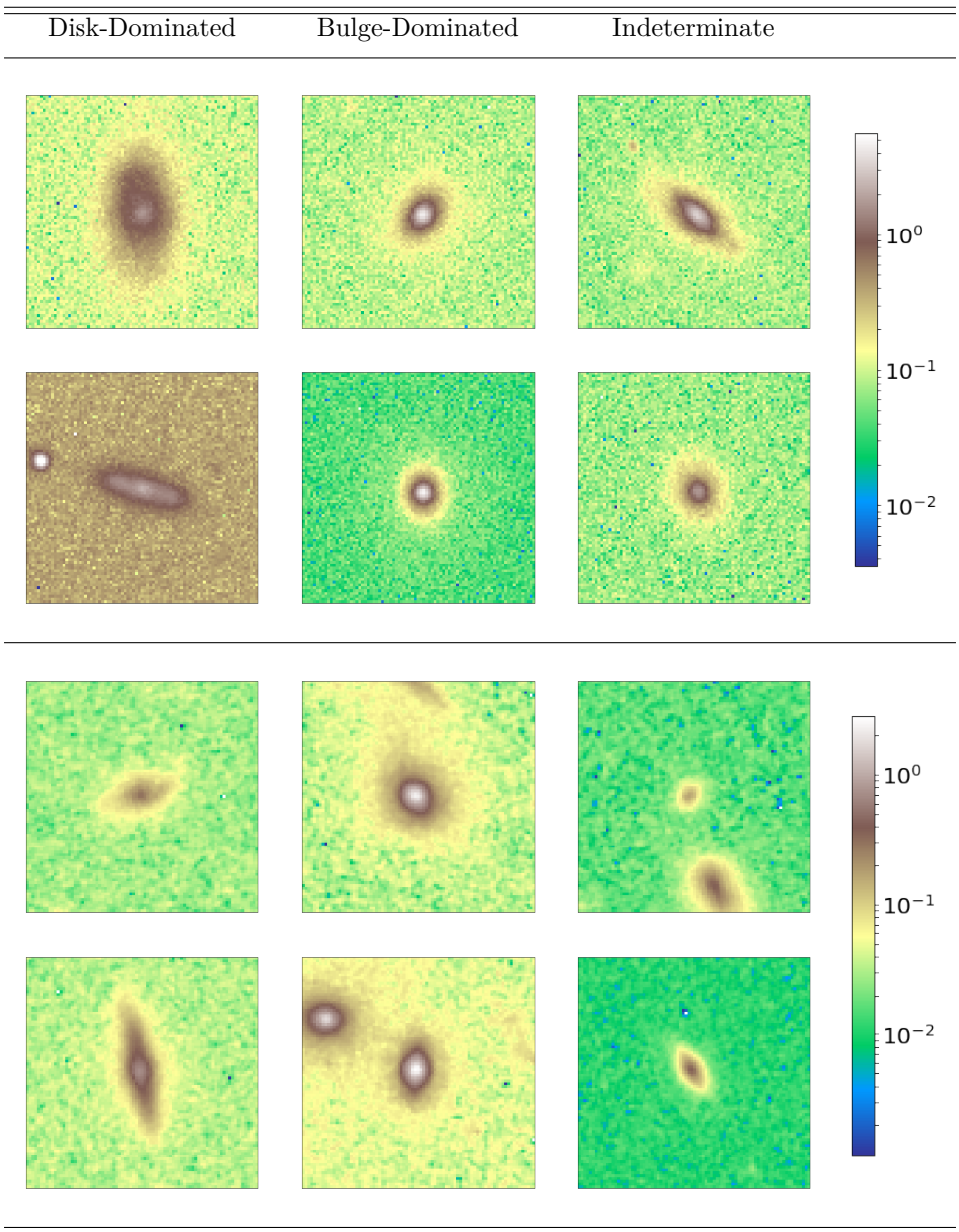


Figure 2.1: The above figure contains randomly chosen galaxies from both our data sets classified by GaMorNet as being disk-dominated (left column panels), bulge-dominated (middle column panels) or indeterminate (right column panels). Refer to § 2.4.1 for the definitions of these categories. The top two rows show SDSS cutouts, which are 33.07" × 33.07" (83 pixels × 83 pixels) and the bottom two rows show CANDELS cutouts, which are 4.98" ×4.98" (83 pixels × 83 pixels). During training, GaMorNet focuses on galaxies located at the center of the image and, thus, can process cutouts with other objects in the frame besides the central galaxy, as is evident from the images above.

major mergers, but these are still less than half the galaxy population. However, this study was done with a sample of only 2651 disks and 126 spheroids. Much larger studies, across a broader redshift range, will better illuminate the effect of mergers and AGN on galaxy evolution.

The two traditional ways of obtaining morphological classifications — visual classification and fitting light profiles — are not easily scalable to the large data volumes expected from The Large Synoptic Survey Telescope (LSST), the Wide Field Infrared Survey Telescope (WFIRST) and Euclid. The most popular galaxy light profile fitting program, GALFIT (Peng et al., 2002), and automated versions of it like GALAPAGOS (Hauler et al., 2022), suffer from the fact that the quality of the fit depends heavily on the input parameters, and when dealing with hundreds of thousands of galaxies, such hand-refinement of input parameters is an impossible task. There have been attempts to employ visual classifications on large galaxy samples via citizen science projects like Galaxy Zoo (Lintott et al., 2011, 2008a), but even these will fail to keep up with the coming data volume. Moreover, reliable visual classifications require a decent signal-to-noise ratio (S/N), take time to set up and execute, and require an extremely careful de-biasing of the vote shares obtained (Lintott et al., 2008a; Simmons et al., 2017).

For these reasons, using machine learning to classify galaxy morphology is particularly attractive. Data available from the Sloan Digital Sky Survey (SDSS) inspired early attempts at using machine learning to classify galaxies morphologically on a large scale (e.g., Banerji et al., 2010; **?**; **?**). These methods required the user to select proxies for morphology (such as color, concentration index, and spectral features) as inputs to the models. However, as the proxies could have an unknown and biased relation with galaxy morphology, these early networks were not ideal substitutes for the traditional classification methods.

In the last few years, convolutional neural networks (CNNs) have revolutionized the field of image processing (Lecun et al., 2015a; Schmidhuber, 2015a). They are ideal for galaxy morphology classification as they do not require selection of morphological proxies by hand and the network itself decides on which features of the image best discriminate among the different classes. The first serious attempt at using a CNN to classify galaxies morphologically came out of the "Galaxy Challenge" organized by Galaxy Zoo, where teams competed to reproduce the vote shares of each question in Galaxy Zoo 2 using a CNN (the top entry was by Dieleman et al., 2015). This was followed by the work of Huertas-Company et al. (2015), who used a CNN to reproduce visual classifications for galaxies in the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS). Tuccillo et al. (2018) used domain adaptation combined with one-component Seŕsic simulations to reproduce morphological classifications for $\sim$ 5000 CANDELS galaxies. There have also been

attempts at using CNNs for measuring photometric redshifts from galaxy images (Hoyle, 2016), doing star/galaxy separation (Kim & Brunner, 2017), detecting bars in galaxies (Abraham et al., 2018) and detecting mergers (Ackermann et al., 2018).

Most of the previous work involving the use of CNNs to study galaxy morphology has depended on the availability of a large training set of galaxies with known properties. However, if CNNs are to truly replace traditional methods for morphology classification, then there needs to be a single prescription/network that works across multiple data sets and does not require an already classified large training set.

In this paper, we introduce Galaxy Morphology Network (GaMorNet), a CNN that can classify galaxies according to their bulge-to-total ratio (L_B/L_T) for very different data sets without the need for a large, pre-classified training set of real galaxies. We first trained our network on simulated galaxies with both bulge and disk components and then transfer learned on a small part of our real sample to produce bulge/disk classifications for $\sim 80,000$ ($z \sim 0$) SDSS g-band galaxies and $\sim 20,000$ CANDELS ($z \sim 1$) H-band galaxies. A collection of 12 randomly chosen galaxy image cutouts from both data sets with their GaMorNet classifications is shown in Figure 2.1. Using the morphology classifications, we then examine the color-mass diagrams of the two samples, separated by morphology, in order to study the quenching of star formation at $z \sim 0$ and 1.

We describe the details of the SDSS and CANDELS data that we use in § 2.2. In § 2.3, we describe our simulations, the CNNs we use, and our transfer learning algorithm. In § 2.4, we present the results of the morphology classification, including the color-mass diagrams, and in § 2.5, we summarize our results and discuss future applications of GaMorNet.

We make all of the source code used in this work public along with the trained CNN models. We also release the GaMorNet morphological predictions for all of the SDSS and CANDELS galaxies in our data sets. All of the code is being made available under a GNU General Public License v3.0 and more details of the public data release are summarized in Appendix 2.A.

2.2 Data Sets Used

One of the primary aims of this paper is to demonstrate how GaMorNet can be used to identify bulge- and disk-dominated galaxies in different data sets without requiring extensive training on real data. Here, we work with two data sets: the SDSS (York et al., 2000), for nearby galaxies ($z \sim 0$), and CANDELS (Grogin et al., 2011; Koekemoer et al., 2011), for galaxies at $z \sim 1$. Together, these data allow us to probe galaxy evolution at different epochs of star formation and black hole growth.

We first created galaxy samples with which we train and test GaMorNet. Specifically, we identified galaxies in each survey for which bulge/disk decomposition had already been done or which had already been morphologically classified in some other way.

For the SDSS sample, we used 112,547 galaxies in the redshift range $0.02 \leq z \leq 0.07$ that were imaged in the g band and had bulge fractions determined by Simard et al. (2011), who fitted double Sérsic profiles with fixed indices $n = 4$ (pure bulge) and $n = 1$ (pure disk). For each galaxy, we prepared square cutouts of 167 pixels on a side, centered on the galaxy, with a resolution of $0.396''$ per pixel. We used 30,000 of these for the process of transfer learning, described in § 2.3.4 and the remaining 82,547 galaxies to test the performance of the network. In order to calculate the u-r color for each galaxy, we used extinction-corrected model SDSS magnitudes from the NYU-VAGC (Blanton et al., 2005) and adopted K corrections to $z = 0.0$. We obtained aperture and extinction-corrected specific star formation rates (sSFR) and stellar masses from the MPA-JHU DR7 catalog (Brinchmann et al., 2004; Kauffmann et al., 2003), which are calculated using SDSS spectra and broadband photometry.

For CANDELS reference data, we used Sérsic indices from Wel et al. (2012), who fitted the galaxy surface brightness profiles using GALFIT (Peng et al., 2002) with a single (free) Sérsic component. From this catalog, we selected galaxies with redshifts $0.7 \leq z \leq 1.3$ and "good" fits (defined by Wel et al., 2012 as matching the galaxy total magnitude, and having fits that converged, with parameters within an acceptable range). The ensuing sample of 28,946 $z \sim 1$ galaxies from the five CANDELS fields includes 6276 from the Great Observatories Origins Deep Survey–North (GOODS-N), 3942 from the Great Observatories Origins Deep Survey–South (GOODS-S), 7425 from the Cosmic Evolution Survey (COSMOS), 4911 from the Ultra Deep Survey (UDS) and 6392 from the All Wavelength Extended Growth Strip International Survey (AEGIS). We downloaded WFC3/IR F160W(H) mosaics from the CANDELS website[1], then for each galaxy, we made square cutouts of 83 pixels×83 pixels with a resolution of $0.06''$ per pixel. We used 7200 galaxy images for transfer learning and the remaining 21,746 for testing the performance of GaMorNet. We took the rest-frame U-R color, stellar mass, and sSFR of each galaxy from the 3D-HST catalog (Brammer et al., 2012); the stellar masses are based on spectral energy distribution (SED) fits to stellar population models with the FAST code (Kriek et al., 2009) as described in Skelton et al. (2014). The star formation rates used are from Whitaker et al. (2014) and assume that UV light from massive stars is re-radiated in the far-infrared.

1. http://arcoiris.ucolick.org/candels/data_access/Latest_Release.html

Table 2.1: Parameter Ranges for Simulated Galaxies

Component Name	Sérsic Index	Radius (Pixels)	Magnitude (AB)	Axis Ratio	Position Angle (degrees)
SDSS sample at $z \sim 0$					
Disk	1.0	10.0 - 30.0	15.0 - 22.0	0.3 - 1.0	-90.0 - 90.0
Bulge	4.0	4.0 - 17.0	Disk $\pm\,(0, 3.2)^a$	0.3 - 1.0	Disk $\pm\,(0, 15)^b$
CANDELS sample at $z \sim 1$					
Disk	1.0	12.0 - 25.0	17.0 - 27.8	0.3 - 1.0	-90.0 - 90.0
Bulge	4.0	4.0 - 14.0	Disk $\pm\,(0, 3.2)^a$	0.3 - 1.0	Disk $\pm\,(0, 15)^b$

[a]The bulge magnitude differs from the disk magnitude by a randomly chosen value between -3.2 and $+3.2$

[b]The bulge position angle differs from the disk position angle by a randomly chosen value between -15 and $+15$

`NOTE-` The above table shows the ranges of the various Sérsic profile parameters used to simulate the training data. Each simulated galaxy has an $n = 1$ disk and an $n = 4$ bulge component, where n is the Sérsic index. The distributions of all the simulated parameters are uniform except those for the bulge magnitude and bulge position angle. See § 2.3.1 for more details.

It is well known that dust extinction can redden galaxies, and significant reddening has been observed for high-redshift galaxies (Brammer et al., 2009; Cardamone et al., 2010; Williams et al., 2009). For the SDSS sample, we make no reddening correction since Schawinski et al. (2014) showed that dust correction has a negligible effect on the color-mass diagram for local galaxies. However, for the higher redshift CANDELS sample, we corrected the U-R colors using the Calzetti et al. (2000) extinction law:

$$\Delta(U - R) = 0.65 A_V. \tag{2.1}$$

The A_V values, taken from the 3D-HST catalog, come from SED fits to stellar population models (Brammer et al., 2012).

For both data sets, we used only a fraction of the available sample for transfer learning, leaving a much larger fraction for testing the performance of GaMorNet. This demonstrates that GaMorNet can effectively be trained initially on (more extensive) simulations, then re-trained using a small set of real data. Thereafter, GaMorNet can successfully classify a much larger set of real images because it learns to generalize beyond the training galaxies.

2.3 Training our Convolutional Neural Network — GaMorNet

The first hurdle in training a neural network to do morphological classifications is finding a large data set that has already been accurately classified. However, if neural networks are to be used

widely for astronomical analysis, we need a more flexible approach — one that does not require extensive analysis by old, slow (legacy) methods during the training phase and that can be adapted easily to new data sets. Here, we describe how to use simulated galaxies for the initial training of the classification network, followed by the application of a machine learning technique known as "transfer learning", wherein a much smaller set of galaxies, classified using a legacy method, is used to fine tune a partially trained network. This ensures that the network becomes adept at classifying real galaxies without requiring too many of them for the training process.

The process of training GaMorNet to classify galaxy morphologies consists of the following steps:

1. Simulating galaxies corresponding to the desired data set (here, SDSS or CANDELS).
2. Initial training of the neural network on those simulated images.
3. Retraining the neural network using a small part of the real data at hand; this process is known as transfer learning.
4. Testing a similar amount of real data to validate the results.
5. Processing the remainder of the real data through the trained network to obtain morphological classifications.

The galaxy simulations are described in § 2.3.1. § 2.3.2 contains a brief introduction to CNNs and describes the architecture of GaMorNet, while § 2.3.3 describes the initial training of GaMorNet on the simulations. In § 2.3.4, we describe how we perform transfer learning to produce the final trained state of GaMorNet.

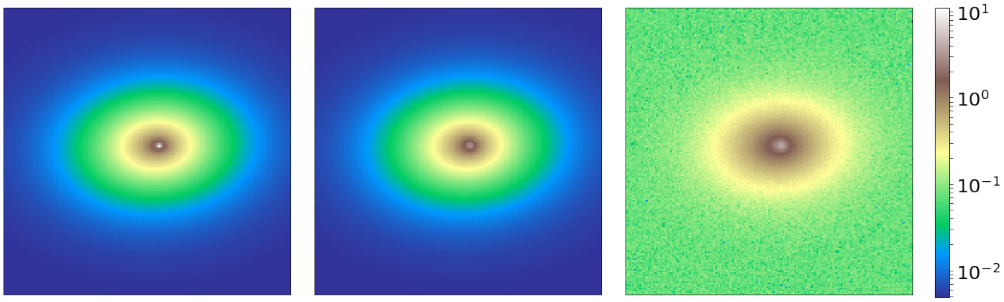
2.3.1 Simulations

We simulated galaxies using the GALFIT program (Peng et al., 2002), which is usually used to fit two-dimensional light profiles of galaxies. Here, we use it instead to create two-dimensional light profiles appropriate for the data sets we are interested in analyzing with GaMorNet.

For each data set, we simulated 100,000 galaxies consisting of a bulge (Sérsic component with fixed index $n = 4$; de Vaucouleurs (1948)) and disk (Sérsic index $n = 1$). The surface brightness for a galaxy with a Sérsic profile is given by

$$\Sigma(r) = \Sigma_e \exp\left[-\kappa\left(\left(\frac{r}{r_e}\right)^{1/n} - 1\right)\right], \tag{2.2}$$

where Σ_e is the pixel surface brightness at the effective radius r_e, n is the Seŕsic index, which controls



(a) Initial Simulated Image (b) Convolved with PSF (c) Noise Added

Figure 2.2: Three stages in simulating an SDSS galaxy. Left (a): Light profile generated by GALFIT with a bulge-to-disk ratio of 0.24. Center (b): The left image convolved with the SDSS PSF. Right (c): SDSS noise added to the middle image. See § 2.3.1 for details of the PSF convolution and noise addition.

the concentration of the light profile, and κ is a parameter coupled to n that ensures that half of the total flux is enclosed within r_e.

The parameters required to generate the Sérsic profiles are drawn from uniform distributions (except the bulge magnitude and position angle) and the ranges of the distributions used for both sets of simulations are summarized in Table 2.1. The galaxy size parameters were chosen to be representative of bright, local galaxies (Binney & Merrifield, 1998); bulges were chosen to have a half-light radius between 3.0 kpc and 6.0 kpc and disks were assigned half-light radii between 6.0 kpc and 10.0 kpc. To obtain the corresponding pixel sizes, we placed the samples at $z = 0.05$ and $z = 1.0$ (corresponding to the mean redshifts of the two samples described in § 2.2) using WMAP7 cosmology (Komatsu et al., 2011) and using the pixel scale for the appropriate data set. We ensured that the number of simulated galaxies was sufficiently large such that even when we consider subsets of galaxies with similar sizes, they not only span the entire range of L_B/L_T values but also mimic the overall bulge-to-total light ratio distribution.

The disk magnitudes were drawn from a uniform distribution chosen so as to include most galaxies at these redshifts, and the magnitude of each corresponding bulge is such that it differs from the disk magnitude by a randomly chosen value between -3.2 and 3.2. This was done to ensure that the bulge-to-total ratio varies between $\sim 5\% - 95\%$. Not enforcing this condition and allowing the bulge magnitude to be independent of the disk magnitude causes most galaxies in the training set to have a very high or a very low bulge-to-total ratio, which is not the case for most galaxies and, in any case, is not detectable. Instead, we want to train the network on a sufficient number of galaxies with intermediate bulge-to-total ratios.

To make the two-dimensional light profiles generated by GALFIT more closely resemble the actual data, we convolved them with a representative point-spread function (PSF), then added noise. For the SDSS simulations, we selected the coordinates of one of the real galaxies in our sample – R.A.: 213.26064353, Decl.: 0.14637573 and then reconstructed the PSF at the corresponding location in the detector using the PSF information stored in the relevant psField file that we obtained from SDSS. To generate the representative noise, we randomly selected 1000 cutouts from our SDSS sample, masked the sources in each cutout using SourceExtractor (Bertinl, 1996) and then read-in the non-masked pixel values to generate a large sample of noise pixels. We sampled this collection of noise pixels randomly to make two-dimensional arrays of the same size as that of the simulated images and then added them to the images. To make sure that the PSF chosen is representative, we reconstructed the PSFs for 12 more randomly chosen galaxies in our sample and convolved each one with a simulated SDSS galaxy, before adding noise. By inspecting the difference images between each image created using one of the new PSFs and the image created using the originally used PSF, we found the average pixel value of each of these difference images to be at least three orders of magnitude lower than the average pixel value of the galaxy image created using the original PSF.

For the CANDELS sample, we used the model PSF generated by Wel et al. (2012) for the COSMOS field and added noise following the same method as for the SDSS simulations. To make sure that the COSMOS PSF is representative, we followed a procedure similar to what we did for SDSS using the GOODS-S and UDS PSFs. We again found the average pixel value of the difference images to be at least three orders of magnitude lower than the average pixel value of the galaxy image created using the original PSF.

The effect of convolving the simulated galaxies with PSF and adding noise is depicted in Figure 2.2.

The goal behind convolving with the PSF and adding noise is not to recreate perfect replicas of the real galaxies in our samples but rather to train the network on realistic simulated images for which we know the intrinsic morphologies. This is why we arbitrarily selected the COSMOS PSF instead of making simulations for each field separately and used only one random SDSS PSF. If we were to make more of an effort to recreate exactly the real data in our sample, then the whole purpose having a CNN is lost. In that case, the neural network ends up having a low variance but an extremely high bias, as it is too closely tied to the training set. Instead, here, the CNN learns to generalize from fewer examples.

Since the galaxies were independently simulated, the simulation code could be trivially parallelized, and we make the simulation code available as a part of our public data release (see Appendix

2.A.4).

2.3.2 The Network

Artificial neural networks, consisting of many connected units called artificial neurons, have been studied for more than five decades now. The neurons are arranged in multiple layers as shown in the schematic representation in Figure 2.3; each network has an input layer via which the data is fed into the network and an output layer that contains the result of propagating the data through the network, with additional hidden layer(s) in between. Each neuron is characterized by a weight vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ and a bias b. The input to a neuron (coming from the outputs in the previous layer) is usually written as $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and the output of the neuron is given by

$$y = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b), \tag{2.3}$$

where σ is the chosen activation function of the neuron. The process of "training" an artificial neural network involves finding out the optimum set of weights and biases of all the neurons such that for a given vector of inputs, the output vector from the network, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, resembles the desired output vector $\hat{\boldsymbol{y}} = (\hat{y_1}, \hat{y_2}, \ldots, \hat{y_n})$ as closely as possible. The process of optimization is usually performed by minimizing a loss function, such as the popular cross-entropy loss function,

$$L = -\frac{1}{N} \sum_{j=1}^{N} \sum_{c=1}^{M} I_{j,c} \log(p_{j,c}) \tag{2.4}$$

where $I_{j,c}$ is a binary indicator function depicting whether class label c is the correct classification for the j^{th} observation. Also, p is the predicted probability (by the network) that observation j is from class c, M is the total number of classes and N is the total number of samples.

Out of the various algorithms available to minimize the loss function, one that is used very widely is stochastic gradient descent (SGD) and its different variants (Nielsen, 2015). In SGD, we estimate the gradient of L using a mini-batch of training samples and update the weights and biases according to

$$\begin{aligned} w' &= w - \eta \frac{\partial L}{\partial w} \\ b' &= b - \eta \frac{\partial L}{\partial b} \end{aligned} \tag{2.5}$$

where η is a small positive constant known as the learning rate. Calculation of the gradient is done using the back-propagation algorithm, and we refer the interested reader to Rumelhart et al. (1986)

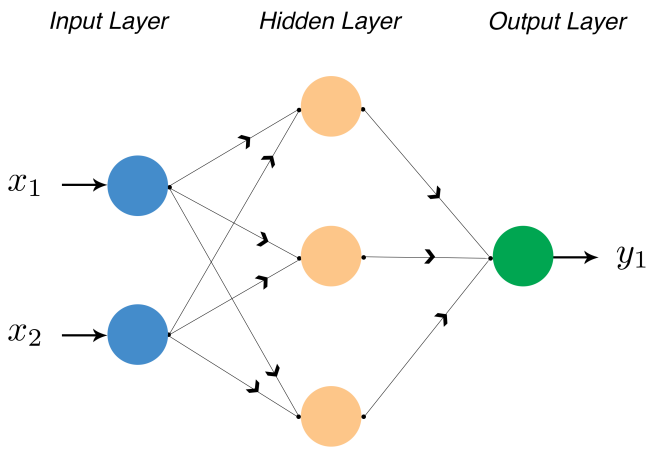


Figure 2.3: A schematic diagram showing a simple artificial neural network with a single hidden layer.

for details.

The artificial neural network that we use for this work is a Convolutional Neural Network (CNN; Fukushima, 1980; LeCun et al., 1998). This is a type of deep artificial neural network that has become extremely popular for image processing in recent years. The input to the network is the two-dimensional vector representation of an image, and in a convolutional layer, each unit receives input from a local image patch of the previous layer known as the receptive field. Convolution involves taking a filter of a particular size and repeatedly applying it (by moving it with a specific stride) to each part of the input image, resulting in a two-dimensional output map of activations called a feature map. The different units in the feature map share the same weight matrix, and hence, each feature map can be interpreted as trying to locate a particular feature at different locations in the image. Each convolutional layer is typically followed by a max-pooling layer wherein the dimensionality of the feature maps are reduced by only preserving the maximum value in a small patch and thus making the network invariant to minor distortions. The convolutional and max-pooling layers are usually followed by a few fully connected layers that use the output of the convolutional layers to infer the correct output for the input image. We refer an interested reader to Nielsen (2015) for a more detailed overview of the above concepts.

The architecture of GaMorNet is based on AlexNet (Krizhevsky et al., 2012), a CNN that won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILVRS), wherein different teams compete to classify about 14 million hand-annotated images. Very broadly speaking, the architecture of GaMorNet consists of five convolutional layers and three fully connected layers. Interspersed

Table 2.2: Structure of GaMorNet

Order	Type of Layer	Layer Description	Activation Function
1	Input	167×167(SDSS) | 83×83(CANDELS)	–
2	Convolutional	Filters: 96 | Filter Size: 11 | Strides: 4	ReLU[a]
3	Max-Pooling	Kernel Size: 3 | Strides: 2	–
4	Loc. Response Norm.	–	–
5	Convolutional	Filters: 256 | Filter Size: 5 | Strides: 1	ReLU[a]
6	Max-Pooling	Kernel Size: 3 | Strides: 2	–
7	Loc. Response Norm.	–	–
8	Convolutional	Filters: 384 | Filter Size: 3 | Strides: 1	ReLU[a]
9	Convolutional	Filters: 384 | Filter Size: 3 | Strides: 1	ReLU[a]
10	Convolutional	Filters: 256 | Filter Size: 3 | Strides: 1	ReLU[a]
11	Max-Pooling	Kernel Size: 3 | Strides: 2	–
12	Loc. Response Norm.	–	–
13	Fully Connected	No. of neurons: 4096	tanh
14	Dropout	Dropout probability: 50%	–
15	Fully Connected	No. of neurons: 4096	tanh
16	Dropout	Dropout probability: 50%	–
17	Fully Connected	No. of neurons: 3	softmax

[a]Rectified Linear Unit

`NOTE-` The various layers of GaMorNet along with the important parameters of each layer and the corresponding activation functions are shown in the table above. The architecture of GaMorNet is based on AlexNet and, broadly speaking, consists of five convolutional layers followed by three fully connected layers. The source code for GaMorNet is made public as described in Appendix 2.A.1

between these are local response normalization, max-pooling and dropout layers. The dropout layers help to prevent over-fitting by randomly ignoring or "dropping out" some number of layer outputs. The size of the input layer corresponds to the size of the images being fed-in, and the output layer corresponds to the three classes into which the galaxies are separated, which are defined in § 2.3.3. The output layer happens to have the softmax activation function and thus, the output value of the three output neurons can be interpreted as the network's prediction probability that the input galaxy is in the corresponding category. In total, GaMorNet has 17 layers, the details of which are summarized in Table 2.2. Figure 2.4 shows a schematic diagram of GaMorNet.

We implemented GaMorNet using TFLearn[2][3], which is a high-level Application Program Interface for TensorFlow[4], an open source library widely used for large-scale machine learning applications. We make the source code of GaMorNet available as a part of our public data release (see Appendix 2.A.1 for more details).

2. http://tflearn.org

3. Also available in Keras now as a part of the data release

4. https://tensorflow.org

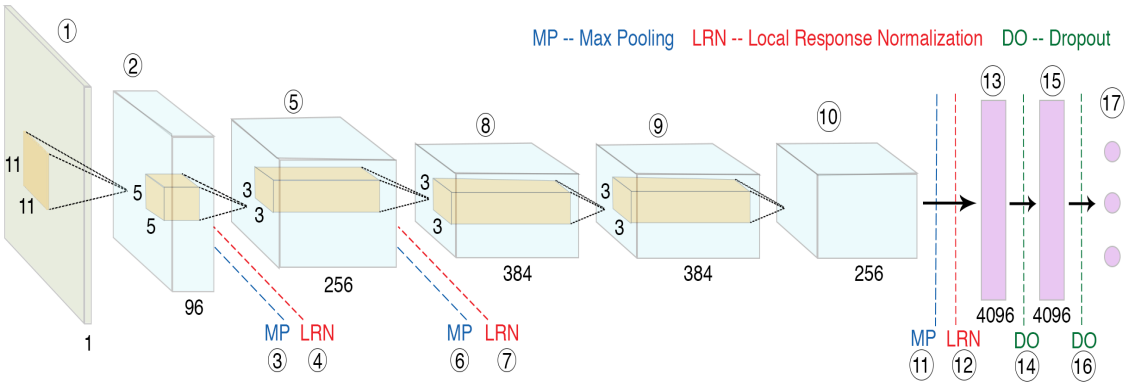


Figure 2.4: Schematic diagram of GaMorNet, a CNN optimized to identify whether galaxies are bulge-dominated or disk-dominated. Its architecture, which is based on AlexNet (Krizhevsky et al., 2012), consists of five convolutional layers and three fully connected layers. Between these layers are max-pooling, local response normalization, and dropout layers.t The numbers inside the circles refer to the layer number and corresponding details for each layer can be found by looking up the corresponding layer order number in Table 2.2.

2.3.3 Initial Training

Using the two sets of simulations corresponding to the SDSS and CANDELS data sets, we trained two different networks, both with the same structure as described in § 2.3.2. Henceforth, we refer to the networks trained on SDSS and CANDELS simulations as GaMorNet-S and GaMorNet-C respectively. During the training process, we trained the networks to separate galaxies into three different categories:

1. Galaxies with $L_B/L_T < 0.45$, i.e., disk-dominated.
2. Galaxies with $L_B/L_T > 0.55$, i.e., bulge-dominated.
3. Galaxies with $0.45 \leq L_B/L_T \leq 0.55$, i.e., indeterminate.

Here, L_B is the luminosity of the bulge component, and L_T is the total luminosity of the galaxy. Since these galaxies are simulated, we used our knowledge of the actual L_B/L_T for each galaxy to train the network.

Of the 100,000 galaxies simulated for each data set, we used 90% for training and the rest for validation. The validation set was used to tune the different hyper-parameters in the network (like the learning rate described in § 2.3.2). We use a learning rate of 0.0001 and a batch size of 64, as these lead to $> 95\%$ accuracy on the validation set and run-times of $\sim \mathcal{O}(1\,\mathrm{hour})$ on Tesla P100 GPUs. The batch size refers to the number of training samples the network works through before the model's internal parameters are updated

During the training process, we used the categorical cross-entropy loss function and minimized it

using the momentum optimizer, which is a variant of SGD and accelerates SGD in the relevant direction besides dampening oscillations during the minimization process. Both SGD and the categorical cross-entropy loss function are described in § 2.3.2.

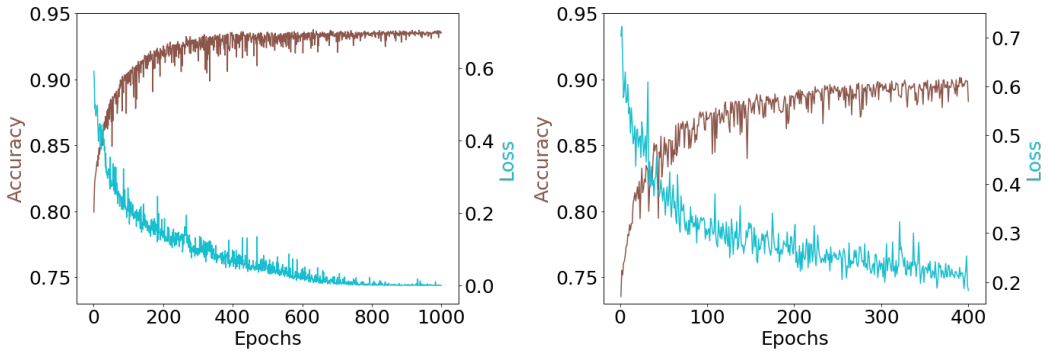
An "epoch" of training refers to running all of the training images through the network once. After each epoch of training, we evaluated the value of the loss function and calculated the accuracy on the validation set. The process of calculating the accuracy involves running all of the images in the validation set through the network. Since the output layer in our network is a softmax layer, the output value of each neuron can be interpreted as the network's predicted probability of the galaxy image to belong to the particular category corresponding to that neuron. A galaxy is said to belong to the L_B/L_T category for which the predicted probability is the highest, and the accuracy was calculated as the number of galaxies classified correctly divided by the total number of galaxies. It is important to note here that we used an additional criterion for classifying the real images later on, as described in § 2.4.1.

We trained both the networks until the values of the accuracy and the loss function stabilized and a significant gain in accuracy did not seem probable with further training. This constituted training GaMorNet-S for 1000 epochs and GaMorNet-C for 400 epochs. Both learning curves are shown in Figure 2.5, which shows the accuracy as well as the value of the loss function after each epoch of training. GaMorNet-S and GaMorNet-C achieved net accuracies of 93.55% and 88.33%, respectively, on the simulated images being used for validation; note that these are simulated images that the network did not "see" during the process of training.

2.3.4 Transfer Learning

CNNs have an extremely large number of free parameters (weights and biases) that need to be tuned during the process of training, and thus, if the size of the training set is not sufficiently large, there is a chance of "over-fitting" after a certain number of epochs of training. That is, with further training, the accuracy of the network increases on the training data but not on the test data, and hence, the network fails to generalize.

Transfer learning involves taking a network trained on a particular data set and optimized for a particular task, and re-tuning the weights and biases for a slightly different task or data set. The advantage here is that a much smaller training set can be used to re-tune the network than to train it from scratch. Transfer learning as a data-science concept has been around since the 1990s (Pan & Yang, 2010), and has been applied to a wide variety of tasks, including image classification (Kulis



(a) SDSS (b) CANDELS

Figure 2.5: Learning curves for the process of training GaMorNet on simulated galaxy images. The accuracy evaluated on the validation set (brown curves, left axes) and the value of the loss function after each epoch of training (blue curves, right axes) are shown for both GaMorNet-S and GaMorNet-C (left and right panels, respectively). GaMorNet-S achieves an accuracy of 93.55% after 1000 epochs of training, and GaMorNet-C achieves an accuracy of 88.33% after 400 epochs of training. For more details about the training process, see § 2.3.3.

et al., 2011; Li et al., 2014; Zhu et al., 2011). As an example, transfer learning was recently applied to detect galaxy mergers (Ackermann et al., 2018), starting from a network that could accurately identify images of everyday objects like cars, cats, dogs, etc.

In the present work, since we want to enable morphological classification even in the absence of a large training set, we use only a small fraction of the SDSS and CANDELS data sets for training. Specifically, we take the network trained on simulations and then re-train it by transfer learning on $\sim 25\%$ of the real SDSS and CANDELS galaxy images.

In transfer learning, it is common to freeze the weights and biases in the initial layers of the network (i.e., those close to the input layer), while allowing variations in layers close to the output layer. The logic behind this approach is that, in a CNN, the deeper feature maps identify more complicated features while the earlier layers identify more basic features (like lines, shapes, and edges). Since transfer learning re-trains a network to do a slightly different task than it was initially trained to do, it is the last few layers that need to be re-tuned for the task at hand. At the same time, since the earlier layers correspond to more basic features, we do not expect that they will need re-tuning. We heuristically tested a combination of the various options mentioned above, and chose the one that maximized accuracy, while not showing any signs of over-training. The details of the transfer learning method used in both cases are summarized in Table 2.3.

For the SDSS data, we have access to estimates of L_B/L_T for each galaxy from Simard et al. (2011), wherein each galaxy was fitted with an $n = 4$ bulge and an $n = 1$ disk component. We

Table 2.3: Transfer Learning Parameters

Network	Non-Trainable Layers	Layers Trained from Previous Training	Layers Trained from Scratch	Learning Rate
GaMorNet-S	None	All Convolutional Layers (2,5,8,9,10)	Last 3 Fully Connected Layers (13,15,17)	0.00001
GaMorNet-C	First 3 Convolutional Layers (2,5,8)	Last 2 Convolutional Layers + First Fully Connected Layer (9,10,13)	Last 2 fully Connected Layers (15,17)	0.00001

[a] These layers were optimized during the initial training on simulations and then frozen at those values for the transfer learning step.

NOTE- Details of the transfer learning algorithm used for both the SDSS and CANDELS networks. The numbers in parentheses refer to the layer numbers according to Table 2.2. The above parameters were chosen by heuristically testing various options and choosing the ones that maximized accuracy, while not showing any signs of over-training.

used this as the "ground truth" for separating galaxies into the three categories defined in § 2.3.3. We randomly selected 10,000 galaxies from each category to make up our transfer learning training data set; this constitutes about a quarter of the full SDSS sample. We found that during transfer learning, it is important to have an equal number of galaxies from each category in the training set because otherwise, the network attempts to maximize accuracy in the category with more samples at the cost of other categories. Since both our samples have many more disk-dominated than bulge-dominated galaxies, a randomly selected training set would result in a very high accuracy in classifying disk-dominated galaxies but a very low accuracy in classifying bulge-dominated galaxies. Using the configuration given in Table 2.3, we trained GaMorNet-S for 300 epochs.

For the CANDELS data, no two-component bulge-disk decompositions were available in the literature. Thus, we translated the Sérsic indices from Wel et al. (2012) into the three classifications used by GaMorNet using results from Simmons & Urry (2008), who analyzed CANDELS-depth HST ACS simulations of bulge+disk galaxies. The authors fitted single Sérsic profiles to their simulations in order to find the correspondence between Sérsic index and actual L_B/L_T. Guided by their result in the redshift bin $z = 1.075$ (see their Fig. 19), appropriate for the CANDELS galaxies we wish to classify, we define galaxies with $n < 2.0$ as disk-dominated, $n > 2.5$ as bulge-dominated, and $2.0 \leq n \leq 2.5$ as indeterminate.

To illustrate these choices, we reproduce in Figure 2.6 the Simmons & Urry (2008) results, specifically, the range in Sérsic index corresponding to different L_B/L_T values for the simulated galaxies. The three broad classifications assigned by GaMorNet-C — disk-dominated, indeterminate, and bulge-dominated — are shown as shaded regions. There is no unique or perfect way to go from

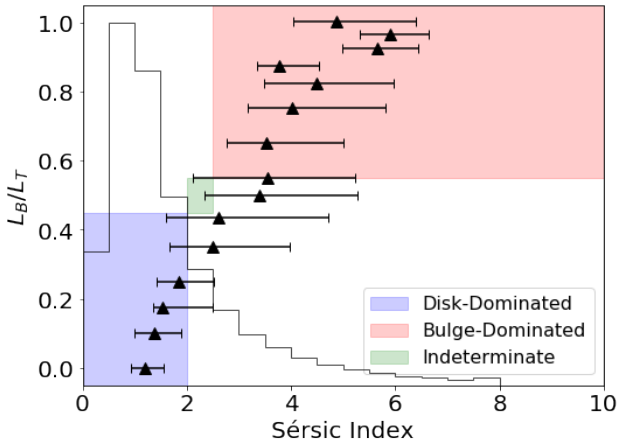


Figure 2.6: The triangles show the input bulge-to-total ratio (L_B/L_T) versus fitted Sérsic index for the galaxies simulated by Simmons & Urry (2008; adapted from the lowest panel in their Figure 19). The plotted points are the median of each bin's distribution, and the error bars mark the central 68% of sources in the bin. The shaded regions correspond to our definitions of the three output classes used by GaMorNet-C. The histogram shows the distribution of the Sérsic index for all the galaxies in our CANDELS sample, most of which are disk-dominated (see § 2.2). Clearly, all galaxies with $n < 2$ are truly disk-dominated (i.e., have $L_B/L_T < 0.45$) but, because of the spread in Sérsic indices, some disk-dominated or intermediate galaxies may get misclassified as bulge-dominated. Although a higher n threshold (for, e.g., $n \sim 6$) would lead to a purer bulge-dominated sample, for reasons mentioned in § 2.3.4, it would make the transfer learning sample insufficiently small. Note that readers can choose different bin boundaries, doing their own transfer learning step on the simulation-trained network made available via § 2.A.2

Sérsic index to L_B/L_T; although, the choice of $n < 2$ is pretty clean, i.e., all such galaxies have $L_B/L_T < 0.45$ and are disk-dominated. For $n > 2.5$, most galaxies are bulge-dominated (i.e., have $L_B/L_T > 0.55$) as is evident from the top-right portion of the figure; although, a few disk-dominated galaxies with $L_B/L_T \sim 0.4$ may be incorrectly included in that category.

A higher n threshold (for, eg., $n \sim 6$) leads to a purer bulge-dominated sample, but drastically reduces the number of bulge-dominated galaxies available for transfer learning, as is evident from the histogram shown in Figure 2.6. As mentioned previously, we need roughly equal numbers of galaxies in each bin for the training process during transfer learning, and thus, the upper limit on the total number of galaxies available for training is set by the size of the least populous bin. The above choices ensure a sufficient number of galaxies in each category (needed for the transfer learning step) and produce statistically acceptable classifications. Readers can set these boundaries differently, as appropriate to their science goals, using GaMorNet-C models trained only on the bulge + disk simulations made publicly available via § 2.A.2. Instructions on how to train these models for transfer learning are available in the GitHub repository.

Using the above definitions of the three classes, we re-trained the simulation-trained GaMorNet-

C for 75 epochs using the Transfer Learning configuration in Table 2.3; note that only the weights and biases in the last two of the total five convolutional layers are adjusted during the transfer learning step. For this process, we used 2400 galaxies from each of the three morphological categories, or about a quarter of the total CANDELS sample.

Table 2.4: Classification Probabilities for 82,547 SDSS Galaxies

ObjID[a]	R.A. (deg)	Decl. (deg)	Disk Prob.	Bulge Prob.	Indeterminate Prob.	Classification
587722953...	237.4210352	0.2367580	0.1356	0.4439	0.4205	indeterminate
587722981...	202.6811651	-1.0804622	1.0000	0.0000	0.0000	disk-dominated
587722982...	219.5687676	-0.3497467	0.9384	0.0001	0.0615	disk-dominated
587722983...	213.2606435	0.1463757	0.9977	0.0000	0.0023	disk-dominated
587722983...	216.5747982	0.1543351	0.0590	0.7170	0.2240	indeterminate
⋮	⋮	⋮	⋮	⋮	⋮	⋮

[a]These are pre-DR8 ObjIDs. Complete object IDs are not mentioned here for stylistic purposes.
NOTE- GaMorNet-S classification probabilities (of being disk-dominated, bulge-dominated, or indeterminate) and final classification for all the galaxies in our SDSS test sample. This table is published in its entirety as a part of the public data release (Appendix 2.A.3). The first five entries are shown here for guidance regarding its form and content.

Table 2.5: Classification Summary for 82,547 SDSS Galaxies

Predictions:- Disks: 47,656 | Bulges: 7963 | Indeterminate: 26,928			
		GaMorNet-S classifications	
		Disks	Bulges
Simard et al. (2011) classifications	Disks	47,526	329
	Bulges	94	7552

Percentages	
GaMorNet-classified disks that SC[a] also classified as disks	99.73%
GaMorNet-classified disks that SC[a] classified as bulges	0.20%
GaMorNet-classified bulges that SC[a] also classified as bulges	94.84%
GaMorNet-classified bulges that SC[a] classified as disks	4.13%
Total percentage of galaxies misclassified	0.7%

[a]SC refers to Simard et al. (2011) classifications.
NOTE- Results of running the entire SDSS test set of 82,547 galaxies through GaMorNet-S. Values in the top section refer to the number of galaxies in each category as predicted by GaMorNet-S with respect to the Simard et al. (2011) classifications. For example, the top-left cell value of 47,526 means that out of the 47,656 predicted disks, 47,526 are also classified as disks by Simard et al. (2011).

2.4 Results

2.4.1 Morphology Results

After using about a quarter of the images for transfer learning, the remaining 82,547 galaxies in the SDSS sample were used as our test set. Since GaMorNet's output layer consists of three softmax neurons whose output values sum to 1, each value can be interpreted as the probability that a galaxy belongs to that L_B/L_T category. These probability values are the primary output of GaMorNet. However, in order to compare our results with previous classifications and keeping in mind situations that necessitate rigid classifications, we transform the probability values into classifications.

After some experimentation, we arrived at this decision tree for classification:

1. Disk-dominated when GaMorNet-S reports $\geq 80\%$ probability that $L_B/L_T < 0.45$.
2. Bulge-dominated when GaMorNet-S reports $\geq 80\%$ probability that $L_B/L_T > 0.55$.
3. Otherwise, indeterminate.

This is slightly different than the criterion we used for the initial training, as those galaxies were idealized, and the classifications were unambiguous. For the real galaxies, simply taking the highest probability neuron, including probabilities below 80%, made the classifications far less accurate. Requiring a threshold of 80% greatly improved the classification accuracy at the expense of increasing the number of indeterminate galaxies.

For each galaxy, we have access to its bulge-to-total ratio, i.e., L_B/L_T value from Simard et al. (2011), which we consider to be the "true" value. For mapping L_B/L_T to a classification of being bulge- or disk-dominated, we used the same criterion as during the initial training, outlined at the beginning of § 2.3.3.

Individual morphological classifications by GaMorNet-S are reported in Table 2.4 and Table 2.5 compares the GaMorNet-S and Simard et al. (2011) classifications of SDSS galaxies. Assuming the latter are "true", for disk-dominated galaxies, we achieved an accuracy of 99.7% and for bulge-dominated galaxies, we achieved an accuracy of 94.8%, resulting in a net misclassification rate of 0.7%. A total of 26,928 galaxies, or $\sim 32\%$ of the SDSS test set, were found to have indeterminate morphologies.

For the CANDELS data set, there were 21,746 galaxies in the test set. We classified these using GaMorNet-C and, again, experimented with thresholds for the final neuron values in order to arrive at an acceptable balance between accuracy and fraction with indeterminate morphologies. The thresholds for the CANDELS classification (which are different from those adopted for the

Table 2.6: Classification Probabilities for 21,746 CANDELS galaxies

Field	ID[a]	R.A. (deg)	Decl. (deg)	Disk Prob.	Bulge Prob.	Indeterminate Prob.	Classification
GOODSN	19	189.146484	62.095764	0.3356	0.3372	0.3272	indeterminate
GOODSN	32	189.131485	62.097328	0.3762	0.2722	0.3516	disk-dominated
GOODSN	63	189.117432	62.101723	0.3709	0.2877	0.3414	disk-dominated
GOODSN	68	189.149979	62.101768	0.4039	0.1798	0.4163	indeterminate
GOODSN	72	189.143295	62.102295	0.3006	0.3312	0.3683	indeterminate
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

ID refers to the IDs assigned by the CANDELS team (Grogin et al., 2011; Koekemoer et al., 2011)
NOTE- GaMorNet-C classification probabilities (of being disk-dominated, bulge-dominated, or indeterminate) and final classification for all the galaxies in our CANDELS test sample. This table is published in its entirety as a part of the public data release (Appendix 2.A.3). The first five entries are shown here for guidance regarding its form and content.

Table 2.7: Classification Summary for 21,746 CANDELS Galaxies

Predictions:- Disks: 12,549 | Bulges: 580 | Indeterminate: 8617			
		GaMorNet-C classifications	
		Disks	Bulges
Wel et al. (2012) classifications	Disks	11,524	121
	Bulges	992	456

Percentages	
GaMorNet-classified disks that VdwC[a] also classified as disks	91.83%
GaMorNet-classified disks that VdwC[a] classified as bulges	7.90%
GaMorNet-classified bulges that VdwC[a] also classified as bulges	78.62%
GaMorNet-classified bulges that VdwC[a] classified as disks	20.86%
Total percentage of galaxies misclassified	5.3%

[a]Wel et al. (2012) Classifications
NOTE-Results of running the entire CANDELS test set of 21,746 galaxies through GaMorNet-C. Values in the top section refer to the number of galaxies in each category as predicted by GaMorNet-C with respect to the Wel et al. (2012) classifications. For example, the top-left cell value of 11,524 means that out of the 12,549 predicted disks, 11,524 are also classified as disks by Wel et al. (2012).

SDSS data) are:

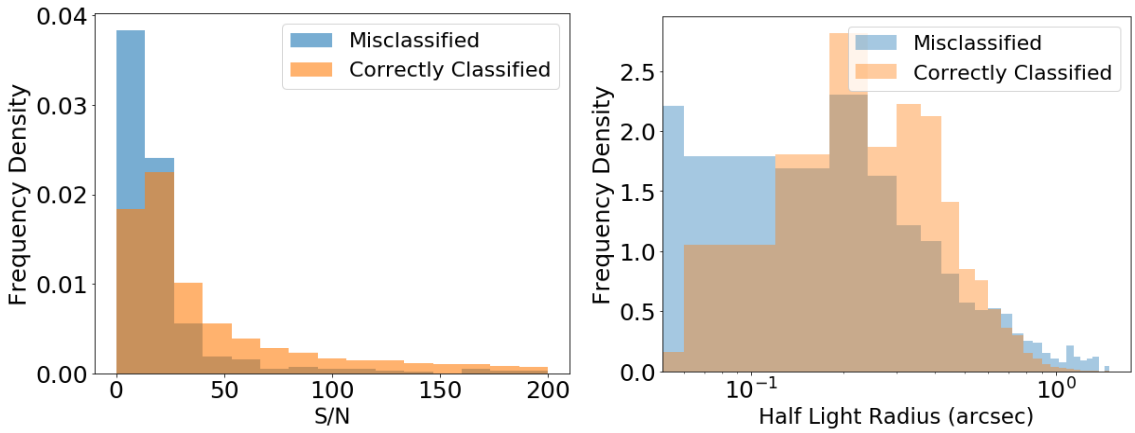
1. Bulge-dominated if GaMorNet-C reports $\geq 55\%$ probability that $L_B/L_T > 0.55$.
2. Disk-dominated if GaMorNet-C reports $\geq 36\%$ probability that $L_B/L_T < 0.45$ *and* this probability exceeds the probabilities of $L_B/L_T > 0.55$, $0.45 \leq L_B/L_T \leq 0.55$.
3. Otherwise, indeterminate.

The choice of these confidence thresholds and their impact on the results is discussed later in this section.

Table 2.6 reports the individual morphological classifications by GaMorNet-C, and Table 2.7 compares these to the results of Wel et al. (2012). From the Sérsic index of each galaxy (Wel et al., 2012), we derive its L_B/L_T following Simmons & Urry (2008) as described in § 2.3.4. Thereafter, we map these values to a classification of being bulge- or disk-dominated using the same criterion as we did during initial training, as described in § 2.3.3. Assuming these as the "true" classifications, GaMorNet-C has an accuracy of 91.8% for disk-dominated galaxies and 78.6% for bulge-dominated galaxies, or a net misclassification rate of 5.3%. A total of 8617 galaxies were classified in the indeterminate category, which is $\sim 39\%$ of the CANDELS test set.

The misclassification rate of CANDELS galaxies is higher than that of SDSS galaxies. To find out why, we investigated various relevant statistics for the misclassified galaxies. The two most significant variables were the S/N and the half-light radius (taken from Wel et al., 2012). Figure 2.7 shows the distribution of both the correctly classified and misclassified galaxies over these parameters. Although both the misclassified and correctly classified galaxies are distributed similarly over S/N, the misclassified population peaks much more sharply at a lower S/N, showing that a much larger fraction of the misclassifed sample has a low S/N compared to the correctly classified fraction. Similarly, a much larger fraction of the misclassified sample has low values of r_e compared to the correctly classified galaxies. Therefore, we conclude that the misclassified galaxies are essentially galaxies with a small half-light radius comparable to the PSF and/or a low S/N, and thus, it is inherently difficult for GaMorNet-C to correctly classify these galaxies. The misclassified population in the SDSS data set also peaks more sharply at a lower value of r_e compared to the correctly classified galaxies; although, we have poor statistics for this, as the misclassification rate is $< 1\%$.

The choice of the confidence threshold values to classify a galaxy as bulge- or disk-dominated primarily affects two parameters: the misclassification rate and the number of indeterminate galaxies. Having a high confidence threshold results in a low misclassification rate but a high number of indeterminate galaxies, and vice-versa. We show in Figure 2.8 how changing the value of the con-



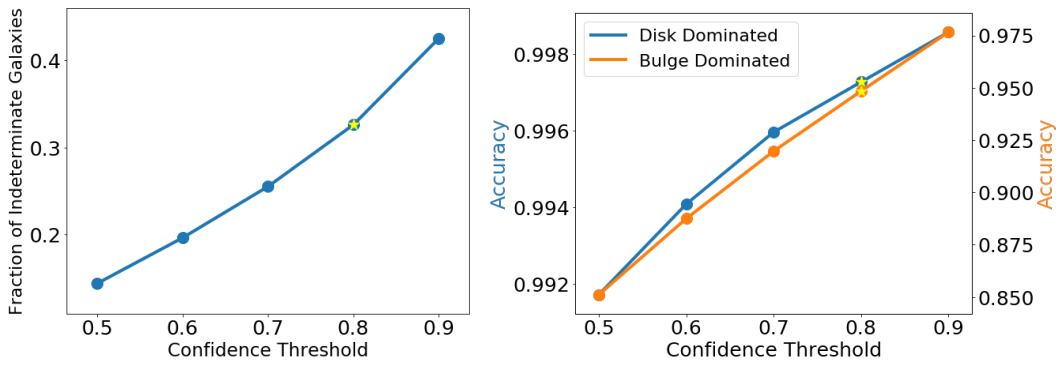
(a) (b)

Figure 2.7: The normalized distribution of correctly classified and misclassifed CANDELS galaxies in the test set as a function of the signal-to-noise ratio (S/N) and half-light radius (r_e). Both plots show that compared to the correctly classified galaxies, a higher fraction of the misclassified galaxies have a low S/N ratio and/or small r_e. 'Frequency density' refers to the number counts normalized to form a probability density.

fidence threshold affects the number of indeterminate galaxies and the accuracy of both the bulge- and disk-dominated galaxies for the SDSS sample. We chose a threshold value of 0.8 or 80% but as the figure shows, even with a threshold of 60%, it is possible to get $> 85\%$ accuracy for both bulge- and disk-dominated galaxies with an indeterminate fraction as low as $\sim 20\%$.

For the CANDELS data set, setting a common/joint threshold as high as we did for the SDSS data led to most of the data being classified as indeterminate. Thus, we use separate confidence thresholds for the disk and bulge classifications, and the variation of the indeterminate fraction and accuracy with both thresholds is shown in Figure 2.9. We chose the final threshold values of 0.36 and 0.55 for the disk- and bulge-dominated galaxies, respectively, as a compromise between the two competing requirements of having a low indeterminate fraction and high accuracy.

For our choice of confidence thresholds, the indeterminate fraction is $> 25\%$ of the test set for both SDSS and CANDELS. This indeterminate fraction consists of two kinds of galaxies: those with intermediate bulge-to-total ratios (i.e., $0.45 \leq L_B/L_T \leq 0.55$) and those for which the network is not confident enough to make a prediction, because of low S/Ns and/or small sizes. For comparison, Powell et al. (2017) used GALFIT to do single Sérsic fits to 4479 GOODS-S and GOODS-N galaxies; they found that $\sim 38\%$ of the population could not be classified due to poor fits ($\chi^2 > 1.5$) or galaxies having $2.0 < n < 2.5$. Similarly, large fractions of Galaxy Zoo classifications have $\lesssim 80\%$ agreement among classifiers (Land et al., 2008). Thus, even with stringent confidence threshold



(a) (b)

Figure 2.8: Relation of confidence threshold to completeness and accuracy of classification, for the SDSS data set. Left (a): The fraction of indeterminate galaxies increases with increasing confidence threshold. Right (b): The accuracy of both disk-dominated (blue line, left axis) and bulge-dominated (orange line, right axis) classifications increases with increasing confidence threshold. We decided on a confidence threshold of 0.8 (or 80%) for GaMorNet-S (star in both plots) as the optimal compromise between accuracy and completeness.

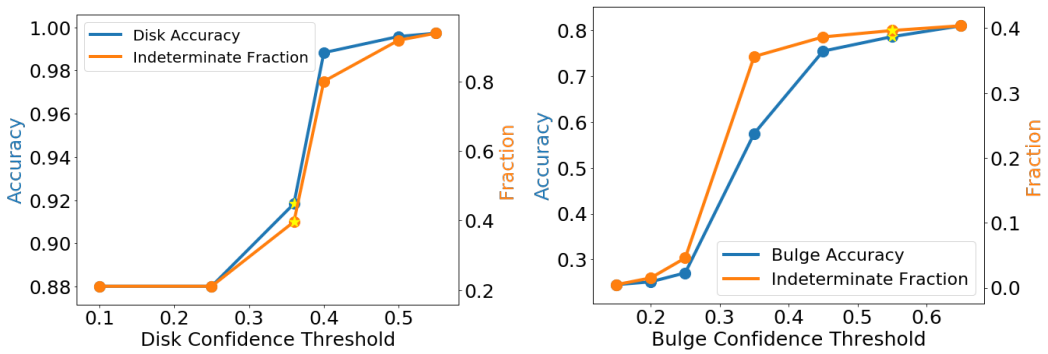
values, GaMorNet is able to match the indeterminate fraction of traditional studies.

The choice of the confidence threshold is arbitrary and should be chosen appropriately for the particular task at hand. Toward this end, Figures 2.8 and 2.9 can be used to asses the trade-off between accuracy and completeness for both the samples. We have emphasized accuracy over completeness, since we have very large samples already and can show that the misclassified objects simply have lower S/Ns and/or are too compact to classify accurately.

2.4.2 Color - Mass Results

In this section, we study the quenching of star formation in $z \sim 0$ (SDSS) and $z \sim 1$ (CANDELS) galaxies by examining their color-mass diagrams constructed using the morphological classifications obtained in § 2.4.1. Refer to § 2.2 for details about the calculation of colors, masses, and sSFR for both samples.

Figure 2.10 shows the *u-r* color-mass diagram for the $z \sim 0$ SDSS test set separated by disk- and bulge-dominated morphologies. The color of each point in panels (a) and (b) refer to the specific star formation rate of each galaxy. The contours in all plots refer to the linear number density of galaxies, and the straight lines in panels (c) and (d) mark the location of the green valley, which we define to be the region between the colors mentioned below:



(a) (b)

Figure 2.9: Relation of confidence threshold to the accuracy (blue lines, left axes) and completeness (orange lines, right axes) of GaMorNet-C classification of the CANDELS data set. Stars denote the adopted confidence thresholds. Left (a): For the chosen disk confidence threshold of 0.36, provided the probability of being disk-dominated exceeds the probabilities of being bulge-dominated or indeterminate, the classification accuracy is better than 92% and the indeterminate fraction <40%. Right (b): For the chosen bulge confidence threshold of 0.55, we obtain an accuracy of >80% and indeterminate fraction <40%.

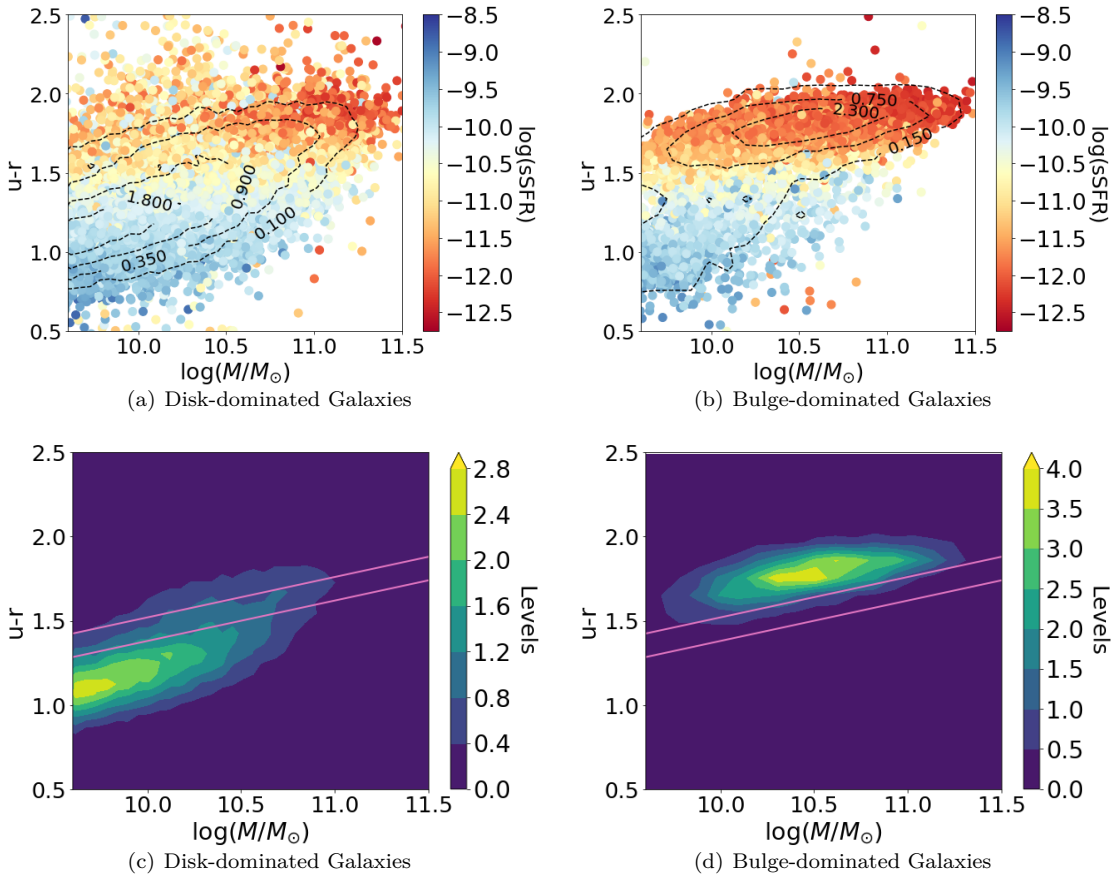
$$u - r(M) = -1.02 + 0.24 \times \log(M/M_{\odot}) \tag{2.6}$$

$$u - r(M) = -0.88 + 0.24 \times \log(M/M_{\odot}). \tag{2.7}$$

The *U*-*R* color-mass diagram for the $z \sim 1$ CANDELS data is shown in Figure 2.11 and this figure is arranged in the same way as Figure 2.10. We define the green valley, in this case, as the region between *U*-*R* colors 1.0 and 1.5.

The demographics of galaxies by color and morphology for both samples is summarized in Table 2.8. Note that the total number of galaxies in the table does not match that in § 2.4.1 as we have omitted galaxies that lack estimates of either mass or sSFR. The omitted fraction is $\sim 0.7\%$ and $\sim 3.4\%$ for the SDSS and CANDELS samples, respectively.

For both the samples, we see that both bulge- and disk-dominated galaxies span the entire range of colors (i.e., we see examples of red disk-dominated galaxies as well blue bulge-dominated galaxies). As expected, the disk-dominated galaxies peak in the blue cloud while the bulge-dominated galaxies dominate the red sequence. The green valley is not a feature for either morphology; that is, there is no bimodality. Rather, the number density of galaxies declines monotonically from a red or blue peak. Thus, the green valley only arises when plotting the color-mass diagram of all galaxies together, as was first pointed out for $z \sim 0$ galaxies by Schawinski et al. (2014).



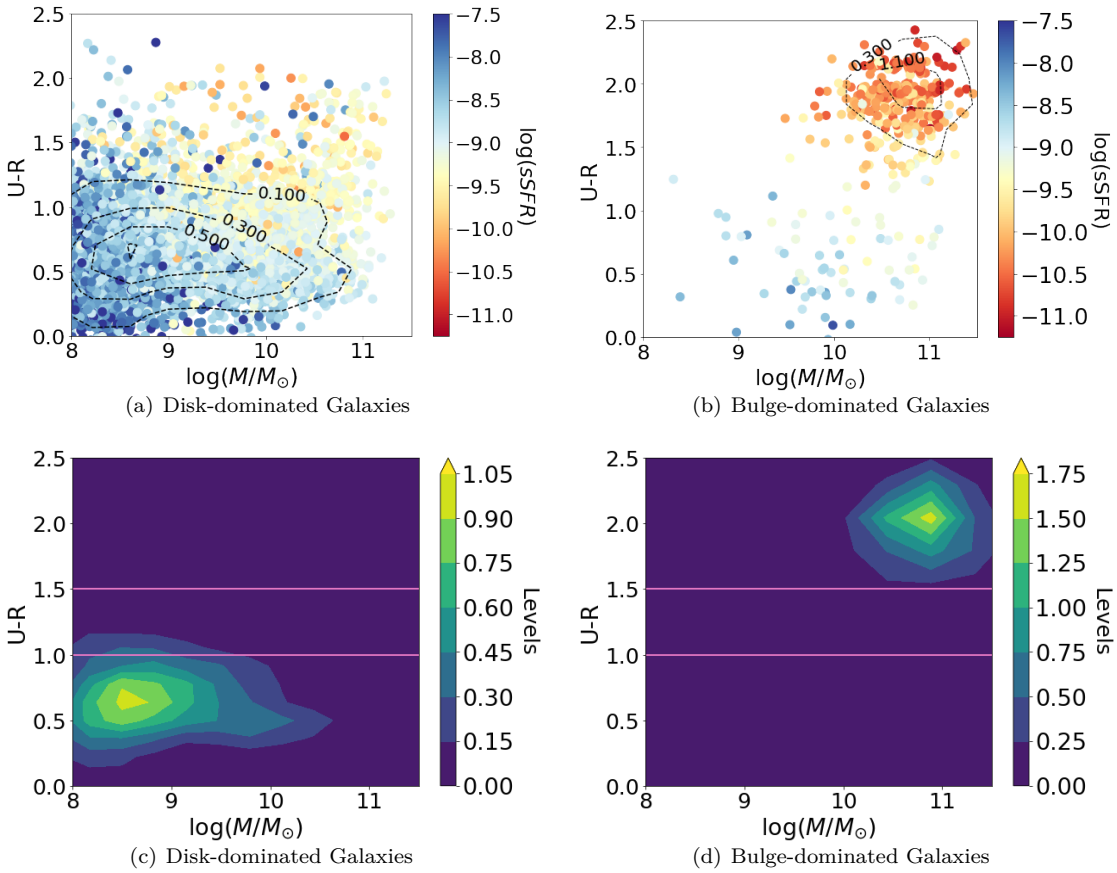
(a) Disk-dominated Galaxies

(b) Bulge-dominated Galaxies

(c) Disk-dominated Galaxies

(d) Bulge-dominated Galaxies

Figure 2.10: Color-mass diagrams for the galaxies in the SDSS test set, separated by morphology. Disk-dominated galaxies (panels (a) and (c)) are mostly blue until they reach high masses (and presumably high halo masses), at which point they evolve to the red. In contrast, bulge-dominated galaxies (panels (b) and (d)) are predominately red, and appear to evolve rapidly from a short-lived population of rare, blue ellipticals that likely formed from major mergers of disky star-forming galaxies. Panels (a) and (b) show individual data points, with color indicating the specific star formation rates (sSFR) for each galaxy in units of yr^{-1}. Contours show the linear density of galaxies in this plot, and the numbers refer to the levels of the contours. Panels (c) and (d) are the same data plotted in terms of galaxy density. The lines mark the position of the green valley.



(a) Disk-dominated Galaxies

(b) Bulge-dominated Galaxies

(c) Disk-dominated Galaxies

(d) Bulge-dominated Galaxies

Figure 2.11: Color-mass diagrams for the galaxies in the CANDELS test set, separated by morphology. Similar to Fig. 2.10, disk-dominated galaxies (panels (a) and (c)) show signs of secular evolution, while bulge-dominated galaxies (panels (b) and (d)) appear to evolve rapidly from a short-lived population of rare, blue ellipticals. Panels (a) and (b) show individual data points, with color indicating the specific star formation rates (sSFR) for each galaxy in units of yr^{-1}. Contours show the linear density of galaxies in this plot and the numbers refer to the levels of the contours. Panels (c) and (d) are the same data plotted in terms of galaxy density. The lines mark the position of the green valley.

Table 2.8: Statistics of the Color-Mass Diagrams

Galaxy Sample		SDSS		CANDELS	
		Number	Fraction	Number	Fraction
Disk-dominated	Blue Cloud	32870	69.16%	10614	87.10 %
	Green Valley	7814	16.44%	1330	10.91%
	Red Sequence	6845	14.40%	242	1.99%
	Total	47529	100%	12186	100%
Bulge-dominated	Blue Cloud	995	12.53%	80	16.19%
	Green Valley	633	7.97%	39	7.89%
	Red Sequence	6313	79.50%	375	75.91%
	Total	7941	100%	494	100%

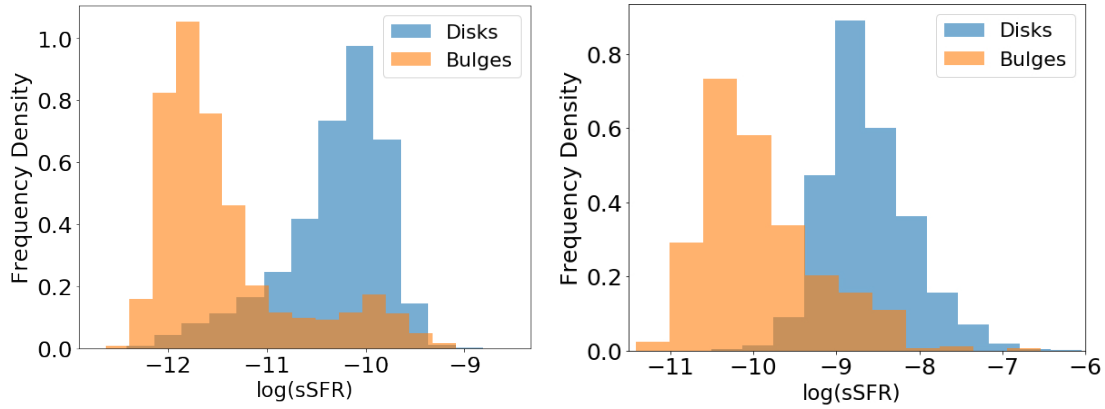
NOTE- The demographics of SDSS and CANDELS galaxies disaggregated by morphology and color. The green valley for both samples is defined in § 2.4.2 and the three zones are shown in Figs. 2.10 and 2.11. We omit galaxies used in training GaMorNet (∼25% of each sample) as well as galaxies lacking estimates for the mass or sSFR (∼ 0.7% for SDSS and ∼ 3.4% for CANDELS).

Figs. 2.10(c) and 2.11(c) show that the disk-dominated galaxies peak in the blue cloud and decline gradually to the red sequence, in a unimodal way. This suggests that the disks undergo a gradual decline in star formation as opposed to being rapidly quenched through the green valley into the red sequence. At high masses, there are relatively more red disk-dominated galaxies, suggesting that high halo masses may play a role in shutting off the gas supply and quenching star formation. These conclusions agree with other studies of star formation in local galaxies (Lopes et al., 2016; Powell et al., 2017; Schawinski et al., 2014; Tojeiro et al., 2013).

Conversely, bulge-dominated galaxies in both samples show a unimodal peak in the red sequence, with very few precursors at green and blue colors. This is consistent with a scenario in which bulge-dominated galaxies form from major mergers of disk-dominated blue galaxies and then are rapidly quenched through the green valley (Schawinski et al., 2014).

The morphology-sorted color-mass diagrams we obtained using GaMorNet classifications largely agree with the previous results of Schawinski et al. (2014) at $z \sim 0$ and Powell et al. (2017) at $z \sim 1$; although, in the latter case, we present an order of magnitude more galaxies. For both samples, the galaxy fractions in the three zones of the color-mass diagram differ at the few percent level with respect to Schawinski et al. (2014) and Powell et al. (2017). It is important to note here that our definition of the green valley is slightly different from that used by Schawinski et al. (2014) due to their use of reddening corrected colors. Besides, Schawinski et al. (2014) and Powell et al. (2017) used visual classification and GALFIT, respectively, compared to our use of GaMorNet. Finally, our sample sizes are much larger: at $z \sim 0$, we have twice as many galaxies as Schawinski et al. (2014), and at $z \sim 1$, we have six times the galaxies analyzed by Powell et al. (2017). Larger

samples are particularly important for bins with low statistics. For example, Powell et al. (2017) identified only 5 bulge-dominated galaxies in the green valley, whereas we find 39, so the statistical uncertainties on that fraction are lower.



(a) $z \sim 0$ SDSS sample

(b) $z \sim 1$ CANDELS sample

Figure 2.12: The normalized distribution of the specific star formation rate (sSFR), separated by morphology, for the SDSS and CANDELS data sets as obtained from the MPA-JHU and 3D-HST catalogs, respectively. 'Frequency density' refers to the number counts normalized to form a probability density.

Figure 2.12 shows the distribution of sSFR separated by morphology. For both samples, the distribution of bulge-dominated galaxies peaks at a lower sSFR, showing the association of disk-dominated galaxies with consistent secular star formation and bulge-dominated galaxies with recent quenching.

2.5 Summary and Discussion

In this article, we introduced GaMorNet, a convolutional neural network (CNN) that can classify galaxies morphologically. We first trained GaMorNet on simulations of galaxies with a bulge and a disk component (§ 2.3.1) to separate galaxies according to their bulge-to-total ratio (L_B/L_T). To make the network better at handling real galaxies, we then transfer learned (§ 2.3.4) on $\sim 25\%$ of both the SDSS $z \sim 0$ and CANDELS $z \sim 1$ samples and thereafter tested the network on the remaining $\sim 75\%$ of both the samples. The net misclassification rate (calculated by weighting the disk- and bulge-dominated accuracies appropriately) achieved for both samples is $\lesssim 5\%$. For the SDSS test set of 82,547 galaxies, we achieved accuracies of 99.7% for disk-dominated galaxies and 94.8% for bulge-dominated galaxies. The corresponding numbers for the CANDELS test set of 21,746 galaxies are 91.8% and 78.6%. We showed in § 2.4.1 that the misclassified CANDELS galaxies are dominated

by galaxies with a half-light radius comparable to the PSF and galaxy images with low S/Ns.

Although it has previously been shown that CNNs can be used to recover single-component Sérsic fits of galaxies and visual morphologies (eg. Huertas-Company et al., 2015; Tuccillo et al., 2018), according to our knowledge, this is the first time it has been demonstrated that CNNs can be used to classify galaxies according to their bulge-to-total ratios.

More importantly, this work demonstrates that GaMorNet can be applied across different data sets to perform morphological classification without the need for a large training set of real galaxies. By using a roughly 25-75 train-test split during transfer learning, we have clearly demonstrated that even when training on 25% of the total sample, GaMorNet can generalize beyond the training data and classify galaxies with high accuracy. This has very important consequences, as the applicability of CNNs to future data-intensive surveys like LSST, WFIRST, and Euclid will depend on their ability to perform without the need for a large training set of real data.

We make the source code of GaMorNet, the trained network models, as well the morphological classifications of all the galaxies in our sample available to the public (Appendix 2.A). Although GaMorNet-S and -C were tuned for g-band and H-band images, respectively, the networks should perform with comparable accuracies in other nearby bands for all SDSS $z \sim 0$ and CANDELS $z \sim 1$ galaxies. We also make available the weights and biases of GaMorNet before transfer learning, i.e., after training with simulations only, so that additional data sets can be used for transfer learning. Our general prescription of training on simulations and then transfer learning should work for morphological classifications of any data set.

In § 2.4.2, we used the morphological classifications obtained using GaMorNet (§ 2.4.1) to study the quenching of star formation using the color-mass diagrams of our samples at $z \sim 0$ and $z \sim 1$.

For both samples, the morphology-separated color-mass diagrams do not show any bimodality. The disk-dominated galaxies peak in the blue cloud and then gradually extend to the red sequence, suggesting that quenching in disks is a secular process. Conversely, bulge-dominated galaxies in both samples peak in the red sequence, with very few precursors in the green valley and blue cloud. This is consistent with a scenario in which bulge-dominated galaxies form from major mergers of disk-dominated blue galaxies and then are rapidly quenched through the green valley.

Our results largely agree with previous similar studies performed at these redshifts. Our sample sizes are twice and six times as large, respectively, as those in the two previous studies done using visual classifications (Schawinski et al., 2014) and using GALFIT (Powell et al., 2017). The reason that we were able to use such large sample sizes is that GaMorNet, once trained, can process large data sets very quickly and easily compared to more traditional methods.

In the future, we aim to use GaMorNet to study the correlation of AGN with host galaxy morphology. We also plan to take GaMorNet beyond bulge/disk classification and use it to derive different properties of AGN host galaxies.

Chapter Acknowledgements

We would like to thank the anonymous referee for a thorough review of the manuscript and suggesting changes that greatly improved the quality and clarity of our manuscript.

This work used data from SDSS. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

This work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

This work is based on observations taken by the 3D-HST Treasury Program with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

This material is based upon work supported by the National Science Foundation under grant No. 1715512

C.M.U would like to acknowledge support from National Aeronautics and Space Administration via ADAP Grant 80NSSC18K0418.

Appendix

2.A Public Release of Code, Models, and Galaxy Morphological Classifications

Here, we provide an outline of all the material that we make public as a part of this work. An up-to-date record of this public data release will also be maintained at http://gamornet.ghosharitra.com and http://www.astro.yale.edu/aghosh/gamornet.html in case any of the URLs below stop working over time.

2.A.1 GaMorNet Source Code

GaMorNet was implemented using TFLearn (http://tflearn.org), which is a high-level Application Program Interface for TensorFlow (https://tensorflow.org), an open source library widely used for large-scale machine learning applications.

The source code of GaMorNet is maintained as a GitHub Repository and is available at https://github.com/aritra ghsh09/GaMorNet. Instructions for installing TFLearn and using GaMorNet are available in the above GitHub repository. An implementation of GaMorNet in Keras (https://keras.io/) is also available at the above repository.

2.A.2 GaMorNet Trained Models

Trained Models for both GaMorNet-S and -C are being made available as a part of this data release.

For more details about the various stages of training, refer to § 2.3.3 & 2.3.4. All of the models below are being made available via Yale Astronomy's Public FTP service `ftp://ftp.astro.yale.edu/pub/aghosh/gamornet/trained_models`.

You can copy and paste the above link into a browser window to download the files, or you can also issue the following commands from a terminal to login to the ftp server

```
ftp ftp.astro.yale.edu
```

Use the username 'anonymous' and keep the password field blank. After logging-in, do the following:

```
cd pub/aghosh/gamornet/<appropriate_subdirectory>
get <file_name>
quit
```

To list the files at your current location, you can use the 'ls' command.

The various subdirectories are named as follows in the list below:

1. GaMorNet-S model trained only on simulations → `/trained_models/SDSS/sim_rained/`
2. GaMorNet-S model trained on simulations and transfer learned on real data → `/trained_models/SDSS/tl/`
3. GaMorNet-C model trained only on simulations → `/trained\_models/CANDELS/sim_trained/`
4. GaMorNet-C model trained on simulations and transfer learned on real data → `/trained_models/CANDELS/tl/`

Models 2 and 4 can be applied directly to SDSS g-band data at $z \sim 0$ and CANDELS H-band data at $z \sim 1$ (or data in other nearby bands), respectively, without any further training. However, if you plan to apply GaMorNet to data that is different from the above mentioned data sets, we recommend using any of the models above and then transfer learning on your new data. The exact nature of the data will decide which of the models above is the best starting point for the transfer learning process.

For more information on how to load these models in TFLearn and use them, refer to the documentation of the GaMorNet GitHub repository mentioned in Sec. 2.A.1.

2.A.3 Tables with predicted probabilities and classifications

The predicted probabilities (of being disk-dominated, bulge-dominated, or indeterminate) and the final classifications for all the galaxies in our SDSS and CANDELS test sets, as determined by GaMorNet-S and -C, are made available below as .txt files. These tables are the full versions of Tables 2.4 & 2.6.

Both the tables are being made available via Yale Astronomy's Public FTP service `ftp://ftp.astro.yale.edu/pub/aghosh/gamornet/pred_tables`. Instructions for accessing the service from the command line can be found in § 2.A.2. The two files are located according to the list below:

- Full version of Table 2.4 corresponding to the SDSS data set → `/pred_tables/pred_table_sdss.txt`
- Full version of Table 2.6 corresponding to the CANDELS data set → `/pred_tables/pred_table_candels.txt`

2.A.4 GalaxySim Source Code

The code that was used to simulate the galaxies described in § .2.3.1 is available as a GitHub repository at https://git hub.com/aritraghsh09/GalaxySim

This code makes use of GALFIT (Peng et al., 2002) to simulate idealized double component galaxies. Since the simulations of galaxy surface brightness profiles are independent of each other, the code could be trivially parallelized. Instructions for using GalaxySim are available in the above GitHub repository.

Chapter 3

GaMPEN: A Machine Learning Framework for Estimating Bayesian Posteriors of Galaxy Morphological Parameters

Originally published by the American Astronomical Society in *The Astrophysical Journal*, Volume 935, Issue 2, pp. 138, DOI:10.3847/1538-4357/ac7f9e

Aritra Ghosh, C. Megan Urry, Amrit Rau, Laurence Perreault-Levasseur, Miles Cranmer, Kevin Schawinski, Dominic Stark, Chuan Tian, Ryan Ofman, Tonima Tasnim Ananna, Connor Auge, Nico Cappelluti, David B. Sanders, and Ezequiel Treister

We introduce a novel machine learning framework for estimating the Bayesian posteriors of morphological parameters for arbitrarily large numbers of galaxies. The Galaxy Morphology Posterior Estimation Network (GaMPEN) estimates values and uncertainties for a galaxy's bulge-to-total light ratio (L_B/L_T), effective radius (R_e), and flux (F). To estimate posteriors, GaMPEN uses the Monte Carlo Dropout technique and incorporates the full covariance matrix between the output parameters in its loss function. GaMPEN also uses a Spatial Transformer Network (STN) to automatically crop input galaxy frames to an optimal size before determining their morphology. This will allow it to be applied to new data without prior knowledge of galaxy size. Training and testing GaMPEN on galaxies simulated to match $z < 0.25$ galaxies in Hyper Suprime-Cam Wide g-band images, we demonstrate that GaMPEN achieves typical errors of 0.1 in L_B/L_T, 0.17 arcsec ($\sim 7\%$) in R_e, and 6.3×10^4 nJy ($\sim 1\%$) in F. GaMPEN's predicted uncertainties are well-calibrated and accurate ($< 5\%$ deviation) – for regions of the parameter space with high residuals, GaMPEN correctly predicts correspondingly large uncertainties. We also demonstrate that we can apply categorical labels (i.e., classifications such as "highly bulge-dominated") to predictions in regions with high residuals and verify that those labels are $\gtrsim 97\%$ accurate. To the best of our knowledge, GaMPEN is the first machine learning framework for determining joint posterior distributions of multiple morphological

parameters and is also the first application of an STN to optical imaging in astronomy.

3.1 Introduction

For almost a century, starting with Hubble in 1926, astronomers have linked the morphology of galaxies to the physics of galaxy formation and evolution. Morphology has been shown to be related to many fundamental properties of the galaxy and its environment, including galaxy mass, star formation rate, stellar kinematics, merger history, cosmic environment, the influence of supermassive black holes (e.g., Powell et al., 2017; Pozzetti et al., 2010; Wuyts et al., 2011; **?**; **?**; **?**; **?**; **?**; **?**). Studying the morphology of large samples of galaxies at different redshifts is crucial in order to understand the physics of galaxy formation and evolution.

Over the last decade, machine learning (ML) has been increasingly employed by astronomers for a wide variety of tasks – from identifying exoplanets to studying black holes (e.g., Hoyle, 2016; Kim & Brunner, 2017; **?**; **?**; **?**). Especially, Convolutional Neural Networks (CNNs)[1] have revolutionized the field of image processing and have become increasingly popular for determining galaxy morphology (e.g., **????????**). Previously, we have developed a publicly available CNN, called GaMorNet (**?**), that classifies galaxies morphologically with minimal real training data, and has been demonstrated to achieve accuracy $\gtrsim 95\%$ across multiple datasets.

This use of CNNs has been driven by the fact that traditional methods of classifying morphologies—visual classification and template fitting to the surface brightness profile of a galaxy—are not scalable to the data volume expected from future surveys such as The Vera Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezic et al., 2019), the Nancy Grace Roman Space Telescope (NGRST; Spergel et al., 2013), and Euclid (Racca et al., 2016). The quality of fits obtained using template fitting depends significantly on the initial input parameters, and when dealing with millions of galaxies, such hand-refinement of input parameters is an intractable task. Although large citizen science projects like Galaxy Zoo (Lintott et al., 2008b) have been successful in processing many surveys in the past, even these will fail to keep up with the upcoming data glut. Moreover, reliable visual classifications require a decent signal-to-noise ratio, take time to set up and execute, and require an extremely careful de-biasing of the vote shares obtained (e.g., Lintott et al., 2008b; Simmons et al., 2017).

From early attempts at using a CNN to classify galaxies morphologically (e.g., **?**) to the largest

1. CNNs are a specific form of machine learning algorithm that specializes in processing data with a grid-like topology, such as an image. See §3.3 for more details.

CNN produced morphology catalogs currently available (**??**), most CNNs have provided broad, qualitative classifications, rather than numerical estimates of morphological parameters. Such studies typically entail classifying galaxies based on their morphological properties (e.g., based on whether the galaxy has a disk or a bulge or a bar, etc.) as opposed to predicting values of relevant morphological parameters that help characterize the galaxy (such as bulge-to-total light ratio, radius, etc.). By contrast, **?** used a CNN to estimate the parameters of a single-component Sérsic fit, though without uncertainties. Meanwhile, the computation of full Bayesian posteriors for different morphological parameters is crucial for drawing scientific inferences that account for uncertainty and thus are indispensable in the derivation of robust scaling relations (e.g., **??**) or tests of theoretical models using morphology (e.g., **?**). Thus, producing posterior estimates will significantly increase the scientific potential of morphological catalogs produced using CNNs.

In this work, we introduce GaMPEN (the Galaxy Morphology Posterior Estimation Network), a novel machine learning framework that estimates the Bayesian posteriors for three morphological parameters: the bulge-to-total light ratio (L_B/L_T), the effective radius (R_e), and the total flux (F). GaMPEN uses a CNN module to estimate the joint posterior probability distributions of these parameters. This is done by using the negative log-likelihood of the output parameters as the loss function combined with the Monte Carlo Dropout technique (Gal & Ghahramani, 2016). We also used the full covariance matrix in the loss function, using a series of algebraic manipulations (see §3.4). The full covariance matrix accounts for dependencies among different output parameters and ensures that the posterior distributions for all three output variables are well calibrated.

Although the use of CNNs in the recent past has allowed astronomers to process large data volumes quickly, some challenges related to data pre-processing have remained. One of these challenges has to do with making cutouts of proper sizes. Most trained CNNs require input images of a fixed size—thus, most previous work (e.g., **??**) has resorted to selecting a large cutout size for which "most galaxies" would remain in the frame. However, this means that for many galaxies in the dataset, especially smaller ones, typical cutouts contain other galaxies in the frame, often leading to less accurate results. This problem is aggravated when designing a CNN applicable over an extensive range in redshift, which corresponds to a large range of galaxy sizes. Lastly, most previous work has used computations of R_e from previous catalogs to estimate the correct cutout size to choose. This is, of course, not possible when one is trying to use a CNN on a new, unlabeled dataset.

To address these challenges, GaMPEN automatically crops the input image frames using a Spatial Transformer Network (STN) module upstream of the CNN. STNs are self-consistent modules that can be used for the spatial manipulation of data within machine learning frameworks. In GaMPEN,

based on the input image, the STN predicts the parameters of an affine transformation which is then applied to the input image. The transformed image is then passed onto the downstream CNN. The inclusion of the STN in the framework greatly reduces the amount of time spent on data pre-processing as it trains simultaneously with the downstream CNN without additional supervision. We later show in §3.3.1 how the STN automatically learns to make appropriate affine transformations (such as cropping) on the input data, which are helpful in the downstream task of morphological parameter estimation.

To the best of our knowledge, GaMPEN is the first ML framework to apply an STN to optical imaging data and is the first to estimate full Bayesian posteriors for galaxy morphological parameters. In order to have a robust understanding of the performance, bias, and limitations of GaMPEN, we train and test GaMPEN on simulations of galaxy images—the only situation where we have access to the "ground truth" morphological parameters of the galaxies. We match our simulations to the observations of the Hyper Suprime-Cam (HSC) Wide survey (Aihara et al., 2018a), as this is an obvious application (to be described in a forthcoming paper). We use real HSC Wide images, with their multiples galaxies, to validate the STN performance.

In §3.2, we describe the simulated data used to train and test GaMPEN. We describe the structure and code of GaMPEN in §3.3 and outline the entire mechanism behind the prediction of posteriors in §3.4. In §3.5 we describe GaMPEN's training procedure. We present our results in §3.6, and summarize our findings along with future applications of GaMPEN in §**??**. GaMPEN's data-access policies are described in Appendix 3.A.

3.2 Simulated Galaxies

We train and test GaMPEN using mock galaxy image cutouts simulated to match g-band data from the Hyper Suprime-Cam (HSC) Subaru Strategic Program wide-field optical survey (Aihara et al., 2018a). The Subaru Strategic Program, ongoing since 2014, uses the HSC prime-focus camera, which provides extremely high sensitivity and resolving power due to the large 8.2 meter mirror of the Subaru Telescope. Its g-band seeing, with median FWHM of $0.85''$, is a large improvement over the Sloan Digital Sky Survey (SDSS; York et al., 2000), which has a median g-band seeing of $1.4''$.

To generate mock images, we used GalSim (**?**), the modular galaxy image simulation toolkit. GalSim has been extensively tested and shown to yield very accurate rendered images of galaxies. We simulated 150,000 galaxies in total, with a mixture of both single and double components, in order to have a diverse training sample. To be exact, 75% of the simulated galaxies consisted of

both bulge and disk components, while the remaining 25% had either a single disk or a bulge.

For both the bulge and disk components, we used the Sérsic profile, the surface brightness of which is given by

$$\Sigma(R) = \Sigma_e \exp\left[-\kappa\left(\left(\frac{R}{R_e}\right)^{1/n} - 1\right)\right], \tag{3.1}$$

where Σ_e is the pixel surface brightness at the effective radius R_e, n is the Seŕsic index, which controls the concentration of the light profile, and κ is a parameter coupled to n that ensures that half of the total flux is enclosed within R_e. The standard formula for an exponential disk corresponds to $n = 1$, and a de Vaucouleurs profiles is $n = 4$.

Table 3.1: Parameter Ranges of Simulated Galaxies

Component Name	Sérsic Index	Half-Light Radius (arcsec)	Flux (nJy)	Axis Ratio	Position Angle (degrees)
Single Component Galaxies					
	0.8 - 1.2 or 3.5 - 5.0[a]	0.1 - 5.0	10^3 - 5×10^6	0.25 - 1.0	-90.0 - 90.0
Double Component Galaxies					
Disk	0.8 - 1.2	0.1 - 5.0	0.0 - 1.0[b]	0.25 - 1.0	-90.0 - 90.0
Bulge	3.5 - 5.0	0.1 - 3.0	1.0 - Disk [b]	0.25 - 1.0	Disk $\pm [0, 15]$[c]

[a] The single component galaxies are equally divided between galaxies with a Sérsic index between 0.8 - 1.2 and galaxies with a Sérsic index between 3.5 - 5.0.
[b] Fractional fluxes are noted here. The bulge flux fraction is chosen such that for each simulated galaxy it is added with the disk flux fraction to give 1.0. The total flux of the galaxies is varied between 10^3 and 5×10^6 nJy.
[c] The bulge position angle differs from the disk position angle by a randomly chosen value between -15 and $+15$ degrees.
`NOTE-` The above table shows the ranges of the various Sérsic profile parameters used to simulate training and testing data. 75% of the simulated galaxies have both disk and bulge components, and the remainder have either a disk or a bulge component. The distributions of all the simulation parameters are uniform except for the position angle and flux of the double-component galaxies. For more details about these choices, please refer to § 3.2.

The parameters required to generate the Sérsic profiles were drawn from uniform distributions, over ranges given in Table 3.1. For the disk and bulge components, we let the Sérsic index vary between $0.8 - 1.2$ and $3.5 - 5.0$, respectively. We chose to have varying Sérsic indices as opposed to fixed values for each component in order to have a training set with diverse light profiles. The parameter ranges for fluxes and half-light radii are quite expansive, such that the simulations represent most local galaxies (Binney & Merrifield, 1998) at $z \leq 0.25$ (i.e., the simulation parameters are chosen to match HSC $z < 0.25$ galaxies).

Specifically, single-component galaxies were assigned a half-light radius between 0.25 kpc and

11.5 kpc. In the double-component galaxies, the disk half-light radius was varied across the same range, and the bulge half-light radius was varied between 0.25 kpc and 7.0 kpc. To obtain the corresponding angular sizes for simulation, we placed the sample at $z = 0.125$ using the Planck18 cosmology ($H_0 = 67.7$ km/s/Mpc, Aghanim et al., 2018) and the appropriate pixel scale.

For the single-component galaxies, the total flux was varied between 10^3 and 5×10^6 nJy ($m_{\rm AB} \sim 14 - 23$). For the double component galaxies, we first draw L_B/L_T from a uniform distribution between 0 and 1. Thereafter, the total flux of the galaxy is chosen from a uniform distribution with a range of $10^3 - 5 \times 10^6$ nJy. To assign fluxes to the bulge and disk components, we multiply L_B/L_T and $(1 - L_B/L_T)$ respectively by the total flux. Not following this procedure and drawing the bulge and disk fluxes independently causes most galaxies in the training set to have a very high or a very low L_B/L_T, which is not the case for most galaxies, and in any case we already have single-component galaxies in our sample. For the double component sample, we wanted to have a sufficient number of galaxies with intermediate values of L_B/L_T.

What matters in training a CNN is not matching the observed distributions of the simulation parameters; rather, it is spanning the full range of those parameters. Having too many of any one type—even if that is the reality in real data—can result in lower accuracy for minority populations (e.g., **?**). By not weighting the simulated galaxy sample in any specific regions of the parameter space, we are able to optimize GaMPEN for the full range of galaxy morophologies.

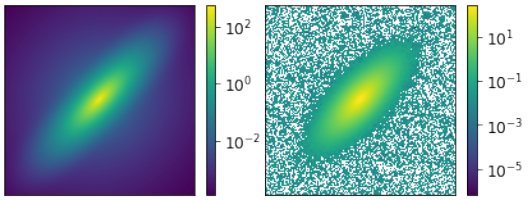


Figure 3.1: Two stages of simulating an HSC galaxy. (*Left*): A randomly chosen two-dimensional light profile generated by GalSim. (*Right*): The same image after PSF convolution and noise addition. The white pixels represent (small) negative values that arise from the process of noise addition.

To make the two-dimensional light profiles generated by GalSim realistic, we convolved these with a representative point-spread function (PSF), and added appropriate noise. Figure 3.1 shows a randomly chosen simulated light profile, as well as the corresponding image cutout generated after PSF convolution and noise addition.

To curate a collection of representative PSFs, we first selected 100 galaxies at random from the HSC PDR2 Wide field (Aihara et al., 2021) with $z \leq 0.25$ and $m_g \leq 23$, and that did not have any quality flags set to True (the quality flags check for cosmic ray hits, interpolated pixels etc.).

We then used the HSC PSF Picker Tool[2] to obtain the PSF at the location of these 100 galaxies. Each simulated light profile was convolved with a randomly chosen PSF out of these 100. To make sure that the PSFs are representative (i.e., do not contain any outliers), we ran a test where we convolved each one with a simulated galaxy light profile, before adding noise. We then inspected all possible difference images for each convolved galaxy, to make sure the average pixel value of the difference image was always at least three orders of magnitude lower than the average pixel value of the convolved galaxy image.

To generate representative noise, we used one-thousand 2×2 arcsec "sky objects" from the HSC PDR2 Wide field. Sky-objects are empty regions identified by the HSC pipeline that are outside object footprints and are recommended for being used in blank-sky measurements. We visually verified that our sky objects did not contain any sources. We then read in the pixel values of these sky objects to generate a large sample of noise pixels. We randomly sampled this collection of noise pixels to make two-dimensional arrays of the same size as that of the simulated images and then added them to the images.

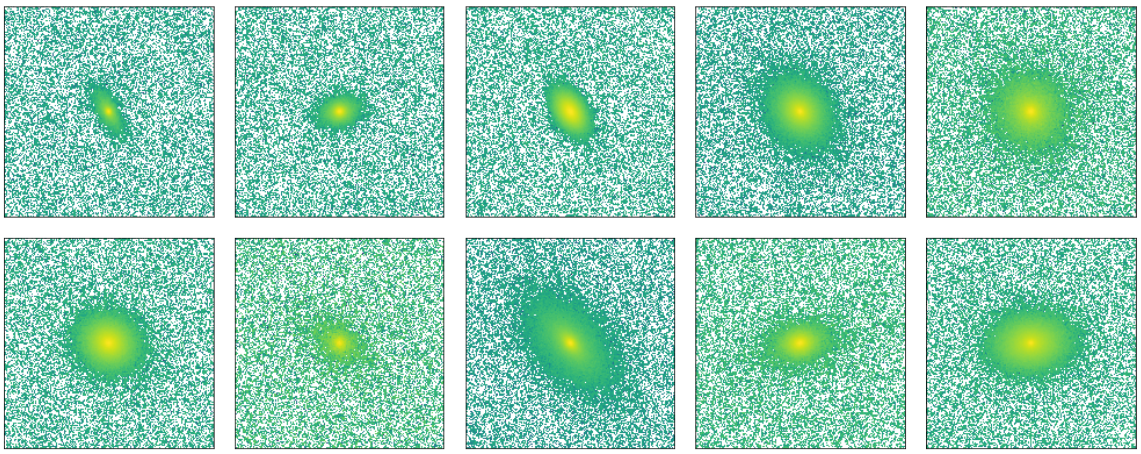


Figure 3.2: Ten randomly selected galaxies from our simulated dataset. The simulation parameters are chosen such that the simulated galaxies represent a diverse range of light profiles and include most bright, local galaxies at $z \lesssim 0.25$.

All the simulated galaxy cutouts were chosen to have a size of 239×239 pixels, which translates to roughly 40×40 arcsecs given HSC's pixel scale of 0.168 arcsecs/pixel. Ten randomly chosen simulated galaxies from our dataset are shown in Figure 3.2.

2. https://hsc-release.mtk.nao.ac.jp/psf/pdr3/

3.3 GaMPEN Architecture

Artificial neural networks, consisting of many connected individual units called artificial neurons, have been studied for more than five decades. These artificial neurons are usually arranged in multiple layers and such networks typically have (a) an input layer to feed data into the network; (b) an output layer that contains the result of propagating the data through the network. In between, there are additional hidden layer(s). Each neuron is characterized by a weight vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ and a bias b. The output of a single neuron in the network is given by

$$y = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b), \tag{3.2}$$

where σ is the chosen activation function of the neuron and $\boldsymbol{x}$ is the vector of inputs to the neuron. The process of training an artificial neural network involves finding the optimum set of weights and biases of all neurons such that, for a given set of inputs, the output of the network resembles the desired outputs as closely as possible. The optimization is usually performed by minimizing a loss function using stochastic gradient descent.

The backbone of GaMPEN is a Convolutional Neural Network (Fukushima, 1980; LeCun et al., 1998). Without convolutional layers, neural networks learn global patterns, whereas CNNs learn to identify thousands of local patterns in their input images that are translation invariant. Additionally, CNNs learn the spatial hierarchies of these patterns, allowing them to process increasingly complex and abstract visual concepts. These two key features have allowed deep CNNs to revolutionize the field of image processing in the last decade (Lecun et al., 2015b; Schmidhuber, 2015b).

The architecture of GaMPEN is shown in Figure 3.3. It consists of a Spatial Transformer Network module followed by a downstream CNN module, described in § 3.3.1 and 3.3.2, respectively. The design of GaMPEN is based on our previously successful classification CNN, GaMorNet (**?**), as well as different variants of the Visual Geometry Group (VGG) networks (Simonyan & Zisserman, 2014), which are highly effective at large-scale visual recognition. We tried different architectures of these "base" models by varying the depth of the entire network and the sizes of the various layers. To quickly and systematically search this model-design space, we use ModulosAI's[3] AutoML platform, which uses a Bayesian optimization strategy. The said strategy involves using the current model's performance to determine which variant to try next. When choosing new configurations, the optimizer balances the exploitation of well-performing search spaces and the exploration of unknown

3. https://www.modulos.ai

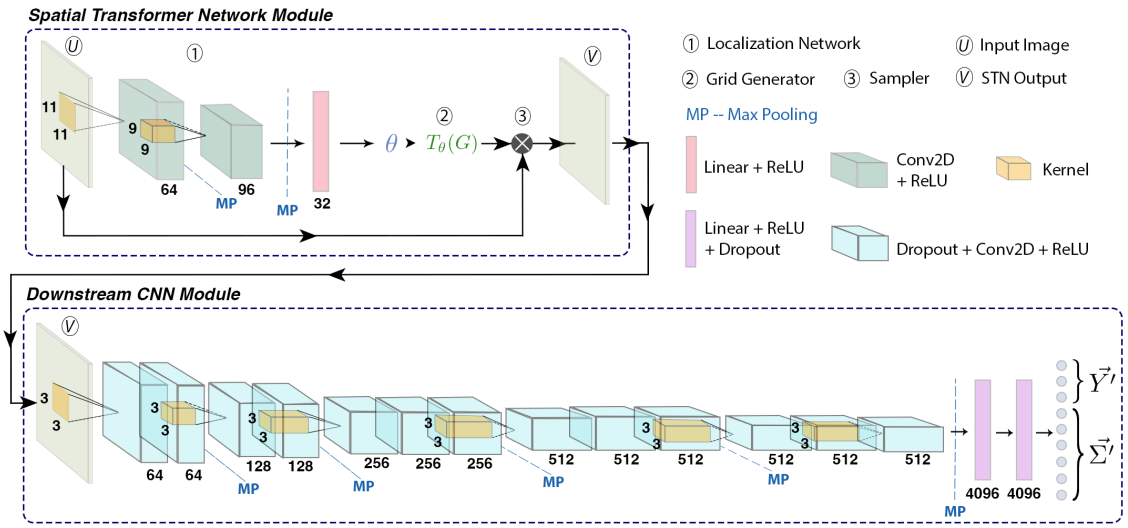


Figure 3.3: A schematic diagram of the Galaxy Morphology Posterior Estimation Network. GaMPEN's architecture consists of a downstream CNN module preceded by an upstream STN module. The CNN module empowers GaMPEN to estimate posterior distributions of galaxy morphology parameters. The upstream STN module trains without any extra supervision and learns to apply appropriate cropping transformations to the input image before passing it on to the CNN (for more details about these modules, see §§ 3.3.1, 3.3.2). The numbers below each layer refer to the number of filters/neurons in each layer. The yellow boxes inside the convolutional layers show the kernel and the number beside it refers to the corresponding kernel size. Only one kernel is shown per set of convolutional layers; all other layers in the set have kernels of the same size. Conv2D and ReLU refer to Convolutional Layers and Rectified Linear Units, respectively (described in §.3.3.2).

regions.

To implement GaMPEN, we use PyTorch, which is an open-source machine learning framework, written in Python.

3.3.1 The Spatial Transformer Network Module

Spatial Transformer Networks (STNs) were introduced by Jaderberg et al. (2015) as a learnable module that can be inserted into CNNs and explicitly allows for the spatial manipulation of data within the CNN. In the astronomical context, STNs have only been used by Wu et al. (2019) previously in morphological analysis of radio data.

In GaMPEN, the STN is upstream of the CNN, where it applies a two-dimensional affine transformation to the input image, and the transformed image is then passed to the CNN. Each input image is transformed differently by the STN, which learns the appropriate cropping during the training of the downstream CNN without additional supervision. As shown in the upper part of Figure 3.3, the STN consists of (1) a localization network, (2) a parameterized grid generator, and (3) a

sampler, as described below.

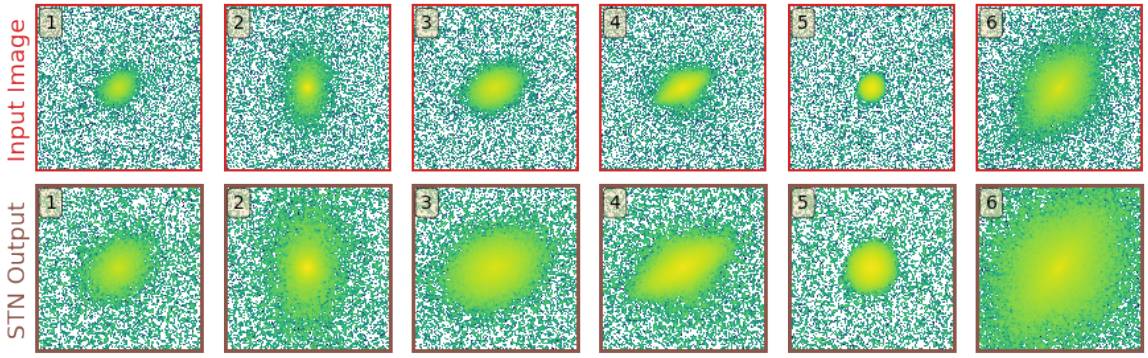


Figure 3.4: Examples of the transformation applied by the STN to six randomly selected input galaxy images. The top row shows the input galaxy images, and the bottom row shows the corresponding output from the STN. The numbers in the top-left yellow boxes help correspond the output images to the input images. As can be seen, the STN learns to apply an optimal amount of cropping for each input galaxy.

The localization network takes the input image, U ($U \in \mathbb{R}^{H \times W \times C}$, with height H, width W, and C channels), and outputs $\boldsymbol{\theta}$, the six-parameter matrix of the affine transformation, $\mathcal{T}_\theta$, to be applied to the input image. The localization network in the STN is a CNN with two convolutional layers followed by two fully connected layers at the end.

To perform the transformation, $\mathcal{T}_\theta$, the values of the output pixels are computed by applying a sampling kernel on the input image. As the first step in this process, the parameterized grid generator is used to generate a grid, G, of target coordinates, $G_i = (x_i^t, y_i^t)$, forming the output of the STN. For our case, $\mathcal{T}_\theta$ is a 2D affine transformation A_θ, and the pointwise transformation is given by

$$\begin{aligned} \begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta\left(G_i\right) = \mathrm{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \\ = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}, \end{aligned} \tag{3.3}$$

where (x_i^s, y_i^s) are the source coordinates in the input image that define the sample points (Jaderberg et al., 2015). The transformation shown in Equation 3.3 allows for cropping, translation, rotation, and skewing to be applied to the input image. However, the simulated galaxy images in our dataset

are already centered, and our primary aim of using the STN is to achieve optimal cropping; thus, we constrain the type of affine transformations allowed by modifying A_θ such that

$$\mathrm{A}_\theta = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix}. \tag{3.4}$$

The localization network predicts the optimal value of s for each input image. As can be seen from Eq. 3.4, $s = 1$ results in an identity transformation (i.e., the image output by the STN and the input image are the same). For values of $s < 1$, lower fractions of the input image are retained in the output image. For example, when $s = 0.7$, 70% of each side (length/width) of the input image is retained in the output image.

Note that although GaMPEN's STN does not perform rotations, we are able to induce rotational invariance using our training procedure. Since our simulated training set is very large, it happens to be that there are many galaxies with different position angles, but similar (other) structural parameters.

In the final step, the sampler takes the set of sampling points $\mathcal{T}_\theta(G)$ along with the input image, U, to produce the output image, V. Each (x_i^s, y_i^s) coordinate in $\mathcal{T}_\theta(G)$ defines the spatial location in the input where a bilinear sampling kernel is applied to get the value at a particular pixel in the output image. This can be written as

$$\begin{aligned} V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max\left(0, 1 - |x_i^s - m|\right) \times \\ \max\left(0, 1 - |y_i^s - n|\right), \end{aligned} \tag{3.5}$$

where U_{nm}^c is the value at location (n,m) in channel c of the input, and V_i^c is the output value for pixel i at location (x_i^t, y_i^t) in channel c. To allow the backpropagation of the loss through this sampling mechanism, we can define the gradients with respect to U and G. This allows loss gradients to flow back to the sampling grid coordinates and therefore back to the transformation parameter s and the localization network.

Placing the STN upstream in the GaMPEN framework allows the network to learn how to actively transform the input image to minimize the overall loss function during training. Because the transformation we use is differentiable with respect to the parameters, gradients can be backpropagated through the sampling points $\mathcal{T}_\theta(G)$ to the localization network output $\boldsymbol{\theta}$. This crucial property allows the STN to be trained using standard backpropagation along with the downstream

CNN, without any additional supervision.

Figure 3.4 shows examples of the transformations applied by the STN of a trained GaMPEN framework to simulated HSC data. As can be seen, the STN learns to apply an optimal amount of cropping for each input galaxy.

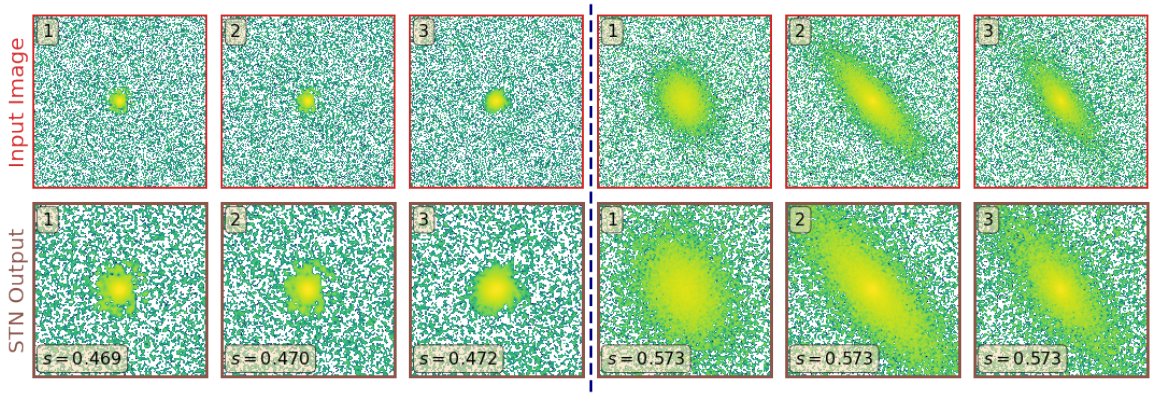


Figure 3.5: (*Left*): Galaxies in the testing dataset with the lowest values of s (i.e., the most aggressive crops) (*Right*): Galaxies in the testing dataset with the highest values of s (i.e., the least aggressive crops). As can be seen, the STN correctly learns to apply the most aggressive crops to small galaxies; and the least aggressive crops to large galaxies.

To further validate the performance of our STN, we process all images in our testing dataset through the STN module of a trained GaMPEN framework. After that, we sort all the processed images using the value of the parameter s (from Eq. 3.4) predicted by the localization network. Higher values of s denote that a more significant fraction of the input image was retained in the output image produced by the STN (i.e., minimal cropping). In Figure 3.5, we show the images from our testing dataset with the highest and lowest values of s. Figure 3.5 demonstrates that the STN correctly learns to apply the most aggressive crops to smallest galaxies in our dataset, and the least aggressive crops to the largest galaxies.

Lastly, in order to demonstrate the purpose of including an STN in GaMPEN, we show its performance on real HSC galaxies. We apply the STN module of a trained GaMPEN framework to three randomly chosen g-band galaxies in the HSC-Wide survey with $z \leq 0.25$ and $m_g \leq 23$. Each input image is 40×40 arcsec. (Note that, for this demonstration, we did not retrain GaMPEN on real galaxies in any way.) The results are shown in Figure 3.6. The STN learns to systematically crop out secondary galaxies in the cutouts and focus on the galaxy of interest at the center of the cutout. At the same time, the STN also correctly applies minimal cropping to the largest galaxies, making sure the entirety of these galaxies remains in the frame.

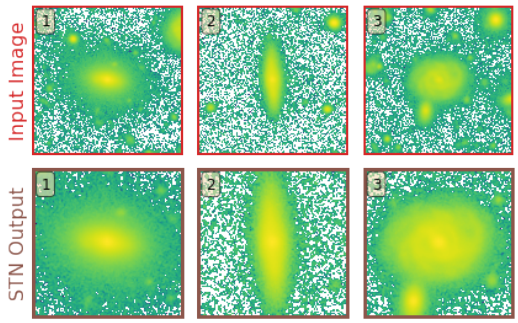


Figure 3.6: Examples of the transformation applied by a trained STN to real HSC-Wide g-band galaxies. The STN helps the downstream CNN to focus on the galaxy of interest at the center of the cutout by cropping out most secondary galaxies present in the input frame.

3.3.2 The Convolutional Neural Network Module

The input image, once transformed by the STN, is passed to the downstream CNN module, as depicted in Figure 3.3. This downstream module predicts the posterior distribution of the bulge-to-total light ratio, effective radius, and total flux for each input galaxy.

The architecture of this downstream CNN is based on the design of VGG-16 (Simonyan & Zisserman, 2014), a CNN that performed well in the 2014 ImageNet Large Scale Visual Recognition Challenge, wherein different teams compete to classify about 14 million hand-annotated images. The primary feature of the VGG class of networks is that they use tiny convolutional filters combined with significantly deep networks, which have been shown to be highly successful in computer vision. Broadly speaking, GaMPEN's downstream CNN consists of thirteen convolutional layers, followed by three fully connected layers. The convolutional layers are arranged in five blocks, and in between each block is a max-pooling layer. Note that one of the primary differences between our network and VGG-16 is that all the convolutional layers in GaMPEN are preceded by a dropout layer in order to facilitate the prediction of epistemic uncertainties, as described further in §3.4.1.

The convolutional layers (Conv2D in Fig. 3.3) work in unison to identify hierarchies of translational invariant spatial patterns in the images. Each convolutional layer does this by using a collection of 3×3 pixel windows (called "filters"), wherein each filter is a specific pattern that the CNN is looking for in the image. These windows slide around the input to generate a "response-map" or "feature-map", which quantifies the presence of the filter's pattern at different locations of the input. Each convolutional layer is preceded by a dropout layer, which is one of the most effective and commonly used regularization techniques that prevent over-fitting by randomly "dropping" (i.e., setting to zero) several output features of the layer during training. The "dropout rate" defines the fraction of features that are zeroed out. For GaMPEN, our choice of the dropout rate is guided by

calibration of the predicted uncertainties and is described in §3.5.

The goal of the max-pooling layers (MP in Fig. 3.3) is to aggressively down-sample the outputs of the convolutional layer that they follow. Simply speaking, max-pooling is dividing the output of the convolutional layer into a collection of windows and then using the maximum value in each window as the output. Max-pooling can be thought of as a technique for detecting a given feature in a broad region of the image and then throwing away the exact positional information. The intuition is that once a feature has been found, its exact location is not as crucial as its rough location relative to other features. An additional advantage is that by aggressively down-sampling, max-pooling forces successive convolutional layers to look at increasingly large windows as a fraction of the input to the layer. This helps to induce spatial-filter hierarchies.

Throughout the network, we use the rectified linear unit (ReLU in Fig. 3.3) as the activation function, except for the output layer, which is linear. The output of a ReLU unit with input $\mathbf{x}$, weight $\mathbf{w}$, and bias b is given by $\max(0, \boldsymbol{w} \cdot \boldsymbol{x} + b)$. The application of the ReLU activation function makes the network non-linear.

At the end of the network are three fully connected layers. They use the output of the convolutional layers, denoting the presence of various features in the image, to predict the correct output variables given an image. The output layer predicts nine parameters. Three of these construct the vector of means of the output variables ($\hat{\boldsymbol{\mu}}$), and the remaining six are used to construct the covariance matrix $\hat{\boldsymbol{\Sigma}}$. In § 3.4, we describe more about how these two variables are used to generate the predicted distributions.

Table 3.3 in the Appendix gives extended descriptions of each GaMPEN layer. For more technical details about the various layers and functions described there, we refer the reader to Chollet (2021); Goodfellow et al. (2016); Nielsen (2015).

3.4 Prediction of Posteriors

Traditional CNNs consist of neurons with fixed, deterministic values of weights and biases, resulting in deterministic outputs. However, if the weights in such a network are probability distributions, then the calculation can be defined within a Bayesian framework (Denker & Lecun, 1991). Such CNNs can then be used to capture the posterior probabilities of the outputs, resulting in well-defined estimates of uncertainties. The key distinguishing property of the Bayesian approach is marginalization over multiple networks rather than a single optimization run.

Two primary sources of error contribute to the uncertainties in the parameters predicted by

GaMPEN. The first arises from errors inherent to the input imaging data (e.g., noise and PSF blurring), and this is commonly referred to as aleatoric uncertainty. The second error comes from the limitations of the model being used for prediction (e.g., the number of free parameters in GaMPEN, the amount of training data, etc.); this is referred to as epistemic uncertainty. It is important to note that while epistemic uncertainties can be reduced with proper changes to the model (e.g., more training data, more flexible model), aleatoric uncertainties are determined by the input images and thus cannot be reduced. There has been much recent work on how to estimate uncertainties efficiently in deep learning (e.g., Gal & Ghahramani, 2016; **?**; **?**; **?**) and some of these techniques have also been applied to astrophysical problems (e.g., **????**). The following two sections describe how we arrange for GaMPEN to estimate both parameter values and their uncertainties.

3.4.1 Bayesian Implementation of GaMPEN and Epistemic Uncertainties

To create a Bayesian framework while predicting morphological parameters, we have to treat the model itself as a random variable—or more precisely, the weights of our network must be probabilistic distributions instead of single values. For a network with weights, $\boldsymbol{\omega}$, and a training dataset, $\mathcal{D}$, of size N with input images $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ and output parameters $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$, the posterior of the network weights, $p(\boldsymbol{\omega} \mid \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}, \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}) \equiv p(\boldsymbol{\omega} \mid \mathcal{D})$ represents the plausible network parameters. To predict the probability distribution of the output variable $\hat{\boldsymbol{Y}}$ given a new test image $\hat{\boldsymbol{X}}$, we need to marginalize over all possible weights $\boldsymbol{\omega}$:

$$p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \mathcal{D}) = \int p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathcal{D}) d\boldsymbol{\omega}. \tag{3.6}$$

In order to calculate the above integral, we need to know $p(\boldsymbol{\omega} \mid \mathcal{D})$, i.e., how likely is a particular set of weights given the available training data, $\mathcal{D}$. Since we have trained only the one model, it does not tell us how likely different sets of weights are. Different approximations have been introduced in order to calculate this distribution, with variational inference (**?**) being the most popular.

Now, the dropout technique was introduced by **?** in order to prevent neural networks from overfitting; they temporarily removed random neurons from the network according to a Bernoulli distribution, i.e., individual nodes were set to zero with a probability, p, known as the dropout rate. This dropout process can also be interpreted as taking the trained model and permuting it into a different one (**?**).

Using variational inference and dropout, we can approximate the integral in Equation 3.6 as

$$\int p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathcal{D}) d\boldsymbol{\omega} \approx \frac{1}{T} \sum_{t=1}^{T} p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \boldsymbol{\omega}_t), \tag{3.7}$$

wherein we perform T forward passes with dropout enabled and $\boldsymbol{\omega}_t$ is the set of weights during the t^{th} forward pass. This procedure is what is referred to as Monte Carlo Dropout. For a detailed derivation of Equation 3.7, please refer to Appendix 3.B

In order to obtain epistemic uncertainties for GaMPEN, we insert a dropout layer before every weight layer in the network. Each forward pass through GaMPEN samples the approximate parameter posterior. Thus, in order to obtain epistemic uncertainties, we feed every test image $\hat{\boldsymbol{X}}_i$ to the trained GaMPEN framework T times and collect the outputs. In implementing the Monte Carlo Dropout technique, an often-ignored key step is tuning the dropout rate (i.e., the rate at which neurons are set to zero). We discuss the tuning of the dropout rate for GaMPEN in §3.5.

3.4.2 Likelihood Calculation and Aleatoric Uncertainties

Our simulated training data consists of noisy input images by design, but we know the corresponding morphological parameters with perfect accuracy. However, due to the different amounts of noise in each image, the predictions of GaMPEN at test time should have different levels of uncertainties. Thus, in this situation, we want to use a heteroscedastic model – a model that can capture different levels of uncertainties in its output predictions. We achieve this by training GaMPEN to predict the aleatoric uncertainties.

As outlined in §3.5, GaMPEN predicts a multivariate log-normal distribution of output variables for any given input image. Thus, for a given set of network weights $\boldsymbol{\omega}$, the likelihood $p(\mathcal{D} \mid \boldsymbol{\omega})$ is simply the product of the probabilities that the GaMPEN output for each image is drawn from the associated multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in $\mathbf{R^3}$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Although we would like to use GaMPEN to predict aleatoric uncertainties, the covariance matrix, $\boldsymbol{\Sigma}$, is not known *a priori*. Instead, we train GaMPEN to learn these values by minimizing the negative log-likelihood of the output parameters for the training set, which can be written as

$$
\begin{aligned}
-\log \mathcal{L}_{VI} \propto \sum_n \frac{1}{2} \left[\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_n\right]^\top \hat{\boldsymbol{\Sigma}_n}^{-1} \left[\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_n\right] \\
+ \frac{1}{2} \log[\det(\hat{\boldsymbol{\Sigma}_n})] + \lambda \sum_i \left\|\boldsymbol{\omega_i}\right\|^2 .
\end{aligned}
\tag{3.8}
$$

where $\hat{\boldsymbol{\mu}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ are the mean and covariance matrix of the multivariate Gaussian distribution predicted by GaMPEN for an image, $\boldsymbol{X}_n$. λ is the strength of the regularization term, and $\boldsymbol{\omega}_i$ are sampled from $q(\boldsymbol{\omega})$. For a detailed derivation of Equation 3.8, we refer an interested reader to Appendix 3.C.

The covariance matrix here represents the uncertainties in the predicted parameters arising from inherent corruptions to the input or the output data. Note that using the full covariance matrix in Equation 3.8 instead of just the diagonal terms (i.e., assuming the output variables to be independent), helps GaMPEN to incorporate the structured relationship between the different output parameters. We further outline the effects of this in §3.5.

3.4.3 Practical Implementation Details

In order to predict $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the final layer of GaMPEN contains nine output nodes (see Fig. 3.3). Three of these nodes are used to characterize $\boldsymbol{\mu}$. Now, although $\boldsymbol{\Sigma}$ is a 3×3 matrix, we are able to characterize it with just six parameters due to its special properties. Because $\boldsymbol{\Sigma}$ is a symmetric, positive-definite matrix, we can use the LDL decomposition, a variant of the Cholesky decomposition (Cholesky, 1924), to represent

$$
\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}
\tag{3.9}
$$

as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top$, where

$$
\boldsymbol{L} = \begin{pmatrix} 1 & 0 & 0 \\ \sigma_{21} & 1 & 0 \\ \sigma_{31} & \sigma_{32} & 1 \end{pmatrix}
\tag{3.10}
$$

and

$$D = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}. \tag{3.11}$$

Thus, three of GaMPEN's output nodes are used to predict the off-diagonal elements in Equation 3.10, and three more are used to predict ${}^{3}_{i=1}s_i$ where $s_i = \log(\sigma_i^2)$. We predict s_i instead of σ_i^2 in order to achieve better numerical stability during training.

The loss function, outlined in Equation 3.8, contains the determinant and the inverse of $\boldsymbol{\Sigma}$. Calculation of the determinant and the inverse of a matrix are potentially numerically unstable and slow operations. Thus, in order to achieve the maximum speed possible on our GPUs and for numerical stability, we replace these operations using the Cholesky decomposition outlined above and standard linear algebra. That is, $\boldsymbol{\Sigma}^{-1}$ can be written as $\boldsymbol{\Sigma}^{-1} = \left(\boldsymbol{L}^{-1}\right)^{\top} \boldsymbol{D}^{-1}\boldsymbol{L}^{-1}$, where

$$\boldsymbol{D}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} \end{pmatrix} \tag{3.12}$$

because $\boldsymbol{D}$ is a diagonal matrix. Because $\boldsymbol{L}$ is a lower triangular matrix, we can also write its inverse as

$$\boldsymbol{L}^{-1} = (\boldsymbol{I} + \boldsymbol{N})^{-1} = \boldsymbol{I} + \sum_{k=1}^{2} (-1)^k \boldsymbol{N}^k, \tag{3.13}$$

where $\boldsymbol{I}$ is a 3×3 identity matrix and $\boldsymbol{N}$ is a strictly lower triangular and nilpotent matrix such that $\boldsymbol{N} = \boldsymbol{L} - \boldsymbol{I}$.

Finally, we can write the $\log(\det(\boldsymbol{\Sigma}))$ as

$$\begin{aligned} \log(\det(\boldsymbol{\Sigma})) &= \log(\det(\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\top})) \\ &= \log(\prod_i D_{ii}) \\ &= \sum \log D_{ii}. \end{aligned} \tag{3.14}$$

By combining Equations 3.12, 3.13, and 3.14, we can calculate the log-likelihood outlined in Equation 3.8 without having to calculate the inverse or determinant of any matrix, allowing us to fully utilize the capabilities of a GPU and avoiding any numerical instabilities.

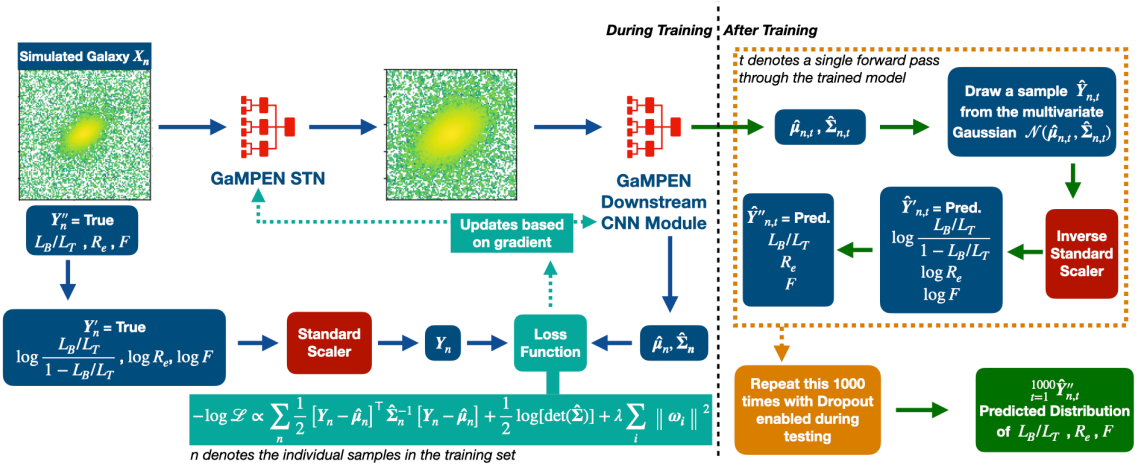


Figure 3.7: Diagram outlining the training (*left*) and inference (*right*) phases of the GaMPEN workflow. Training consists of feeding 105,000 simulated images (with known parameter values) through the STN and CNN modules, minimizing the loss function (Eqn. 3.8) using Stochastic Gradient Descent. During this process, we re-scale the variables as described in the text, and return them to the original variable space during inference. After the STN+CNN are trained, the inference step consists of 1000 forward passes with dropout enabled for each galaxy image. We draw a sample from the predicted multivariate Gaussian distribution during each forward pass, and the collection of these samples gives us the predicted posterior distribution.

3.4.4 Combining Aleatoric and Epistemic Uncertainties

To obtain the posterior distribution of the output variables, we need to combine the aleatoric and epistemic uncertainties. After training a model by maximizing the log-likelihood outlined in Equation 3.8, we perform Monte Carlo Dropout. To do this, as outlined in Figure 3.7, we feed each input image, $\hat{\boldsymbol{X}}_n$, in the test set 1000 times into GaMPEN with dropout enabled. During each iteration, we collect the predicted set of $\left(\hat{\boldsymbol{\mu}}_{n,t}, \hat{\boldsymbol{\Sigma}}_{n,t}\right)$ for the t^{th} forward pass. Then, for each forward pass, we draw a sample $\hat{\boldsymbol{Y}}_{\boldsymbol{n,t}}$ from the multivariate normal distribution $\mathcal{N}\left(\hat{\boldsymbol{\mu}}_{n,t}, \hat{\boldsymbol{\Sigma}}_{n,t}\right)$.

The distribution generated by the collection of all 1000 forward passes, $\hat{\boldsymbol{Y}}_{\boldsymbol{n}}$, represents the predicted posterior distribution for the test image $\hat{\boldsymbol{X}}_n$. The different forward passes capture the epistemic uncertainties, and each prediction in this sample also has its associated aleatoric uncertainty represented by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n,t}}$. Thus the above procedure allows us to incorporate both aleatoric and epistemic uncertainties in the prediction of posteriors of morphological parameters by GaMPEN.

3.5 Training GaMPEN

We split the dataset of 150,000 simulated galaxies into training, validation, and testing sets with 70%, 15%, and 15% of the total sample, respectively. We train GaMPEN using the training set,

and set the values of various hyper-parameters (e.g., learning rate, batch size; see below) using the validation set. Finally, we evaluate the trained model on the testing set (which has never been seen before by the network) and report the results in §3.6.

We pass all the images in the simulated dataset through the arsinh function to reduce the dynamic range of pixel values in the images. This function behaves linearly around zero and logarithmically for large values. Reducing the dynamic range of pixel values has been found to be helpful in neural network convergence (e.g., Tanaka et al., 2022; Walmsley et al., 2021; Zanisi et al., 2021), hence this approach.

The three output variables that we predict with GaMPEN have quite different ranges, by orders of magnitude. Thus, we re-scale these ground-truth training values before feeding them into the network in order to prevent variables with larger numerical values from making a disproportionate contribution to the loss function. Additionally, we also need to make sure that none of the values predicted by GaMPEN happen to be unphysical; that is, we require all output values to adhere to the following ranges: $0 \leq L_B/L_T \leq 1$; $R_e > 0$; $F > 0$.

Therefore, we first apply the logit transformation to L_B/L_T and log transformations to R_e and F:

$$\boldsymbol{Y'_n} = f''(\boldsymbol{Y''}_n) = \left(\log \frac{L_B/L_T}{1 - L_B/L_T}, \log R_e, \log F\right), \tag{3.15}$$

where $\boldsymbol{Y''}_n = [L_B/L_T, R_e, F]$ is the set of ground-truth parameters corresponding to the simulated image, $\boldsymbol{X}_n$, and f'' is how we will refer to the transformation in Equation 4.1. Note that the uniformity of these transformations allows us to write the likelihood in terms of a multivariate Gaussian distribution. Next we apply the standard scaler to each parameter (calibrated on the training data), which amounts to subtracting the mean value of each parameter and scaling its variance to unity:

$$\boldsymbol{Y}_n = f'(\boldsymbol{Y'}_n) \quad \text{where} \quad Y_{n,i} = \frac{Y'_{n,i} - \overline{Y'_i}}{\sqrt{\text{var}(Y'_i)}}, \tag{3.16}$$

where the i subscript refers to each of the three parameters. Combining the above transformations, f' and f'', ensures that all three variables have similar numerical ranges.

Note that effectively GaMPEN is trained in the $\boldsymbol{Y}_n$ variable space, and the predictions made by GaMPEN are also in this space. Thus, post training, during inference, we need to apply the inverse of the standard scaler function, f'^{-1} (with no re-tuning of the mean or variance), followed by the inverse of the logit and log transformations, f''^{-1}, as indicated in Figure 3.7. These final

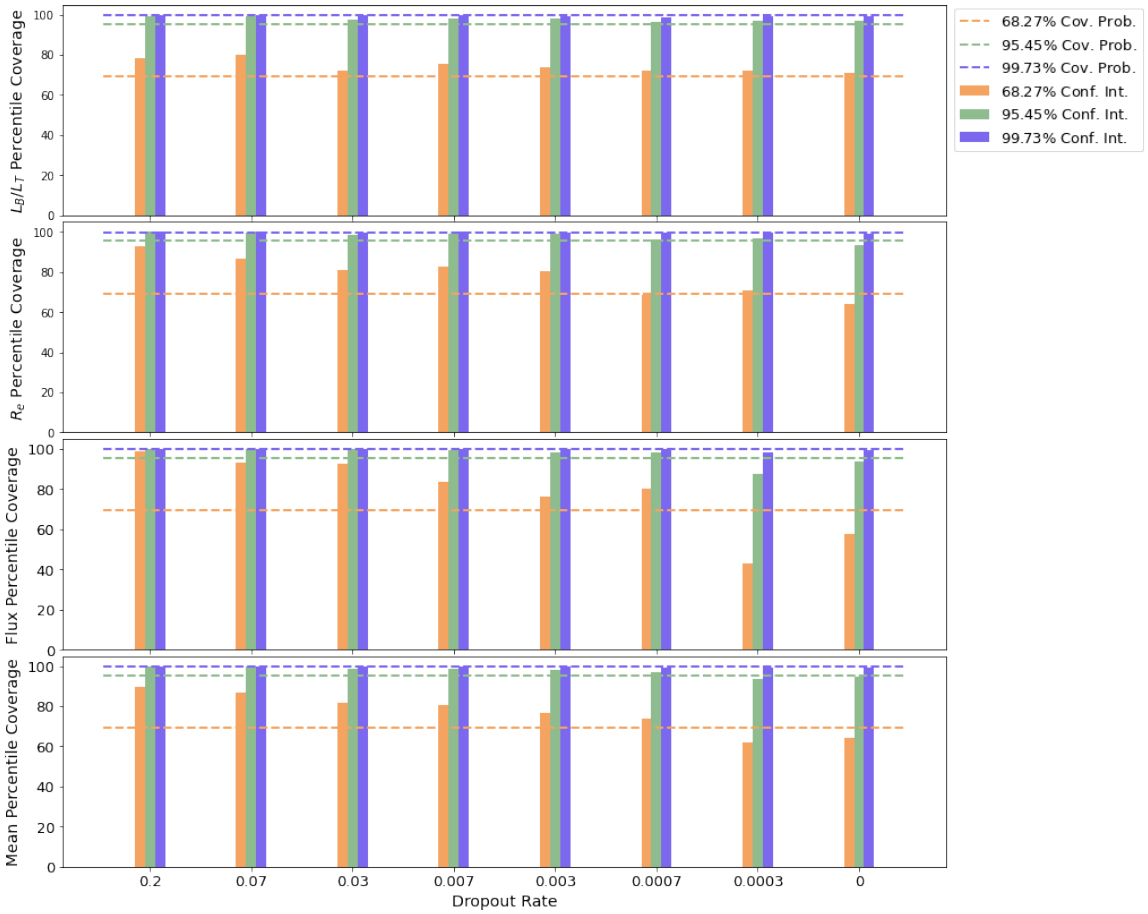


Figure 3.8: The calculated percentile coverage probabilities for different dropout rates. The top three rows show coverage probabilities for each output variable individually, while the bottom row shows the probabilities averaged over the three variables. The coverage probabilities are defined as the percentage of the total test examples where the true value of the parameter lies within a particular confidence interval of the predicted distribution. A dropout rate of 7×10^{-4} leads to coverage probabilities very close to their corresponding confidence levels.

transformations also ensure that the predicted values are all within the physical ranges mentioned earlier.

We train GaMPEN by minimizing the loss function in Equation 3.8 using Stochastic Gradient Descent, wherein we estimate the gradient of the loss function using a mini-batch of training samples and update the network weights and biases accordingly. Calculation of the gradient is done using the back-propagation algorithm, and we refer an interested reader to Rumelhart et al. (1986) for details.

The training process involves hyper-parameters that must be chosen: the learning rate (the step-size during gradient descent), momentum (acceleration factor used for faster convergence), strength of L2 regularization (λ in the loss function in Eqn. 3.8), and batch size (the number of

images processed before weights and biases are updated). To choose these hyper-parameters, we trained GaMPEN with a given set of hyper-parameters for forty epochs, then verified convergence by checking whether the value of the loss function and the mean-absolute-error on the validation set had stabilized over at least the last ten epochs. An epoch of training refers to running all of the images in the training set through the network once. We chose final hyper-parameters that resulted in the lowest value for the loss function. This resulted in the following values: Learning Rate, 5×10^{-7}; Momentum, 0.99; Strength of L2 regularization $\lambda = 10^{-4}$, and Batch Size, 16. The grid of values we used for the hyper-parameter search is as follows:- Learning Rate - 10^{-5}, 5×10^{-5}, 10^{-6}, 5×10^{-6}, ..., 5×10^{-8}; Momentum - 0.8, 0.9, 0.95, 0.99; λ - 10^{-5},10^{-4}, ..., 10^{-2}; Batch Size: 8, 16, 32, 64.

One of the most critical adjustable parameters is the dropout rate, as it directly affects the calculation of the epistemic uncertainties (as described in §3.4.1). On average, higher dropout rates lead networks to estimate higher epistemic uncertainties. To determine the optimal value for the dropout rate, we trained variants of GaMPEN with dropout rates from 0 to 0.2, all with the same optimized values of momentum, learning rate, and batch size given above. After that, we performed inference using each model as outlined in Figure 3.7.

To compare these models, we calculated the percentile coverage probabilities associated with each model, defined as the percentage of the total test examples where the true value of the parameter lies within a particular confidence interval of the predicted distribution. We calculate the coverage probabilities associated with the 68.27%, 95.45%, and 99.73% central percentile confidence levels, corresponding to the 1σ, 2σ, and 3σ confidence levels for a normal distribution. For each distribution predicted by GaMPEN, we define the 68.27% confidence interval as the region on the x-axis of the distribution that contains 68.27% of the most probable values of the integrated probability distribution. In order to estimate the probability distribution function from the GaMPEN predictions (which are discrete), we use kernel density estimation, which is a non-parametric technique to estimate the probability density function of a random variable.

We calculate the 95.45% and 99.73% confidence intervals of the predicted distributions in the same fashion. Finally, we calculate the percentage of test examples for which the true parameter values lie within each of these confidence intervals. An accurate and unbiased estimator should produce coverage probabilities equal to the confidence interval for which it was calculated (e.g., the coverage probability corresponding to the 68.27% confidence interval should be 68.27%).

Figure 3.8 shows the coverage probabilities for the three output parameters individually (top three panels), as well as the coverage probabilities averaged over the three output variables (bottom panel).

As can be seen, higher values of the dropout rate lead to GaMPEN over-predicting the epistemic uncertainties, resulting in too high coverage probabilities. In contrast, extremely low values lead to GaMPEN under-predicting the epistemic uncertainties. For a dropout rate of 7×10^{-4}, the calculated coverage probabilities are very close to their corresponding confidence levels, resulting in accurately calibrated posteriors. The dropout rate is clearly a variational parameter of GaMPEN, and all the results shown hereafter correspond to a GaMPEN model trained with a dropout rate of 7×10^{-4}.

It is important to note that the inclusion of the full covariance matrix in the loss function allowed us to incorporate the relationships between the different output variables in GaMPEN predictions. This allowed us to achieve simultaneous calibration of the coverage probabilities for all three output variables. In contrast, using only the diagonal elements of the covariance matrix resulted in substantial disagreement, for a fixed dropout rate, among the coverage probabilities of the different parameters. Additionally, when we used three different neural networks to predict each output variable, we achieved a poorer overall accuracy. Thus, using the full covariance matrix, facilitated by the linear algebraic tricks outlined in §3.4.3, allows GaMPEN to predict accurate, calibrated posteriors.

3.6 Results

After training GaMPEN and tuning its hyper-parameters on the training and validation sets, as outlined in §3.5, we perform inference using the testing set of 22,500 galaxies.

3.6.1 Inspecting the Predicted Posteriors

As outlined in Figure 3.7 and §3.4.4, during the inference phase, we pass each image in the testing set 1000 times through GaMPEN. Note that due to our use of Monte-Carlo Dropout, each of these forward passes happens through a slightly different network because of how the technique drops out (sets to zero) randomly selected neurons according to a Bernoulli distribution. This technique allows us to effectively factor in the uncertainty about our predictive model into GaMPEN predictions. For each forward pass t and for each test image $\hat{\boldsymbol{X}}_n$, GaMPEN predicts a vector of means $\hat{\boldsymbol{\mu}}_{n,t}$ and a covariance matrix $\hat{\boldsymbol{\Sigma}}_{n,t}$. These two parameters are used to define a multivariate Gaussian distribution $\mathcal{N}(\hat{\boldsymbol{\mu}}_{n,t}, \hat{\boldsymbol{\Sigma}}_{n,t})$ from which we draw a sample $\hat{\boldsymbol{Y}}_{n,t}$. These are then processed through two sets of transformations (f'^{-1} and f''^{-1}) outlined in §3.5 resulting in the transformed prediction $\hat{\boldsymbol{Y}''}_{n,t}$. The collection of these samples for the 1000 forward passes $\overset{1000}{t=1}\hat{\boldsymbol{Y}''}_{n,t}$ represents the posterior distribution

predicted by GaMPEN for the test image $\hat{\boldsymbol{X}}_n$.

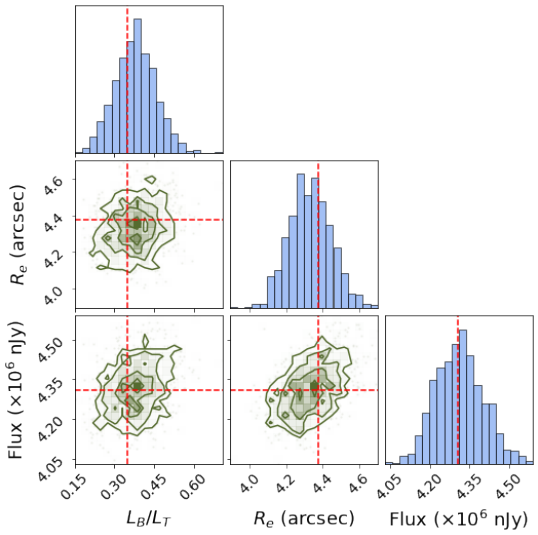


Figure 3.9: Joint and marginalized probability distributions predicted by GaMPEN for a randomly chosen galaxy in our testing set. The red dotted lines show the true values of the parameters.

Using the above process, we extract the joint probability distribution of all the output parameters for each of the 22,500 galaxies in our test set. Figure 3.9 shows the two-dimensional joint distributions of the output parameters, as well the marginalized distributions, for a randomly chosen galaxy in our test set. The same galaxy is shown in the second row of Figure 3.10, which illustrates the predicted posterior distributions for a few more cases. Figure 3.10 also shows the image of each galaxy at the left. As expected, all the predicted distributions are unimodal, smooth, and resemble Gaussian or truncated Gaussian distributions. For each predicted distribution, the figure also shows the parameter space regions that contain 68.27%, 95.45%, and 99.73% of the most probable values of the integrated probability distribution. We use kernel density estimation to estimate the probability distribution function (PDF; shown by a blue line in the figure) from the predicted values. The mode of this PDF is what we refer to as the predicted value when calculating residuals. In the figure, for most cases, the true value lies within the most probable 68.27% percentile region. We also visually inspected the distributions predicted by GaMPEN for ~ 200 galaxies to ensure that there were no systematic or catastrophic errors (e.g., substantial errors for a specific parameter only, or bi-modal or irregular distributions for specific kinds of galaxies, etc.).

By design, GaMPEN predicts only physically possible values. This is especially apparent in the L_B/L_T column of rows 1 and 4 of Figure 3.10. Note that to achieve this, we do not artificially truncate these distributions. Instead, we use the inverse of the logit transformation on the prediction space of GaMPEN as outlined in §3.5. This ensures that the predicted L_B/L_T values are always

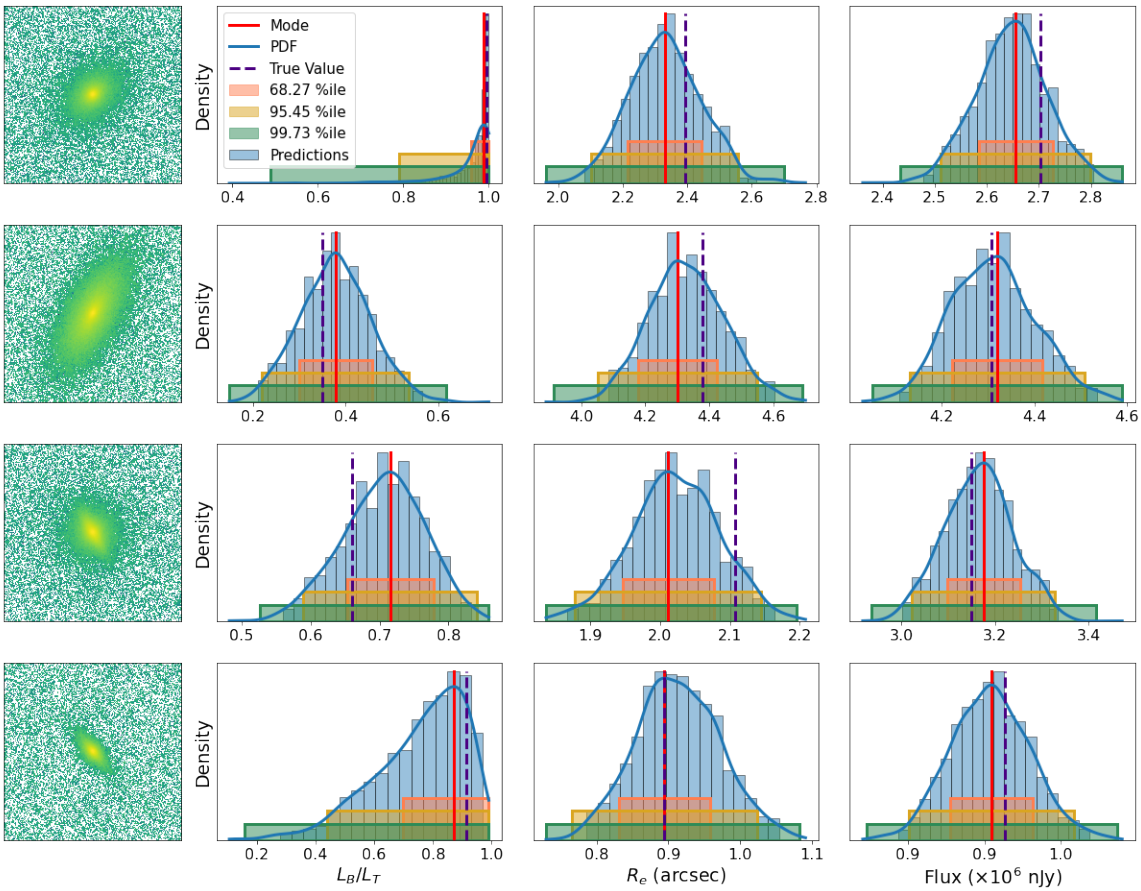


Figure 3.10: Examples of predicted posterior distributions for four randomly chosen simulated galaxies. The blue shaded histogram shows the predictions from GaMPEN and the blue solid lines show the associated probability distribution functions estimated by kernel density estimation. These are used to calculate the confidence intervals shown in the figure with pink, yellow, and green shading. The mode (red line) shows the most probable value of each morphological parameter. As expected, in most cases, the true value (purple line) lies within the 68.27% confidence interval.

between 0 and 1. Similarly, we also ensure that the R_e and F values predicted by GaMPEN are positive through appropriate transformations.

3.6.2 Evaluating the Accuracy of GaMPEN

In §3.6.1, we explored the predicted distributions for a handful of cases, where the true values of the parameters mostly lay within the densest parts of the probability distribution predicted by GaMPEN. In order to evaluate the accuracy of GaMPEN, we now report summary statistics that outline the framework's performance on the entire testing set.

In Table 3.2, we report the coverage probabilities that GaMPEN achieves on the test set. In the ideal situation, they would perfectly mirror the confidence levels; that is, 68.27% of the time, the

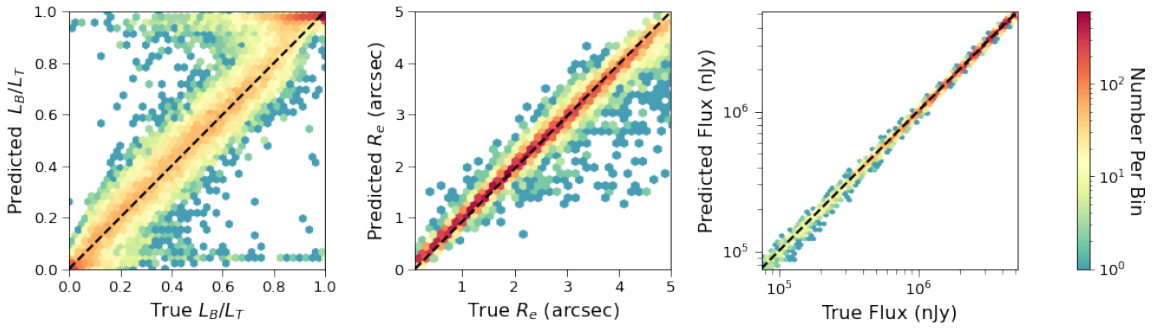


Figure 3.11: The true values of the galaxy parameters plotted against the most probable values predicted by GaMPEN. The black dashed line marks the $y = x$ diagonal on which perfectly recovered parameters should lie. The color of each hexagon corresponds to the number of galaxies it contains, as indicated by the colorbar at right.

Table 3.2: Coverage Probabilities on the Test Set

Parameter Name	68.27% Conf. Level	95.45% Conf. Level	99.73% Conf. Level
L_B/L_T	71.8%	96.9%	98.9%
R_e	68.1%	95.9%	98.3%
F	78.7%	98.2%	99.9%
Mean	72.9%	97.0%	99.0%

NOTE- The coverage probabilities are defined as the percentage of the total test samples where the true value of the parameter lies within a particular confidence interval of the predicted distribution.

true value would lie within 68.27% of the most probable volume of the predicted distribution. (Note that in § 3.5 we tuned the dropout rate so they coincide over the validation set, whereas Table 3.2 is calculated on the testing set.) Clearly, GaMPEN produces well calibrated and accurate posteriors, consistently close to the claimed confidence levels. Additionally, we note that even for the flux, for which the coverage probabilities are most discrepant, the uncertainties predicted by GaMPEN are in any case overestimates (i.e., conservative). If GaMPEN were used in a scenario that requires perfect alignment of coverage probabilities, users could employ techniques such as importance sampling (Kloek & van Dijk, 1978) on the distributions predicted by GaMPEN.

Having defined the overall percentage of cases where the true values are within particular confidence levels of the predicted distributions, we now quantify the difference between the most probable values of the predicted parameters (i.e., modes of these predicted distributions) and the true values.

Figure 3.11 shows the most probable values predicted by GaMPEN for the testing set versus the true values, in hexagonal bins of roughly equal size, with the number of galaxies represented according to the colorbar on the right. We have used a logarithmic colorbar to visualize even small

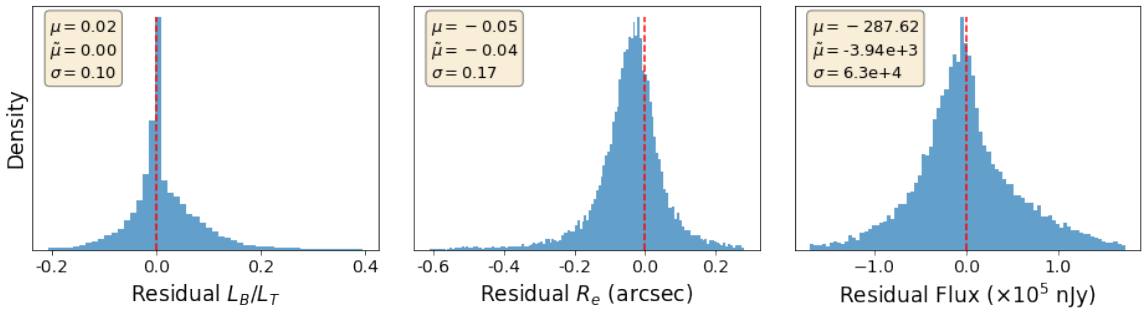


Figure 3.12: Histograms of residuals for all galaxies in the testing set. We define the residuals as the difference between the true value and the most probable value predicted by GaMPEN. The dashed vertical line represents $x = 0$, denoting cases with perfectly recovered parameter values. The mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution are listed in each panel.

clusters of galaxies in this plane. Most galaxies are clustered around the line of equality, showing that the most probable values of the distributions predicted by GaMPEN closely track the true values of the parameters.

The middle panel of Figure 3.11 shows a small bias (note that the color scale is logarithmic) towards low predictions of R_e, especially for true $R_e > 4$ arcsec. Features like this have been seen before in other machine learning studies and are typically referred to as an "edge effect" – that is, sometimes the model performs poorly at the edges of the parameter space on which it was trained. Here, since $R_e = 5$ arcsec is the largest radius present in our training data, for some galaxies with R_e close to 5 arcsec, the network is hesitant to predict the highest value it has ever seen and predicts a slightly lower value. This results in the small bias toward lower predicted values of R_e. However, note that despite this effect for a small number of galaxies; even at a large radius, GaMPEN accurately estimates R_e for the large majority of galaxies. Among some other larger deviations evident in Figure 3.11, are predictions near the limits of L_B/L_T. We explore this further below.

In Figure 3.12, we show the residual distribution for the three parameters predicted by GaMPEN. We define the residual for each parameter as the difference between the most probable predicted value and the true value, i.e., Mode($\hat{\boldsymbol{Y}}_{\boldsymbol{n}}$) − $\boldsymbol{Y}_{\boldsymbol{n}}$. The box in the upper left corner gives the mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution. All three distributions are normally distributed (verified using the Shapiro Wilk test), and have $\mu \sim \tilde{\mu} \sim 0$. The GaMPEN prediction of the bulge-to-total ratio is, in $\sim$ 68.27% of cases, within 0.1 of the true value. The typical error in effective radius is 0.17 arcsec. Typical uncertainties in the flux are at the 0.1-1% level.

Although Figures 3.11 and 3.12 indicate the overall accuracy of GaMPEN, they do not reveal how

those errors depend on location in the parameter space. This is critical information as this enables us to potentially ignore predictions for regions of parameter space that have large errors (according to the validation set). Figure 3.13 shows the residuals for the three output parameters plotted against the true values. As in Figure 3.11, we have split the parameter space into hexagonal bins and used a logarithmic color scale to denote the number of galaxies in each bin. The purpose of this plot is to identify regions of parameter space where GaMPEN performs especially well or badly, so that, in the future, we can flag predictions in these regions as "very secure" or "unreliable". Note that because we are performing the test here on simulated galaxies, we have access to the ground-truth values. However, in a scenario where GaMPEN is being used on real galaxies which have not been morphologically studied before, we won't have access to the ground-truth values, and any such cuts on the X-axis would need to be made based on the values predicted by GaMPEN. Thus, we created Figure 3.14, where we replaced the X-axis with the predicted values of the parameters instead of the true values.

In both Figures 3.13 and 3.14, for most of the panels, the large majority of galaxies are clustered uniformly around the black dashed line, $y = 0$, which denotes the ideal case of perfectly recovered parameters.

There are a few other notable features in these two figures. In the top left panel, the L_B/L_T residuals are highest near the limits of L_B/L_T. This is another manifestation of the edge-effect mentioned earlier, wherein sometimes machine learning algorithms perform poorly at the edges of the parameter space on which they were trained. To delve deeper, we looked at the L_B/L_T residuals separately for double and single component galaxies, as shown in Figure 3.15. For single-component galaxies, the typical L_B/L_T residual ($\sigma = 0.06$) is roughly half as large as for double-component galaxies ($\sigma = 0.11$); among the latter, the residuals are especially high when $L_B/L_T > 0.85$ or $L_B/L_T < 0.1$. In other words, accurately determining L_B/L_T is challenging when both a bulge and a disk are present, and becomes even more difficult when one component strongly dominates the other. Larger residuals in the predictions near the limits of L_B/L_T leads to the features seen in the top left panel of both figures.

This edge effect also results in the top and bottom streaks seen in the left panel of Figure 3.11. Given the logarithmic color bar used in this figure, note that most of the galaxies in the upper streak have true $L_B/L_T > 0.75$ and those in the bottom streak have true $L_B/L_T < 0.25$. For these cases, when one component completely dominates over the other component, precisely determining L_B/L_T is challenging. For some of the galaxies with $0.25 >$ True $L_B/L_T > 0.75$, GaMPEN assigns almost the entirety of the light to the dominant component, resulting in the streaks in the left panel. We

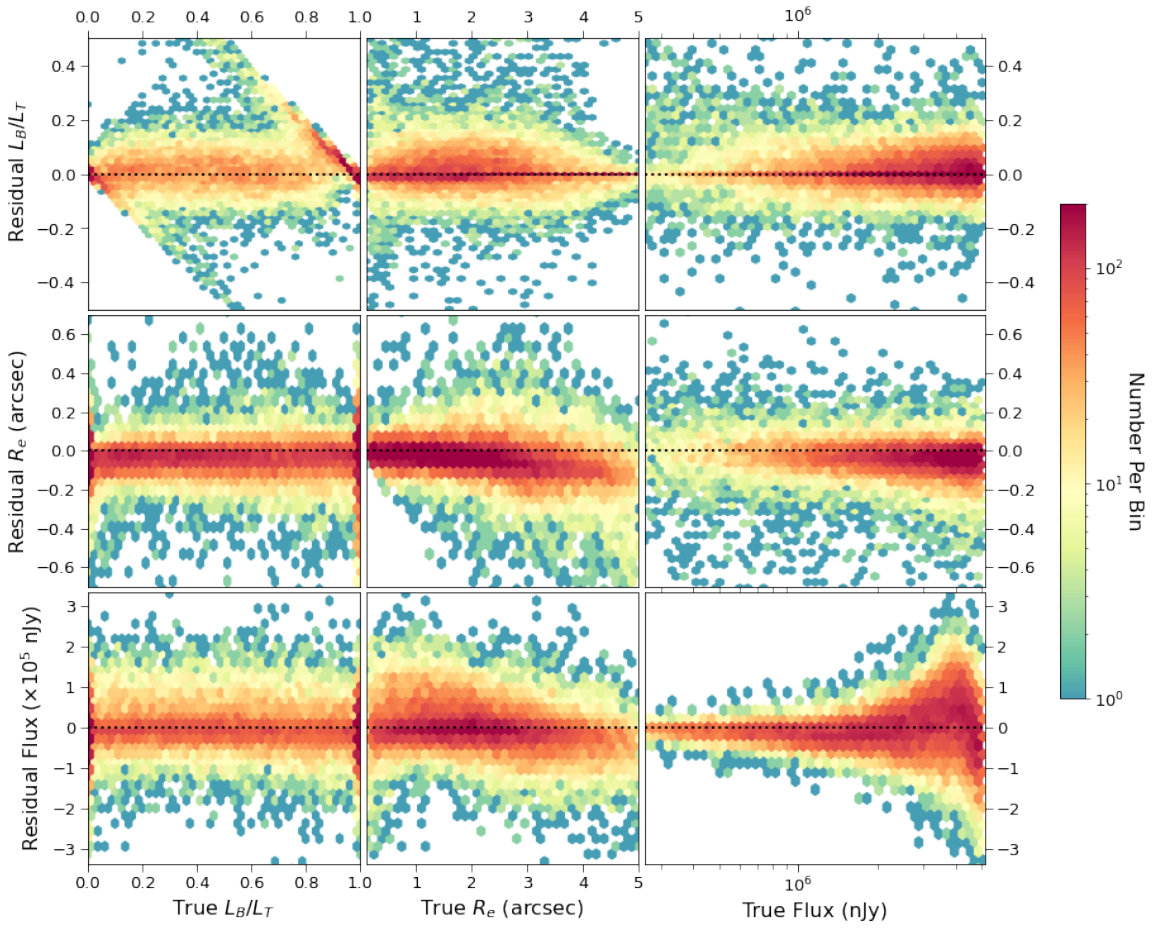


Figure 3.13: Residuals of GaMPEN predicted parameter values plotted against the true values. The residual for each parameter is defined as the difference between the most probable predicted value and the true value, i.e., Mode($\hat{\boldsymbol{Y}}_{\boldsymbol{n}}$) $- \boldsymbol{Y}_{\boldsymbol{n}}$. The color of each hexagonal bin corresponds to the number of galaxies it contains, as shown by the colorbar on the right. The black dotted line ($y = 0$) represents perfectly recovered parameters.

use a parameter transformation to mitigate this edge effect, as described in § 3.6.4.

In the top-middle panels of Figures 3.13 and 3.14, there is a slight broadening of the residuals at low values of the effective radius. This result also makes sense: smaller galaxies are challenging to analyze for any image processing algorithm. Somewhat surprising features appear in the panels showing residuals of effective radius (mid-row, mid-column) and flux (bottom-row, right-column). The R_e residuals increase in magnitude (with a bias towards negative values) toward increasing values of R_e, and the residual flux also grows rapidly with higher flux values. However, these increases are simply the result of the increasing numerical values of the parameters. To show this, Figure 3.16 plots the dimensionless fractional R_e and F residuals (note that L_B/L_T is inherently dimensionless), using their absolute values, such that the ideal scenario (i.e., zero residuals) is at the

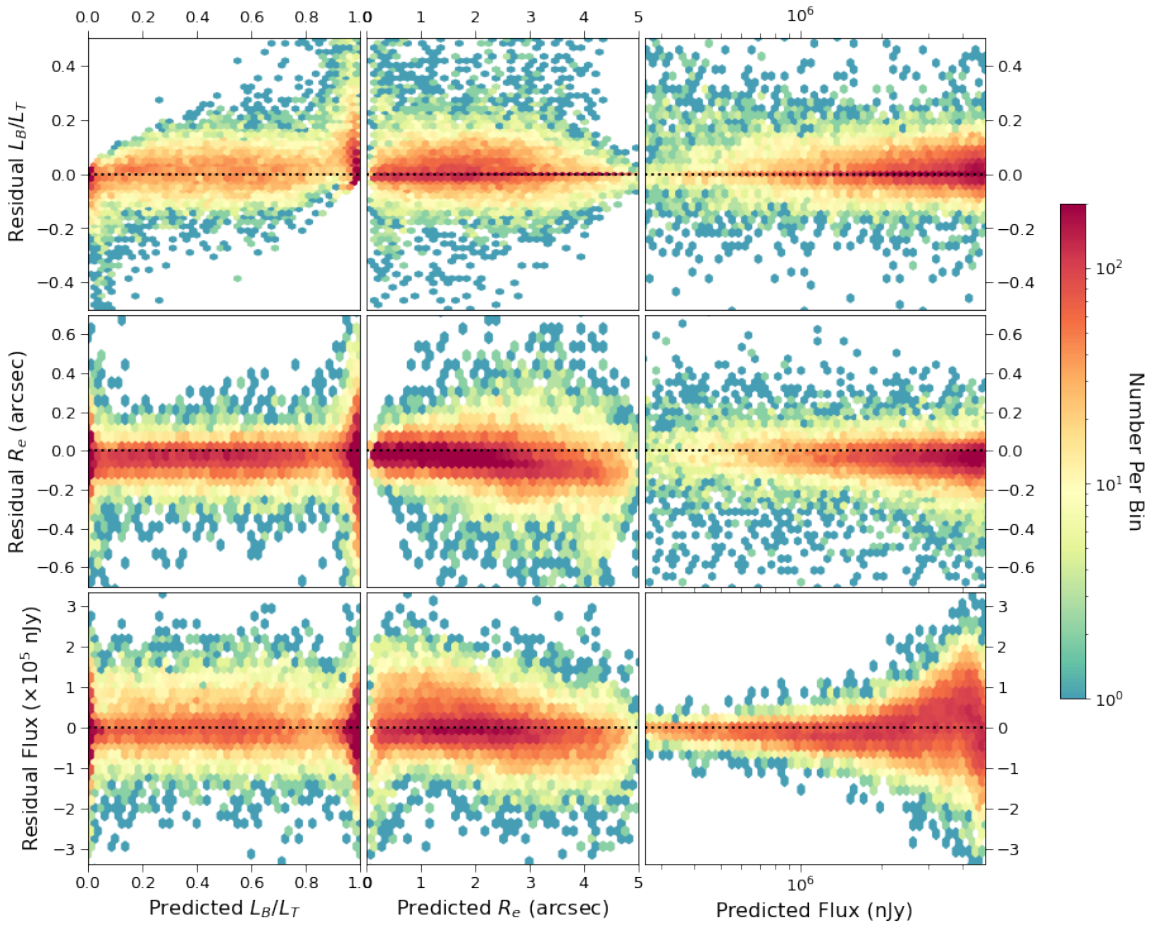


Figure 3.14: Residuals of the output parameters plotted against the predicted values. This figure allows us to assign quality labels to GaMPEN predictions (e.g., flagging parameters that are unreliable) based on the output values. See § 3.6.4 for details.

bottom of each panel instead of in the middle. In this presentation, both features noted above not only disappear but reverse. For small values of effective radius, $R_e < 1.0$ arcsec, there is an increase in the magnitude of the residuals. Similarly, the right two panels show that the residuals of R_e and F are systematically higher for faint galaxies, $F < 10^6$ nJy.

In other words, GaMPEN systematically becomes less accurate at predicting the radii of galaxies when their sizes become comparable to the seeing of the HSC-Wide Survey (g-band median FWHM $\sim 0.85''$). Similarly, GaMPEN finds it more challenging to predict the sizes and fluxes of fainter galaxies, just as one would expect. With our previously published classification network, GaMorNet (**?**), we observed a similar reduction in prediction accuracy for smaller and fainter galaxies. Figures 3.13, 3.14, and 3.16 help quantify the errors in GaMPEN predictions in different regions of parameter space. These will be essential in order to interpret results appropriately when applying GaMPEN to real galaxies.

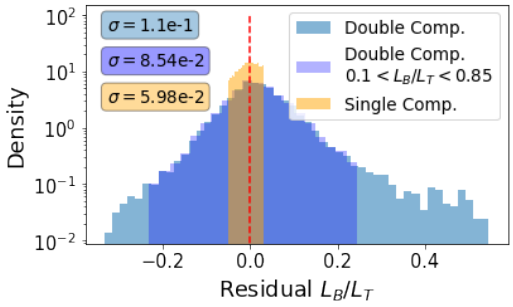


Figure 3.15: Histograms of L_B/L_T residuals shown separately for single component galaxies, all double component galaxies, and double component galaxies with $0.1 < L_B/L_T < 0.85$. The standard deviation (σ) for each distribution is also shown in the top left. The dashed vertical line represents $x = 0$, denoting cases with perfectly recovered L_B/L_T. The apparent hard cutoffs in the distributions of the single component, and the restricted range double-component galaxies arise from the fact that the y-scale is logarithmic. We have verified that when plotted on a linear scale, the apparent hard cutoffs disappear.

3.6.3 Inspecting the Predicted Uncertainties

The primary advantage of a Bayesian ML framework like GaMPEN is its ability to predict the full posterior distributions of the output parameters instead of just point estimates. Thus, we would expect such a network to inherently produce wider distributions (i.e., larger uncertainties) in regions of the parameter space where residuals are higher. Here we delineate regions of the parameter space for which GaMPEN predicts broader distributions and we see that these generally coincide with those that have the largest residuals (Figs. 3.13, 3.14, 3.16).

Figure 3.17 shows the uncertainties for the three predicted parameters plotted against the true values of the different parameters. We define the uncertainty predicted for each parameter as the width of the 68.27% confidence interval (i.e., the parameter interval that contains 68.27% of the most probable values of the predicted distribution; see Fig. 3.10). The lower two panels have been normalized, so that all three panels show dimensionless fractional uncertainties.

The uncertainties in the predicted values of the radius increase sharply for galaxies with $R_e < 2$ arcsec and/or $F < 10^6$ nJy. Similarly, the uncertainties in flux increase for galaxies with $R_e < 1$ arcsec and/or $F < 10^6$ nJy. This aligns perfectly with what we expect: the sizes and fluxes of small galaxies and/or faint galaxies are not well constrained. Just as the residuals for GaMPEN predictions were larger for small and/or faint galaxies (Figs. 3.13, 3.14, 3.16), the uncertainties predicted by GaMPEN are also larger for these galaxies.

The top left panel of Figure 3.17 shows that GaMPEN is reasonably certain of its predicted bulge-to-total ratio across the full range of values but appears slightly more certain when $L_B/L_T \leq 0.2$

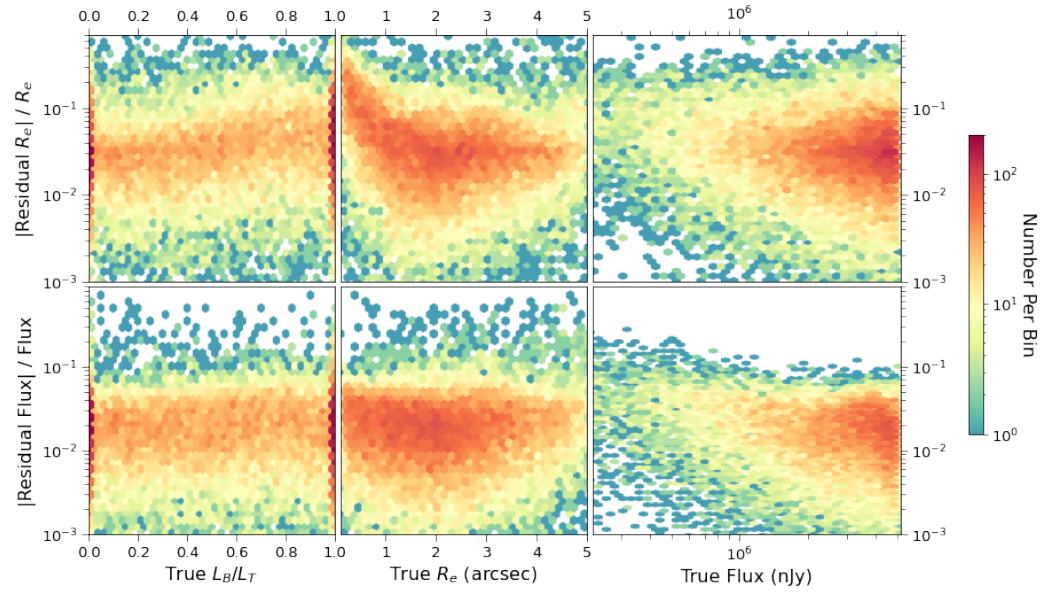


Figure 3.16: Fractional residuals for the effective radius and flux plotted against their corresponding true values. Note that since we are plotting the absolute values, the ideal situation of perfectly recovered parameters is at the bottom of each panel. The right two panels show that both the residuals increase for fainter galaxies, while the top-middle panel shows that the radius residuals increase for smaller galaxies.

or $L_B/L_T \geq 0.8$. It turns out that the smaller uncertainties at the limits correspond to the single-component galaxies, while for the double-component galaxies, the edge effect is less pronounced, in agreement with the residuals observed in Figure 3.15. Not surprisingly, the predicted uncertainty in R_e decreases with decreasing values of L_B/L_T (i.e., galaxies with more dominant disks, which are on average larger than bulge-dominated galaxies in our simulation sample).

In Figure 3.18, we further assess the estimated uncertainties by investigating their relation to the measured residuals. Note that while the uncertainties represent the widths of the central 68.27% confidence intervals, the residuals are the difference between the modes of the predicted distributions and the true values. Thus, we do not expect the two values to be linearly correlated; rather, on average, GaMPEN should predict larger uncertainties for galaxies with larger residuals, as is seen in all three panels of the figure. According to a Spearman's rank correlation test (see Dodge, 2008, for more details), there is a positive correlation between the residuals and uncertainties for all three variables, and the null hypothesis of non-correlation can be rejected at extremely high significance ($p < 10^{-200}$).

The results shown in this section outline the primary advantage of using a Bayesian framework like GaMPEN – even in situations where the network is not perfectly accurate, it is able to predict the right level of precision, allowing its predictions to be reliable and well-calibrated.

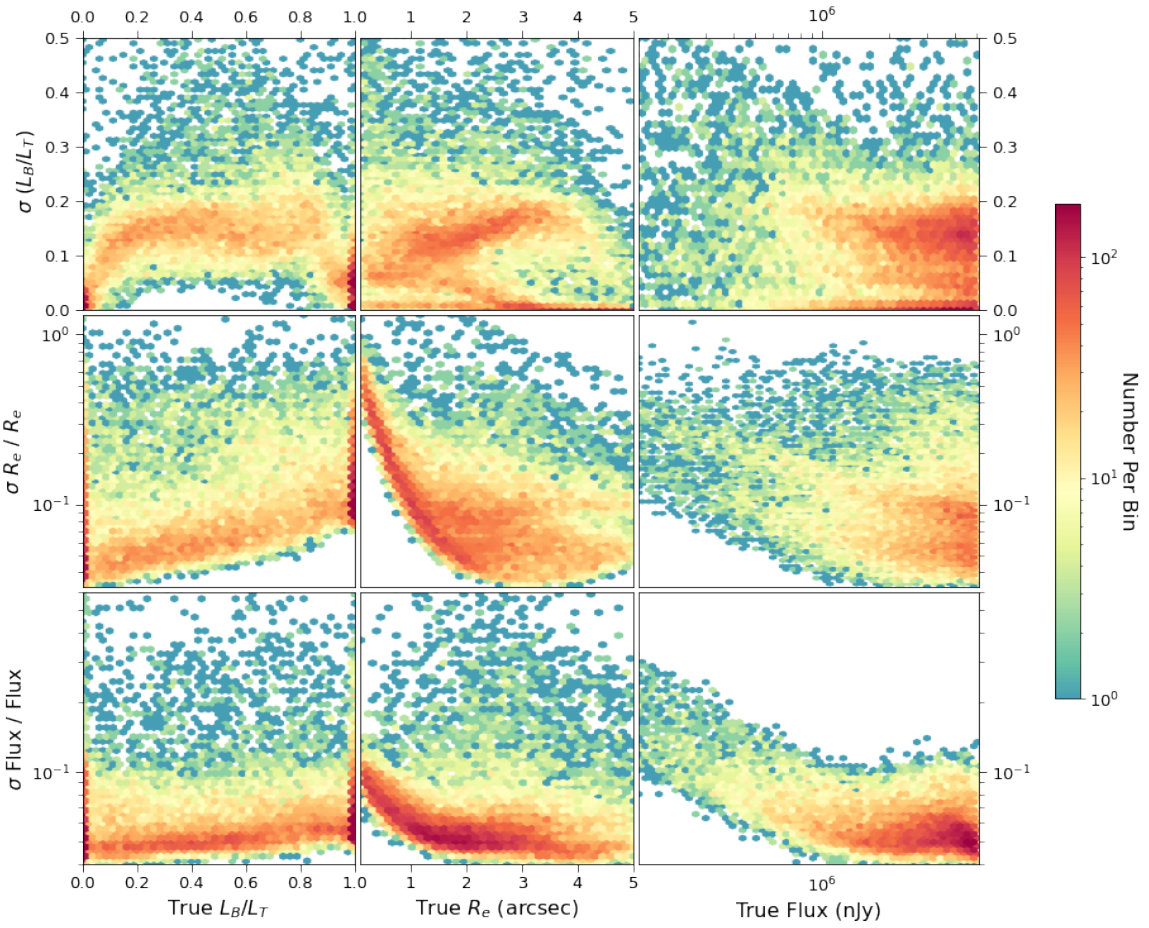


Figure 3.17: Uncertainties predicted by GaMPEN for each parameter plotted against the true values. The σ for each parameter is defined as the width of the 68.27% confidence interval. Note that we plot fractional uncertainties for radius and flux in order to make the y-axis dimensionless for all three rows.

3.6.4 Qualitative Transformation of GaMPEN Predictions

Given that we know GaMPEN residuals are higher for certain regions of the parameter space, we explore how using only qualitative labels in those regions (instead of quantitative predictions) affects the overall residual values. The labeling is informed by the results of §3.6.2 and the labels are assigned by us based on the parameter values predicted by GaMPEN. The labels are applied based on the predicted values of GaMPEN because we will not have access to true values of the parameters when applying GaMPEN to previously unanalyzed real galaxies. This is crucial given that when we apply GaMPEN to real data, techniques like this will provide us practical tools to deal with predictions in regions of the parameter-space where we know GaMPEN to be less accurate.

For the bulge-to-total ratio, we retain GaMPEN's numerical predictions for $0.1 < L_B/L_T < 0.85$, but label the more extreme galaxies as "highly bulge-dominated" ($L_B/L_T \geq 0.85$) or "highly disk-

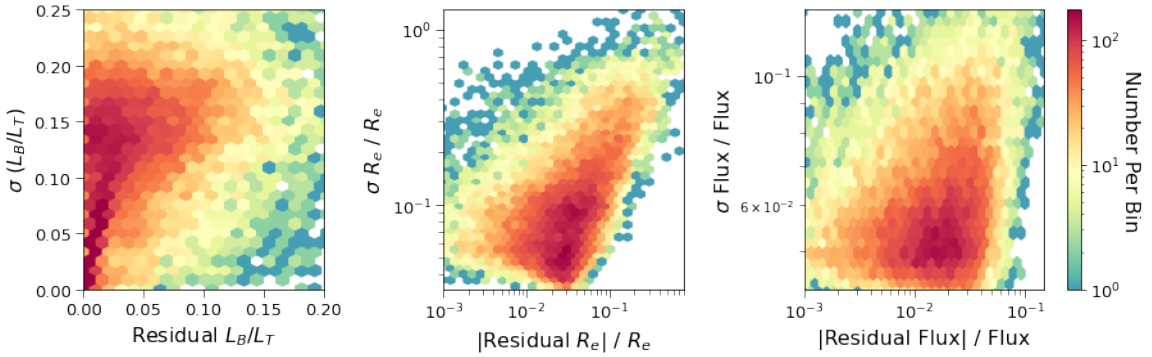


Figure 3.18: Uncertainties (widths of the 68.27% confidence intervals) predicted by GaMPEN for each parameter versus the corresponding residuals (predicted mode minus true value). Fractional uncertainties and residuals are plotted for radius and flux in order to make all the quantities dimensionless. The trend in all three cases is that GaMPEN-estimated uncertainties increase for cases where its predictions are less accurate. The coverage probabilities reported on the test set (Table 3.2) confirm that the predicted uncertainties are well-calibrated and correspond well to the quoted confidence intervals.

dominated" ($L_B/L_T \leq 0.1$). The top left panel of Figure 3.19 shows the two labeled regions (black-shaded grid), which is where the residuals are highest. The right panel of the top row shows the residual distributions including and excluding the extreme cases. As indicated by the standard deviation (top right corner), removing these extreme cases eliminates the largest errors in the predicted values of L_B/L_T. We also checked the accuracy of our assigned labels, and show the confusion matrix in Figure 3.20. From this, we calculate the net accuracy of our extreme L_B/L_T labels to be $\gtrapprox 99\%$.

We apply similar labels to small predicted values of the effective radius. As shown in the bottom row of Figure 3.19, we flag galaxies with $R_e < 1.0$ arcsec with the label "galaxy with $R_e < 1$ arcsec" in place of the exact numerical value. This reduces the typical error for R_e, as shown in the histogram on the right. We calculate the accuracy of this label to be $\sim 97\%$.

Thus, replacing GaMPEN's quantitative predictions in certain small regions of the parameter space with qualitative flags results in a reduction of the typical residuals as well as highly accurate qualitative predictions.

3.7 Discussion & Conclusions

In this work, we introduced the Galaxy Morphology Posterior Estimation Network (GaMPEN), a machine learning framework that can estimate posterior distributions for a galaxy's bulge-to-total light ratio, effective radius, and flux. Although GaMPEN was trained to estimate these specific parameters, it can be adapted by users easily to predict other/additional morphological parameters

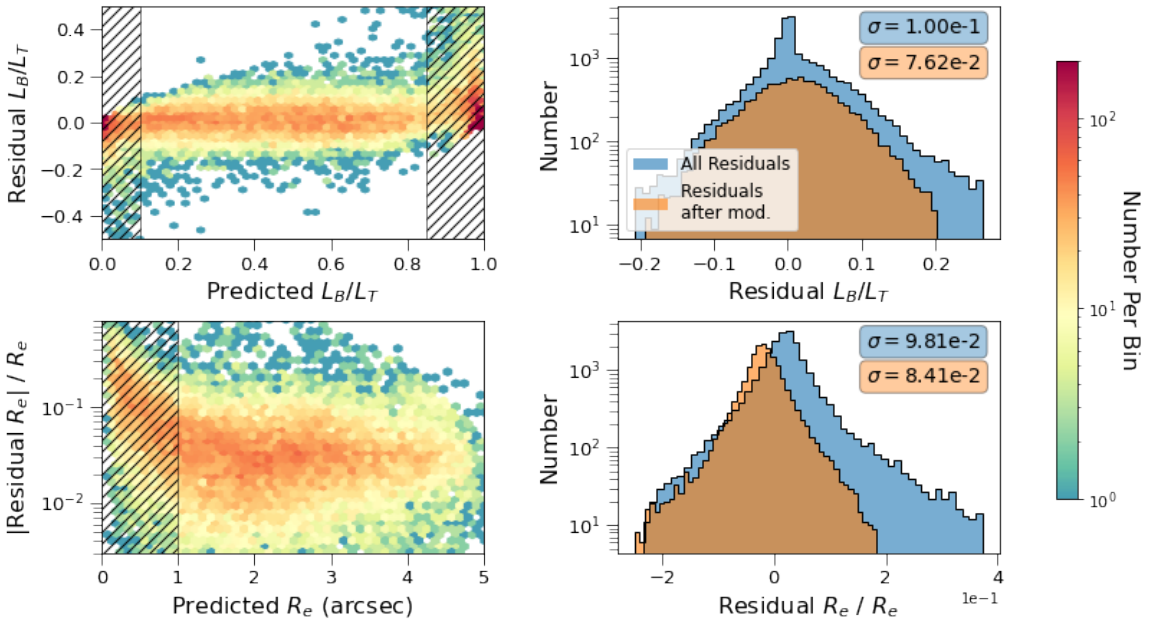


Figure 3.19: The left panels show the residuals for bulge-to-total light ratio and radius plotted against their predicted values. The black dashed regions show the parameter-space where we replace the quantitative predictions with qualitative flags. Each corresponding histogram on the right shows the distribution of residuals before and after the transformation of output values.

(e.g., axis ratio, position angle, etc.). One important consideration while choosing how many parameters to predict using a single GaMPEN framework is that the number of terms in the covariance matrix will increase as $\mathcal{O}(n^2)$, where n is the number of output variables. Although the computation time will increase at a much less steep rate, the exact nature of the increase will depend on the specifics of the hardware being used.

We trained GaMPEN on two NVIDIA Tesla P100/V100 GPUs with each training run taking about $\sim 12-16$ hours. GaMPEN is designed to use multiple GPUs during training and using more GPUs can reduce this training time even further. Our hyperparameter search required ~ 30 runs. Given that we expect $\sim 100,000$ images to always be enough to train GaMPEN, our framework can easily be trained on other datasets within a similar reasonable timescale. Once trained, it takes GaMPEN less than a millisecond to process each input galaxy image. Thus, to predict distributions (~ 1000 inference runs for each image) for a million galaxies on two GPUs, GaMPEN needs ~ 5 days of runtime. Therefore, GaMPEN can be used to process data from future large surveys like LSST, NGRST, and Euclid within a reasonable timescale.

Training and testing GaMPEN on galaxies simulated to match Hyper Suprime-Cam Wide g-band $z < 0.25$ data, we found excellent agreement between the coverage probabilities and the corresponding confidence thresholds (Table 3.2). This demonstrates that GaMPEN predicted posterior

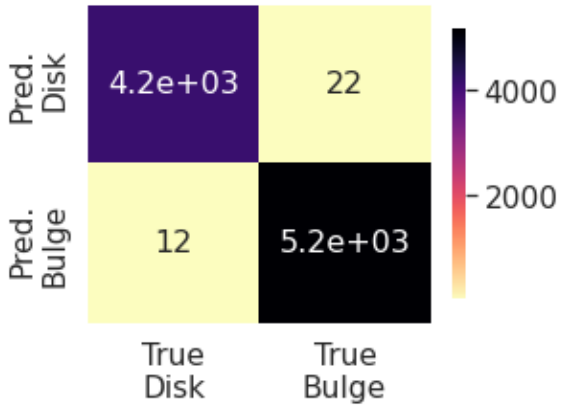


Figure 3.20: Confusion matrix between the labels we assign when GaMPEN predicts extreme bulge-to-total ratios, $L_B/L_T < 0.1$ or > 0.85, and their true L_B/L_T values. The number in each block shows how many galaxies correspond to that panel, resulting in an overall accuracy $> 99\%$.

distributions are calibrated and accurate.

To account for both aleatoric and epistemic uncertainties in GaMPEN predictions, we incorporated the full covariance matrix in our loss function and used the Monte Carlo Dropout technique. Using the covariance matrix also allowed us to incorporate the structured relationships among the output parameters into GaMPEN predictions. This made it possible to achieve the simultaneous calibration of the posteriors of all three output variables (Fig. 3.8). In order to incorporate the covariance matrix in the loss function, we used the Cholesky decomposition and a set of linear algebraic tricks (§3.4.3).

The typical values of errors in GaMPEN predictions are 0.10, 0.17 arcsec ($\sim 7\%$), and 6.3×10^4 nJy ($\sim 1\%$), for L_B/L_T, R_e, and F respectively. The error in GaMPEN predictions of R_e increases when $R_e < 1$ arcsec (i.e., when R_e becomes comparable to the seeing of the HSC-Wide survey) and/or $F < 10^6$ nJy. Galaxies fainter than 10^6 nJy also result in higher flux residuals. These trends result from the inherent challenge in analyzing small and faint galaxies. GaMPEN accounts for these high residuals by correctly predicting higher uncertainties in R_e and F for smaller and fainter galaxies. In other words, GaMPEN predicts broader distributions in regions where it is less precise.

The residuals in GaMPEN predictions of L_B/L_T are high for $L_B/L_T \sim 0$ and ~ 1. We demonstrate that by applying qualitative labels for $0.1 \geq L_B/L_T \geq 0.85$ instead of quantitative values, we can reduce the typical error in L_B/L_T to 0.076. The produced qualitative labels (in the regions with high residuals) have extremely high accuracies of $\gtrapprox 99\%$. Similarly, by labeling predictions for $R_e < 1$ arcsec, we achieve a similar reduction in the typical R_e residual, and the produced labels are highly accurate. Thus, the qualitative transformation of the output values gives us tools to deal

with regions of the parameter space where residuals in GaMPEN predictions are high.

It is difficult to accurately compare GaMPEN's performance to existing morphology parameter estimation pipelines (such as GALFIT (Peng et al., 2002), GIM2D (Simard et al., 2002), or **?**'s neural network) primarily due to the fact that none of them estimate Bayesian posteriors for the predicted parameters. GALFIT and GIM2D do include analytical estimates of errors, but Haussler et al. (2007) found that both these algorithms severely underestimate the true uncertainties by an extremely large factor ($\geq 70\%$ for most galaxies). In contrast, GaMPEN's predicted uncertainties are well-calibrated and accurate ($< 5\%$ deviation). Although GaMPEN's predictions are best used and interpreted in a probabilistic context, we compare below GaMPEN's residuals (assuming the most probable value to be the predicted value) to the residuals achieved by other frameworks.

Meert et al. (2013) used GALFIT to fit two-component light profiles to simulated Sloan Digital Sky Survey (SDSS; York et al., 2000) galaxies and found the typical R_e error to be 10% and the typical magnitude error to be 0.075 mag. GaMPEN achieves a typical R_e error of $\sim 7\%$ and a typical magnitude error of 0.051 mag. Haussler et al. (2007) used single-component fits to analyze simulated galaxies in the Hubble Space Telescope (HST) Galaxy Evolution from Morphology and SEDs (GEMS; Rix et al., 2004) survey using GALFIT and GIM2D. They found the typical error in magnitude to be 0.05 mag using GALFIT and 0.10 mag using GIM2D. They found the typical ratio between the predicted and true R_e to be 0.98 ± 0.06 using GALFIT and 1.01 ± 0.11 using GIM2D. The same value for GaMPEN is 0.98 ± 0.08. **?** used a CNN to obtain predictions for the parameters of a single component Sérsic fit using simulations of HST Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS; Grogin et al., 2011) galaxies and reported the degree of regression accuracy defined as

$$R^2 = 1 - \frac{\sum_i^n (y_i - f_i)^2}{\sum_i^n (y_i - \bar{y})^2} \tag{3.17}$$

where f_i is the predicted value of the true variable y_i and $\bar{y}$ is the mean over all n samples. The R^2 for magnitude and R_e were reported to be 0.997 and 0.972 respectively. The authors also analyzed the same galaxies using GALFIT and found the corresponding R^2 to be 0.983 and 0.877. The R^2 achieved by GaMPEN for magnitude and R_e are 0.998 and 0.980 respectively. While Haussler et al. (2007) and **?** did not have estimates of L_B/L_T (as they used single-component fits), Meert et al. (2013) did not report their residuals for L_B/L_T. Thus, it was not possible to compare GaMPEN's L_B/L_T residuals with these previous works. Although none of the above represent absolutely equivalent comparisons, they indicate that GaMPEN's prediction accuracy is comparable

to the most popular state-of-the-art morphology prediction tools.

GaMPEN contains a Spatial Transformer Network that enables it to crop the input image automatically. We demonstrated that GaMPEN does this based on galaxy size, without any need for specific instruction. Because the transformation is differentiable, loss gradients can be backpropagated, and thus the STN can be trained along with the rest of the framework without any additional supervision. The STN in GaMPEN will empower us to apply it to future large datasets over a broad range of redshifts without having to worry about optimal cutout sizes.

Although in recent years there has been a significant increase in the use of CNNs for morphological determination, GaMPEN is the first machine learning framework that can robustly estimate posterior distributions of multiple morphological parameters. GaMPEN is also the first application of an STN to optical imaging in astronomy.

By testing GaMPEN on simulated HSC g-band galaxies, where we have access to robust ground-truth values, we demonstrated its effectiveness in recovering morphological parameters and we quantified errors/uncertainties in GaMPEN predictions across different regions of the parameter space.

Note that, for this work, we trained and tested GaMPEN on single-band images. However, we have tested and verified that both the CNN and STN in GaMPEN can be easily adapted to intake an arbitrary number of channels, with each channel being a different band. To obtain separate morphological parameters for each band, the number of output parameters would need to be increased appropriately. We will perform a detailed evaluation of GaMPEN's performance on multi-band images in future work.

In this work, we performed a thorough analysis of GaMPEN's performance using HSC $z < 0.25$ simulations. However, GaMPEN can be applied to a wide variety of other datasets – including real HSC images, imaging from other ground and space-based observatories as well as higher redshift data. However, in order to apply GaMPEN to real data, one would need to perform appropriate transfer-learning (i.e., fine-tuning the simulation trained GaMPEN models using a small amount of data from the application dataset). We refer an interested reader to **?**, where we performed transfer-learning and demonstrated the application of our classification framework, GaMorNet, to SDSS $z \sim 0$ and CANDELS $z \sim 1$ data.

Just like other image analysis methods, we expect GaMPEN's performance to change based on the quality of the images being used (e.g., the pixel-scale of the survey, the noise, the redshift of the object). We will explore how GaMPEN performs in each of these above situations in future work.

Chapter Acknowledgments

The authors would like to thank the anonymous referee for their insightful, encouraging, and extremely thorough comments about our manuscript. Their constructive criticism has greatly assisted us in improving the manuscript, extending the discussion section, and making it more accessible to readers.

This material is based upon work supported by the National Science Foundation under Grant No. 1715512

CMU and AG would like to acknowledge support from the National Aeronautics and Space Administration via ADAP Grant 80NSSC18K0418.

AG would like to acknowledge support received from the Yale Graduate School of Arts & Sciences through the Dean's Emerging Scholars Research Award.

AG would like to acknowledge computing grants received through the Amazon Cloud Credits for Research Program and the Yale Center for Research Computing (YCRC) Resarch Credits Program. AG would also like to acknowledge computing support from YCRC and Yale Information Technology Services staff members and scientists.

ET acknowledges support from FONDECYT Regular 1190818 and 1200495, ANID grants CATA-Basal AFB-170002, ACE210002, and FB210003, and Millennium Nucleus NCN19_058.

The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

This paper makes use of software developed for the Large Synoptic Survey Telescope. We thank the LSST Project for making their code available as free software at http://dm.lsst.org.

The Pan-STARRS1 Surveys (PS1) have been made possible through contributions of the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the

Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation under Grant No. AST-1238877, the University of Maryland, and Eotvos Lorand University (ELTE) and the Los Alamos National Laboratory.

Based, in part, on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center at National Astronomical Observatory of Japan.

GaMPEN Appendix

3.A Early Data Access

Currently, we are applying GaMPEN to real data, and will make the source code public in the Fall of 2022, along with documentation, tutorials, and trained models. Readers interested in using GaMPEN before the full public release can access the source code and trained models of GaMPEN by emailing the corresponding author of this paper.

The public data release for GaMPEN will be hosted at the following two locations:

- http://www.ghosharitra.com/
- http://www.astro.yale.edu/aghosh/

3.B Extended Derivation for Bayesian Implementation of GaMPEN

In variational inference, the posterior, $p(\boldsymbol{\omega} \mid \mathcal{D})$ in Equation 3.6, is replaced by an approximate variational distribution with an analytic form $q(\boldsymbol{\omega})$. Now, Equation 3.6 can be written as

$$p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}) \approx \int p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \boldsymbol{\omega}) q(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{3.18}$$

The choice of the variational distribution is arbitrary. One such choice, introduced by Gal & Ghahramani (2016), involves dropping different neurons from some layers in order to assess the impact on the model. The dropout technique was introduced by **?** in order to prevent neural networks from overfitting; they temporarily removed random neurons from the network according to a Bernoulli distribution, i.e., individual nodes were set to zero with a probability, p, known as the dropout rate.

In the variational application, we use dropouts to interrogate the model. Specifically, if p_i is the probability of a neuron being turned off, and $[z_{i,j}]_{j=1}^{J_{i-1}}$ is a vector of length J_{i-1} containing the Bernoulli-distributed random variables for unit $j = 1, \ldots, J_{i-1}$ in the $(i-1)^{th}$ layer with probabilities

p_i, then

$$\boldsymbol{\omega}_i = \boldsymbol{M}_i \cdot \text{diag}\left(\left[z_{i,j}\right]_{j=1}^{J_{i-1}}\right), \tag{3.19}$$

where $\boldsymbol{M}_i$ is the $J_i \times J_{i-1}$ matrix of variational parameters to be optimized.

Thus, sampling from $q(\boldsymbol{\omega})$ is now equivalent to using dropouts on a set of layers, with weights $\boldsymbol{M}$ (i.e., $\boldsymbol{M}_i$ for the i^{th} layer). We perform inference on the trained network by approximating Equation 3.18 with a Monte Carlo integration:

$$\int p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \boldsymbol{\omega}) q(\boldsymbol{\omega}) d\boldsymbol{\omega} \approx \frac{1}{T} \sum_{t=1}^{T} p(\hat{\boldsymbol{Y}} \mid \hat{\boldsymbol{X}}, \boldsymbol{\omega}_t), \tag{3.20}$$

wherein we perform T forward passes with dropout enabled and $\boldsymbol{\omega}_t$ is the set of weights during the t^{th} forward pass.

3.C Extended Derivation of the Loss Function

As outlined in Equation 3.6, we seek the most likely set of model parameters given our training data, i.e., we maximize

$$p(\boldsymbol{\omega} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega}). \tag{3.21}$$

In Equation 3.21, $p(\boldsymbol{\omega})$ is the prior on the neural networks weights. The weight prior here is unimportant and what matters is the prior induced on the output parameters of GaMPEN. And as outlined above, we use an uninformative multivariate Gaussian prior to induce an uninformative prior on the output. Please refer to Wilson (2020) for a detained discussion on priors in Bayesian deep learning.

For a regression task using a standard CNN, wherein the network outputs predictions $\hat{\boldsymbol{Y}}_n\left(\hat{\boldsymbol{X}}_n, \boldsymbol{\omega}\right)$ for true values $\boldsymbol{Y}_n$, one popular choice is to minimize the squared-error loss function $\sum_n \left\| \boldsymbol{Y}_n - \hat{\boldsymbol{Y}}_n\left(\hat{\boldsymbol{X}}_n, \boldsymbol{\omega}\right)\right\|^2$, where the sum over n denotes a sum over the training set. However, in contrast to the traditional approach, for each new test image $\hat{\boldsymbol{X}}$, GaMPEN needs to predict the parameters of a multivariate Gaussian distribution, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Now, as discussed in §3.4.1, we replace $p(\boldsymbol{\omega} \mid \mathcal{D})$ in Equation 3.6 with an approximating variational distribution $q(\boldsymbol{\omega})$. This is performed by minimizing their Kullback-Leibler (KL) divergence, a measure of the similarity between two distributions. Since minimizing the KL divergence is equivalent to maximizing the log-evidence lower bound,

$$\log \mathcal{L}_{\mathrm{VI}} = \int q(\omega) \log p(\left\{ \boldsymbol{Y}_{n=1}^{N} \right\} \mid \left\{ \boldsymbol{X}_{n=1}^{N} \right\}, \boldsymbol{\omega}) d\boldsymbol{\omega} - \mathrm{KL}(q(\omega) \| p(\omega)). \tag{3.22}$$

The first term in Equation 3.22 is the log-likelihood for the output parameters for the training set, and as shown in Gal & Ghahramani (2016), the KL term can be approximated as an L_2 regularization. Therefore, Equation 3.22 can be written as

$$\log \mathcal{L}_{\mathrm{VI}} \sim \sum_{n=1}^{N} \log \mathcal{L} \left(\boldsymbol{Y}_n, \hat{\boldsymbol{Y}}_n \left(\boldsymbol{X}_n, \boldsymbol{\omega} \right) \right) - \lambda \sum_i \left\| \boldsymbol{\omega_i} \right\|^2, \tag{3.23}$$

where $\log \mathcal{L} \left(\boldsymbol{Y}_n, \hat{\boldsymbol{Y}}_n \left(\boldsymbol{X}_n, \boldsymbol{\omega} \right) \right)$ is the log-likelihood of the network predictions $\hat{\boldsymbol{Y}}_n \left(\boldsymbol{X}_n, \boldsymbol{\omega} \right)$ for training input $\boldsymbol{X}_n$ with true values $\boldsymbol{Y}_n$, λ is the strength of the regularization term, and $\boldsymbol{\omega}_i$ are sampled from $q(\boldsymbol{\omega})$.

For the multivariate Gaussian distribution predicted by GaMPEN, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can write the log-likelihood of the network predictions (first-term on the right side in Equation 3.23) as

$$\log \mathcal{L} \propto \sum_n -\frac{1}{2} \left[\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_n \right]^\top \hat{\boldsymbol{\Sigma}}_n^{-1} \left[\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_n \right] - \frac{1}{2} \log[\det(\hat{\boldsymbol{\Sigma}}_n)], \tag{3.24}$$

where $\hat{\boldsymbol{\mu}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ are the mean and covariance matrix of the multivariate Gaussian distribution predicted by GaMPEN for an image, $\boldsymbol{X}_n$.

We train GaMPEN by minimizing the negative log-likelihood of the output parameters for the training set, which by combining Eqs. 3.23 and 3.24, can be written as

$$-\log \mathcal{L}_{VI} \propto \sum_n \frac{1}{2} \left[\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_n \right]^\top \hat{\boldsymbol{\Sigma}_n}^{-1} \left[\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_n \right] + \frac{1}{2} \log[\det(\hat{\boldsymbol{\Sigma}_n})] + \lambda \sum_i \left\| \boldsymbol{\omega_i} \right\|^2. \tag{3.25}$$

3.D Additional Technical Details on GaMPEN

In Table 3.3, we have outlined the various layers of the GaMPEN framework along with the important parameters of each layer and the corresponding activation functions.

Table 3.3: Structure of GaMPEN

Order	Type of Layer	Layer Description	Activation Function
		Upstream Spatial Transformer Network	

1	Input	Size: 239×239	–
2	Convolutional	Filters: 64 | Size: 11	ReLU[a]
3	Max-Pooling	Kernel Size: 3 | Strides: 2	–
4	Convolutional	Filters: 96 | Size: 9	ReLU
5	Max-Pooling	Kernel Size: 3 | Strides: 2	–
6	Fully Connected	No. of neurons: 32	ReLU
7	Fully Connected	No. of neurons: 1	Linear

Downstream Morphological Estimation Network			
1	Input	Size: 239×239	–
2	Convolutional	Filters: 64 | Size: 3 | Strides: 1	ReLU
3	Dropout	-	-
4	Convolutional	Filters: 64 | Size: 3 | Strides: 1	ReLU
5	Max-Pooling	Kernel Size: 2 | Strides: 2	–
6	Dropout	-	-
7	Convolutional	Filters: 128 | Size: 3 | Strides: 1	ReLU
8	Dropout	-	-
9	Convolutional	Filters: 128 | Size: 3 | Strides: 1	ReLU
10	Max-Pooling	Kernel Size: 2 | Strides: 2	–
11	Dropout	-	-
12	Convolutional	Filters: 256 | Size: 3 | Strides: 1	ReLU
13	Dropout	-	-
14	Convolutional	Filters: 256 | Size: 3 | Strides: 1	ReLU
15	Dropout	-	-
16	Convolutional	Filters: 256 | Size: 3 | Strides: 1	ReLU
17	Max-Pooling	Kernel Size: 2 | Strides: 2	–
18	Dropout	-	-
19	Convolutional	Filters: 512 | Size: 3 | Strides: 1	ReLU
20	Dropout	-	-
21	Convolutional	Filters: 512 | Size: 3 | Strides: 1	ReLU
22	Dropout	-	-

23	Convolutional	Filters: 512 | Size: 3 | Strides: 1	ReLU
24	Max-Pooling	Kernel Size: 2 | Strides: 2	–
25	Dropout	-	-
26	Convolutional	Filters: 512 | Size: 3 | Strides: 1	ReLU
27	Dropout	-	-
28	Convolutional	Filters: 512 | Size: 3 | Strides: 1	ReLU
29	Dropout	-	-
30	Convolutional	Filters: 512 | Size: 3 | Strides: 1	ReLU
31	Max-Pooling	Kernel Size: 2 | Strides: 2	–
32	Fully Connected	No. of neurons: 4096	ReLU
33	Dropout	-	-
34	Fully Connected	No. of neurons: 4096	ReLU
35	Dropout	-	-
36	Fully Connected	No. of neurons: 9	Linear

[a] Rectified Linear Unit

NOTE- The dropout rate of the various layers are set according to the calibration step described in §3.5

Chapter 4

Morphological Parameters and Associated Uncertainties for 8 Million Galaxies in the Hyper Suprime-Cam Wide Survey

Accepted for publication by the American Astronomical Society in *The Astrophysical Journal* | Currently in press | Also available on arXiv: 2212.00051

Aritra Ghosh, C. Megan Urry, Aayush Mishra, Laurence Perreault-Levasseur, Priyamvada Nataranjan, David B. Sanders, Daisuke Nagai, Chuan Tian, Nico Cappelluti, Jeyhan S. Kartaltepe, Meredith C. Powell, Amrit Rau, and Ezequiel Treister

We use the Galaxy Morphology Posterior Estimation Network (GaMPEN) to estimate morphological parameters and associated uncertainties for ~ 8 million galaxies in the Hyper Suprime-Cam (HSC) Wide survey with $z \leq 0.75$ and $m \leq 23$. GaMPEN is a machine learning framework that estimates Bayesian posteriors for a galaxy's bulge-to-total light ratio (L_B/L_T), effective radius (R_e), and flux (F). By first training on simulations of galaxies and then applying transfer learning using real data, we trained GaMPEN with $< 1\%$ of our data-set. This two-step process will be critical for applying machine learning algorithms to future large imaging surveys, such as the Rubin-Legacy Survey of Space and Time (LSST), the Nancy Grace Roman Space Telescope (NGRST), and Euclid. By comparing our results to those obtained using light-profile fitting, we demonstrate that GaMPEN's predicted posterior distributions are well-calibrated ($\lesssim 5\%$ deviation) and accurate. This represents a significant improvement over light profile fitting algorithms which underestimate uncertainties by as much as $\sim 60\%$. For an overlapping sub-sample, we also compare the derived morphological parameters with values in two external catalogs and find that the results agree within the limits of uncertainties predicted by GaMPEN. This step also permits us to define an empirical relationship between the Sérsic index and L_B/L_T that can be used to convert between these two parameters. The catalog presented here represents a significant improvement in size ($\sim 10\times$), depth

($\sim$ 4 magnitudes), and uncertainty quantification over previous state-of-the-art bulge+disk decomposition catalogs. With this work, we also release GaMPEN's source code and trained models, which can be adapted to other data sets.

4.1 Introduction

The morphology of galaxies has been shown to be related to various other fundamental properties of galaxies and their environment, including galaxy mass, star formation rate, stellar kinematics, merger history, cosmic environment, and the influence of supermassive black holes (e.g., Powell et al., 2017; Pozzetti et al., 2010; Shimakawa et al., 2021; Wuyts et al., 2011; **?**; **?**; **?**; **?**). Therefore, quantitative measures of the morphological parameters for large samples of galaxies at different redshifts are of fundamental importance in understanding the physics of galaxy formation and evolution.

Distributions of morphological quantities alone can place powerful constraints on possible galaxy formation scenarios. And when combined with other physical quantities, they can provide key insights into evolutionary processes at play or even reveal the role of new physical mechanisms that impact evolution (e.g., **?????**). However, such studies often involve subtle correlations or hidden variables within strong correlations that demand greater statistics and measurement precision than what has been available in the preceding decades. An often-overlooked factor in such studies has been the computation of robust uncertainties. The computation of full Bayesian posteriors for different morphological parameters is crucial for drawing scientific inferences that account for uncertainty and, thus are indispensable in the derivation of robust scaling relations (e.g., **??**) or tests of theoretical models using morphology (e.g., **?**).

A quantitative description of galaxy morphology is typically expressed in terms of its structural parameters — brightness, shape, and size — all of which can be determined by fitting a single two dimensional analytic light profile to the galaxy image (e.g., Tarsitano et al., 2018; Wel et al., 2012). However, moving beyond single-component determinations by using separate components to analyze galaxy sub-structure (e.g., disk, bulge, bar, etc.) can provide us additional insights into the formation mechanisms of these components: bulges, disks, and bars may be formed as a result of secular evolution (e.g., Genzel et al., 2008; Kormendy, 1979; Kormendy & Kennicutt, 2004; Sellwood, 2014) or due to the interaction of disk instabilities with smooth and clumpy cold streams (e.g., Dekel et al., 2009a,b). In this sense, contrary to what is often expected, bulges can also be formed without major galaxy mergers.

Over the last decade, machine learning (ML) has been increasingly used for a wide variety of

tasks—from identifying exoplanets to studying black holes (e.g., Hoyle, 2016; Kim & Brunner, 2017; **?**; **?**; **?**). Unsurprisingly, these algorithms have become increasingly popular for determining galaxy morphology as well. (e.g., Ghosh et al., 2020; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**). The use of these techniques has been driven by the fact that traditional methods of analyzing morphologies—visual classification and template fitting—are not scalable to the data volume expected from future surveys such as the Vera Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezic et al., 2019), the Nancy Grace Roman Space Telescope (NGRST; Spergel et al., 2013), and Euclid (Racca et al., 2016).

Most previous applications of ML to galaxy morphology produced broad, qualitative classifications rather than numerical estimates of morphological parameters. **?** did estimate parameters of single-component Sérsic fits. However, they did not analyze galaxy sub-structures or provide uncertainties. In order to address these challenges, in Ghosh et al. (2022), we introduced The Galaxy Morphology Posterior Estimation Network (GaMPEN). GaMPEN is a machine learning framework that estimates full Bayesian posteriors for a galaxy's bulge-to-total light ratio (L_B/L_T), effective radius (R_e), and flux (F). GaMPEN takes into account covariances between the different parameters in its predictions and has been shown to produce calibrated, accurate posteriors.

GaMPEN can also automatically crop input galaxy cutouts to an optimal size before determining their morphology. This feature is critical given that morphology-determination ML frameworks typically require input cutouts of a fixed size, and thus, cutouts of a "typical" size often contain secondary objects in the frame. By cropping out most secondary objects in the frame, GaMPEN can make more accurate predictions over wide ranges of redshift and magnitude.

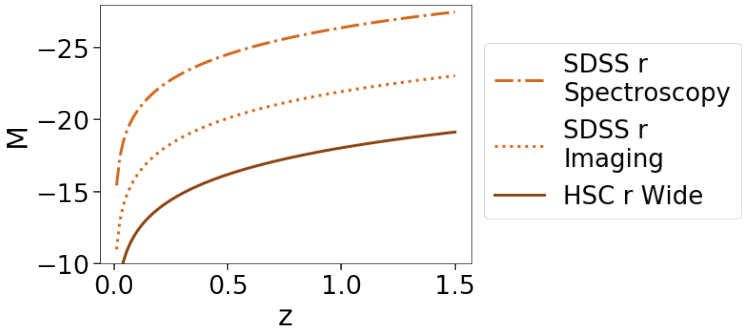


Figure 4.1: The limiting absolute magnitudes probed by the Hyper Suprime-Cam (HSC) Wide Survey and Sloan Digital Sky Survey (SDSS) at different redshifts.

In this paper, we use GaMPEN to estimate Bayesian posteriors for L_B/L_T, R_e, and F for $\sim$ 8 million galaxies with $z \leq 0.75$ from the Hyper Suprime-Cam (HSC) Wide survey (Aihara et al., 2018b). A few recent works have studied the morphology of smaller subsets of HSC galaxies:

Shimakawa et al. (2021) classified $\sim 2\times10^5$ massive HSC galaxies into spiral and non-spiral galaxies, while Kawinwanichakij et al. (2021) analyzed $\sim 1.5\times10^6$ HSC galaxies using single-component Sérsic fits. To date, the state-of-the-art morphological catalog that provided bulge+disk decomposition parameters at low redshift has been that of Simard et al. (2011), which used Sloan Digital Sky Survey (SDSS; York et al., 2000) imaging to estimate morphological parameters of $\sim$ 1 million $m < 18$ galaxies, most with $z < 0.2$. As Figure 4.1 shows, HSC imaging allows us to probe much fainter magnitudes than SDSS with significantly better seeing (0.″85 for HSC-W g compared to 1.″4 for SDSS g). The catalog presented in this paper builds on Simard et al. (2011) by providing an order of magnitude increase in sample size and probing four magnitudes deeper, to a higher redshift threshold. Along with estimates of parameters, this catalog also estimates robust uncertainties of the predicated parameters, which have typically been absent from previous large morphological catalogs. This catalog represents a significant step forward in our capability to quantify the shapes and sizes of galaxies in our universe.

This paper also demonstrates that ML techniques can be used to study morphology in new surveys, which do not have already-classified large training sets available. Most previous works involving ML to study galaxy morphology have depended on the availability of an extensive training set of real galaxies with known properties from the same survey or a similar one. However, if Convolutional Neural Networks (CNNs) are to replace traditional methods for morphological analysis, we must be able to use them on new surveys that do not have a morphological catalog readily available to be used as a training set. In this paper, we demonstrate that by first training on simulations of galaxies and then using real data for transfer learning (fine-tuning the simulation-trained network), we can fully train GaMPEN, while having to label $< 1\%$ of our dataset. This work outlines an easy pathway to apply morphological ML techniques to upcoming large imaging surveys.

In §4.2, we describe the HSC data used in this study, along with the simulated two-dimensional light profiles. §4.3 and §4.4 provide a brief introduction to GaMPEN and how we train it. In §4.5, we outline how we determined the morphological parameters of $\sim 60,000$ HSC-Wide galaxies for transfer learning and validating the results of our ML framework. §4.6 provides detailed information on the accuracy of GaMPEN's predictions. §4.7 compares our predictions to that of other catalogs. We end with a summary and discussion of our results in §4.8.

The full Bayesian posteriors for all $\sim$ 8 million galaxies are being released with the publication of this work. We are also releasing the source code of GaMPEN, along with the trained models and extensive documentation and tutorials on how to use GaMPEN. The public data release is described in Appendix 4.A.

4.2 Data

4.2.1 Hyper Suprime-Cam Data

We apply GaMPEN to g,r,i-band data from the Hyper Suprime-Cam (HSC) Subaru Strategic Program Public Data Release 2 (PDR2; Aihara et al., 2019). The Subaru Strategic Program, ongoing since 2014, uses the HSC prime-focus camera, which provides extremely high sensitivity and resolving power due to the large 8.2 meter mirror of the Subaru Telescope.

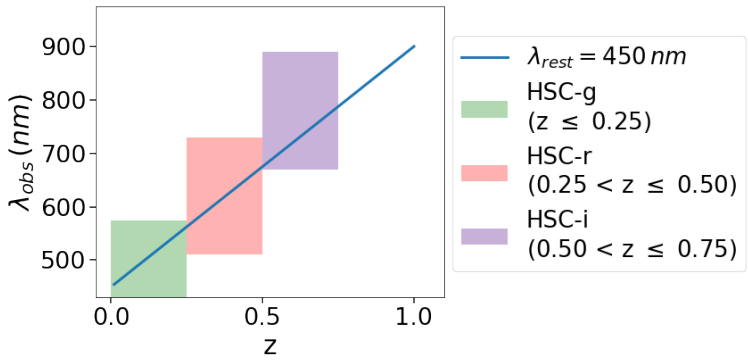


Figure 4.2: The filter used for each redshift bin is shown along with the wavelength range sampled by each filter. The blue line shows where rest-frame 450 nm emission falls for redshifts labeled on the x-axis. As this figure shows, the chosen filters allow us to consistently perform morphology determination in the rest-frame g-band.

In order to have a large but uniform sample of HSC PDR2 galaxies, we focus on the largest volume HSC survey, namely, the Wide layer, which covers 1400 deg^2 to the nominal survey depth in all filters and contains over 450 million primary objects. To consistently perform morphology determination in the rest-frame g-band across our entire sample, we use different filters for galaxies at different redshifts, as shown in Figure 4.2. We use g-band for $z \leq 0.25$, r-band for $0.25 < z \leq 0.50$, and i-band for $0.50 < z \leq 0.75$. Given HSC's typical seeing, all objects with sizes $\lesssim 5$ kpc cannot be resolved beyond $z = 0.75$. Therefore, given that the HSC-Wide light profiles of the large majority of galaxies beyond this redshift will be dominated by the PSF, we restrict this work till $z = 0.75$ and will explore the application of GaMPEN to higher redshift deeper HSC data in future work.

In order to select galaxies, we used the PDR2 galaxy catalog produced using forced photometry on coadded images. The HSC data has been reduced using a pipeline built on the prototype pipeline developed by the Rubin Observatory's Large Synoptic Survey Telescope's Data Management system, and we refer the interested reader to Bosch et al. (2018) for more details. We use the `extendedness_value` flag only to select extended sources. The `extendedness_value` flag relies on the difference between the Composite Model (CModel) and PSF magnitudes to separate galaxies

from stars, and contamination (from stars) increases sharply for $m > 23$ for median HSC seeing in the Wide layer as outlined here[1]. Thus, for each redshift bin, we select galaxies with magnitude $m < 23$ (in the appropriate band). The full query to download the data in each redshift bin is available in Appendix 4.B.

Table 4.1: Data Characteristics

Sample	Redshift	Number	Imaging	Spectra
Low-z	$z \leq 0.25$	820,566	g	7.4%
Mid-z	$0.25 < z \leq 0.50$	2,937,871	r	2.4%
High-z	$0.50 < z \leq 0.75$	4,247,208	i	1.6%

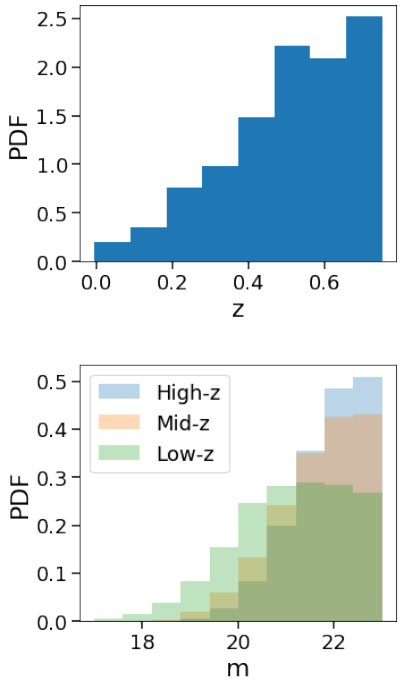


Figure 4.3: Redshift (*top*) and magnitude (*bottom*) distributions for the $\sim$ 8 million galaxies used in this study. We used spectroscopic redshifts when available and high-quality photometric redshifts otherwise. The spectroscopic completeness of each sub-sample is shown in Table 4.1.

We use spectroscopic redshifts when available ($\sim$ 2.5%) and high-quality photometric redshifts otherwise. The spectroscopic redshifts were collated by the HSC Subaru Strategic Program (SSP) team from a wide collection of spectroscopic surveys – zCOSMOS DR3 (Lilly et al., 2009), UDSz (Bradshaw et al., 2013), 3D-HST (Momcheva et al., 2016), FMOS-COSMOS (Silverman et al., 2015), VVDS (Fevre et al., 2013), VIPERS PDR1 (Garilli et al., 2014), SDSS DR12 (Alam et al., 2015), GAMA DR2 (Liske et al., 2015), WiggleZ DR1 (Drinkwater et al., 2010), DEEP2 DR4

1. https://hsc-release.mtk.nao.ac.jp/doc/index.php/stargalaxy-separation-2/

(Newman et al., 2013), PRIMUS DR1 (Cool et al., 2013), and VANDELS DR2 (Pentericci et al., 2018). The photometric redshifts in HSC PDR2 were calibrated based on the above spectroscopic sample and were calculated using the Direct Empirical Photometric code (DEmP; Hsieh & Yee 2014), an empirical quadratic polynomial photometric redshift fitting code, and Mizuki (Tanaka, 2015), a Bayesian template fitting code. For an extended description, we refer the interested reader to Nishizawa et al. (2020). To remove galaxies with unsecure photometric redshifts, we set the `photoz_risk_best` parameter < 0.1, which controls the risk of the photometric redshift being outside the range $z_{\rm true} \pm 0.15(1 + z_{\rm true})$, with 0 being extremely safe to 1 being extremely risky. Using the `cleanflags_any` flag, we further excluded objects flagged to have any significant imaging issues (in the relevant band) by the HSC pipeline. This flag can be triggered by a wide range of issues; however, for $\sim 80\%$ of cases, the flag was triggered by cosmic ray hits, as shown in Appendix 4.B. We checked and confirmed that none of the above cuts significantly modified the redshift or magnitude distribution of our galaxy sample.

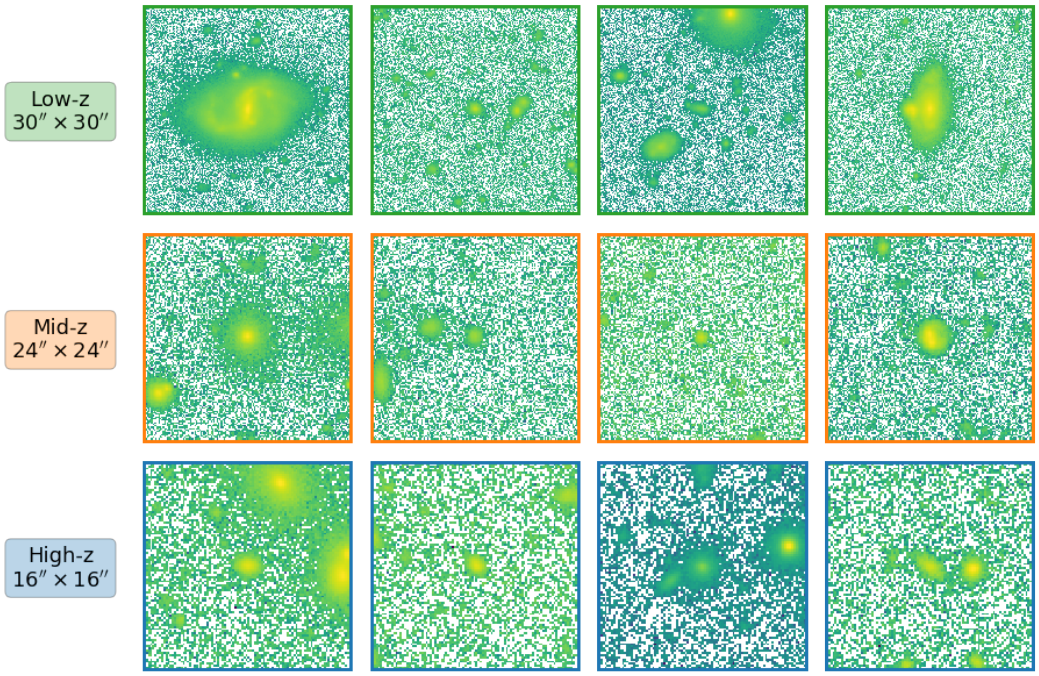


Figure 4.4: Four randomly chosen galaxy cutouts are shown here for each redshift bin, with the object of interest at the center of each cutout. Note that most of these cutouts have secondary objects in the frame, which can often cause ML algorithms to produce spurious classifications. GaMPEN uses a Spatial Transformer Network to crop most secondary objects out of the frame (see §3).

The above process resulted in the selection of ~ 8 million galaxies, with ~ 1 million, ~ 3 million, ~ 4 million galaxies in the low-z, mid-z, and high-z bins, as shown in Table 4.1. The magnitude

and redshift distribution of the data is shown in Figure 4.3. Using the HSC Image Cutout Service[2], we downloaded cutouts for each galaxy with sizes of $30''$, $24''$, $16''$ for the low-, mid-, and high-z bins, respectively. These sizes are large enough that they should capture all objects in the relevant bin. Using the results of our light-profile fitting analysis on a sub-sample, as outlined in §4.5, we expect $\sim 99.5\%$ of our sample to have a downloaded cutout size $\geq 10\times$(size of the major axis of the galaxy). Figure 4.4 shows randomly chosen examples of galaxy cutouts for each redshift bin.

4.2.2 Simulated Galaxies for Initial Training

As outlined later in §4.4.2, we perform GaMPEN's initial round of training on mock galaxy image cutouts, simulated to match HSC observations in the appropriate band. To generate mock images, we used GalSim (**?**), the modular galaxy image simulation toolkit. GalSim has been extensively tested and shown to yield very accurate rendered images of galaxies. We simulated 150,000 galaxies in each redshift bin, with a mixture of both single and double-component galaxies, in order to have a diverse training sample. To be exact, 75% of the simulated galaxies consisted of both bulge and disk components, while the remaining 25% had either a single disk or a bulge.

For both the bulge and disk components, we used the Sérsic profile, and the parameters required to generate the Sérsic profiles were drawn from uniform distributions over ranges given in Table 4.2. For the disk and bulge components, we allow the Sérsic index to vary between $0.8-1.2$ and $3.5-5.0$, respectively. We chose to have varying Sérsic indices as opposed to fixed values for each component in order to have a training set with diverse light profiles. Note that the single-component galaxies were included in the simulations to have some examples of galaxies that are purely disk-dominated (i.e., no bulge component) and some that are purely bulge-dominated (i.e., no disk component). Thus, the Sérsic indices chosen for the single-component galaxies mirror the values chosen for the disk and bulge components in the double-component galaxies.

The parameter ranges for fluxes, and half-light radii are quite expansive and are representative of most galaxies at the appropriate redshift range. To obtain these parameter ranges, we first start with a set of parameters that represent most local galaxies (Binney & Merrifield, 1998). Thereafter, we redshift these parameters for each galaxy using Planck18 cosmology ($H_0 = 67.7$ km/s/Mpc, Aghanim et al., 2018) and the appropriate pixel scale.

To make the two-dimensional light profiles generated by GalSim realistic, we convolved these with representative point-spread functions (PSFs) downloaded from the HSC survey. We also added

2. https://hsc-release.mtk.nao.ac.jp/das_cutout/pdr2/

Table 4.2: Parameter Ranges of Simulated Galaxies

Sample Name	Component Name	Sérsic Index	Half-Light Radius (arcsec)	Flux (ADUs)	Axis Ratio	Position Angle (degrees)
Low-z			Single-Component Galaxies			
		0.8 - 1.2 or 3.5 - 5.0[a]	0.1 - 5.0	30 - 1.35×10^5 ($m_g \sim 14$ - 23)	0.25 - 1.0	-90.0 - 90.0
			Double-Component Galaxies			
	Disk	0.8 - 1.2	0.1 - 5.0	0.0 - 1.0[b]	0.25 - 1.0	-90.0 - 90.0
	Bulge	3.5 - 5.0	0.1 - 3.0	1.0 - Disk[b]	0.25 - 1.0	Disk $\pm [0, 15]$[c]
Mid-z			Single-Component Galaxies			
		0.8 - 1.2 or 3.5 - 5.0[a]	0.1 - 3.0	30 - 3×10^4 ($m_r \sim 15.5$ - 23)	0.25 - 1.0	-90.0 - 90.0
			Double-Component Galaxies			
	Disk	0.8 - 1.2	0.1 - 3.0	0.0 - 1.0[b]	0.25 - 1.0	-90.0 - 90.0
	Bulge	3.5 - 5.0	0.1 - 2.0	1.0 - Disk[b]	0.25 - 1.0	Disk $\pm [0, 15]$[c]
High-z			Single-Component Galaxies			
		0.8 - 1.2 or 3.5 - 5.0[a]	0.1 - 2.0	30 - 2×10^4 ($m_i \sim 16$ - 23)	0.25 - 1.0	-90.0 - 90.0
			Double-Component Galaxies			
	Disk	0.8 - 1.2	0.1 - 2.0	0.0 - 1.0[b]	0.25 - 1.0	-90.0 - 90.0
	Bulge	3.5 - 5.0	0.1 - 1.5	1.0 - Disk[b]	0.25 - 1.0	Disk $\pm [0, 15]$[c]

[a] The single-component galaxies are equally divided between galaxies with a Sérsic index between 0.8 - 1.2 and galaxies with a Sérsic index between 3.5 - 5.0.

[b] Fractional fluxes are noted here. The bulge flux fraction is chosen such that for each simulated galaxy it is added with the disk flux fraction to give 1.0. The total flux of the galaxies is varied between the values given in the top-row for each sample

[c] The bulge position angle differs from the disk position angle by a randomly chosen value between -15 and $+15$ degrees.

`NOTE-` The above table shows the ranges of the various Sérsic profile parameters used to simulate mock HSC cutouts. 75% of the simulated galaxies have both disk and bulge components, and the remainder has either a disk or a bulge component. All the simulation parameters are drawn from uniform distributions.

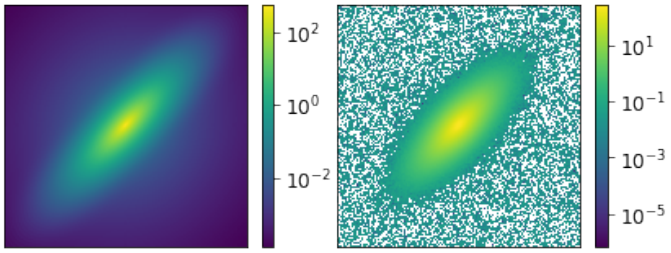


Figure 4.5: Two stages of simulating an HSC galaxy. (*Left*): A randomly chosen two-dimensional light profile generated by GalSim. (*Right*): The same image after PSF convolution and noise addition. The white pixels represent (small) negative values that arise from the process of noise addition.

realistic noise using one-thousand $2'' \times 2''$ "sky objects" from the HSC PDR2 Wide field. Sky objects are empty regions identified by the HSC pipeline that are outside object footprints and are recommended for being used in blank-sky measurements. For PSF convolution and noise addition, we follow the procedure detailed in Ghosh et al. (2022) and refer the interested reader to that work for more details. Figure 4.5 shows a randomly chosen simulated light profile and the corresponding image cutout generated after PSF convolution and noise addition. All the simulated galaxy images in each redshift bin were chosen to have cutout sizes equal to their real data counterparts, as outlined in §4.2.1.

We would like to note that even after PSF convolution and noise addition, our simulated galaxies are only semi-realistic and do not account for many specific features seen in real data (e.g., spiral arms, knots, non-classical bulges, etc.). The primary goal of the simulation dataset is to provide a large corpus of images on which the initial training can be done. The second step of fine-tuning GaMPEN on real data (described in §4.4.3) ensures that the framework also learns about the existence of features in the real data that are missed by the simulations.

4.3 Brief Introduction to GaMPEN

The Galaxy Morphology Posterior Estimation Network (GaMPEN; Ghosh et al., 2022) is a novel machine learning framework that can predict posterior distributions for a galaxy's bulge-to-total light ratio (L_B/L_T), effective radius (R_e), and flux (F). In this section, we provide a brief introduction to GaMPEN; however, for a complete understanding of GaMPEN's architecture and how it predicts posteriors, we refer the reader to Appendix 4.C and Ghosh et al. (2022).

The architecture of GaMPEN consists of an upstream Spatial Transformer Network (STN) module followed by a downstream Convolutional Neural Network (CNN) module. The upstream STN is used to automatically crop each input image frame to an optimal size before morphology determina-

tion. Based on the input image, the STN predicts the parameters of an affine transformation which is then applied to the input image. The transformed image is then passed onto the downstream CNN, which estimates the joint posterior distributions of the morphological parameters. Because the transformation we use is differentiable, the STN can be trained using standard backpropagation along with the downstream CNN without any additional supervision.

The inclusion of the STN in the framework greatly reduces the amount of time spent on data pre-processing as we do not have to worry about making cutouts of the "proper size"—an inherently difficult task since most of the galaxies in our sample have not been morphologically analyzed before. More importantly, the STN greatly reduces the chance of spurious results by cropping most secondary objects out of frame (see §4.6.2).

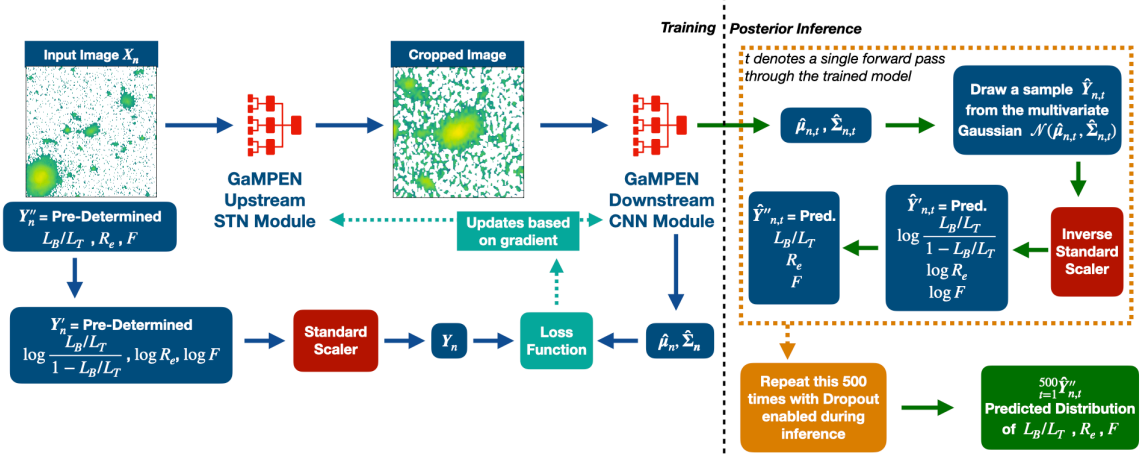


Figure 4.6: Diagram outlining the training (*left*) and posterior inference (*right*) phases of the GaMPEN workflow. Training consists of feeding galaxies (with pre-determined parameter values) through the STN and CNN modules, minimizing the loss function using Stochastic Gradient Descent. During this process, we re-scale the variables described in the text, and return them to the original variable space during inference. After the STN+CNN networks are trained, the posterior inference step consists of 500 forward passes with dropout enabled for each galaxy image. We draw a sample from the predicted multivariate Gaussian distribution during each forward pass, and the collection of these samples gives us the predicted posterior distribution.

Two primary sources of error contribute to the uncertainties in the parameters predicted by GaMPEN. The first arises from errors inherent to the input imaging data (e.g., noise and PSF blurring), and this is commonly referred to as aleatoric uncertainty. The second source of error comes from the limitations of the model being used for prediction (e.g., the number of free parameters in GaMPEN, the amount of training data, etc.); this is referred to as epistemic uncertainty.

For every given input image, GaMPEN predicts a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Although we would like to use GaMPEN to predict aleatoric

uncertainties, the covariance matrix, $\boldsymbol{\Sigma}$, is not known *a priori*. Instead, we train GaMPEN to learn these values by minimizing the negative log-likelihood of the output parameters for the training set. The covariance matrix here represents the aleatoric uncertainties in GaMPEN's predictions.

In order to obtain epistemic uncertainties, we use the Monte-Carlo Dropout technique (**?**), wherein during inference, each image is passed through the trained networks multiple times. During each forward pass, random neurons from the network are removed according to a Bernoulli distribution, i.e., individual nodes are set to zero with a probability, p, known as the dropout rate.

The entire procedure used to estimate posteriors is summarized in Figure 4.6. Once GaMPEN has been trained, we feed each input image, $\hat{\boldsymbol{X}}_n$, 500 times into the trained model with dropout enabled. During each iteration, we collect the predicted set of $\left(\hat{\boldsymbol{\mu}}_{n,t}, \hat{\boldsymbol{\Sigma}}_{n,t}\right)$ for the t^{th} forward pass. Then, for each forward pass, we draw a sample $\hat{\boldsymbol{Y}}_{\boldsymbol{n,t}}$ from the multivariate normal distribution $\mathcal{N}\left(\hat{\boldsymbol{\mu}}_{n,t}, \hat{\boldsymbol{\Sigma}}_{n,t}\right)$. The distribution generated by the collection of all 500 forward passes, $\hat{\boldsymbol{Y}}_{\boldsymbol{n}}$, represents the predicted posterior distribution for the test image $\hat{\boldsymbol{X}}_n$. The different forward passes capture the epistemic uncertainties, and each prediction in this sample also has its associated aleatoric uncertainty represented by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n,t}}$. Thus the above procedure allows us to incorporate both aleatoric and epistemic uncertainties in GaMPEN's predictions.

4.4 Training GaMPEN

Since most of the galaxies described in §4.2.1 have not been morphologically analyzed before, we devise a method to train GaMPEN with only 0.5% of our sample of real galaxies. In order to achieve this, we first train GaMPEN using the simulated galaxies described in §4.2.2. Thereafter, we apply "transfer learning", wherein we fine-tune the already trained network using a small sample of real galaxies, analyzed using GALFIT (Peng et al., 2002).

The process of training GaMPEN consists of the following steps, summarized in Figure 4.7.

- Simulating galaxies corresponding to the target data set (described previously in §4.2.2).
- Initial training of GaMPEN on the above simulated images (described further in §4.4.2).
- Fine-tuning GaMPEN using a small fraction of the real dataset. For this, we used $\sim 0.5\%$ of the HSC data described in §4.2.1. This process is known as transfer learning and is described further in §4.4.3.
- Testing the performance of the fine-tuned network on a test set of real galaxies. For this, we used $\sim 0.25\%$ of the HSC data described in §4.2.1.

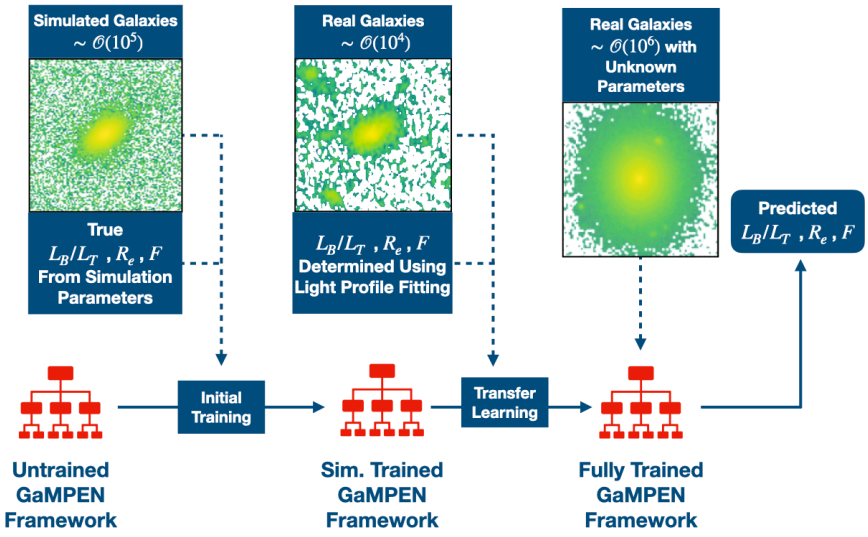


Figure 4.7: Diagram outlining the different stages of training GaMPEN. We first train GaMPEN using simulated light profiles described in §4.2.2. Thereafter, we fine-tune the simulation-trained framework using 0.5% of our real data sample, for which we pre-determined the morphological parameters using light-profile fitting, as described in §4.5. Finally, we process all the $\sim$ 8 million galaxies in our dataset through the trained GaMPEN framework to obtain estimates of their morphological parameters and associated uncertainties.

- Processing the remainder of the real data ($\sim$ 99% of the HSC data described in §4.2.1) through the trained framework.

We describe each of the above steps in detail below.

4.4.1 Data Transformations

To make the training process more robust against numerical instabilities, we transform the input images and target variables following the steps outlined in Ghosh et al. (2022).

Since reducing the dynamic range of pixel values has been found to be helpful in neural network convergence (e.g., Tanaka et al., 2022; Walmsley et al., 2021; Zanisi et al., 2021), we pass all images in our dataset through the arsinh function. For all the target variables, we first apply the logit transformation to L_B/L_T and log transformations to R_e and F:

$$\boldsymbol{Y_n'} = f''(\boldsymbol{Y}_n'') = \left(\log \frac{L_B/L_T}{1 - L_B/L_T}, \log R_e, \log F\right), \tag{4.1}$$

where $\boldsymbol{Y}_n'' = [L_B/L_T, R_e, F]$ is the target set of variables before transformation and f'' is how we will refer to the transformation in Equation 4.1. Next, we apply the standard scalar transformation to each parameter (calibrated on the training data), which amounts to subtracting the mean value

of each parameter and scaling its variance to unity. These two transformations ensure that all three variables have similar numerical ranges and prevent variables with larger numerical values from making a disproportionate contribution to the loss function.

Post training, during inference, we apply the inverse of the standard scalar function (with no re-tuning of the mean or variance), followed by the inverse of the logit and log transformations, f''^{-1}, as indicated in Figure 4.6. Besides transforming the variables back to their original space, these final transformations also ensure that the predicted values always conform to physically meaningful ranges ($0 \leq L_B/L_T \leq 1$; $R_e > 0$; $F > 0$).

4.4.2 Initial Training of GaMPEN on simulated galaxies

The purpose of training GaMPEN initially on simulations is two-fold. Firstly, it greatly reduces the amount of real data needed for training the framework. Secondly, since simulated galaxies are the only situation where we have access to the "ground-truth" morphological parameters, they provide the perfect dataset to assess GaMPEN's typical accuracy for the different output parameters.

In Ghosh et al. (2022), we extensively tested and reported on GaMPEN's performance on simulated HSC $z \leq 0.25$ g-band galaxies. Here, we extend that to include simulated r-band $0.25 < z \leq 0.5$ and i-band $0.50 < z \leq 0.75$ galaxies. Out of the 150,000 galaxies simulated in each z-bin, we use **70**% to train the framework and 15% as a validation set to determine the optimum value for various hyper-parameters (such as learning rate, batch size, etc.). Thereafter, we use the remaining 15%, which the framework has never encountered before, to evaluate the performance of the trained framework.

We train GaMPEN by minimizing its loss function using Stochastic Gradient Descent. The different hyper-parameters that need to be tuned are: the learning rate (the step size during gradient descent), momentum (acceleration factor used for faster convergence), the strength of L2 regularization (the degree to which larger weight values are penalized), and batch size (the number of images processed before weights and biases is updated). To choose these hyper-parameters, we trained GaMPEN with different sets of hyper-parameters and chose the ones that resulted in the lowest value for the loss function on the validation set. The final hyper-parameters for the trained models are given in Appendix 4.D.

In order to test the robustness of our simulation-trained framework, we compare the predictions made by GaMPEN on the test set to the true values determined from the simulation parameters. The results are similar across all redshift bins and closely follow what we determined previously in

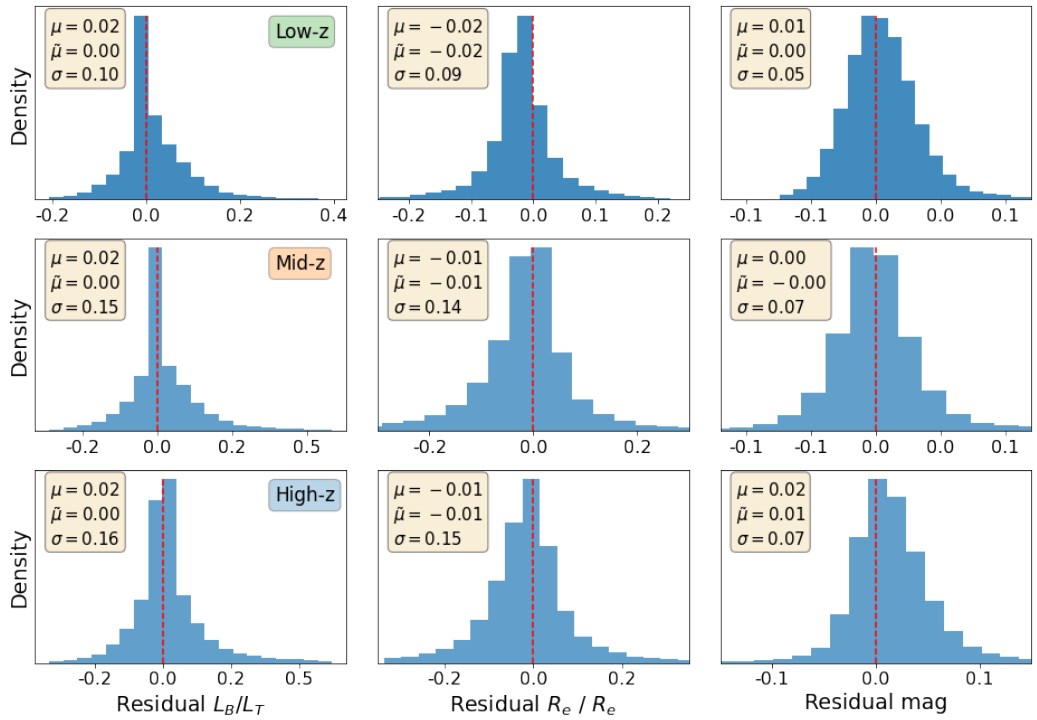


Figure 4.8: Histograms of residuals for simulated galaxies (in the test set) across all three redshift bins. We define the residuals as the difference between the true value and the most probable value predicted by GaMPEN. The dashed vertical line represents $x = 0$, denoting cases with perfectly recovered parameter values. The mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution are listed in each panel.

Ghosh et al. (2022). The histograms of residuals for GaMPEN's output parameters are shown in Figure 4.8 across all three redshift bins. Note that to make all three parameters dimensionless, we report the (Residual R_e)/R_e, instead of simply Residual R_e. For each histogram, we also show the mean (μ), median ($\tilde{\mu}$), and standard deviation (σ). The mean and median help demonstrate that the distributions are all centered around zero, and the standard deviation indicates the value of the "typical error" made by the framework (i.e., 68.27% of the time, GaMPEN's prediction errors are less than this value).

Note that for the simulated data, the typical error made by GaMPEN increases with redshift. This is expected given that in our simulations, galaxies in the higher redshift bins are preferentially smaller, fainter, and have lower signal-to-noise ratios than their lower redshift counterparts—thus, these galaxies are harder to analyze morphologically for any image processing algorithm. However, for all the parameters across all redshift bins, the GaMPEN error is typically always $\lesssim 15\%$ for the simulated sample.

We would like to note that GaMPEN does not explicitly predict the number of components a

galaxy has, but we performed an analysis of its relative performance on single- and double-component galaxies in Ghosh et al. (2022) (see Figure 15 therein). We found that accurately determining L_B/L_T is more challenging for double-component galaxies (compared to single-component galaxies); and becomes even more difficult when one of the components strongly dominates the other.

4.4.3 Transfer Learning using Real Data

Transfer Learning as a data-science concept has been around since the late 1990s (e.g., Blum & Mitchell, 1998) and involves taking a network trained on a particular dataset and optimized for a specific task, and re-tuning/fine-tuning the weights and biases for a slightly different task or dataset. In Ghosh et al. (2020), we introduced the concept of training on simulated light-profiles and then transfer learning using real data for galaxy morphology analysis. Here, we employ the same idea, taking the networks trained on the simulated galaxies and then employing transfer learning using $\sim 15,000$ real HSC-Wide galaxies in each redshift bin.

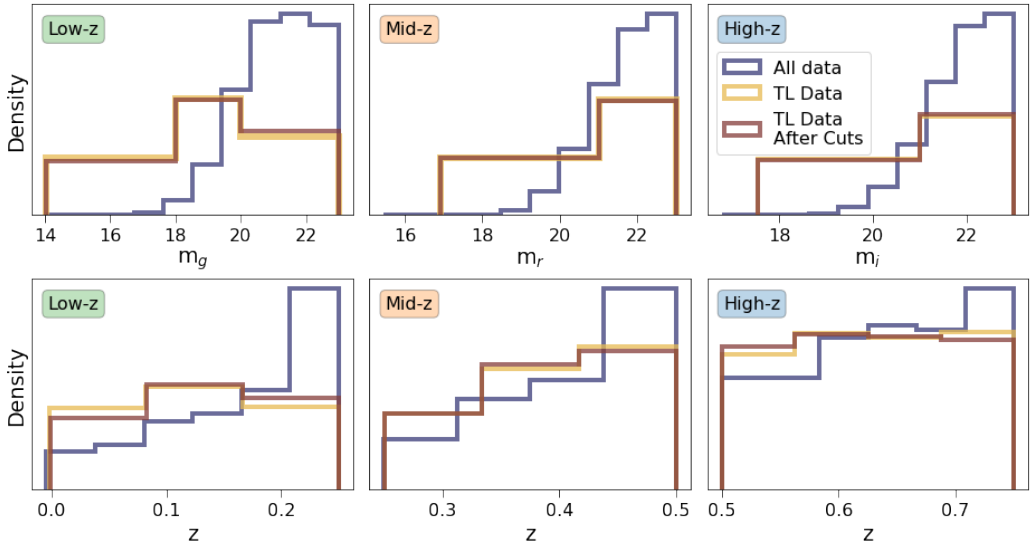


Figure 4.9: Magnitude and redshift distributions for all galaxies in each redshift bin, plotted along with the galaxies selected for transfer learning (before and after applying quality cuts, as described in §4.5). Note that we plot density on the y-axis, not the number of samples. The total number of galaxies used for transfer learning is $\sim 0.5\%$ of all the galaxies in our dataset. The relative density of some magnitude bins is higher than others (e.g., $18 < m \leq 20$ for low-z) because they span a smaller range while having roughly the same number of galaxies as the other bins.

In order to select a sample of galaxies to use for transfer learning, we start with the galaxies summarized in Table 4.1. Note that what matters in training a CNN is not matching the observed distributions of the simulation parameters; rather, it is spanning the full range of those parameters with high fidelity. Having too many of a particular type—even if that is the reality in real data—can

result in lower accuracy for minority populations (e.g., Ghosh et al., 2020). As seen in Figure 4.3, the sample for all three redshift bins is heavily biased towards fainter galaxies. To ensure that GaMPEN gets to train on enough bright galaxies, we split the mid- and high-z samples into two sub-samples: $m \leq 21$ and $m > 21$. For the low-z sample, since it has a more substantial tail towards the lower m values (compared to the mid- and high-z samples), we use three sub-samples—$m \leq 18$, $18 < m \leq 20$, $m > 20$. We select 20,000 galaxies from each redshift bin, making sure to sample equally across the magnitude bins mentioned above. Thereafter, we determine their morphological parameters using light-profile fitting, as described later in §4.5. The magnitude and redshift distributions of the selected galaxies are shown in Figure 4.9. As seen from the figure, the transfer learning sample has sufficiently large numbers of examples from all parts of the parameter space. This empowers us to optimize GaMPEN for the full range of galaxy morphologies.

Out of the 20,000 galaxies selected in each redshift bin, we use 75% for fine-tuning the simulation trained GaMPEN models and another 5% for selecting the various hyper-parameters to be used during the transfer learning process. The remaining 20% is used to evaluate the performance of the fully trained GaMPEN frameworks in §4.6.

In order to artificially augment the number of galaxies being used for transfer learning, we apply random rotations and horizontal/vertical flips on the galaxies earmarked for training. This takes the effective number of samples used for fine-tuning in each bin from $\sim 15,000$ to $\sim 90,000$. Using the values determined by light-profile fitting as the labels, we fine-tune the three GaMPEN models trained on simulations. We choose the values of the different hyper-parameters based on the loss computed on the validation set, and the final chosen hyper-parameters for transfer learning are reported in Appendix 4.D.

Fine-Tuning the Dropout Rate

Aside from the hyper-parameters mentioned in §4.4.2, there is one more adjustable parameter in GaMPEN—the dropout rate, which directly affects the calculation of the epistemic uncertainties outlined in §4.3. On average, higher dropout rates lead networks to estimate higher epistemic uncertainties. To determine the optimal value for the dropout rate, during transfer learning, we trained variants of GaMPEN with dropout rates from 10^{-3} to 10^{-5}, all with the same optimized values of momentum, learning rate, and batch size mentioned in Appendix 4.D.

To compare these models, we calculate the percentile coverage probabilities associated with each model, defined as the percentage of the total test examples where the parameter value determined using light profile fitting lies within a particular confidence interval of the predicted distribution.

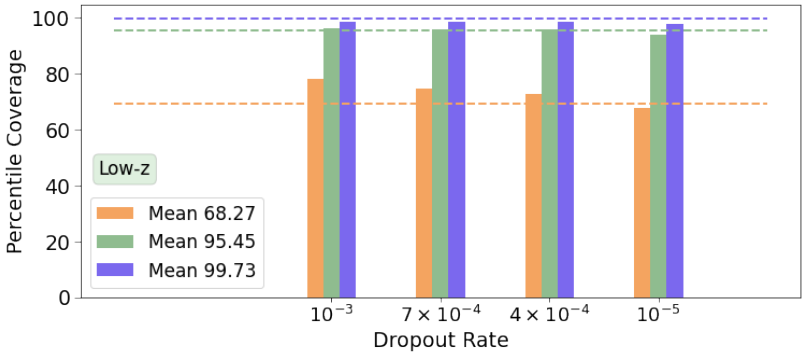


Figure 4.10: The calculated percentile coverage probabilities for different dropout rates for the low-z bin. Note that the coverage probabilities have been averaged over the three output variables. The coverage probabilities are defined as the percentage of the total test examples where the value determined using light profile fitting lies within a particular confidence interval of the predicted distribution. A dropout rate of 4×10^{-4} leads to coverage probabilities very close to their corresponding confidence levels. A similar process of tuning in the mid-z and high-z bins leads to an optimal dropout rate of 2×10^{-4} in both of them.

We calculate the coverage probabilities associated with the 68.27%, 95.45%, and 99.73% central percentile confidence levels, corresponding to the 1σ, 2σ, and 3σ confidence levels for a normal distribution. For each distribution predicted by GaMPEN, we define the 68.27% confidence interval as the region on the x-axis of the distribution that contains 68.27% of the most probable values of the integrated probability distribution. To estimate the probability distribution function from the GaMPEN predictions (which are discrete), we use kernel density estimation, which is a non-parametric technique to estimate the probability density function of a random variable.

We calculate the 95.45% and 99.73% confidence intervals of the predicted distributions in the same fashion. Finally, we calculate the percentage of examples for which the GALFIT-ed parameter values lie within each of these confidence intervals. An accurate and unbiased estimator should produce coverage probabilities equal to the confidence interval for which it was calculated (e.g., the coverage probability corresponding to the 68.27% confidence interval should be 68.27%). For every redshift bin, we choose the dropout rate for which the calculated coverage probabilities are the closest to their corresponding confidence levels. This leads to a dropout rate of 4×10^{-4} for the low-z bin, and 2×10^{-4} for the mid- and high-z bins.

As an example, we show in Figure 4.10 the coverage probabilities (averaged across the three output variables) for different dropout rates for the low-z sample. As can be seen, higher values of the dropout rate lead to GaMPEN over-predicting the epistemic uncertainties, resulting in too high coverage probabilities. In contrast, extremely low values lead to GaMPEN under-predicting the epistemic uncertainties. For a dropout rate of 4×10^{-4}, the calculated coverage probabilities are

very close to their corresponding confidence levels, resulting in accurately calibrated posteriors.

4.5 Galfitting Galaxies for Transfer Learning & Validation

In this section, we describe a semi-automated pipeline that we developed and used to determine the morphological parameters of $\sim 60,000$ galaxies ($\sim 20,000$ in each z-bin), which are used for transfer learning and to test the efficacy of the trained GaMPEN frameworks.

In order to estimate the parameters, we use GALFIT, which is a two-dimensional fitting algorithm designed to extract structural components from galaxy images. However, before running GALFIT, we run Source Extractor (Bertinl, 1996) on all the cutout frames in order to obtain segmentation maps for each input image. We use these segmentation maps to mask all secondary objects present in the cutout frame during light-profile fitting. We also use the Source Extractor estimates of various morphological parameters as the initial starting guesses during light profile fitting. Lastly, we use the Source Extractor estimates to pick a cutout size for each galaxy, which we set to be ten times the effective radius estimated by Source Extractor.

Using GALFIT, we fit each galaxy with two Sérsic components:- a disk-component with Sérsic index, $n = 1$, and a bulge component with $3.5 < n < 5.0$. For each galaxy, we perform two consecutive rounds of fitting:- first with some constraints placed on the values of the different parameters and second round with almost no constraints. This is to help the parameters converge quickly during the initial round while still allowing the full exploration of the parameter-space in the subsequent round. In the initial round, we constrain the radius of each galaxy to be between 0.5 - 90 pixels (0.85 - 15.12 arcsec) and the magnitude to be between ± 7.5 of the initial value guessed by Source Extractor. We also constrain the difference between the magnitudes of the two components to be between -7.5 and $+7.5$ ($0.001 < L_B/L_T < 0.999$), and the relative separation between the centers of the two light profiles to be not more than 30 pixels (5.04 arcsec). These constraints are much more expansive than what we expect galaxy parameters to be in this z-range and are similar to what has been used for many previous studies (e.g., Wel et al., 2012; **?**). During the second round of fitting, we only use the constraints on the Sérsic indices and the relative separation between the two components. After fitting the two components, we calculate the R_e of each galaxy as the radius that contains 50% of the total light in the combined disk + bulge model. Figure 4.11 shows the image cutouts, masks, fitted models, and residuals for two typical galaxies chosen from each redshift bin.

After the two rounds of fitting, we excluded galaxies for which GALFIT failed to converge ($\sim 3.2\%$, $\sim 3.8\%$, and $\sim 2.9\%$ for the low-, mid-, and high-z bins, respectively). Thereafter, we

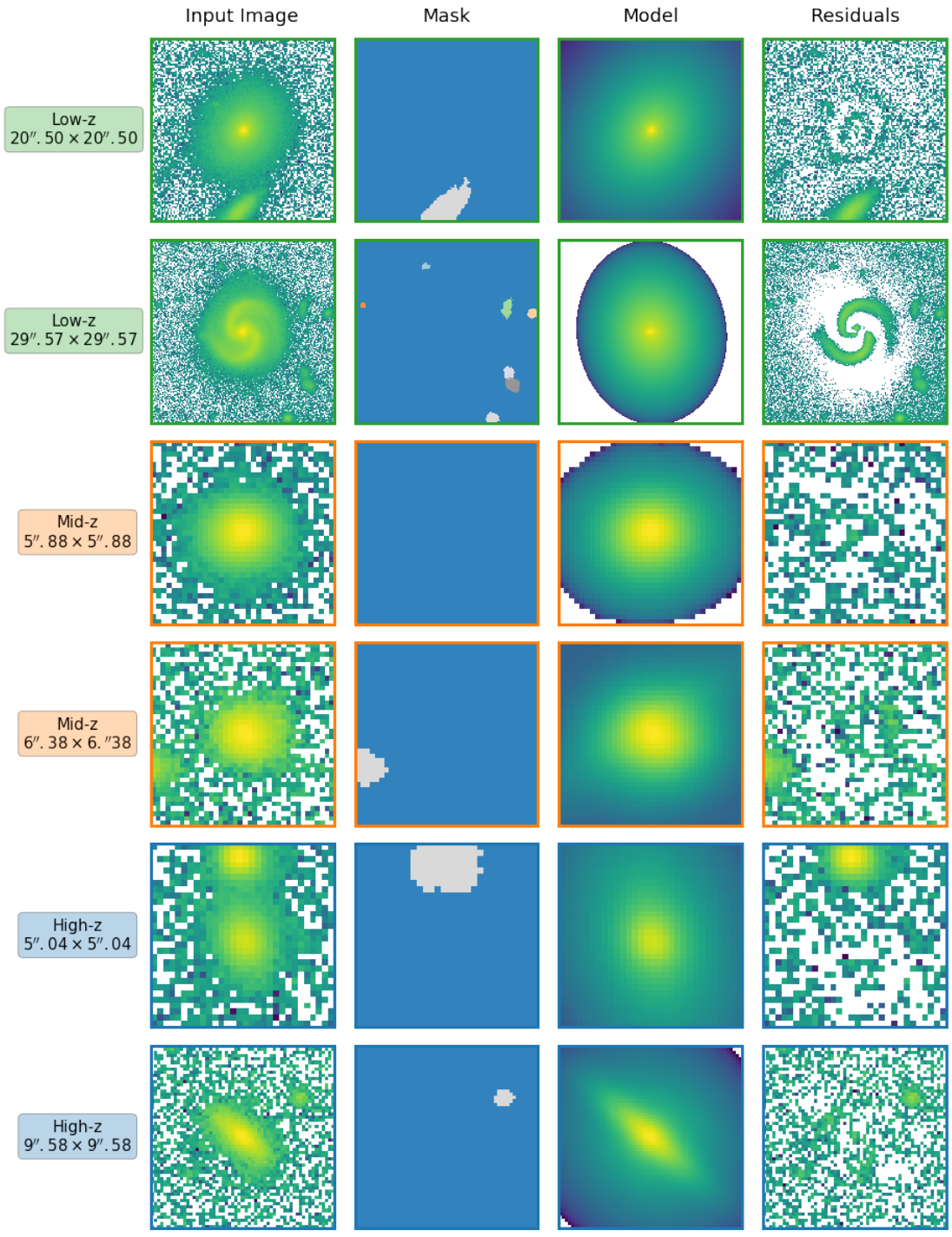


Figure 4.11: Steps used in our light-profile fitting pipeline to determine morphological parameters, for two representative galaxies in each redshift bin. From left to right we show the input image, the mask generated by Source Extractor, the model generated by GALFIT, and the residuals. Note that since we do not explicitly model any Fourier bending modes or coordinate rotations, we expect features like spiral arms to show up in the residuals, as depicted in the second row.

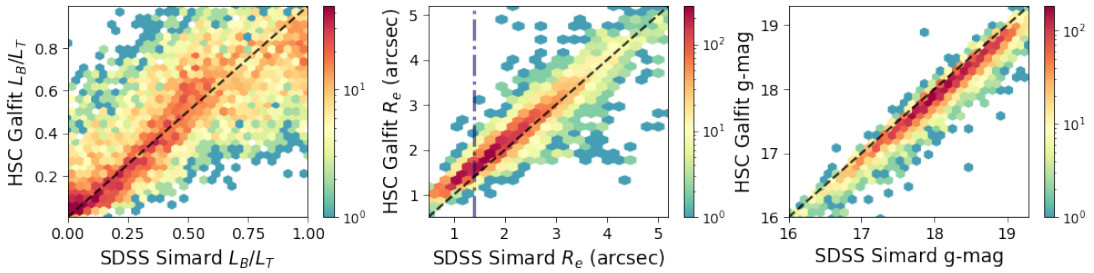


Figure 4.12: Morphological parameters determined for HSC-imaged galaxies using our light-profile fitting pipeline, versus morphological parameters for the same galaxies determined by Simard et al. (2011) based on SDSS imaging. The black dashed diagonal corresponds to perfect agreement. The vertical line in the middle panel shows the median SDSS g-band seeing.

Table 4.3: GALFIT Quality Cuts

Sample	Criteria	Excluded %
Low-z	(`problematic_value_flags == True`) OR (Sep. Dist. $> 2''$) OR (`max_iters_flag == True` AND Red. $\chi^2 > 2.5$)	26.04%
Mid-z	(`problematic_value_flags == True`) OR (Sep. Dist. $> 1''$) OR (`max_iters_flag == True` AND Red. $\chi^2 > 1.25$)	32.10%
High-z	(`problematic_value_flags == True`) OR (Sep. Dist. $> 0''.7$) OR (`max_iters_flag == True` AND Red. $\chi^2 > 1.25$)	34.13%

visually inspected the fits of a randomly selected sub-sample of ~ 300 galaxies from each redshift bin. The process of visually inspecting ~ 900 galaxies helped us to identify three failure modes of our fitting pipeline: i) for some galaxies, GALFIT assigned an extremely small axis ratio to one of the components leading to an unphysical lopsided bulge/disk; ii) for some galaxies, the centroids of the two fitted components were too far away from each other; iii) some galaxies had residuals that were too large. Examples of each of these failure modes are shown in Appendix 4.E.

In order to get rid of these problematic fits from our GALFIT-ed sample, we use a mixture of GALFIT flags and calculated parameters. Specifically, as summarized in Table 4.3 and detailed below:

a) `problematic_value_flags`[3] flags galaxies that have an extremely small axis-ratio (< 0.1) or radius (< 0.5 pixels), or any other parameters that caused issues with numerical convergence.

b) `max_iters_flag` flags galaxies for which GALFIT quit after reaching the maximum number

3. Refer to Bullet 8 at https://users.obs.carnegiescience.edu/peng/work/galfit/TOP10.html

of iterations (100).

c) The reduced χ^2 of the fit is poor.

d) The distance between the centers of the two fitted components is too large.

Table 4.3 shows the criteria used to exclude fits in each redshift bin. The variation in the thresholds with redshift is to account for the fact that galaxies at higher redshift are preferentially smaller, fainter, and have lower signal-to-noise ratios. Figure 4.9 shows that the exclusion criteria do not selectively exclude more galaxies from certain regions of the magnitude/redshift parameter space compared to others. In order to determine appropriate thresholds for the various flags above, we balanced excluding too many galaxies against ensuring only good fits are included in the final transfer learning dataset. The choice of thresholds is arbitrary to a certain extent. In order to empower users to retrain GaMPEN using different criteria for their own scientific analysis, we are making public the entire catalog of GALFIT-ed values as outlined in Appendix 4.A.

To the best of our knowledge, there is no published large catalog of bulge+disk decomposition of HSC galaxies against which we can compare the results of our light-profile fitting pipeline. However, Simard et al. (2011) performed bulge+disk decomposition using g and r-band imaging from the Sloan Digital Sky Survey (SDSS). We should note that HSC-Wide differs from SDSS in a multitude of ways, with the most significant differences being in median seeing [HSC-Wide: $0.''79$ compared to SDSS: $1.''4$ in the g-band] and pixel scale [HSC: 0.168 arcsecs/pixel compared to SDSS: 0.396 arcsecs/pixel]. Thus, we do not expect our analysis to yield the exact same results as that of Simard et al. (2011). However, given that a significant portion of our low-z sample overlaps with that of Simard et al. (2011), it is still useful to compare our results to that of Simard et al. (2011), as the overall trends should agree.

Using an angular cross-match diameter of 0.15 arcsec, we cross-matched our low-z GALFIT sample with that of Simard et al. (2011) to obtain a sample of ~ 6500 galaxies. Note that although our HSC sample extends to $g < 23$, the cross-matched galaxies are mostly $g < 20$ due to the shallower depth of the SDSS data. In Figure 4.12, we compare the results of our light-profile fitting results with that of Simard et al. (2011). The figure shows galaxies in hexagonal bins of roughly equal size, with the number of galaxies in each bin represented according to the colorbar on the right. Note that we are using a logarithmic colorbar to explore the full distribution of galaxies, down to 1 galaxy/bin. Although there are outliers present for all three parameters, and the scatter of the relationship depends on the parameter, our GALFIT-derived parameters strongly correlate with that of Simard et al. (2011). According to the Spearman's rank correlation test (see Dodge, 2008, for more details),

there is a positive correlation for all three variables, and the null hypothesis of non-correlation can be rejected at extremely high significance ($p < 10^{-200}$). The correlation coefficients obtained for L_B/L_T, R_e, and F are 0.85, 0.95, and 0.98, respectively. For both our GALFIT predictions and that of Simard et al. (2011), we define R_e for double-component fits as the radius that encompasses 50% of the light from both components combined.

Note that the higher scatter in the L_B/L_T relation is expected, given that bulge+disk decomposition involves constraining two different imaging components simultaneously and is thus more sensitive to algorithmic differences and the differences in imaging quality between SDSS and HSC mentioned above. These differences in imaging quality, data-reduction pipeline, and filters (Kawanomoto et al., 2018) might also explain the slight offsets in the R_e and flux measurements. It is also important to note that most of the differences in R_e measurements are at effective radii values smaller than the median SDSS g-band seeing, as seen in the middle panel of Figure 4.12.

4.6 Evaluating GaMPEN's performance

After both rounds of training are complete, we apply the final trained GaMPEN models to all the ~ 8 million galaxies outlined in §4.2.1. In this section, we evaluate GaMPEN's performance to assess the reliability of its predictions.

4.6.1 Inspecting the Predicted Posteriors

Using the procedure outlined in §4.3 and Figure 4.6, we use the trained GaMPEN frameworks to obtain joint probability distributions of all the output parameters for each of the ~ 8 million galaxies. Figure 4.13 shows the marginalized posterior distributions for six randomly selected galaxies; along with the input cutouts fed to GaMPEN. All the predicted distributions are unimodal, smooth, and resemble Gaussian/skewed-Gaussian distributions. For each predicted distribution, the figure also shows the parameter space regions that contain 68.27%, 95.45%, and 99.73% of the most probable values of the integrated probability distribution. We use kernel density estimation to estimate the probability distribution function (PDF; shown by a blue line in the figure) from the predicted values. The mode of this PDF is what we refer to as the "predicted value" henceforth. The figure also demonstrates how GaMPEN predicts distributions of different widths based on the input frame (e.g., the L_B/L_T distribution in the second row is wider compared to the first row), and we will explore this in more detail in §4.6.4.

By design, GaMPEN predicts only physically possible values. This is especially apparent in the

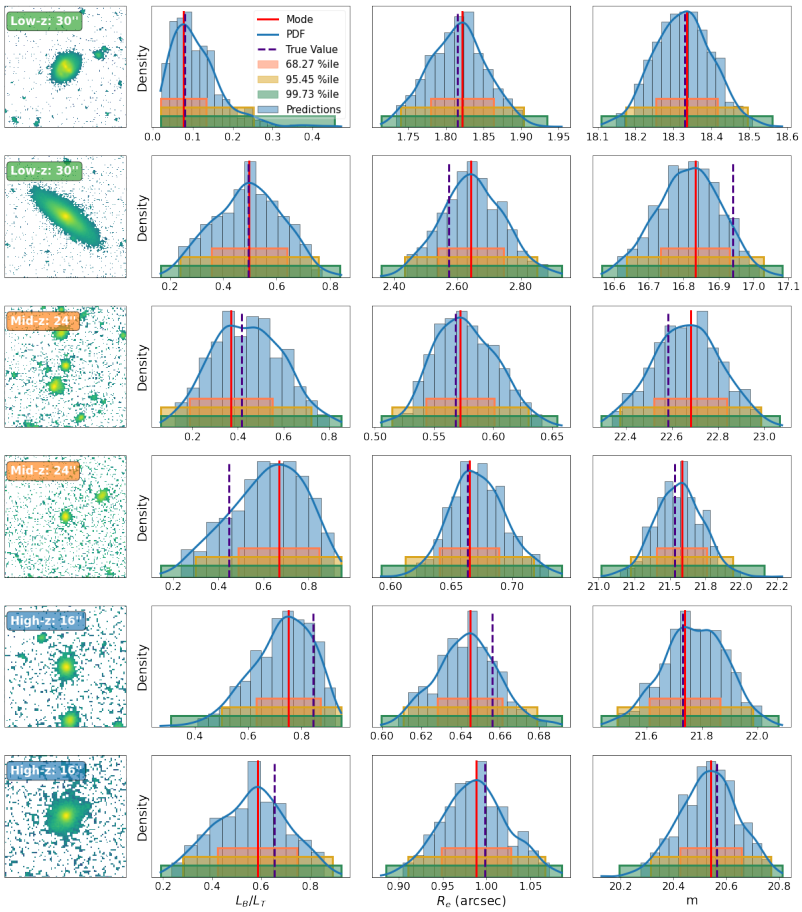


Figure 4.13: Examples of predicted posterior distributions for two randomly chosen galaxies from each redshift bin. The blue shaded histograms show the predictions from GaMPEN, and the solid blue lines show the associated probability distribution functions estimated by kernel density estimation. These are used to calculate the confidence intervals shown in the figure with pink, yellow, and green shading. The mode (solid red line) shows the most probable value of each morphological parameter. As expected, in most cases, the GALFIT-ed value (dashed black line) lies within the 68.27% confidence interval.

L_B/L_T column of rows 1 and 4 of Figure 4.13. Note that to achieve this, we do not artificially truncate these distributions. Instead, we use data transformations, as outlined in §4.4.1. This ensures that the predicted L_B/L_T values are always between 0 and 1. Similarly, we also ensure that the R_e and F values predicted by GaMPEN are positive through appropriate transformations.

While performing quality checks on the predicted posteriors of all the $\sim$ 8 million galaxies, we noticed that sometimes GaMPEN predicts R_e and F values outside the parameter range on which it was trained (e.g., $m > 23$). It has been shown that while machine learning frameworks are excellent at interpolation, one should be extremely cautious while trying to extrapolate too much beyond the training set (e.g., Quionero-Candela et al., 2009; Recht et al., 2019; Taori et al., 2020). Thus, for each redshift bin, we exclude galaxies with predicted R_e and F values which are outside the upper

and lower bounds of the training set by more than 0.5 arcsecs or 0.5 mags, respectively. This led to $\sim 6\%$, $\sim 2.5\%$, and $\sim 1.2\%$ of the data being excluded in the low-, mid-, and high-z bins.

4.6.2 Evaluating the STN performance

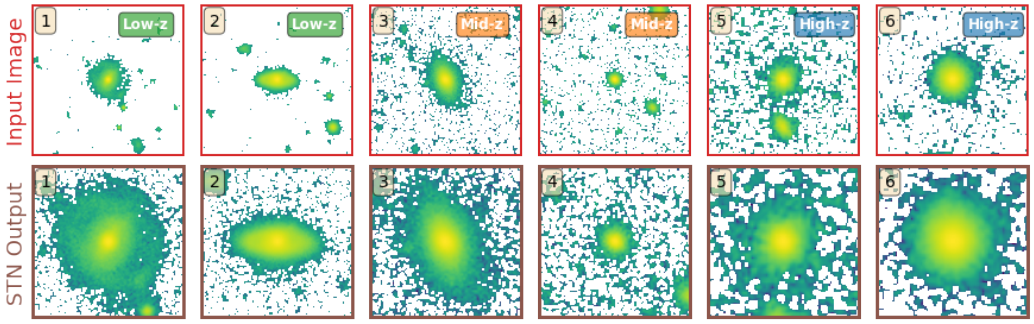


Figure 4.14: Examples of the transformation applied by the STN to two randomly selected galaxies from each redshift bin. The top row shows the input galaxy images, and the bottom row shows the corresponding output from the STN. The numbers in the top-left yellow boxes help correspond the output images to the input images. As can be seen, the STN learns to crop most secondary objects present in the input frame.

As can be seen from Rows 3, 4, and 5 of Figure 4.13, GaMPEN can accurately predict morphological parameters even when the primary galaxy of interest occupies a small portion of the input cutout, and secondary objects are present in the input frame. This is primarily enabled by the upstream STN in GaMPEN which, during training, learns to apply an optimal amount of cropping to each input image.

Figure 4.14 shows examples of the transformations applied by the STN to randomly selected galaxies in all three redshift bins. As can be seen, the STN crops out most secondary galaxies present in the cutouts and helps the downstream CNN to focus on the galaxy of interest at the center.

To further validate the performance of the STN, we measured the amount of cropping applied by the STN for all galaxies in the low-z bin. We chose the lowest redshift bin for this test as it has the largest range of galaxy radii among the different redshift bins. After that, we sorted all the processed images based on the amount of cropping applied to each input image. In Figure 4.15, we show example images from our dataset with extremely high and extremely low values of applied crops. The s parameter shown in the lower left of each panel denotes what fraction of the input image was retained in the STN output—higher values of s denote that a more significant fraction of the input image was retained in the output image produced by the STN (i.e., minimal cropping). Figure 4.15 demonstrates that (without us having to engineer this specifically), the STN correctly

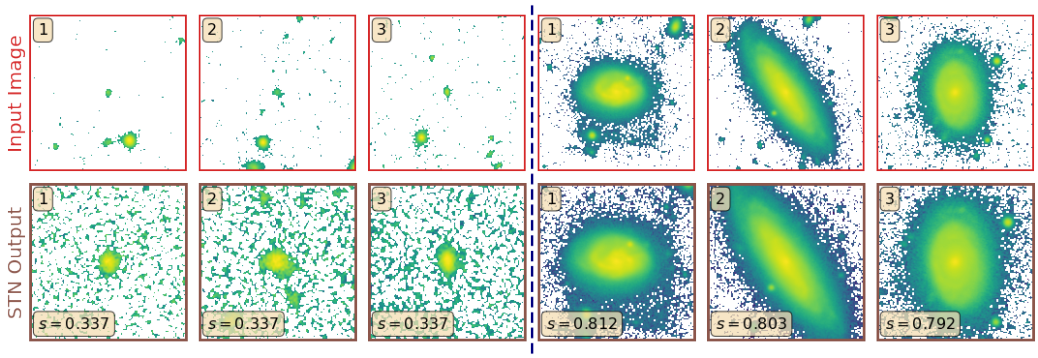


Figure 4.15: (*Left*): Galaxies in the low-z bin with the lowest values of s (i.e., the most aggressive crops) (*Right*): Galaxies in the low-z bin with the highest values of s (i.e., the least aggressive crops). The s parameter denotes the fraction of the input image that was retained in the STN output. As can be seen, the STN correctly learns to apply the most aggressive crops to small galaxies; and the least aggressive crops to large galaxies.

learns to apply the most aggressive crops to the smallest galaxies in our dataset, and the least aggressive crops to the largest galaxies

Thus, GaMPEN's STN learns to systematically crop out secondary galaxies in the cutouts and focus on the galaxy of interest at the center of the cutout. At the same time, the STN also correctly applies minimal cropping to the largest galaxies, making sure the entirety of these galaxies remains in the frame.

4.6.3 Comparing GaMPEN predictions to GALFIT predictions

Out of the $60,000$ galaxies analyzed using GALFIT in §4.5, we use 80% as the training and validation sets. We use the remaining 20%, which the trained GaMPEN frameworks have never seen, to evaluate the accuracy of the predicted parameters. We refer to this as the "test set" henceforth.

In Figure 4.16, we show the coverage probabilities achieved by GaMPEN on the test set. Note that in § 4.4.3, we tuned the dropout rate using the validation set, whereas the values in Figure 4.16 are calculated on the test set. In the ideal situation, they would perfectly mirror the confidence levels; (e.g., 68.27% of the time, the true value would lie within 68.27% of the most probable volume of the predicted distribution). Clearly, the coverage probabilities achieved by GaMPEN are consistently close to the claimed confidence levels, both when averaged over the three output parameters, as well as for each parameter individually. The mean coverage probability never deviates by more than 4.5% from the claimed confidence interval, and when considered for each parameter individually, the coverage probability never deviates by more than 8.7% from the corresponding confidence interval. Additionally, we note that even for the case for which the coverage probabilities

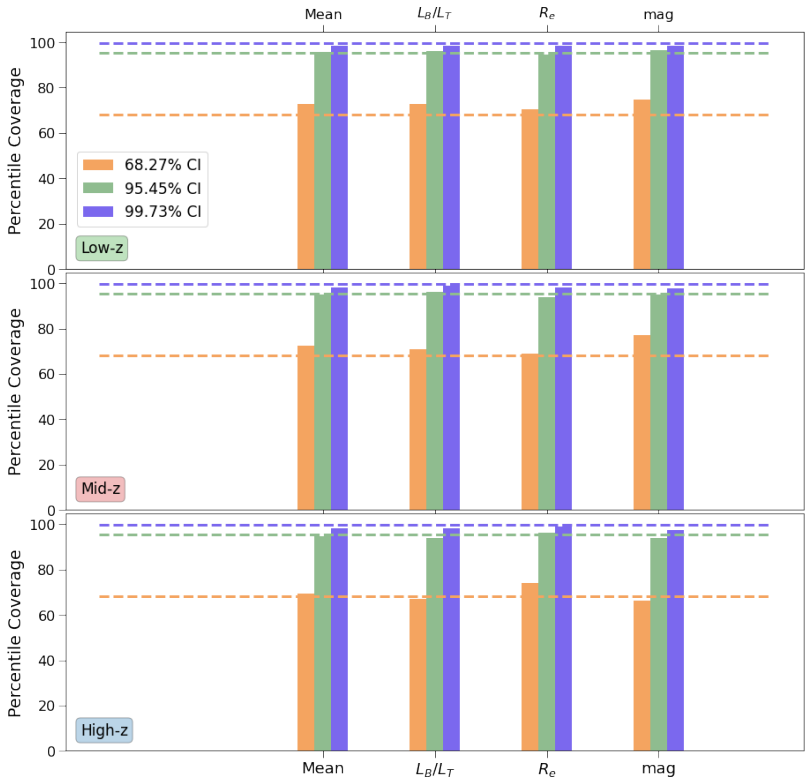


Figure 4.16: Percentile coverage probabilities achieved on the test set shown separately for each redshift bin. The leftmost set of bars in each panel shows the coverage probabilities when averaged over the three output parameters, and the right three sets of bars show the coverage probabilities for each parameter individually. The mean coverage probability never deviates by more than 4.5% from the claimed confidence interval, and when considered for each parameter separately, the coverage probability never deviates by more than 8.7%. This demonstrates that GaMPEN produces well-calibrated accurate uncertainties.

are most discrepant (68% flux confidence interval for the mid-z model), the uncertainties predicted by GaMPEN are in any case overestimates (i.e., conservative). If GaMPEN were used in a scenario that requires perfect alignment of coverage probabilities, users could employ techniques such as importance sampling (Kloek & van Dijk, 1978) on the distributions predicted by GaMPEN. We note here that incorporating the covariances between the predicted parameters into our loss function was key to achieving simultaneous calibration of all three output variables.

Having shown above that the posteriors predicted by GaMPEN are well calibrated, we now investigate how close the modes (most probable values) of the predicted distributions are to the parameter values determined using light-profile fitting. Figure 4.17 shows the most probable values predicted by GaMPEN for the test set plotted against the values determined using GALFIT in hexagonal bins of roughly equal size. The number of galaxies is represented according to the colorbar on the right. Note that we use a logarithmic colorbar to visualize even small clusters of galaxies in

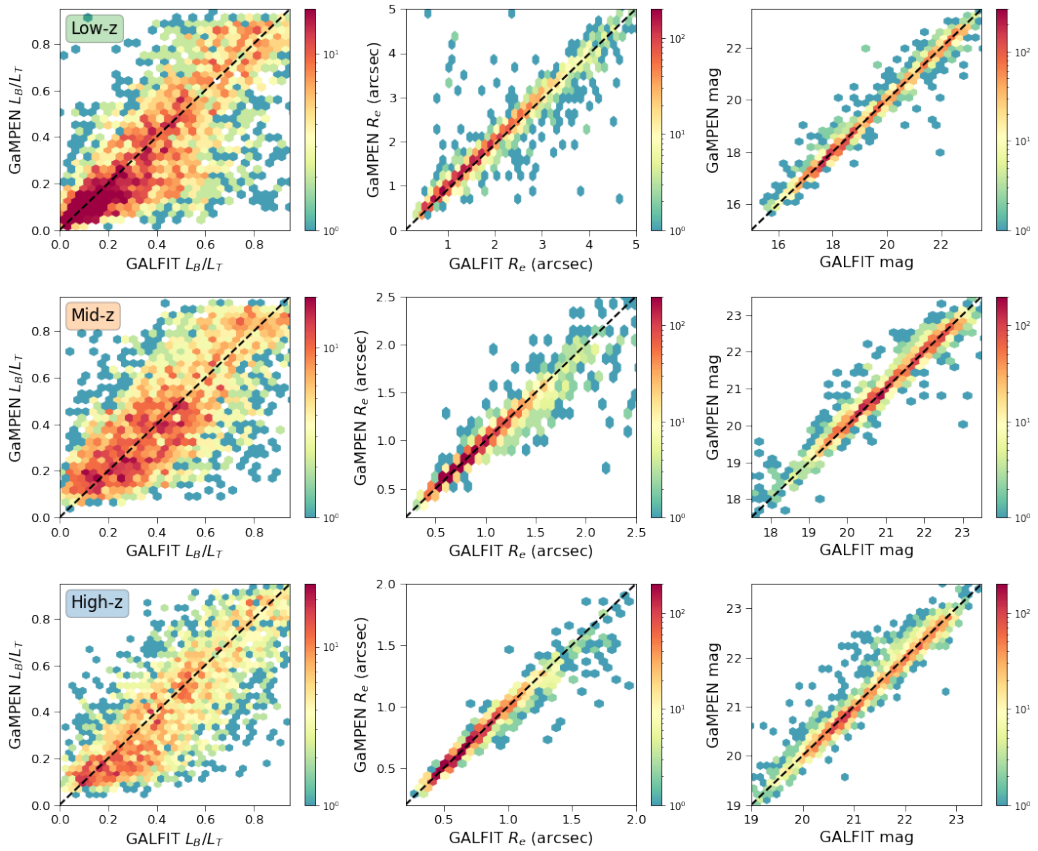


Figure 4.17: The most probable parameter values predicted by GaMPEN for all galaxies in the test set plotted against the values determined using GALFIT. Galaxies are plotted in hexagonal bins of roughly equal size, and the number of galaxies in each bin is represented according to the logarithmic colorbar to the right of each panel. The top, middle, and bottom rows show the results for the low-, mid-, and high-z bins, respectively. The dashed black $y = x$ line represents the line of equality. Across all three redshift bins, values predicted by GaMPEN closely mirror the values obtained using light-profile fitting.

this plane, down to 1 galaxy/bin. Across all three redshift bins, a large majority of all the galaxies are clustered around the line of equality, showing that the most probable values of the distributions predicted by GaMPEN closely track the values obtained using light-profile fitting. The scatter obtained for L_B/L_T is larger compared to the other two output parameters, and we explore that in more detail later in this section.

In Figure 4.18, we show the residual distribution for GaMPEN's output parameters in all three redshift bins. We define the residual for each parameter as the difference between the most probable value predicted by GaMPEN and the value determined using light profile fitting. The box in the upper left corner gives the mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution. All nine distributions are normally distributed (verified using the Shapiro Wilk test), and have $\mu \sim \tilde{\mu} \sim 0$. The σ of each distribution also identifies the typical disagreement for each

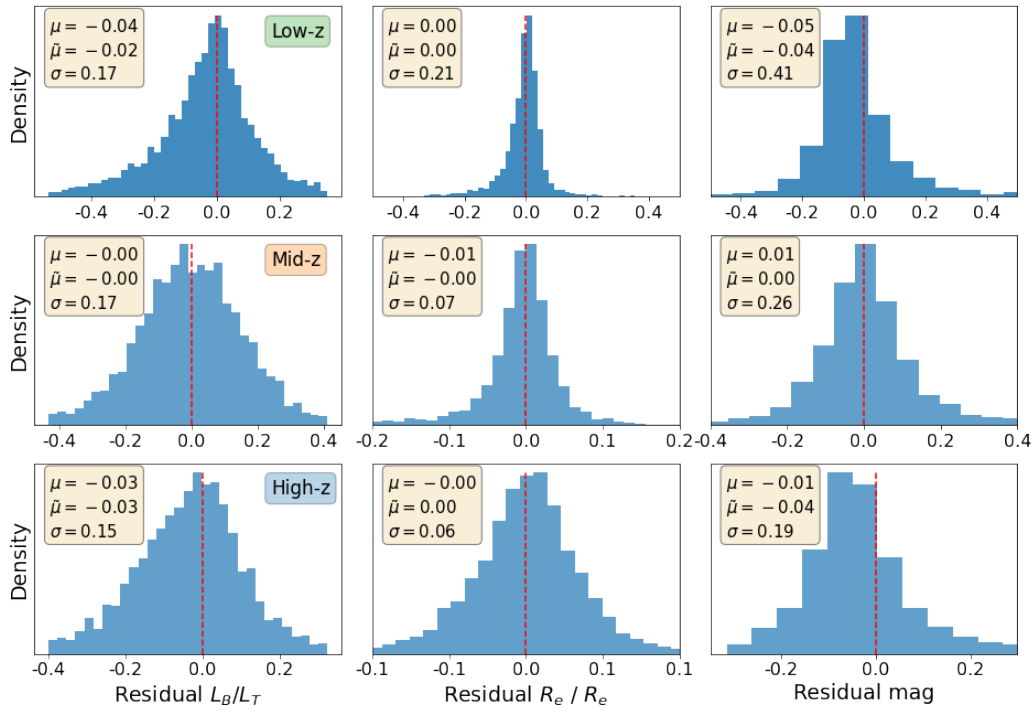


Figure 4.18: Distributions of residuals for all galaxies in the test set; specifically, the differences between the values predicted by GaMPEN and those obtained via light-profile fitting. The top, middle, and bottom rows show the results for the low-, mid-, and high-z bins, respectively. The boxes in the top-left corner of each panel show the mean (μ), median ($\tilde{\mu}$), and standard deviation (σ) of each residual distribution. The σ of each distribution identifies the typical disagreement for each parameter (e.g., for the low-z bin, in 68.27% cases, the predicted magnitude is within $\pm$0.41 of the value determined by light profile fitting). The dashed red vertical line marks $x = 0$.

parameter (e.g., for the low-z bin, in 68.27% cases, the predicted L_B/L_T value is within $\pm$0.17 of the value determined by light profile fitting). The residual R_e, when converted to physical units, correspond to typical disagreements of 0.32 arcsecs, 0.14 arcsecs, and 0.14 arcsecs for the low-, mid-, and high-z bins, respectively. The L_B/L_T residuals are mostly constant across the three redshift bins, while the disagreement between GaMPEN and GALFIT predictions for R_e and F decrease slightly as we go from the low to the higher redshift bins. This could be driven by the fact that the HSC median seeing becomes better as we move from the g-band to the i-band (g-band: 0.″79; r-band: 0.″75; i-band: 0.″61). Additionally, lower redshift galaxies have preferentially more resolved structural features (e.g., spiral arms), which are not accounted for in our disk + bulge GALFIT decomposition pipeline. This could lead to a higher disagreement between the GALFIT and GaMPEN predictions.

Although Figures 4.17 and 4.18 indicate the overall agreement between GaMPEN and GALFIT, they do not reveal the dependence of this agreement on location in the parameter space. This is

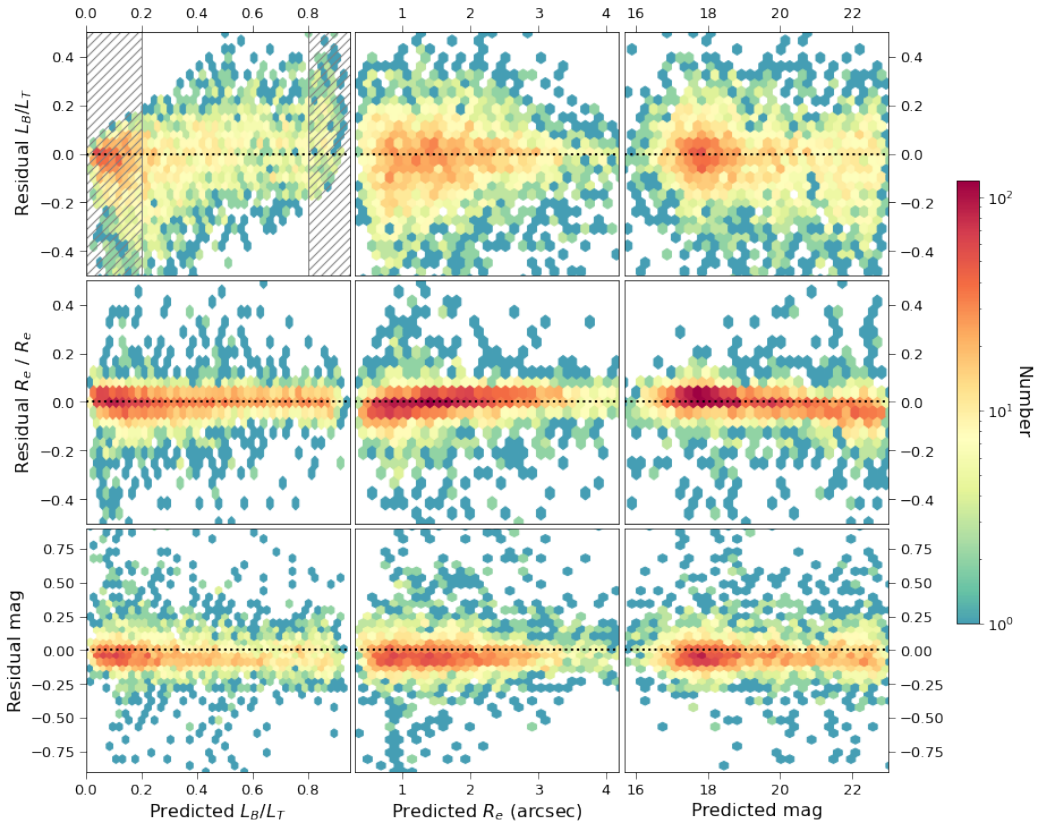


Figure 4.19: Residuals of the output parameters (difference between GaMPEN and GALFIT predictions) plotted against the values predicted by GaMPEN for all galaxies in the low-z test set. To make the y-axis dimensionless for all three parameters, we plot the fractional R_e residuals instead of absolute values. This figure allows us to assign quality labels to GaMPEN's predictions based on the output values (e.g., flagging regions of the parameter space with high levels of disagreement, as shown by the line-shaded region in the top-left panel). See § 4.6.3 for details. The equivalent figures for the mid- and high-z bins are shown in Appendix 4.F.

critical information as this identifies regions of parameter space where GaMPEN agrees especially well or badly with light-profile fitting results, so that future users can flag the reliability of predictions in these regions or use results only from certain regions of the parameter space for specific scientific analyses. Figure 4.19 shows the residuals for the three output parameters for the low-z bin plotted against the values predicted by GaMPEN. Note that in order to make the y-axis dimensionless for all three parameters, we plot the fractional R_e residuals instead of absolute values. As in Figure 4.17, we have split the parameter space into hexagonal bins and used a logarithmic color scale to denote the number of galaxies in each bin. The trends between the residuals and different parameters are very similar across all three redshift bins. Thus, to keep the main text concise, we have shown the plot for the low-z bin here and shown the same plot for the mid- and high-z bins in Appendix 4.F.

For most of the panels, the large majority of galaxies are clustered uniformly around the black

dashed line, $y = 0$, which denotes the ideal case of perfectly recovered parameters (assuming the GALFIT parameters are correct). There are a few notable features on the top row, which depicts the L_B/L_T residuals. In the top left panel, the L_B/L_T residuals are highest near the limits of L_B/L_T. We noticed the same effect when testing GaMPEN with simulated galaxies in Ghosh et al. (2022), and refer to this as the "edge effect" For L_B/L_T values near the edges (i.e., when the disk/bulge component completely dominates over the other component), precisely determining L_B/L_T is challenging for GaMPEN — in fact, this is difficult for any image analysis algorithm). In some of these cases, GaMPEN assigns almost the entirety of the light to the dominant component, resulting in the streaks seen at the edges of the figure. Poor structural parameter determination for galaxies with $L_B/L_T < 0.2$ and $L_B/L_T > 0.8$ have also been independently observed in other studies using different algorithms (e.g., Bretonniere et al., 2022; Hauler et al., 2022)

In order to mitigate this, GaMPEN users can choose to transform GaMPEN's quantitative predictions in the region $0.2 \geq L_B/L_T \geq 0.8$ (demarcated by shaded lines in Figure 4.19) to qualitative values such as "highly bulge-dominated" ($L_B/L_T \geq 0.8$) or "highly disk-dominated" ($L_B/L_T \leq 0.2$). We followed a similar procedure in Ghosh et al. (2022) and found the net accuracy of these labels to be $\gtrapprox 95\%$.

The top-middle panel of Figure 4.19 also shows that L_B/L_T residuals are higher for galaxies with smaller R_e. In other words, GaMPEN and GALFIT systematically disagree more for galaxies with smaller sizes—and this effect becomes more pronounced as the sizes become comparable to the seeing of the HSC-Wide Survey (g-band: 0.79 arcsec).

To comparatively evaluate the accuracy of GaMPEN and GALFIT specifically for smaller galaxies, we ran our GALFIT pipeline on a subset of simulated galaxies with $R_e \leq 2$ arcsec. Thereafter, we compared these results to the predictions made by GaMPEN for the same galaxies. As shown in Appendix 4.G, GaMPEN outperforms GALFIT for these smaller simulated galaxies. This provides preliminary evidence that GaMPEN's predictions on the smaller galaxies referred to in the previous paragraph are more accurate than those obtained using GALFIT. However, we would like to note that our simulated galaxies are semi-realistic and do not represent the full range of complexities present in real data. In a future publication, we will compare GaMPEN and GALFIT's performance on more realistic simulated galaxies generated using radiative transfer from hydrodynamical simulations.

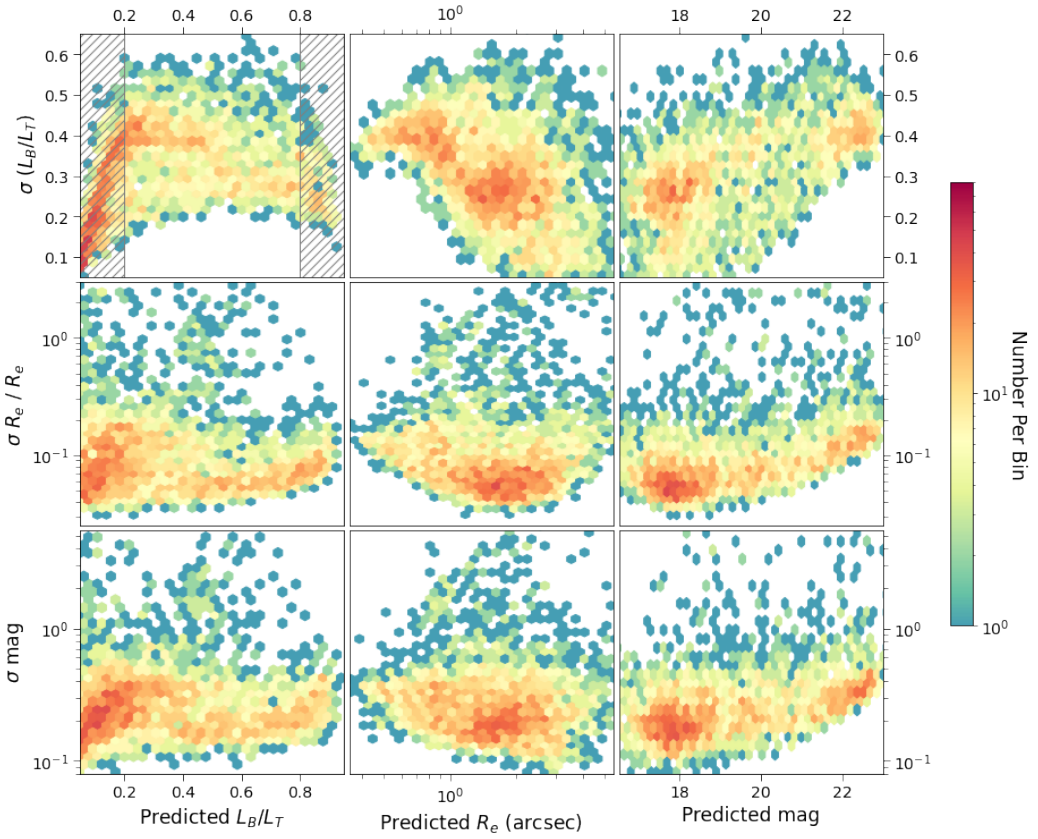


Figure 4.20: Uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the low-z test set. The σ for each parameter is defined as the width of the 68.27% confidence interval. Note that we plot fractional uncertainties for the radius in order to make the y-axis dimensionless for all three rows. The line-shaded region in the top-left panel shows the region where we recommend transforming quantitative L_B/L_T predictions to qualitative labels (see §4.6.3 for details).

4.6.4 Inspecting the Predicted Uncertainties

The primary advantage of a Bayesian ML framework like GaMPEN is its ability to predict the full posterior distributions of the output parameters instead of just point estimates. Thus, we would expect such a network to inherently produce wider distributions (i.e., larger uncertainties) in regions of the parameter space where residuals are higher.

Figure 4.20 shows the uncertainties for the three predicted parameters plotted against the predicted values for the low-z test set. We define the uncertainty predicted for each parameter as the width of the 68.27% confidence interval (i.e., the parameter interval that contains 68.27% of the most probable values of the predicted distribution; see Fig. 4.13). The y-axis of the middle row has been normalized so that all three panels show dimensionless fractional uncertainties. The distributions of uncertainties look very similar across all redshifts; thus, we have shown the uncertainty distributions

for the mid- and high-z bins in Appendix 4.F.

In Figure 4.19, we saw that GaMPEN's L_B/L_T residuals are higher for lower values of R_e. Here, we see that GaMPEN accurately predicts higher L_B/L_T uncertainties for lower values of R_e. This compensatory effect is what allows GaMPEN to achieve the calibrated coverage probabilities shown in Figure 4.16.

In the right column, we see that for all three parameters, the uncertainty in GaMPEN's predictions increase for fainter galaxies. This is in line with what we expect and had seen in Ghosh et al. (2022)—morphological parameters for fainter galaxies are more difficult to constrain compared to brighter galaxies and thus should have higher uncertainties.

The top left panel of Figure 4.20 shows that GaMPEN is reasonably certain of its predicted bulge-to-total ratio across the full range of values but appears slightly more certain when $L_B/L_T \leq 0.2$ or $L_B/L_T \geq 0.8$. We had also seen the same effect in Ghosh et al. (2022) with simulated galaxies. We found that the smaller uncertainties at the limits corresponded to the single-component galaxies, while for the double-component galaxies, the edge effect is less pronounced (see Figure 15 of Ghosh et al. (2022)). Here, we are seeing the same effect— GaMPEN's uncertainties at the edges are systematically lower for galaxies that can be described completely by only a disk or bulge component. These lower uncertainties also contribute to the higher residuals near the edges of L_B/L_T (as wider distributions would reduce the number of galaxies with high residuals at the edges). However, as can be seen from the top left panel of Figure 4.20, most of the galaxies with very low uncertainties lie in the region $0.2 \geq L_B/L_T \geq 0.8$, where we recommend transforming the quantitative L_B/L_T predictions to qualitative labels, as outlined in §4.6.3.

The results shown in this section outline the primary advantage of using a Bayesian framework like GaMPEN— even in situations where the network is not perfectly accurate, it can predict the right level of precision, allowing its predictions to be reliable and well-calibrated.

4.6.5 Comparing GaMPEN's Uncertainty Estimates to Other Algorithms

As §4.6.3 and §4.6.4 demonstrate, GaMPEN predicts well-calibrated uncertainties on our HSC data. Previous studies (e.g., Haussler et al., 2007) have shown that analytical estimates of errors from traditional morphological analysis tools like GALFIT or GIM2D (Simard et al., 2002) are smaller than the true uncertainties by $\geq 70\%$ for most galaxies.

Recently, Bretonniere et al. (2022) reported coverage probabilities obtained by five different light-profile fitting tools—Galapagos-2 (Hauler et al., 2022), Morfometryka (Ferrari et al., 2015), ProFit

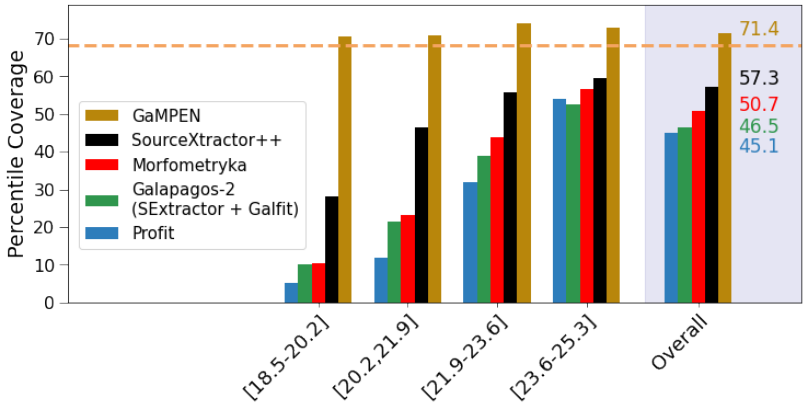


Figure 4.21: Percentile coverage probabilities for the 68.27% confidence interval obtained by GaMPEN on our HSC sample compared to coverage probabilities obtained by various light-profile fitting algorithms on simulated Euclid data (from Bretonniere et al., 2022). The rightmost set of bars shows the values calculated on the entire dataset, while the other sets display values calculated on sub-samples of galaxies with specific magnitude ranges (AB mag, shown on the x-axis). Compared to light-profile fitting tools, the uncertainties predicted by GaMPEN are better calibrated by $\sim 15-25\%$ overall and by as much as $\sim 60\%$ for the brightest galaxies.

(Robotham et al., 2017), and SourceXtractor++(Bertin et al., 2020)—on simulated Euclid data. The Euclid sample consisted of ~ 1.5 million galaxies ranging from $I_E \sim 15$ to $I_E \sim 30$, simulated at $0.''1$ /pixel. The simulations included analytic Sérsic profiles with one and two components, as well as more realistic galaxies generated with neural networks. As we did not have access to the simulated Euclid dataset, we could not test GaMPEN's performance on the same data. Instead, we compared GaMPEN's coverage probabilities for the HSC data set to those reported for the Euclid simulations. Although the latter is significantly different from our HSC sample, coverage probabilities reflect the ability of the predicted uncertainty to capture the true uncertainty and are not necessarily correlated with accuracy (which often varies across different data sets). Moreover, GaMPEN's predicted uncertainties can be tuned for specific data sets, as shown in Figure 4.10, which should only improve the GaMPEN outcome. Therefore, the results presented by Bretonniere et al. (2022) allow us to perform a preliminary comparison of GaMPEN's uncertainty prediction to that of other algorithms.

Figure 4.21 shows the 68.27% coverage probabilities achieved by GaMPEN on the HSC data compared to values for the four light profile fitting codes on the simulated Euclid dataset (averaging over the different structural parameters). When considering all galaxies, GaMPEN's uncertainties are at least $\sim 15-25\%$ better calibrated than the other algorithms. The differences are much larger for brighter galaxies, suggesting that the uncertainties predicted by these algorithms depend primarily on the flux of the object. In contrast, GaMPEN's uncertainty predictions remain well-

calibrated throughout and are better by as much as $\sim 60\%$ for the brightest galaxies. The severe under-prediction of uncertainties seems to hold true even for Bayesian codes like ProFit. It is likely that GaMPEN's robust implementation of aleatoric and epistemic uncertainties, along with a carefully selected transfer learning set spanning the entire magnitude range (see Fig. 4.9), allows it to predict well-calibrated uncertainties across a wide range of magnitudes.

4.7 Comparing Our Predictions to Other Catalogs

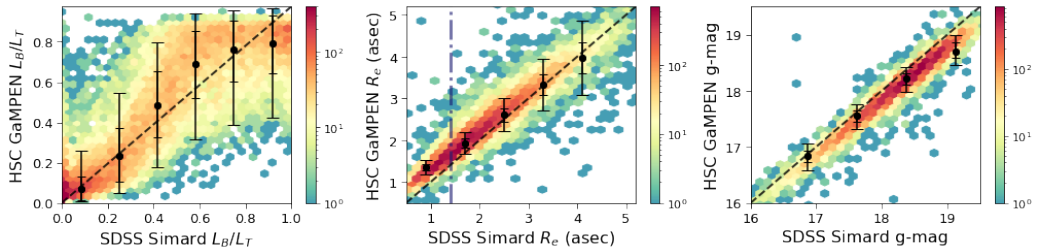


Figure 4.22: GaMPEN predictions plotted against values estimated by Simard et al. (2011) for a cross-matched sample of $\sim 20,000$ galaxies with $z < 0.2$ and $m < 19$. The density of points in each histogram is represented according to the logarithmic colorbar on the right. The black dots show the median y values in bins of equal width along the x-axis, with the error bars depicting the average 68.27% and 95.45% confidence intervals predicted by GaMPEN in that bin. The dash-dotted vertical line in the middle panel shows the median SDSS g-band seeing.

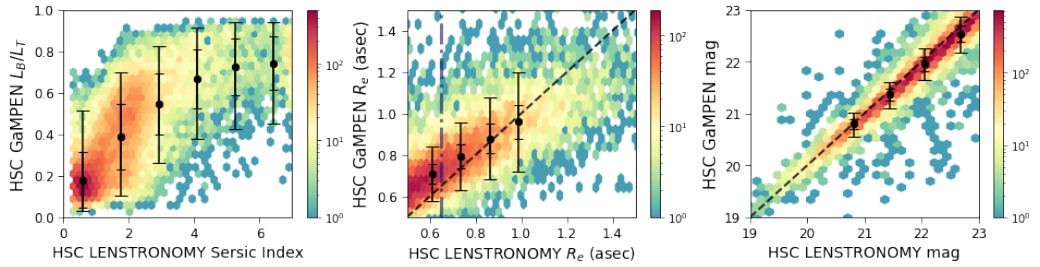


Figure 4.23: GaMPEN predictions plotted against values estimated by Kawinwanichakij et al. (2021) for a cross-matched sample of $\sim 200,000$ galaxies with $0.5 < z \leq 0.75$ and $m \leq 23$. Similar to Figure 4.22, we show the median y-values and associated error bars in bins of equal width along the x-axis. The dash-dotted vertical line in the middle panel shows the median HSC i-band seeing.

In §4.6.3, we compared GaMPEN's predictions to the values determined using our light-profile fitting pipeline. In order to further assess the reliability of GaMPEN's predictions, we now compare our predictions to two other morphological catalogs.

In Figure 4.22, we compare GaMPEN's predictions to the fits of Simard et al. (2011). Using an angular cross-match diameter of $0.''15$, we cross-matched our entire sample to that of Simard et al. (2011) to obtain an overlapping sample of $\sim 20,000$ galaxies. A large majority of these galaxies

have $m < 19.5$ and $z < 0.2$. We have included error bars in this figure to show the typical width of predicted distributions for different regions of the parameter space. We binned the x-axis of each parameter into bins of equal width and plotted the median y-value in each bin as a point, with the error bars showing the average 68.27% and 95.45% confidence intervals for all galaxies in that bin.

Despite the differences in SDSS and HSC imaging quality (with regards to pixel-scale and seeing), our results largely agree with those of Simard et al. (2011) within the ranges of predicted uncertainties. Using Spearman's rank correlation test, we obtain correlation coefficients of 0.87, 0.96, and 0.98 for L_B/L_T, R_e, and F, respectively. It is also interesting to note that the widths of the error bars account for almost the entire scatter in the distribution of points, showing the robustness of GaMPEN's uncertainty estimates.

In the L_B/L_T panel, there is a cluster of points near $L_B/L_T = 0.8$, and this is due to the edge-effect described in §4.6.3. GaMPEN and Simard et al. (2011)'s R_e predictions become discrepant (even when accounting for uncertainty) for $R_e \sim 1''$. This is not unexpected given that these R_e values are almost $0.''4$ smaller than the median SDSS seeing, depicted by the dash-dotted vertical line in the middle panel of Figure 4.22. The slight offset seen in the magnitude estimates is due to the differences in SDSS and HSC imaging quality, data-reduction pipeline, and filters (Kawanomoto et al., 2018), given that this discrepancy disappears when we compare our results to another catalog that uses HSC imaging (see below). Note that some of the scatter in Figure 22 can also be attributed to the fact that Simard et al. (2011) uses a morphology-determination pipeline that is significantly different from GaMPEN.

In Figure 4.23, we compare our results to that of Kawinwanichakij et al. (2021)[4], wherein the authors fitted single Sérsic light-profiles to 1.5×10^6 HSC i-band galaxies using LENSTRONOMY (**?**), a multipurpose open-source gravitational lens modeling Python package. Following Figure 4.22, we also show mean error bars in different bins of the fitted parameter in this figure. To compare results in the same band, we cross-matched our high-z sample with that of Kawinwanichakij et al. (2021) to obtain a sample of $\sim 200,000$ galaxies with $0.50 < z \leq 0.75$ and $i \leq 23$. Note that the magnitude discrepancy present in Figure 4.22 now disappears. GaMPEN and LENSTRONOMY radius measurements are also in agreement within the limit of uncertainties, with the mean trend deviating slightly from the $y = x$ diagonal at R_e values lower than the HSC median seeing, depicted by the dash-dotted vertical line in Figure 4.23. It is also important to note that while we define R_e as the radius that contains 50% of the total light of the combined bulge + disk profile, Kawinwanichakij

4. Catalog obtained from John D. Silverman, private communication.

et al. (2021) defines it as the semi-major axis of the ellipse that contains half of the total flux of the best-fitting Sérsic model. The correlation coefficients for R_e and magnitude are 0.87 and 0.96, respectively.

Since Kawinwanichakij et al. (2021) used single-component fits, we cannot compare our L_B/L_T predictions with their catalog. However, the left panel of Figure 4.23 shows the correlation between fitted Sérsic index (n) and measured bulge-to-total light ratio and can be used empirically to convert one measure into another. We find that, in line with expectations, higher Sérsic indices generally correspond to higher values of L_B/L_T. A large majority of galaxies with $n \leq 1.5$ have most of their light in the disk component. Galaxies with $n \geq 3$ have a large fraction of their light in the bulge component, although this fraction may vary from values as low as 40% to 95%. Note that these trends largely agree with what was reported by Simmons & Urry (2008).

4.8 Conclusions & Discussion

In this paper, we used GaMPEN, a Bayesian machine learning framework, to estimate morphological parameters (L_B/L_T, R_e, F) and associated uncertainties for ~ 8 million galaxies in the HSC Wide survey with $z \leq 0.75$ and $m \leq 23$. Our catalog is one of the largest morphological catalogs and is the first publicly available structural parameter catalog for HSC galaxies. It provides an order of magnitude more galaxies compared to the current state-of-the-art disk+bulge decomposition catalog of Simard et al. (2011) while probing four magnitudes deeper and having a higher redshift threshold. This represents an important step forward in our capability to quantify the shapes and sizes of galaxies and uncertainties therein.

We also demonstrated that by first training on simulations of galaxies and then utilizing transfer learning using real data, we are able to train GaMPEN using $< 1\%$ of our total dataset for training. This is an important demonstration that ML frameworks can be used to measure galaxy properties in new surveys, which do not have already-classified large training sets readily available. Our implemented two-step process provides a new framework that can be easily used for upcoming large imaging surveys like the Vera Rubin Observatory Legacy Survey of Space and Time, Euclid, and the Nancy Grace Roman Space Telescope.

We showed that GaMPEN's STN is adept at automatically cropping input frames and successfully removes secondary objects present in the frame for most input cutouts. Note that the trained STN framework can be detached from the rest of GaMPEN, and can be used as a pre-processing step in any image analysis pipeline.

By comparing GaMPEN's predictions to values obtained using light-profile fitting, we demonstrated that the posteriors predicted by GaMPEN are well-calibrated and GaMPEN can accurately recover morphological parameters. We note that the full computation of Bayesian posteriors in GaMPEN represents a significant improvement over estimates of errors from traditional morphological analysis tools like GALFIT, Galapagos, or ProFit. We demonstrated that while the uncertainties predicted by GaMPEN are well calibrated with $\lesssim 5\%$ deviation across a wide range of magnitudes, traditional light-profile fitting algorithms underestimate uncertainties by $\sim 15 - 60\%$ depending on the flux of the galaxy being analyzed. These well-calibrated uncertainties will allow us to use GaMPEN for the derivation of robust scaling relations (e.g., **??**) as well as for tests of theoretical models using morphology (e.g., **?**).

GaMPEN's residuals increase for smaller galaxies, but GaMPEN correctly accounts for that by predicting correspondingly higher uncertainties for these galaxies. GaMPEN's L_B/L_T residuals are also high when the bulge or disk component completely dominates over the other component. GaMPEN's quantitative L_B/L_T predictions for these galaxies ($0.2 \geq L_B/L_T \geq 0.8$) can be transformed into highly accurate qualitative labels with $\gtrsim 95\%$ accuracy. We did not detect a decline in GaMPEN's performance for fainter galaxies, and GaMPEN's residual patterns were also fairly consistent across all three redshift bins used in this study.

In order to assess the reliability of our catalog using a completely independent analysis, we compared GaMPEN's predictions to those of Simard et al. (2011) and Kawinwanichakij et al. (2021). Within the limit of uncertainties predicted by GaMPEN, our results agree well with these two catalogs. We noticed a slight discrepancy in the flux values determined using SDSS and HSC imaging, which can be attributed to the differences in imaging quality, data-reduction pipeline, and filters between the two surveys. Comparing our L_B/L_T predictions to Kawinwanichakij et al. (2021)'s measured Sérsic indices also allowed us to study the correlation between these two parameters, and this can also be used empirically to switch between these two parameters. We found that although galaxies with $n \geq 3$ have a large fraction of their light in the bulge component, this fraction can be anywhere between 40% to 95%.

Similar to other previous structural parameter catalogs (e.g., Simard et al., 2011; Tarsitano et al., 2018), we did not explicitly exclude merging galaxies from this catalog. We are currently working to incorporate a prediction flag within GaMPEN that will flag merging/irregular galaxies and highly blended sources so that they can be analyzed separately — we will demonstrate this in future publications. However, we would like to note that for the redshift ranges and magnitudes considered in this study, blending and mergers only constitute a limited fraction of the total sample. Using

the `m_(g|r|i)_blendedness_flag` and `m_(g|r|i)_blendedness_abs_flux` parameters available in HSC PDR2, we estimate that $\sim$ 10%, $\sim$ 4%, and $\sim$ 3% of galaxies in the low-z, mid-z, and high-z redshift bins, respectively, have nearby sources within the parent object footprint that can affect structural parameter measurements. For an extended description of these flags, we refer the interested reader to Bosch et al. (2018) and note that users can choose to ignore/exclude these galaxies from their analysis by setting different values/thresholds for these two parameters.

With this work, we are publicly releasing: a) GaMPEN's source code and trained models, along with documentation and tutorials; b) A catalog of morphological parameters for the $\sim$ 8 million galaxies in our sample along with robust estimates of uncertainties; c) The full posterior distributions for all $\sim$ 8 million galaxies. All elements of the public data release are summarized in Appendix 4.A.

Although we used GaMPEN here to predict L_B/L_T, R_e, and F, it can be used to predict any morphological parameter when trained appropriately. Additionally, although GaMPEN was used here on single-band images, we have tested that both the STN and CNN modules in GaMPEN can handle an arbitrary number of channels, with each channel being a different band. We defer a detailed evaluation of GaMPEN's performance on multi-band images for future work. Additionally, GaMPEN can also be used to morphologically analyze galaxies from other ground and space-based observatories. However, in order to apply GaMPEN on these data sets, one would need to perform appropriate transfer learning using data from the target dataset.

Finally, to give readers an estimate of GaMPEN's run-time, we note that once trained, it takes GaMPEN $\sim$ 1 millisecond on a GPU and $\sim$ 150 milliseconds on a CPU to perform a single forward pass on an input galaxy. These numbers, of course, change slightly based on the specifics of the hardware being used. However, as these numbers show, even with access to $\sim$ 1000 CPUs or $\sim$ 10 GPUs, GaMPEN can estimate full Bayesian posteriors for millions of galaxies in just a few days. Therefore, GaMPEN is fully ready for the large samples expected soon from Rubin-LSST and Euclid. Determinations of structural parameters along with robust uncertainties for these samples will allow us to characterize both morphologies as well as other relevant properties traced by morphology (e.g., merger history) as a function of cosmic time, mass, and the environment with unmatched statistical significance.

Chapter Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1715512.

CMU and AG would like to acknowledge support from the National Aeronautics and Space Administration via ADAP Grant 80NSSC18K0418.

AG would like to acknowledge the support received from the Yale Graduate School of Arts & Sciences through the Dean's Emerging Scholars Research Award.

AG would like to acknowledge computing grants received through the Amazon Cloud Credits for Research Program and the Yale Center for Research Computing (YCRC) Research Credits Program. AG would also like to thank the Yale Center for Research Computing and Yale Information Technology Services staff members and scientists, especially Robert Bjorson and Craig Henry, for their guidance and assistance in the vast amount of computation required for this project that was performed on Yale's Grace computing cluster and the Yale Astronomy compute nodes.

We would like to thank John D. Silverman and Lalitwadee Kawinwanichakij for making the Kawinwanichakij et al. (2021) catalog available to us. AG would like to thank Tim Miller and Imad Pasha for helpful discussions.

PN gratefully acknowledges support at the Black Hole Initiative (BHI) at Harvard as an external PI with grants from the Gordon and Betty Moore Foundation and the John Templeton Foundation.

ET acknowledges support from ANID through Millennium Science Initiative Program - NCN19_058, CATA-BASAL ACE210002 and FB210003, and FONDECYT Regular 1190818 and 1200495.

The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

This paper makes use of software developed for the Large Synoptic Survey Telescope. We thank the LSST Project for making their code available as free software at http://dm.lsst.org.

The Pan-STARRS1 Surveys (PS1) have been made possible through contributions of the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation under Grant No. AST-1238877, the University of Maryland, and Eotvos Lorand University (ELTE) and the Los Alamos National Laboratory.

Based, in part, on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center at National Astronomical Observatory of Japan.

HSC Morph Appendix

4.A Data Access

GaMPEN's source code is being publicly released along with the publication of this article. Along with the source code, we are also releasing documentation, tutorials, the trained models, the entire catalog of morphological predictions as well as the estimated PDFs for different morphological parameters of the ~ 8 million galaxies in our sample.

Links to the various components of the public data release mentioned above can be accessed at:-

- Summary – http://gampen.ghosharitra.com/
- Summary (Mirror of Above) –http://www.astro.yale.edu/aghosh/gampen.html
- Source Code – https://github.com/aritraghsh09/GaMPEN
- Documentation & Tutorials – https://gampen.readthedocs.io/en/latest/

We caution users to carefully read the Public Data Release Handbook available at https://gampen.readthedocs.io/en/latest/Public_data.html to understand various aspects of the data release before using the morphological catalogs produced as a part of this paper.

Along with the source code and trained models, this public data release also includes the various parameters obtained using the light profile fitting pipeline described in §4.5.

Since GaMPEN is a living code repository, which we expect to keep changing with time, we have created a "frozen" version of the code at the time of writing this article. This version is tagged as release `v0.1.0` in the above-mentioned Github repository and can also be referred to as Ghosh et al. (2023).

4.B Additional Details About Data

As outlined in §4.2, we used the `cleanflags_any` parameter available as part of HSC PDR2 to exclude objects flagged to have any significant imaging issues by the HSC pipeline. The various

triggers which contribute to the above flag, as well as their prevalence among the galaxies which we excluded from the analysis, are shown in Figure 4.24. As can be seen, $\sim$ 80% of the triggers are caused by cosmic ray hits/interpolated pixels.

In addition, the full SQL queries used to download the low-, mid-, and high-z data are shown in Listings 4.1, 4.2, and 4.3. Note that after downloading the data using these queries, we further excluded data based on the flags referred to in the previous paragraph, as well as the quality of photometric redshift estimates. For an extended description, please refer to §4.2. As noted in §4.8, users may additionally choose to use the various `blendedness` flags available in HSC PDR2 to further exclude merging/blended galaxies.

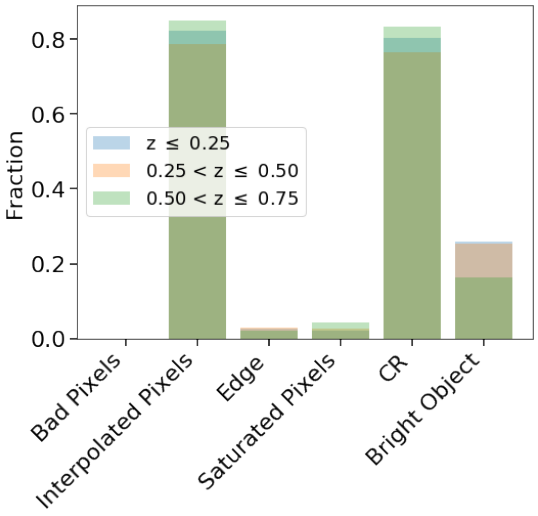


Figure 4.24: We exclude $\sim$ 20% of our downloaded galaxies due to different flags being triggered. The distribution of various flags that contribute to galaxies being excluded from our sample is shown in this figure. As can be seen, the large majority of exclusions are due to cosmic-ray hits (and hence, interpolated pixels).

```
SELECT
    object_id
      , ra
      , dec
FROM
    pdr2_wide.smallcat
    LEFT JOIN pdr2_wide.specz USING (object_id)
    LEFT JOIN pdr2_wide.photoz_mizuki USING (object_id)
    WHERE
    ((photoz_best > -1 AND photoz_best <= 0.25) OR (specz_redshift > -1 AND specz_redshift
     <= 0.25))
    AND (g_kronflux_mag < 23 OR g_cmodel_mag < 23)
```

```
    AND g_extendedness_value = 1
;
```

Listing 4.1: Low-z Sample SQL query

```
SELECT
    object_id
      , ra
      , dec
FROM
    pdr2_wide.smallcat
    LEFT JOIN pdr2_wide.specz USING (object_id)
    LEFT JOIN pdr2_wide.photoz_mizuki USING (object_id)
    WHERE
    ((photoz_best > 0.25 AND photoz_best <= 0.50) OR (specz_redshift > 0.25 AND
    specz_redshift <= 0.50))
    AND (r_kronflux_mag < 23 OR r_cmodel_mag < 23)
    AND r_extendedness_value = 1
;
```

Listing 4.2: Mid-z Sample SQL query

```
SELECT
    object_id
      , ra
      , dec

FROM
    pdr2_wide.smallcat
    LEFT JOIN pdr2_wide.specz USING (object_id)
    LEFT JOIN pdr2_wide.photoz_mizuki USING (object_id)
    WHERE
    ((photoz_best > 0.50 AND photoz_best <= 0.75) OR (specz_redshift > 0.50 AND
    specz_redshift <= 0.75))
    AND (i_kronflux_mag < 23 OR i_cmodel_mag < 23)
    AND i_extendedness_value = 1
;
```

Listing 4.3: High-z Sample SQL query

4.C Additional Details About GaMPEN

To provide readers a visual understanding of how GaMPEN's different architectural components are organized, Figure 4.25 shows a schematic diagram outlining the structure of both the STN and CNN in GaMPEN. For complete details on individual layers in GaMPEN, please refer to Ghosh et al. (2022).

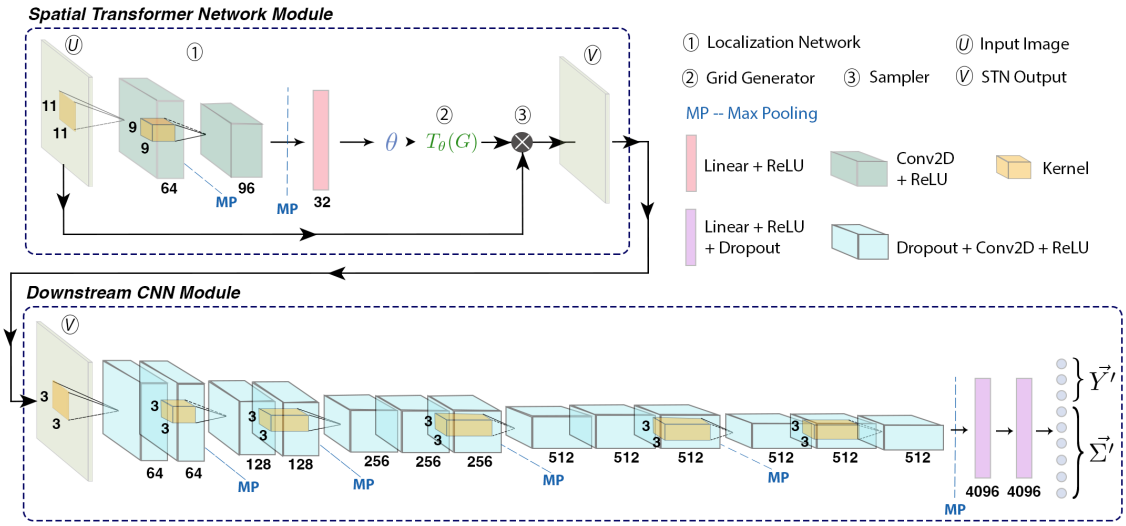


Figure 4.25: A schematic diagram of the Galaxy Morphology Posterior Estimation Network. GaMPEN's architecture consists of a downstream CNN module preceded by an upstream STN module. The CNN module empowers GaMPEN to estimate posterior distributions of galaxy morphology parameters. The upstream STN module trains without any extra supervision and learns to apply appropriate cropping transformations to the input image before passing it on to the CNN (for more details about these modules, see §4.3). The numbers below each layer refer to the number of filters/neurons in each layer. The yellow boxes inside the convolutional layers show the kernel and the number beside it refers to the corresponding kernel size. Only one kernel is shown per set of convolutional layers; all other layers in the set have kernels of the same size. Conv2D and ReLU refer to Convolutional Layers and Rectified Linear Units, respectively.

4.D Additional Details About Trained GaMPEN Models

As noted in §4.4, the training procedure of GaMPEN involves the tuning of various hyper-parameters (e.g., learning rate, batch size, etc.). These hyper-parameters are chosen based on the combination of the values that result in the best performance on the validation data set. The final chosen hyper-parameters for both the simulation trained GaMPEN models, as well as those fine-tuned on real data, are shown in Table 4.4. Please refer to §4.4 for more details on how we train these models.

Table 4.4: Tuned Values of Various Hyper-parameters

Model Name	Learning Rate	Momentum	L2 Regularization (λ)	Batch Size	Dropout Rate
Low-z Sim. Trained	5×10^{-7}	0.99	10^{-4}	16	7×10^{-4}
Mid-z Sim. Trained	5×10^{-7}	0.99	10^{-4}	16	7×10^{-4}
High-z Sim. Trained	5×10^{-7}	0.99	10^{-4}	16	4×10^{-4}
Low-z Final	5×10^{-8}	0.99	10^{-4}	16	4×10^{-4}
Mid-z Final	5×10^{-8}	0.99	10^{-4}	16	2×10^{-4}
High-z Final	5×10^{-6}	0.99	10^{-4}	16	2×10^{-4}

4.E Identifying Issues with our Light Profile Fitting Pipeline

We described in §4.5 a semi-automated pipeline that we used to determine the structural parameters for $\sim 60,000$ galaxies using light-profile fitting. After performing this analysis, we visually inspected the fits of a randomly selected sub-sample of ~ 300 galaxies from each redshift bin, to assess the quality of the fit.

The process of visually inspecting ~ 900 galaxies helped us to identify three failure modes of our fitting pipeline:-

1. for some galaxies, GALFIT assigned an extremely small axis ratio to one of the components leading to an unphysical lopsided bulge/disk
2. for some galaxies, the centroids of the two fitted components were too far away from each other
3. some galaxies had residuals that were too large

Note that for some galaxies, multiple failure modes were applicable. Examples of each of these failure modes are shown in Figure 4.26. We use a combination of different cuts on the fitted dataset to get rid of these failure modes – see §4.5 for an extended discussion.

4.F Additional Two-Dimensional Residual & Uncertainty Plots

Figures 4.27 and 4.28 show the distribution of residuals (difference between GaMPEN and GALFIT predictions) for the mid- and high-z bins.

Figures 4.29 and 4.30 show the uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the mid- and high-z bins.

4.G Comparing GaMPEN and GALFIT's performance on smaller simulated galaxies

As shown in §4.6.3, GaMPEN and GALFIT systematically disagree more for galaxies with smaller sizes. To ascertain their relative performance, specifically for smaller galaxies, we ran our GALFIT pipeline (described in §4.5) on ~ 5000 simulated galaxies from each redshift bin with $R_e \leq 2''$. These galaxies were chosen randomly from the testing set of GaMPEN – thus, none of them were used to train GaMPEN. Thereafter, we compared the results of this fitting procedure to the predictions made by GaMPEN on the same galaxies. The residuals obtained for both GaMPEN and GALFIT are shown in Figure 4.31.

The typical error for each of the parameters is given by $\tilde{\mu} \pm \sigma$, where $\tilde{\mu}$ and σ are the median and standard deviation of the residual distribution respectively. As shown in Figure 4.31, GaMPEN outperforms GALFIT for all three parameters across all redshift bins. This provides preliminary evidence that GaMPEN's predictions on the smaller galaxies referred to in §4.6.3 are more accurate than those obtained using GALFIT.

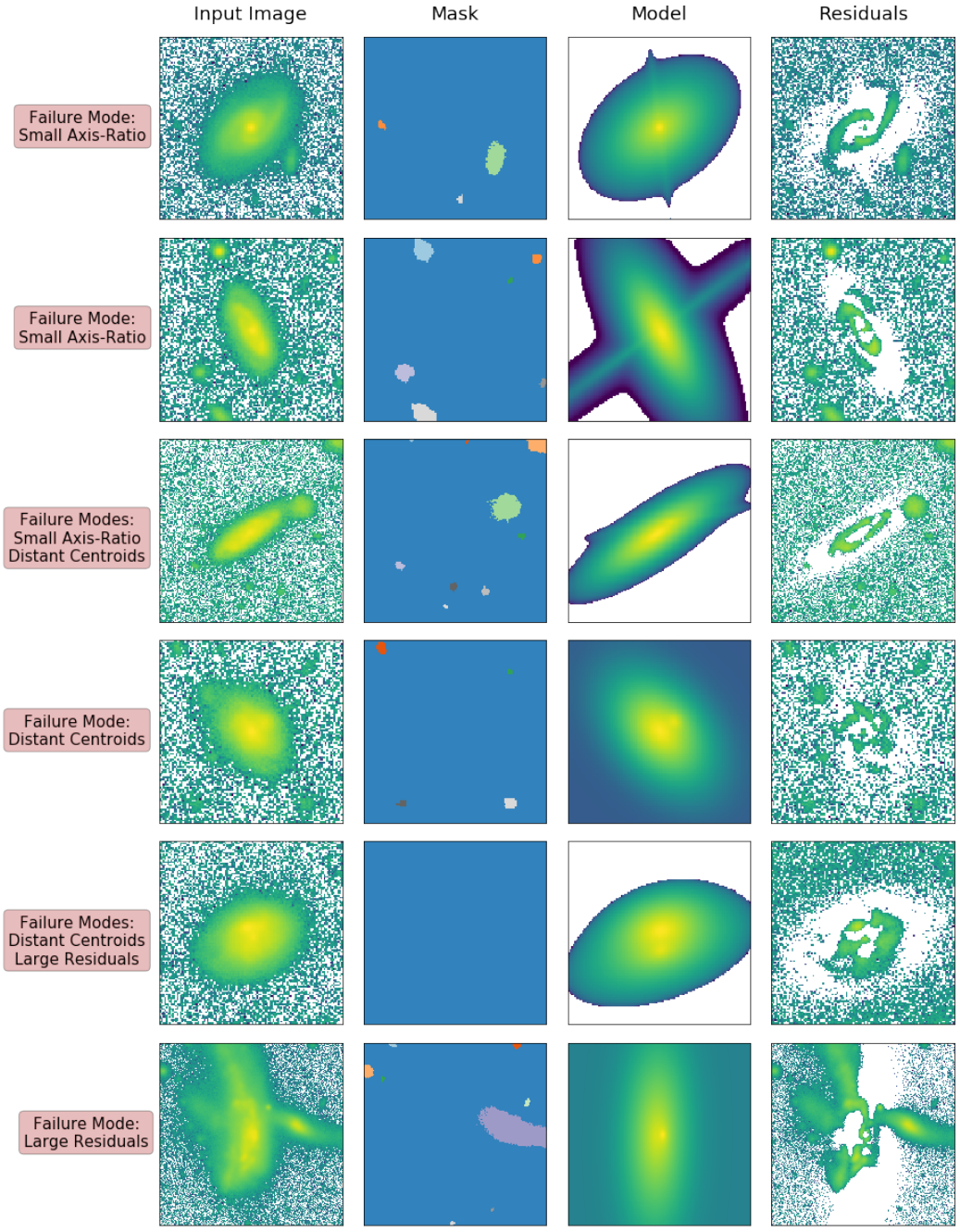


Figure 4.26: The different failure modes of the semi-automated light profile fitting code described in §4.5. From left to right, we show the input image, the mask generated by Source Extractor, the model generated by GALFIT, and the residuals.

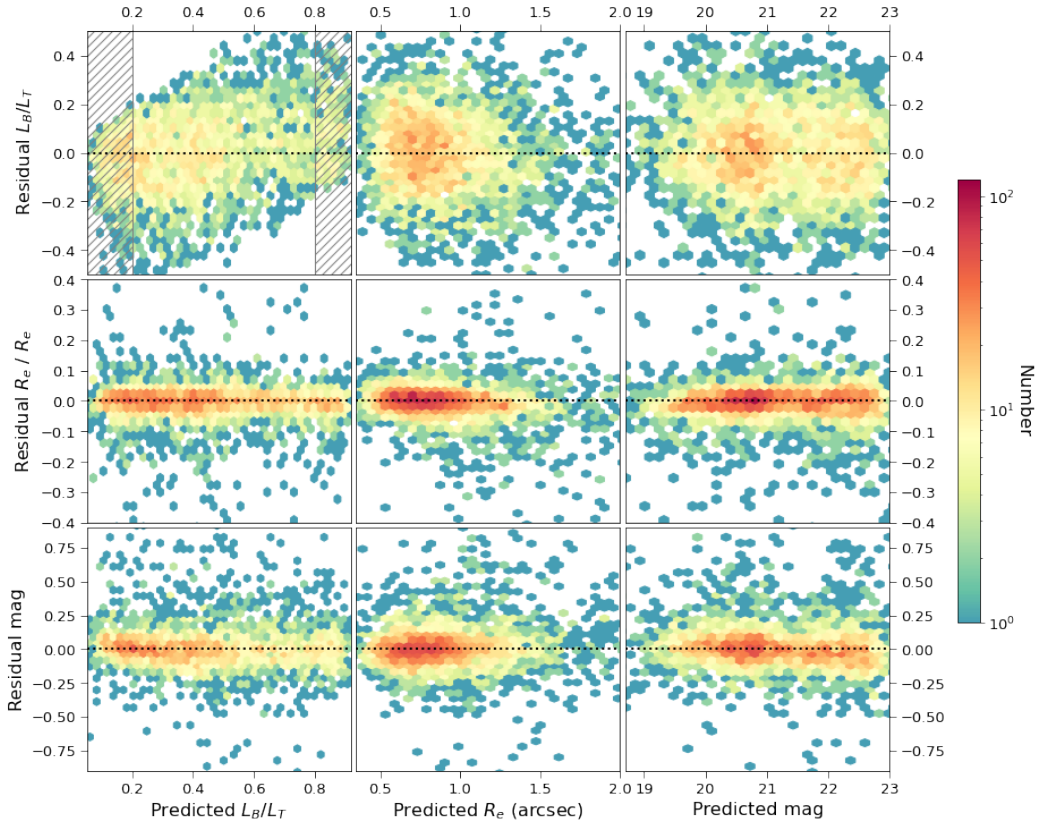


Figure 4.27: Residuals of the output parameters (difference between GaMPEN and GALFIT predictions) plotted against the values predicted by GaMPEN for all galaxies in the mid-z test set. See § 4.6.3 for details.

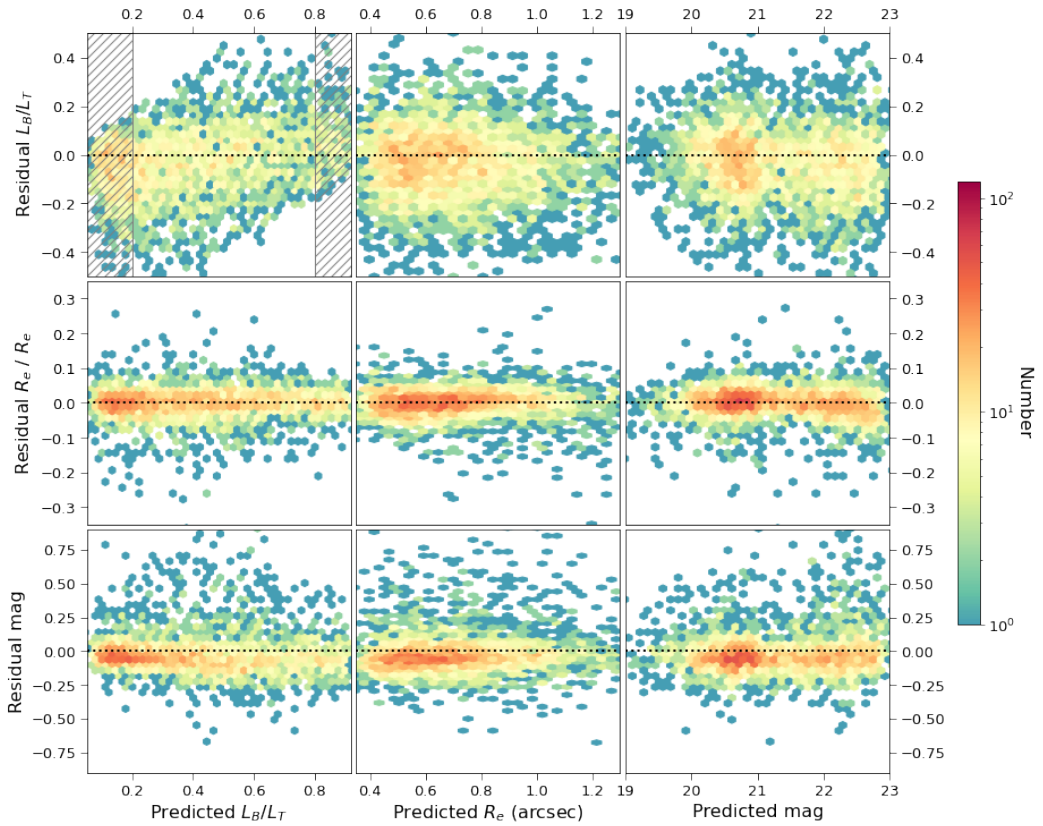


Figure 4.28: Residuals of the output parameters (difference between GaMPEN and GALFIT predictions) plotted against the values predicted by GaMPEN for all galaxies in the high-z test set. See § 4.6.3 for details.

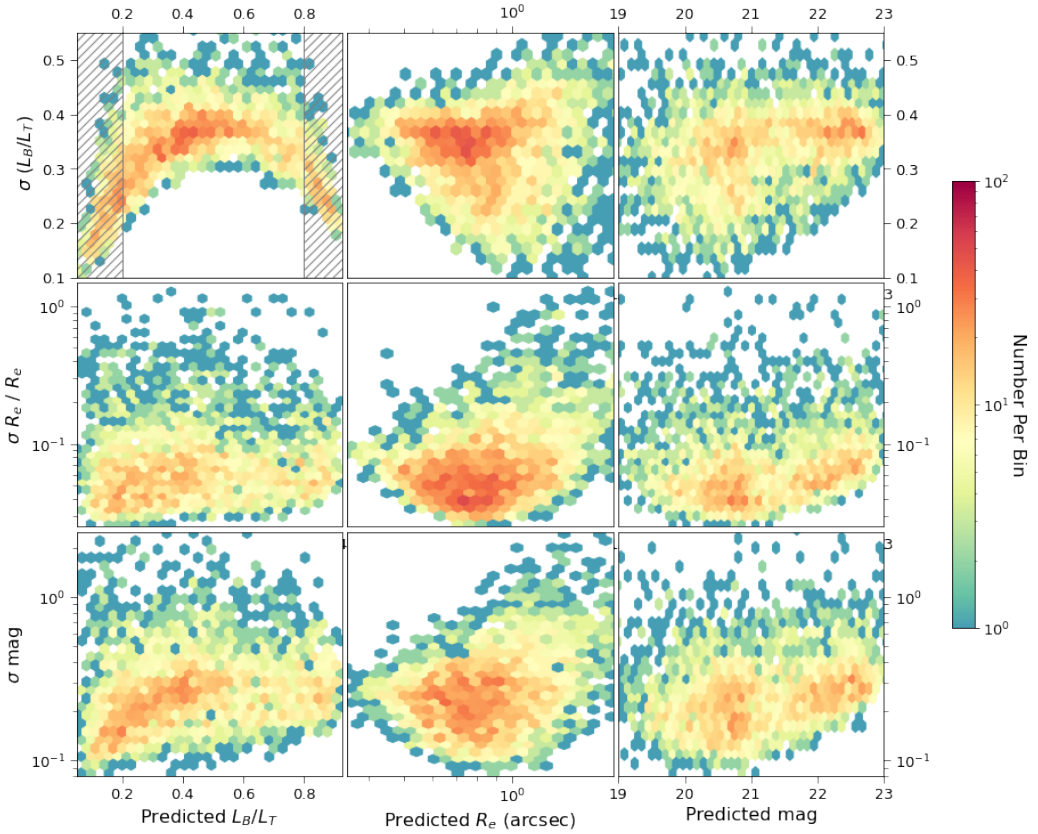


Figure 4.29: Uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the mid-z test set. The σ for each parameter is defined as the width of the 68.27% confidence interval. The line-shaded region in the top-left panel shows the region where we recommend transforming quantitative L_B/L_T predictions to qualitative labels. See §4.6.4 for details.

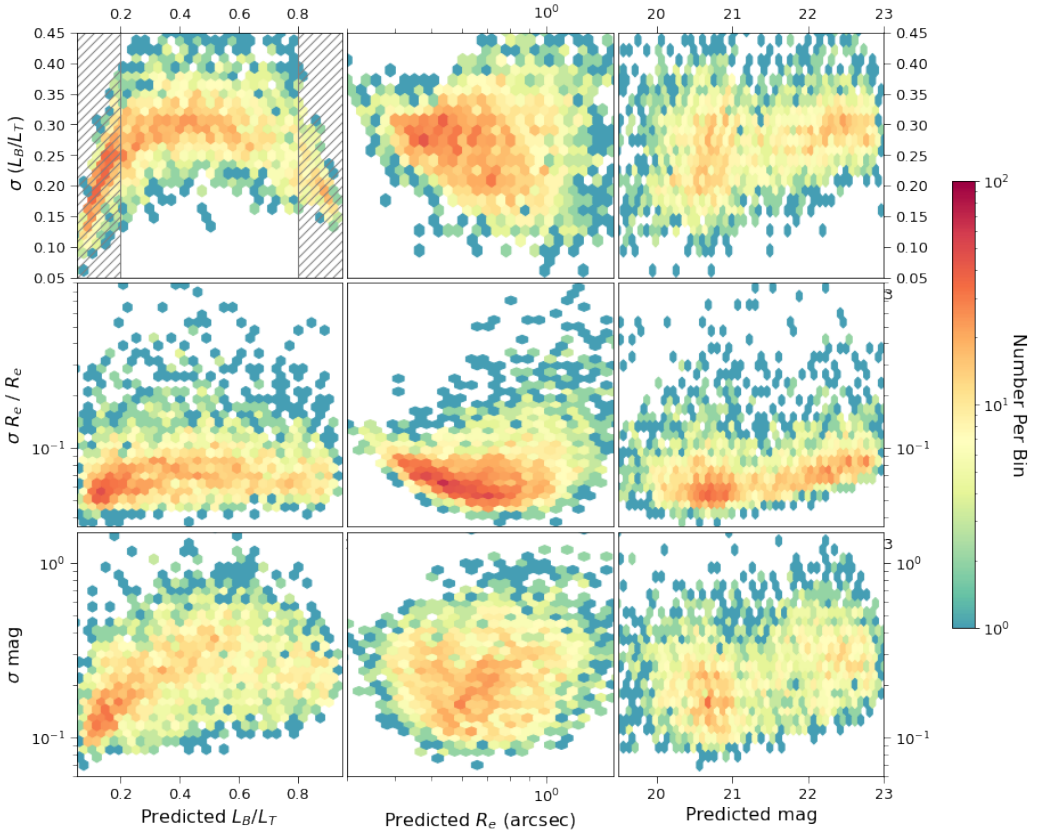


Figure 4.30: Uncertainties predicted by GaMPEN for each parameter plotted against the predicted values for the high-z test set. The σ for each parameter is defined as the width of the 68.27% confidence interval. The line-shaded region in the top-left panel shows the region where we recommend transforming quantitative L_B/L_T predictions to qualitative labels. See §4.6.4 for details.

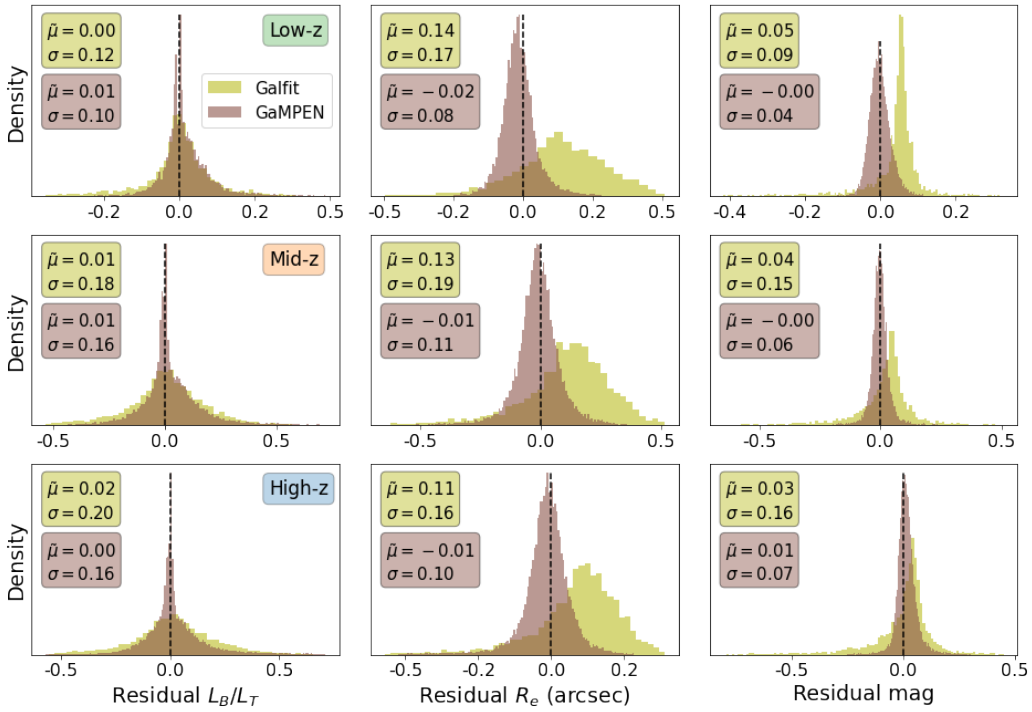


Figure 4.31: Distribution of residuals for ~ 5000 simulated galaxies in each redshift bin with $R_e \leq 2''$. These galaxies were selected randomly from the simulation testing set. The top, middle, and bottom rows show the results for the low-, mid-, and high-z bins respectively. The boxes in the top-left corner of each panel show the median ($\tilde{\mu}$) and standard deviation (σ) of each residual distribution. The dashed black vertical line marks $x = 0$.

Chapter 5

Variation of Morphology With Large Scale Density

This is the chapter based on ...

Chapter 6

Conclusions

This is the conclusions.....

Bibliography

Abraham, S., Aniyan, A. K., Kembhavi, A. K., Philip, N. S., & Vaghmare, K. 2018, Monthly Notices of the Royal Astronomical Society, 477, 894. https://academic.oup.com/mnras/article/477/1/894/4925012

Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K., & Turp, M. D. 2018, Monthly Notices of the Royal Astronomical Society, 479, 415. https://academic.oup.com/mnras/article/479/1/415/5017494

Aghanim, N., Akrami, Y., Ashdown, M., et al. 2018, Astronomy and Astrophysics, 641. http://arxiv.org/abs/1807.06209http://dx.doi.org/10.1051/0004-6361/201833910

Aihara, H., Armstrong, R., Bickerton, S., et al. 2018a, Publications of the Astronomical Society of Japan, 70. https://academic.oup.com/pasj/article/doi/10.1093/pasj/psx081/4494171

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018b, Publications of the Astronomical Society of Japan, 70

Aihara, H., Alsayyad, Y., Ando, M., et al. 2019, Publications of the Astronomical Society of Japan, 71, 114. https://academic.oup.com/pasj/article/71/6/114/5602617

—. 2021, Publications of the Astronomical Society of Japan, 71, 1. https://arxiv.org/abs/2108.13045v1

Alam, S., Albareti, F. D., Prieto, C. A., et al. 2015, The Astrophysical Journal Supplement Series, 219, 12

Baldry, I. K., Balogh, M. L., Bower, R. G., et al. 2006, Monthly Notices of the Royal Astronomical Society, 373, 469. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2006.11081.x

Baldry, I. K., Glazebrook, K., Brinkmann, J., et al. 2004, The Astrophysical Journal, 600, 681. http://stacks.iop.org/0004-637X/600/i=2/a=681

Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, Monthly Notices of the Royal Astronomical Society, 406, 342. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2010.16713.x

Bell, E. F., Wolf, C., Meisenheimer, K., et al. 2004, The Astrophysical Journal, 608, 752. http://stacks.iop.org/0004-637X/608/i=2/a=752

Bertin, E., Schefer, M., Apostolakos, N., et al. 2020 (Astronomical Society of the Pacific), 461

Bertinl, E. 1996, Astronomy and Astrophysics Supplement Series, 117, 393. http://aas.aanda.org/10.1051/aas:1996164

Binney, J., & Merrifield, M. 1998, Galactic astronomy (Princeton University Press), 796. https://press.princeton.edu/titles/6358.html

Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, The Astronomical Journal, 129, 2562. http://stacks.iop.org/1538-3881/129/i=6/a=2562

Blum, A., & Mitchell, T. 1998 (ACM Press), 92–100

Bosch, J., Armstrong, R., Bickerton, S., et al. 2018, Publications of the Astronomical Society of Japan, 70

Bradshaw, E. J., Almaini, O., Hartley, W. G., et al. 2013, Monthly Notices of the Royal Astronomical Society, 433, 194

Brammer, G. B., Whitaker, K. E., Dokkum, P. G. V., et al. 2009, Astrophysical Journal, 706, L173. http://stacks.iop.org/1538-4357/706/i=1/a=L173

Brammer, G. B., Dokkum, P. G. V., Franx, M., et al. 2012, Astrophysical Journal, Supplement Series, 200, 13. http://stacks.iop.org/0067-0049/200/i=2/a=13

Bretonniere, H., Kuchner, U., Huertas-Company, M., et al. 2022. http://arxiv.org/abs/2209.12907

Brinchmann, J., Charlot, S., White, S. D., et al. 2004, Monthly Notices of the Royal Astronomical Society, 351, 1151. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2004.07881.x

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, The Astrophysical Journal, 533, 682. http://stacks.iop.org/0004-637X/533/i=2/a=682

Cardamone, C. N., Urry, C. M., Schawinski, K., et al. 2010, Astrophysical Journal Letters, 721, L38. http://stacks.iop.org/2041-8205/721/i=1/a=L38

Cholesky, A.-L. 1924, Bulletin Geodesique, 2, 67

Chollet, F. 2021, Deep Learning with Python (Simon and Schuster)

Cool, R. J., Moustakas, J., Blanton, M. R., et al. 2013, The Astrophysical Journal, 767, 118

de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247. http://adsabs.harvard.edu/full/1948AnAp...11..247D

Dekel, A., Sari, R., & Ceverino, D. 2009a, The Astrophysical Journal, 703, 785

Dekel, A., Birnboim, Y., Engel, G., et al. 2009b, Nature 2009 457:7228, 457, 451. https://www.nature.com/articles/nature07648

Denker, J. S., & Lecun, Y. 1991, in (Morgan-Kaufmann). https://proceedings.neurips.cc/paper/1990/file/7eacb532570ff6858afd2723755ff790-Paper.pdf

Dieleman, S., Willett, K. W., & Dambre, J. 2015, Monthly Notices of the Royal Astronomical Society, 450, 1441. http://academic.oup.com/mnras/article/450/2/1441/979677/

Dodge, Y. 2008, Spearman Rank Correlation Coefficient, Springer New York. https://doi.org/10.1007/978-0-387-32833-1_379

Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, Monthly Notices of the Royal Astronomical Society, 401, 1429

Faber, S. M., Willmer, C. N. A., Wolf, C., et al. 2007, The Astrophysical Journal, 665, 265. http://stacks.iop.org/0004-637X/665/i=1/a=265

Ferrari, F., de Carvalho, R. R., & Trevisan, M. 2015, The Astrophysical Journal, 814, 55

Fevre, O. L., Cassata, P., Cucciati, O., et al. 2013, Astronomy & Astrophysics, 559, A14

Fukushima, K. 1980, Biological Cybernetics, 36, 193. http://link.springer.com/10.1007/BF00344251

Gal, Y., & Ghahramani, Z. 2016, in (PMLR), 1651–1660. https://proceedings.mlr.press/v48/gal16.html

Garilli, B., Guzzo, L., Scodeggio, M., et al. 2014, Astronomy & Astrophysics, 562, A23

Genzel, R., Burkert, A., Bouche, N., et al. 2008, The Astrophysical Journal, 687, 59

Ghosh, A., Rau, A., & Mishra, A. 2023, GaMPEN: First Stable Release, Zenodo. https://zenodo.org/record/7569024

Ghosh, A., Urry, C. M., Wang, Z., et al. 2020, The Astrophysical Journal, 895, 112. https://iopscience.iop.org/article/10.3847/1538-4357/ab8a47

Ghosh, A., Urry, C. M., Rau, A., et al. 2022, The Astrophysical Journal, 935, 138. https://iopscience.iop.org/article/10.3847/1538-4357/ac7f9e

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (MIT Press)

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, The Astrophysical Journal Supplement Series, 197, 35. http://stacks.iop.org/0067-0049/197/i=2/a=35?key=crossref.7b1f7638768d7cc6a63748c16607a9bf

Harrison, C. M. 2017, Nature Astronomy, 1, 165. https://www.nature.com/articles/s41550-017-0165

Hauler, B., Vika, M., Bamford, S. P., et al. 2022, Astronomy & Astrophysics, 664, A92

Haussler, B., McIntosh, D. H., Barden, M., et al. 2007, The Astrophysical Journal Supplement Series, 172, 615. https://iopscience.iop.org/article/10.1086/518836https://iopscience.iop.org/article/10.1086/518836/meta

Hoyle, B. 2016, Astronomy and Computing, 16, 34. https://www.sciencedirect.com/science/article/pii/S221313371630021X

Hsieh, B. C., & Yee, H. K. C. 2014, The Astrophysical Journal, 792, 102

Hubble, E. P. 1926, The Astrophysical Journal, 64, 321. http://adsabs.harvard.edu/doi/10.1086/143018

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, Astrophysical Journal, Supplement Series, 221, 8. http://stacks.iop.org/0067-0049/221/i=1/a=8

Ivezic, Z., Kahn, S. M., Tyson, J. A., et al. 2019, The Astrophysical Journal, 873, 111

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. 2015, Advances in Neural Information Processing Systems, 28, 2017 . http://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, Monthly Notices of the Royal Astronomical Society, 341, 33. https://academic.oup.com/mnras/article-lookup/doi/10.1046/j.1365-8711.2003.06291.x

Kawanomoto, S., Uraguchi, F., Komiyama, Y., et al. 2018, Publications of the Astronomical Society of Japan, 70, 66. https://academic.oup.com/pasj/article/70/4/66/5045845

Kawinwanichakij, L., Silverman, J. D., Ding, X., et al. 2021, The Astrophysical Journal, 921, 38. https://iopscience.iop.org/article/10.3847/1538-4357/ac1f21

Kim, E. J., & Brunner, R. J. 2017, Monthly Notices of the Royal Astronomical Society, 464, 4463. https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw2672

Kloek, T., & van Dijk, H. K. 1978, Econometrica, 46, 1

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, Astrophysical Journal, Supplement Series, 197, 36. http://stacks.iop.org/0067-0049/197/i=2/a=36

Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, Astrophysical Journal, Supplement Series, 192, 18. http://stacks.iop.org/0067-0049/192/i=2/a=18

Kormendy, J. 1979, The Astrophysical Journal, 227, 714

Kormendy, J., & Kennicutt, R. C. 2004, Annual Review of Astronomy and Astrophysics, 42, 603

Kriek, M., Dokkum, P. G. V., Labbe, I., et al. 2009, Astrophysical Journal, 700, 221. http://stacks.iop.org/0004-637X/700/i=1/a=221

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, ImageNet Classification with Deep Convolutional Neural Networks, ,

Kulis, B., Saenko, K., & Darrell, T. 2011 (IEEE), 1785–1792. http://ieeexplore.ieee.org/document/5995702/

Land, K., Slosar, A., Lintott, C., et al. 2008, Monthly Notices of the Royal Astronomical Society, 388, 1686. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2008.13490.x

Lecun, Y., Bengio, Y., & Hinton, G. 2015a, Nature, 521, 436. http://www.nature.com/articles/nature14539

—. 2015b, Nature, 521, 436. http://www.nature.com/articles/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proceedings of the IEEE, 86, 2278. http://ieeexplore.ieee.org/document/726791/

Li, W., Duan, L., Xu, D., & Tsang, I. W. 2014, IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 1134. http://ieeexplore.ieee.org/document/6587717/

Lilly, S. J., Brun, V. L., Maier, C., et al. 2009, The Astrophysical Journal Supplement Series, 184, 218

Lintott, C., Schawinski, K., Bamford, S., et al. 2011, Monthly Notices of the Royal Astronomical Society, 410, 166. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2010.17432.x

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008a, Monthly Notices of the Royal Astronomical Society, 389, 1179. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2008.13689.x

—. 2008b, Monthly Notices of the Royal Astronomical Society, 389, 1179. https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2008.13689.x

Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, Monthly Notices of the Royal Astronomical Society, 452, 2087

Lopes, P. A., Rembold, S. B., Ribeiro, A. L., Nascimento, R. S., & Vajgel, B. 2016, Monthly Notices of the Royal Astronomical Society, 461, 2559. https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw1497

Meert, A., Vikram, V., & Bernardi, M. 2013, Monthly Notices of the Royal Astronomical Society, 433, 1344

Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, The Astrophysical Journal Supplement Series, 225, 27

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, The Astrophysical Journal Supplement Series, 208, 5

Nielsen, M. A. 2015, Neural Networks and Deep Learning (Determination Press). http://neuralnetworksanddeeplearning.com/

Nishizawa, A. J., Hsieh, B.-C., Tanaka, M., & Takata, T. 2020, Publ. Astron. Soc. Japan, 1. https://arxiv.org/abs/2003.01511v2

Pan, S. J., & Yang, Q. 2010, IEEE Transactions on Knowledge and Data Engineering, 22, 1345. http://ieeexplore.ieee.org/document/5288526/

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, The Astronomical Journal, 124, 266. http://stacks.iop.org/1538-3881/124/i=1/a=266

Pentericci, L., McLure, R. J., Garilli, B., et al. 2018, Astronomy & Astrophysics, 616, A174

Powell, M. C., Urry, C. M., Cardamone, C. N., et al. 2017, The Astrophysical Journal, 835, 22. https://iopscience.iop.org/article/10.3847/1538-4357/835/1/22

Pozzetti, L., Bolzonella, M., Zucca, E., et al. 2010, Astronomy & Astrophysics, 523, A13. http://www.aanda.org/10.1051/0004-6361/200913020

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. 2009, Dataset Shift in Machine Learning (The MIT Press)

Racca, G. D., Laureijs, R., Stagnaro, L., et al. 2016, 99040O. http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2230762

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. 2019. http://arxiv.org/abs/1902.10811

Rix, H., Barden, M., Beckwith, S. V. W., et al. 2004, The Astrophysical Journal Supplement Series, 152, 163. https://iopscience.iop.org/article/10.1086/420885https://iopscience.iop.org/article/10.1086/420885/meta

Robotham, A. S. G., Taranu, D. S., Tobar, R., Moffett, A., & Driver, S. P. 2017, Monthly Notices of the Royal Astronomical Society, 466, 1513

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533. http://www.nature.com/articles/323533a0

Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, Monthly Notices of the Royal Astronomical Society, 440, 889. http://academic.oup.com/mnras/article/440/1/889/1749989/The-green-valley-is-a-red-herring-Galaxy-Zoo

Schmidhuber, J. 2015a, Neural Networks, 61, 85. http://dx.doi.org/10.1016/j.neunet.2014.09.003

—. 2015b, Neural Networks, 61, 85. http://dx.doi.org/10.1016/j.neunet.2014.09.003

Sellwood, J. A. 2014, Reviews of Modern Physics, 86, 1

Shimakawa, R., Tanaka, T. S., Toshikage, S., & Tanaka, M. 2021, Publications of the Astronomical Society of Japan, 73, 1575. https://academic.oup.com/pasj/article/73/6/1575/6406719

Silverman, J. D., Kashino, D., Sanders, D., et al. 2015, The Astrophysical Journal Supplement Series, 220, 12

Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, Astrophysical Journal, Supplement Series, 196, 11. ttp://stacks.iop.org/0067-0049/196/i=1/a=11

Simard, L., Willmer, C. N. A., Vogt, N. P., et al. 2002, The Astrophysical Journal Supplement Series, 142, 1. https://iopscience.iop.org/article/10.1086/341399https://iopscience.iop.org/article/10.1086/341399/meta

Simmons, B. D., & Urry, C. M. 2008, The Astrophysical Journal, 683, 644. http://stacks.iop.org/0004-637X/683/i=2/a=644

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, Monthly Notices of the Royal Astronomical Society, 464, 4420. http://dx.doi.org/10.1093/mnras/stw2587

Simonyan, K., & Zisserman, A. 2014, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. https://arxiv.org/abs/1409.1556v6

Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, Astrophysical Journal, Supplement Series, 214, 24. http://stacks.iop.org/0067-0049/214/i=2/a=24

Spergel, D., Gehrels, N., Breckinridge, J., et al. 2013. http://arxiv.org/abs/1305.5422

Strateva, I., Ivezic, Z., Knapp, G. R., et al. 2001, The Astronomical Journal, 122, 1861. http://stacks.iop.org/1538-3881/122/i=4/a=1861

Tanaka, M. 2015, The Astrophysical Journal, 801, 20

Tanaka, T. S., Shimakawa, R., Shimasaku, K., et al. 2022, Publications of the Astronomical Society of Japan, 74, 1

Taori, R., Dave, A., Shankar, V., et al. 2020, in (Curran Associates, Inc.), 18583–18599. https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

Tarsitano, F., Hartley, W. G., Amara, A., et al. 2018, Monthly Notices of the Royal Astronomical Society, 481, 2018

Tojeiro, R., Masters, K. L., Richards, J., et al. 2013, Monthly Notices of the Royal Astronomical Society, 432, 359. http://academic.oup.com/mnras/article/432/1/359/1128392/

Tuccillo, D., Huertas-Company, M., Decenciere, E., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 894. http://academic.oup.com/mnras/article/475/1/894/4725057

Walmsley, M., Lintott, C., Geron, T., et al. 2021, Monthly Notices of the Royal Astronomical Society, 509, 3966

Wel, A. V. D., Bell, E. F., Haussler, B., et al. 2012, Astrophysical Journal, Supplement Series, 203, 24. http://stacks.iop.org/0067-0049/203/i=2/a=24

Whitaker, K. E., Franx, M., Leja, J., et al. 2014, Astrophysical Journal, 795, 104. http://stacks.iop.org/0004-637X/795/i=2/a=104

Williams, R. J., Quadri, R. F., Franx, M., Dokkum, P. V., & Labbe, I. 2009, Astrophysical Journal, 691, 1879. http://stacks.iop.org/0004-637X/691/i=2/a=1879

Wilson, A. G. 2020, arXiv preprint arXiv:2001.10995. https://arxiv.org/abs/2001.10995v1

Wu, C., Wong, O. I., Rudnick, L., et al. 2019, Monthly Notices of the Royal Astronomical Society, 482, 1211

Wuyts, S., Schreiber, N. M. F., van der Wel, A., et al. 2011, The Astrophysical Journal, 742, 96. https://iopscience.iop.org/article/10.1088/0004-637X/742/2/96

York, D. G., Adelman, J., John E. Anderson, J., et al. 2000, The Astronomical Journal, 120, 1579. http://stacks.iop.org/1538-3881/120/i=3/a=1579

Zanisi, L., Huertas-Company, M., Lanusse, F., et al. 2021, Monthly Notices of the Royal Astronomical Society, 501, 4359

Zhu, Y., Chen, Y., Lu, Z., et al. 2011 (AAAI Publications). https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPaper/3671