Multi-Wavelength Structural Parameter Analysis for 8 Million Galaxies in the HSC Wide Survey

Milind Sarkar [1] Aritra Ghosh [2 3]

[1]Department of Physical Sciences, Indian Institute of Science Education and Research (IISER) Mohali
[2]DiRAC Instutute, University of Washington [3]Department of Astronomy, University of Washington

Why Galaxy Structure & Machine Learning?

The structural parameters of galaxies are connected to various other physical properties of galaxies (e.g., mass, star formation rate, kinematics, cosmic environment, etc.). These relationships have played a fundamental role in furthering our understanding of galaxy formation and evolution. Determining quantitative morphological measures for large numbers of galaxies at different wavelengths is fundamental to this process.

1. Machine learning (ML) techniques are becoming increasingly ubiquitous for structural parameter analysis in large imaging surveys.
2. GaMPEN is a publicly-available ML framework that can estimate full Bayesian posteriors (i.e., values & uncertainties) for millions of galaxies at scale.
3. In this work, we extend GaMPEN's capabilities to do a multi-band structural parameter analysis for ~ 8 million Hyper Suprime-Cam (HSC) Wide galaxies.
4. This is one of the largest multi-band structural parameter catalogs in astronomy, and we will be making it publicly available in the summer/fall of 2025.

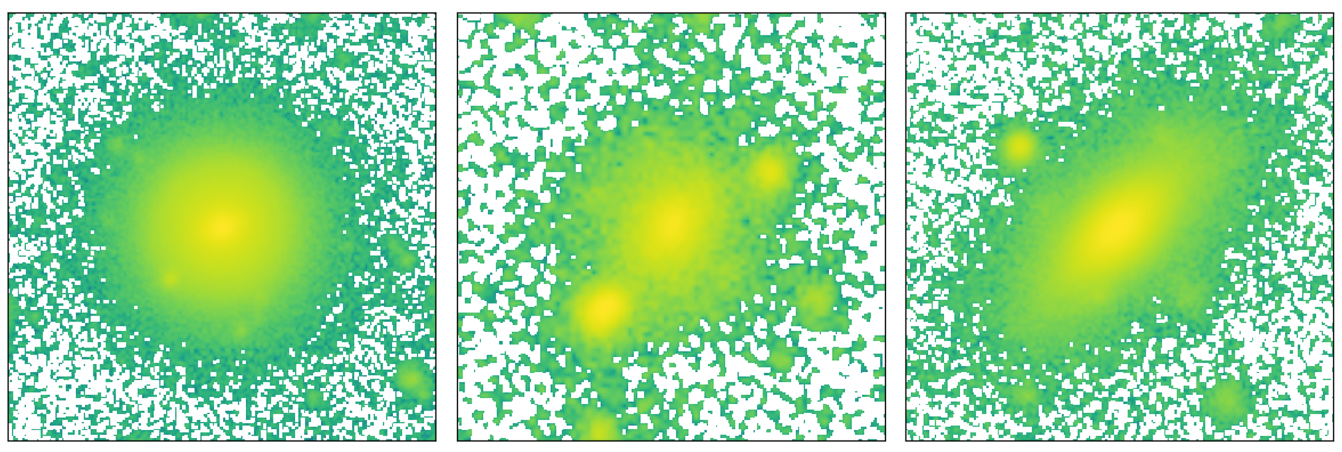


Figure 1. Diagram showing randomly selected galaxy cutouts for the lowest redshift bin in r-band.

Our Dataset

We apply the Galaxy Morphology Posterior Estimation Network (GaMPEN) to g,r,i-band imaging data from the Hyper Suprime-Cam (HSC) Subaru Strategic Program Public Data Release 2

1. We select galaxies from the PDR2 catalog using forced photometry on coadded HSC Wide images, applying the `extendedness_value` flag and a magnitude cut of $m < 23$ for reliable star-galaxy separation.
2. We use spectroscopic redshifts when available ($\sim 2.5\%$) and high-quality photometric redshifts otherwise, selecting galaxies till $z \leq 0.75$.
3. We exclude sources with imaging artifacts (e.g., cosmic ray hits) using the `cleanflags_any` parameter.
4. The final sample consists of ~ 8 million galaxies (~ 1M, ~ 3M, ~ 4M in low-z, mid-z, high-z bins), with magnitude and redshift distributions in Figure 2. Figure 1 shows three randomly selected r-band cutouts for the lowest redshift bin.

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

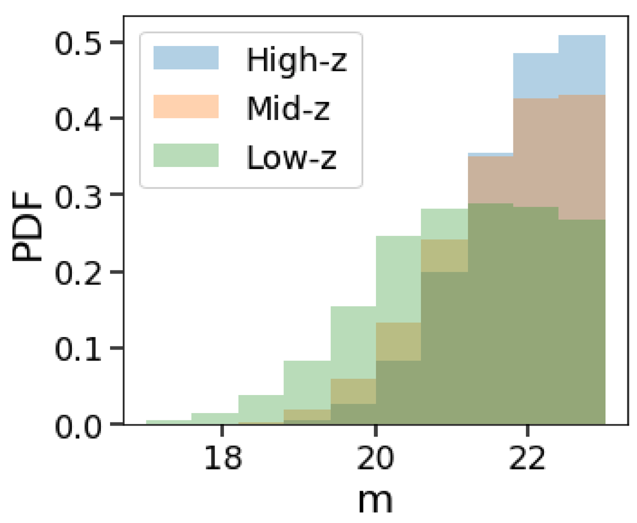
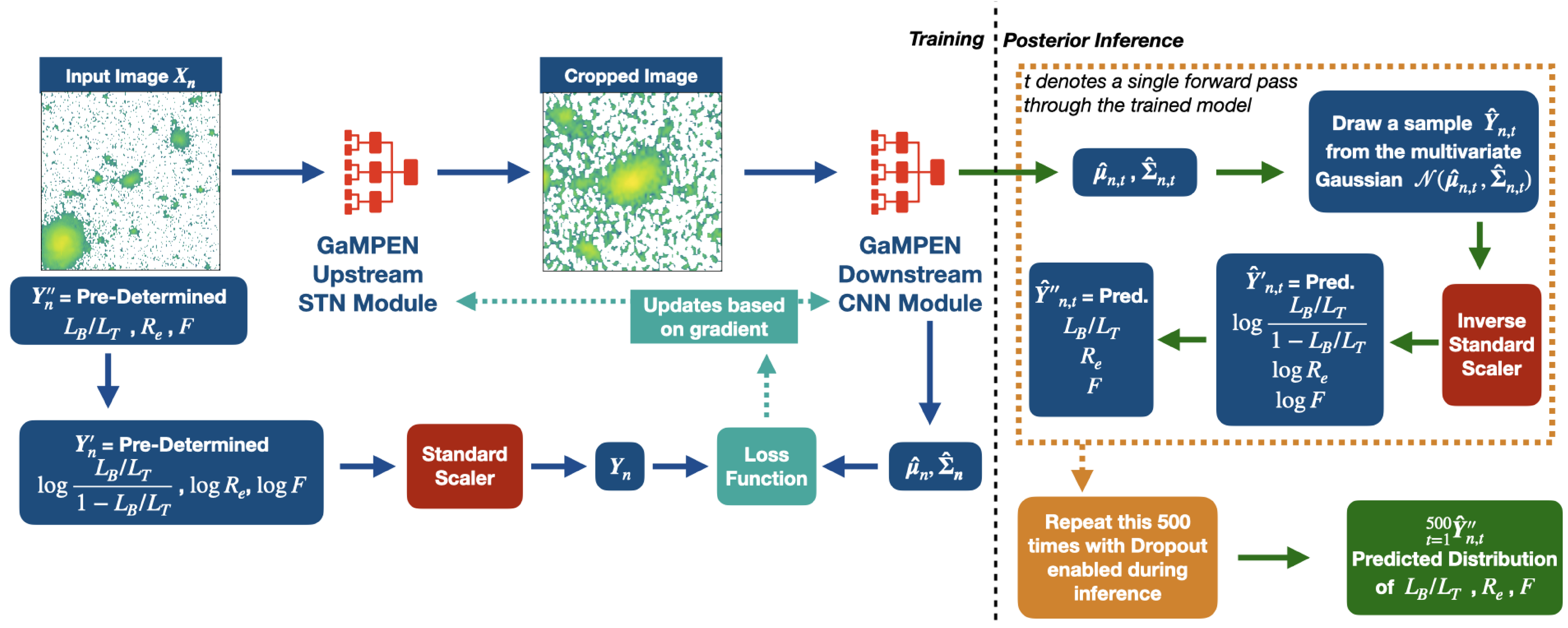


Figure 2. Diagram showing properties of final data

How GaMPEN Predicts Structural Parameters: Training and Posterior Inference in Action

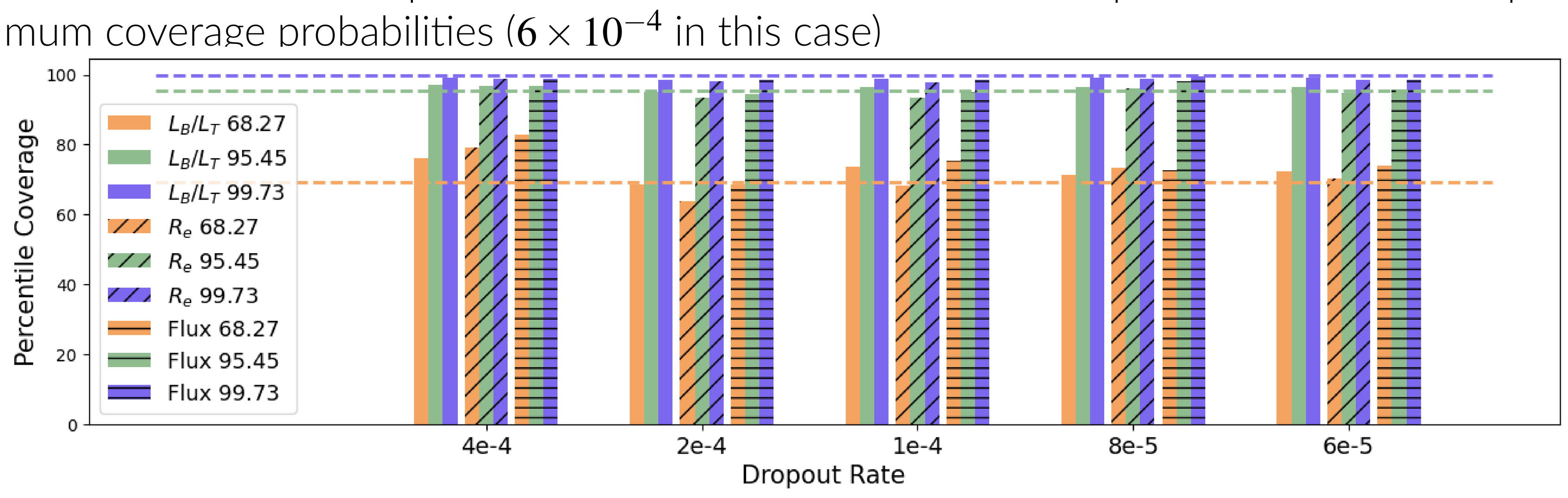


- Diagram outlining the training (*left*) and posterior inference (*right*) phases of the GaMPEN workflow.
- Training consists of feeding galaxies (with pre-determined parameter values) through the STN and CNN modules, minimizing the loss function using Stochastic Gradient Descent.
- For posterior inference, each image is passed 500 times through our trained model with dropout enabled (i.e., Monte Carlo dropout technique). We draw a sample from each of the 500 predicted multivariate Gaussian distributions, and the collection of these samples gives us the predicted posterior distribution.

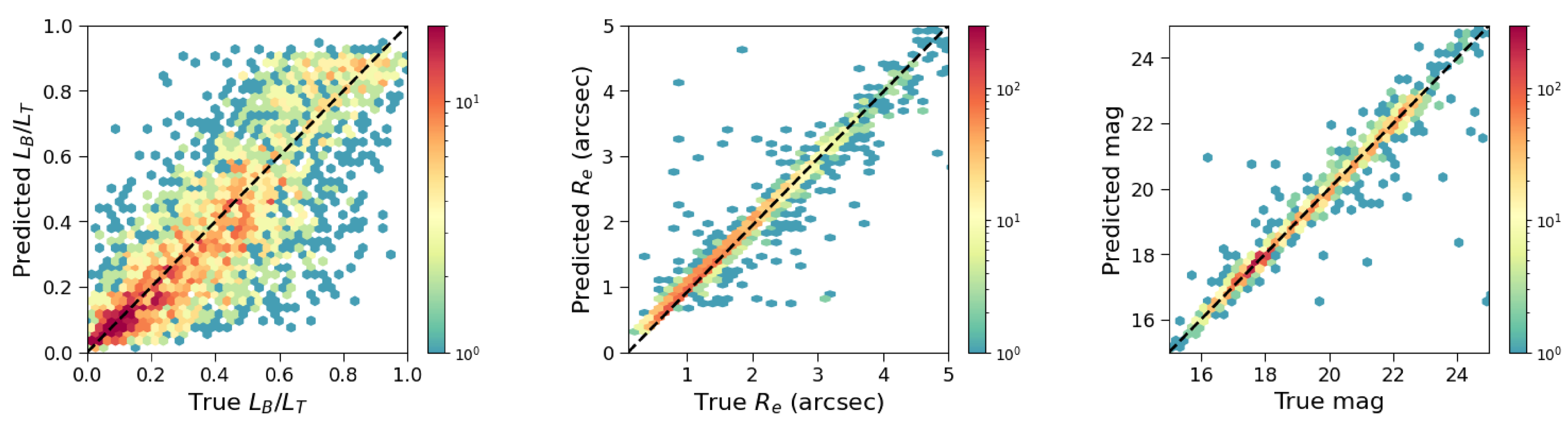
Image reproduced from Ghosh et al. 2023

True vs Predicted Parameters

GaMPEN's predicted uncertainties have been shown to be up to 60% more accurate compared to other light-profile fitting codes. Below, we show the calculated percentile coverage probabilities for different dropout rates for the low-z bin. This shows the percentage of total test examples where the ground-truth value lies within a particular confidence interval of the predicted distribution. We tune the dropout rate to achieve optimum coverage probabilities (6×10^{-4} in this case)

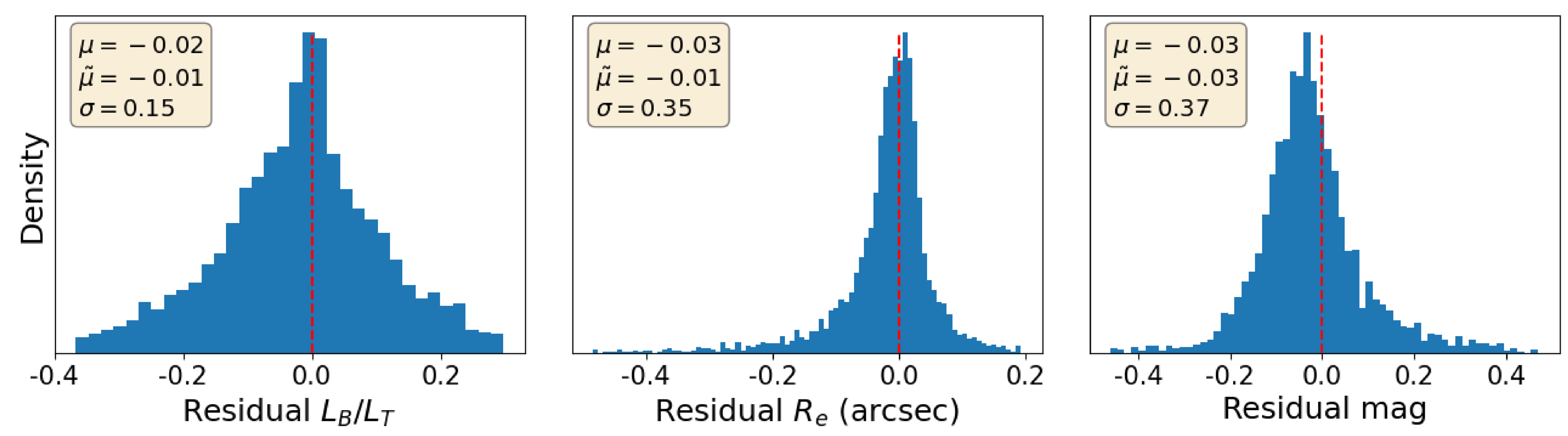


The most probable parameter values predicted by GaMPEN for all galaxies in the test set plotted against the values determined using GALFIT. Galaxies are plotted in hexagonal bins of roughly equal size, and the number of galaxies in each bin is represented according to the logarithmic colorbar to the right of each panel.



Residuals

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 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.



Key Takeaways

1. GaMPEN can accurately predict structural parameters along with robust uncertainties in multi-band imaging data.
2. We successfully trained GaMPEN using $< 1\%$ of the total dataset.
3. GaMPEN's uncertainty estimates deviate by $\lesssim 5\%$, while traditional methods underestimate uncertainties by as much as 60%.
4. **The presented catalogue, when completed, will be one of the largest multi-band optical catalogues in astronomy.** This catalogue along with trained models will be public by Summer '25.
5. GaMPEN Docs, Tutorials & Code ⇒