Assessing the Viability of Generative Modeling in Simulated Astronomical Observations

Patrick Janulewicz [1 2 3 4] **Laurence Perreault-Levasseur** [2 3 4 5 6 7] **Tracy Webb** [1 2]

Abstract

In this paper, we use methods for assessing the quality of generative models and apply them to a problem from the physical sciences. We turn our attention to astrophysics, where cosmological simulations are often used to create mock observations that mimic telescope images. These simulations and their mock observations are often slow and challenging to generate, inspiring some to use generative modeling to enhance the amount of data available to study. In this work, we add realism to simulated images of galaxy clusters and use probability mass estimation to assess their fidelity compared to reality. We find that the simulations show a degree of bias compared to real observations and suggest that researchers applying generative modeling to these systems should proceed with caution.

1. Introduction

With its great success in the past few years, generative modeling has attracted the attention of researchers from a wide range of fields. As performance becomes increasingly strong, many researchers from the natural sciences look to generative models to help accelerate the creation and acquisition of new data. One great strength of these algorithms is the ability to produce additional samples of data that may be difficult to obtain. This is particularly useful in the field of astronomy, where real observations require sophisticated telescopes, and simulated ones require significant computational resources. As a result, some researchers use generative modeling as a way of enhancing available data (Smith et al., 2022; Holzschuh et al., 2022).

In this work, we look at cosmological simulations, where physicists attempt to simulate astrophysical objects and their evolution. More specifically, we study simulated galaxy clusters from The Three Hundred project (Cui et al., 2018). Though the details of cosmological simulations are often complex, the premise is quite simple. Theorists begin by programming a box representing some large volume of physical space. They then assume some physical laws and some initial conditions, and they have the system evolve over time. The results can then be studied and compared with observations. If simulated results agree with observed data, we can have some degree of confidence in our assumptions regarding the physical laws that govern the system. As a result, theorists creating these simulations will look to observations to verify whether their models are correct. On the other hand, observers often look at simulations to explain the physics driving what they see from the telescope.

Though these cosmological simulations are of great use, they tend to be extremely expensive from a computational perspective. Not only can they take millions of CPU hours to run (Nelson et al., 2019), theorists must typically compromise between resolution and box volume. This causes the quantity of data from cosmological simulations to be limited and difficult to scale. For those interested in making inferences based on these simulations, it is natural to look for ways of enhancing the available data. Generative modeling is a promising avenue to achieve this goal. However, it is crucial to first assess the quality of these mock observations and to be cognisant of their limitations. If the simulated observations are biased compared to reality, then any distribution learned from them will likely also be biased. To test this, we study the underlying distribution of mock images and compare it to real data using the sample-based method of probability mass estimation (Lemos et al., 2024). More details on this method can be found in section 2.1.

We begin with simulated observations of galaxy clusters from The Three Hundred project simulation. The Three Hundred project contains 324 large galaxy clusters modeled with full-physics hydrodynamical re-simulations. To test the fidelity of these simulations, we create mock observa-

[1]Department of Physics, McGill University, Montréal, Canada [2]Trottier Space Institute, McGill University, Montréal, Canada [3]Mila - Quebec Artificial Intelligence Institute, Montréal, Canada [4]Ciela Institute, Montréal, Canada [5]Department of Physics, Université de Montréal, Montréal, Canada [6]Center for Computational Astrophysics, Flatiron Institute, New York, USA [7]Perimeter Institute for Theoretical Physics, Waterloo, Canada. Correspondence to: Patrick Janulewicz <patrick.janulewicz@mail.mcgill.ca>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

tions mimicking data from the Sloan Digital Sky Survey (SDSS) (York et al., 2000), focusing on the eighth data release (DR8) (Aihara et al., 2011). Because simulated observations are clean and do not contain the same noise patterns as SDSS, they require post-processing to make them similar in appearance to real data. Real data will generally contain foreground and background objects, as well as noise patterns such as Gaussian noise or Poisson noise. Furthermore, optical instruments respond to point sources and smooth the light via a point spread function (PSF). To mimic these effects, we use an approach for adding realism to simulated images (Bottrell et al., 2017) via a modified version of the RealSim Python package (Bottrell et al., 2019). We then compare these images to real observations of galaxy clusters obtained from the WHL12 galaxy cluster catalogue (Wen et al., 2012). Further details on The Three Hundred project simulations can be found in section 2.2, while further details on SDSS and the WHL12 catalogue can be found in section 2.3.

2. Data

2.1. Comparing two distributions with PQMass

PQMass is a sample-based method designed to assess the quality of generative models, though it can be applied to more general distributions as well. The algorithm takes as input two sets of samples and estimates the probability that the two sets are drawn from the same distribution. This comparison is done by dividing the space into non-overlapping regions and comparing the number of samples in each region. The number of regions n_R is an important parameter. By selecting a large value for n_R, the regions become smaller, allowing us to study the finer details in greater depth. However, this may require more samples and a longer runtime. In the unbiased case, PQMass will output a series of values (henceforth called χ^2_{PQM}) that will follow a chi-squared distribution with $n_R - 1$ degrees of freedom. In the biased case, these χ^2_{PQM} values will be larger.

2.2. Simulated observations

The Three Hundred project simulation used in this paper contains 324 different galaxy clusters that can be viewed at different ages via discrete snapshots in time. These times are expressed by their redshift z. To maximize available data while minimizing evolutionary effects, we select all four available snapshots in the redshift range of $z = 0.15$ to $z = 0.25$. We then randomly generate 10 unit vectors for a given snapshot and cluster, and we project the cluster along each of these axes using a modified version of EzGal (Mancone & Gonzalez, 2012). We create these images in 5 photometric bands which match the filters from the SDSS telescope. These filters are ultraviolet (u), green (g), red (r), near-infrared (i), and infrared (z). We also compare two different codes used to generate the simulations named Gadget-X (Rasia et al., 2015) and GIZMO-SIMBA (Cui et al., 2022), which each make unique assumptions about the physics of the system. The simulated images cover a physical distance of roughly 1 Mpc.

2.3. Real observations

Images of real galaxy clusters are obtained from the WHL12 cluster catalogue. The WHL12 catalogue is a collection of 132,684 galaxy clusters between $z = 0.05$ and $z = 0.8$. It contains celestial coordinates, redshift, and other information for each cluster. To ensure that our real data matches our simulated data, we limit ourselves to the range of $z = 0.15$ to $z = 0.25$ and only select images properly positioned on their central galaxy. Accounting for these differences, we are left with just over 10 000 cluster images.

We also obtain noise that we will use to add realism to the simulated images. To do this, we select images by randomly generating coordinates from areas in the BOSS survey (Dawson et al., 2012) of SDSS. We then cut out a 1 Mpc sky image in each of the five SDSS bands for a given coordinate. We convert each image into standard units of flux (erg/s/cm^2). A sample of a simulated cluster embedded into this SDSS noise can be seen in Figure 1.

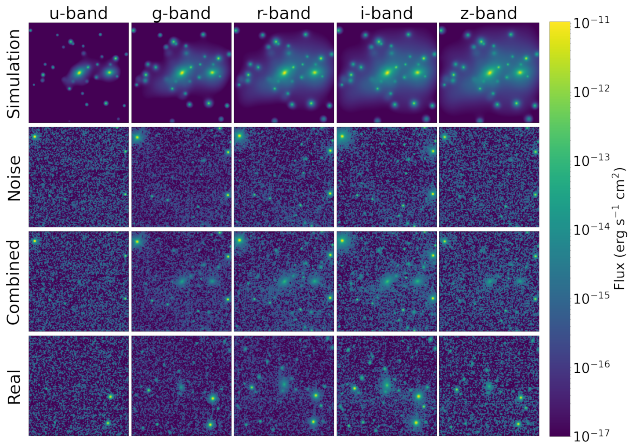

Figure 1. Image samples showing simulated GIZMO-SIMBA clusters in the first row, random SDSS sky images in the second row, and the addition of the two in the third row. A real galaxy cluster is shown in the fourth row for comparison. Columns represent different photometric SDSS bands. Pixel values are in units of flux and are scaled logarithmically.

3. Results

A crucial step before applying PQMass to our data is to show that the null test is passed. To perform the null test, we must split our dataset of real 5-channel galaxy cluster

images into two and compare random samples. Picking 49 degrees of freedom and samples of 1000 images, we repeat the test until we have 100 $\chi^2_{\rm PQM}$ values. We find that the images are in distribution with one another, with the mean $\chi^2_{\rm PQM}$ being very close to the expected value of 49. Results from this test are shown in Figure 2.

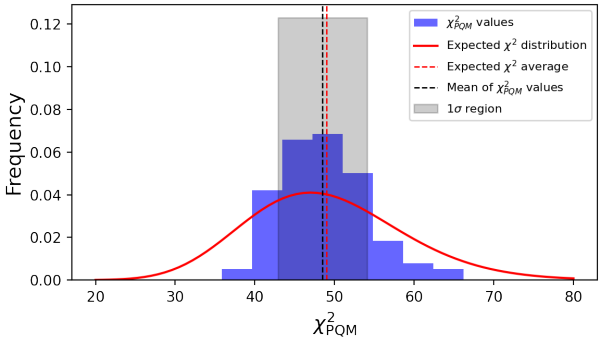

Figure 2. The PQMass null test with 49 degrees of freedom. The real 5-channel cluster images are split into two equal parts and tested against each other over 100 samples. The samples begin to form a χ^2 distribution with their mean approaching 49.

With the null test passed, we can now study different methods of adding realism to these images. We investigate three ways of doing this. The first is by adding a Gaussian sky designed to capture the Gaussian noise present in SDSS. We use RealSim's default standard deviation of 24.2 AB mag/arcsec^2. The remaining two ways both involve combining the clean simulated galaxy cluster images with real SDSS sky images. Because the different images do not necessarily have the same dimensions, we must resize them before combining. One way of doing this is simple interpolation using the Python Imaging Library (Umesh, 2012) or SciPy (Virtanen et al., 2020). Another way is the flux-conserving rebinning method described in RealSim, which is specifically designed to conserve the total flux in an image when resizing. We examine the performance of these three techniques by running a PQMass against real cluster images. We select 49 degrees of freedom and compute the mean $\chi^2_{\rm PQM}$ across 30 samples. PQMass tests are run on all 5 channels together to ensure that the entire distribution is covered at once. The results can be seen in Figure 3.

As expected, the clean images are out of distribution compared to real data. Interestingly, it appears that adding a Gaussian sky does little to bring the simulations closer to being in distribution. It also appears that adding SDSS sky without proper flux conservation performs even worse than the clean images, highlighting the importance of proper flux conservation. Adding SDSS sky with proper flux conservation greatly reduces the bias, bringing the two datasets closer

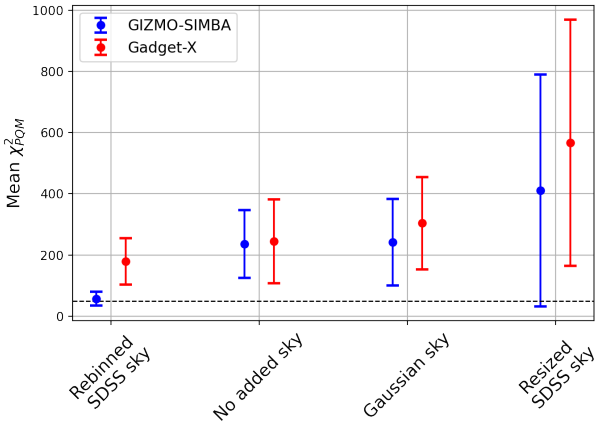

Figure 3. PQMass samples comparing simulations to real data with different methods of embedding the images into a mock sky. Rebinned SDSS sky refers to images resized with flux-conserving rebinning, while resized SDSS sky refers to images resized with standard interpolation techniques. Simulated images are tested against real images resized using the same method.

into distribution with each other. The GIZMO-SIMBA simulation code also appears to produce more faithful results than the Gadget-X code. We now investigate these results further.

Because we know from Figure 3 that the 5-channel images have relatively low bias when using proper rebinning, we now focus on each channel individually. We increase the number of PQMass regions to be 100, allowing us to study the finer details and subtler differences between the two distributions. Once again, we must verify that the null test is passed for every one of the five SDSS bands. To do this, we divide the 5-channel images and separate each channel individually. For a given channel, we compare two sets of 1000 samples to each other and repeat the trial 100 times. We then perform this test on all 5 SDSS bands. Looking at the results shown in Figure 4, we find all mean $\chi^2_{\rm PQM}$ values to be close to the expected value of 99.

We also test whether PSF convolution and the addition of Poisson noise can help reduce the bias of our data. We compare the flux-conserving SDSS sky with three other datasets. The first is with the addition of Poisson noise before adding the SDSS sky, and the second is by convolving with a PSF before adding the SDSS sky. In the third, we convolve with a PSF, add Poisson noise, and then add the SDSS sky. We use the default false PSF from RealSim, which corresponds to a Gaussian with a full width at half maximum (FWHM) of one arcsecond. This also serves as a rough approximation to what has been found in studies (Ross et al., 2011). Results are shown in Figure 5.

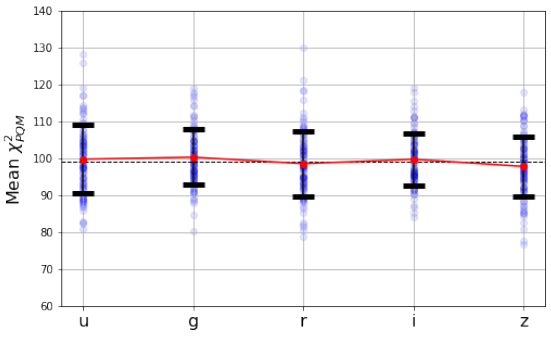

Figure 4. PQMass null test with 99 degrees of freedom for each SDSS band. Red points indicate the mean over 100 $\chi^2_{\rm PQM}$ values, with black error bars indicating the standard deviation. The distribution of individual $\chi^2_{\rm PQM}$ values is shown in blue.

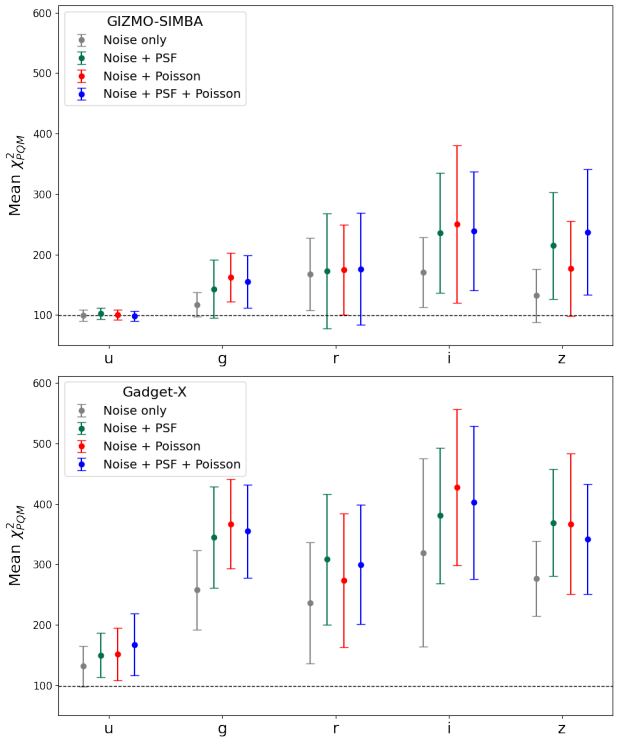

Figure 5. PQMass test with 99 degrees of freedom. Tests are run on individual image channels across different methods of adding realism such as PSF convolution and the addition of Poisson noise.

We find that Poisson noise and PSF convolution seem to have little effect on the the images. In fact, these changes may even bring the images slightly further out of distribution. The evidence on this, however, is not particularly strong, as many results agree within one standard deviation. We highlight two particularly noteworthy results from Figure 5.

The first noteworthy result is that we continue to see a larger bias in the Gadget-X clusters compared to the GIZMO-SIMBA ones, further supporting the trend seen in Figure 3. The second result is that certain bands appear to be more biased than others. For example, the u-band shows consistently small bias compared to the others. This bias then increases for the other bands and peaks at the i-band before it falls slightly at the z-band. This roughly corresponds with the relative visibility of the clusters in each band. If we look back at Figure 1, we find that the galaxy clusters are quite dim in the ultraviolet bands. They become more noticeable in the green, red, and near-infrared bands and appear to slightly dim in the infrared.

4. Discussion and conclusions

We begin by noting the extreme importance of properly replicating the sky in which simulated images are embedded. For those wishing to perform inference on their simulations by adding realism, we stress the importance of ensuring that the added noise is in distribution with real data. Failing to properly model this noise will result in highly biased results and will greatly hurt the ability to perform unbiased inference.

We also highlight that even after carefully adding realism, it may not be sufficient to eliminate bias. Even after we add noise and carefully handle flux conservation, the mock observations continue to be slightly out of distribution. We believe that the remaining bias can be the result of two possibilities. The first is that it comes from differences in data collection or post-processing. For instance, real PSFs are not necessarily Gaussian, which can be a source of uncertainty in our results. The simulated images also have smoothing applied to each particle before they are output. This means that the particles are not treated as point sources like they would be at a telescope, and that a PSF may not be particularly helpful. There is also the fact that SDSS galaxies may have false detections that would not occur in simulations. Another potential source is the fact that real observations were taken from a range of redshifts, while the simulations were only taken from the four discrete snapshots. In sum, there may be many possible systematic uncertainties; we believe that these contribute to some, but not all, of the differences.

The other possibility is that the simulated galaxy clusters are simply biased compared to real ones. This is reinforced by the fact that the two simulation codes showed noticeably different levels of fidelity. Recall that the Gadget-X simulations perform worse than the GIZMO-SIMBA ones;

this aligns with findings that the brightest cluster galaxies in GIZMO-SIMBA are more realistic than in Gadget-X (Cui et al., 2022). This is also supported by the fact that the the bands which showed the worst results are the ones in which the galaxy clusters are the most visible. When noise dominates, the distributions are closer together, indicating that the noise appears to be properly modeled. However, in bands where the galaxies are brighter, the results worsen, suggesting that the differences are coming from the clusters rather than the noise.

We reiterate that this experiment shows quantifiable differences when comparing simulated observations to real ones, and that this bias can be mitigated or worsened depending on the fidelity of the simulation. Though generative modeling is a promising way to scale the number of observations from cosmological simulations, this must be done with caution. Those interested in generating unseen observations based on simulated data must be aware of the limitations of this technique and act accordingly.

Acknowledgements

We acknowledge access to the theoretically modeled galaxy cluster data via The Three Hundred (https:// the300-project.org) collaboration. The simulations used in this paper have been performed in the MareNostrum Supercomputer at the Barcelona Supercomputing Center, thanks to CPU time granted by the Red Española de Supercomputación. As part of The Three Hundred project, this work has received financial support from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement number 734374, the LACEGAL project."

References

Aihara, H., Allende Prieto, C., An, D., et al. The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III. , 193(2):29, April 2011. doi: 10.1088/0067-0049/193/2/29.

Bottrell, C., Torrey, P., Simard, L., and Ellison, S. L. Galaxies in the Illustris simulation as seen by the Sloan Digital Sky Survey - I: Bulge+disc decompositions, methods, and biases. , 467(1):1033–1066, May 2017. doi: 10.1093/mnras/stx017.

Bottrell, C., Hani, M. H., Teimoorinia, H., et al. Deep learning predictions of galaxy merger stage and the importance of observational realism. , 490(4):5390–5413, December 2019. doi: 10.1093/mnras/stz2934.

Cui, W., Knebe, A., Yepes, G., et al. The Three Hundred project: a large catalogue of theoretically modelled galaxy clusters for cosmological and astrophysical applications. , 480(3):2898–2915, November 2018. doi: 10.1093/mnras/sty2111.

Cui, W., Dave, R., Knebe, A., et al. The Three Hundred project: The GIZMO-SIMBA run. *Monthly Notices of the Royal Astronomical Society*, 514(1):977–996, May 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac1402. URL http://dx.doi.org/10.1093/mnras/stac1402.

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. The Baryon Oscillation Spectroscopic Survey of SDSS-III. *The Astronomical Journal*, 145(1):10, December 2012. ISSN 1538-3881. doi: 10.1088/0004-6256/145/1/10. URL http://dx.doi.org/10.1088/0004-6256/145/1/10.

Holzschuh, B. J., O'Riordan, C. M., Vegetti, S., et al. Realistic galaxy images and improved robustness in machine learning tasks from generative modelling. *Monthly Notices of the Royal Astronomical Society*, 515(1):652–677, May 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac1188. URL http://dx.doi.org/10.1093/mnras/stac1188.

Lemos, P., Sharief, S., Malkin, N., Perreault-Levasseur, L., and Hezaveh, Y. PQMass: Probabilistic Assessment of the Quality of Generative Models using Probability Mass Estimation, 2024.

Mancone, C. L. and Gonzalez, A. H. Ezgal: A flexible interface for stellar population synthesis models. *Publications of the Astronomical Society of the Pacific*, 124(916):606–615, June 2012. ISSN 1538-3873. doi: 10.1086/666502. URL http://dx.doi.org/10.1086/666502.

Nelson, D., Pillepich, A., Springel, V., et al. First results from the TNG50 simulation: galactic outflows driven by supernovae and black hole feedback. *Monthly Notices of the Royal Astronomical Society*, 490(3):3234–3261, August 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2306. URL http://dx.doi.org/10.1093/mnras/stz2306.

Rasia, E., Borgani, S., Murante, G., et al. Cool Core Clusters from Cosmological Simulations. , 813(1):L17, November 2015. doi: 10.1088/2041-8205/813/1/L17.

Ross, A. J., Ho, S., Cuesta, A. J., et al. Ameliorating systematic uncertainties in the angular clustering of galaxies: a study using the SDSS-III. , 417(2):1350–1373, October 2011. doi: 10.1111/j.1365-2966.2011.19351.x.

Smith, M. J., Geach, J. E., Jackson, R. A., et al. Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2):1808–1818, January 2022. ISSN

1365-2966. doi: 10.1093/mnras/stac130. URL http://dx.doi.org/10.1093/mnras/stac130.

Umesh, P. Image processing in python. *CSI Communications*, 23, 2012.

Virtanen, P., Gommers, R., Oliphant, T. E., et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wen, Z. L., Han, J. L., and Liu, F. S. A catalog of 132,684 clusters of galaxies identified from Sloan Digital Sky Survey III. *The Astrophysical Journal Supplement Series*, 199(2):34, March 2012. ISSN 1538-4365. doi: 10.1088/0067-0049/199/2/34. URL http://dx.doi.org/10.1088/0067-0049/199/2/34.

York, D. G., Adelman, J., Anderson, John E., J., et al. The Sloan Digital Sky Survey: Technical Summary. , 120(3): 1579–1587, September 2000. doi: 10.1086/301513.