
GAMMA: Galactic Attributes of Mass, Metallicity, and Age Dataset

Ufuk Çakır

Interdisciplinary Center
for Scientific Computing,
University of Heidelberg,
Im Neuenheimer Feld 205,
D-69120 Heidelberg

ufuk.cakir@stud.uni-heidelberg.de

Tobias Buck

Interdisciplinary Center
for Scientific Computing,
University of Heidelberg,
Im Neuenheimer Feld 205,
D-69120 Heidelberg

tobias.buck@iwr.uni-heidelberg.de

Abstract

We introduce the **GAMMA** (Galactic Attributes of Mass, Metallicity, and Age) dataset, a comprehensive collection of galaxy data tailored for Machine Learning applications. This dataset offers detailed 2D maps and 3D cubes of 11 727 galaxies, capturing essential attributes: stellar age, metallicity, and mass. Together with the dataset, we publish our code to extract any other stellar or gaseous property from the raw simulation suite to extend the dataset beyond these initial properties, ensuring versatility for various computational tasks. Ideal for feature extraction, clustering, and regression tasks, **GAMMA** offers a unique lens to explore galactic structures using computational methods and is a bridge between astrophysical simulations and the field of scientific machine learning (ML). As a first benchmark, we applied Principal Component Analysis (PCA) to this dataset. We find that PCA effectively captures the key morphological features of galaxies with a small number of components. We achieve a dimensionality reduction by a factor of ~ 200 (~ 3 650) for 2D images (3D cubes) with a reconstruction accuracy below 5%.

1 Motivation

Galactic structures have long been one of the cornerstones of astrophysics. As computational methods, particularly ML, gain prominence, the need for datasets that bridge astrophysical simulations and computational techniques is growing. The **GAMMA** dataset provides a well-curated collection of 2D and 3D galaxy images, which captures crucial physical attributes for understanding galaxy morphology: stellar age, metallicity, and mass. It combines well-understood physical relations in the data with explicit and controlled image generation processes such as noise level or point-spread function. As such, this dataset not only offers a unique opportunity for feature extraction, clustering, and regression tasks, but also paves the way for a deeper understanding of galactic structures through computational lenses. The dataset is available on Zenodo.¹ Additionally, we publish all of our code on GitHub², which allows one to easily extract images from raw simulations or create higher resolution realizations of the data. In this way, we facilitate the community to go beyond the exploratory benchmarks presented in this work.

¹<https://zenodo.org/records/8375344>

²<https://github.com/ufuk-cakir/GAMMA>

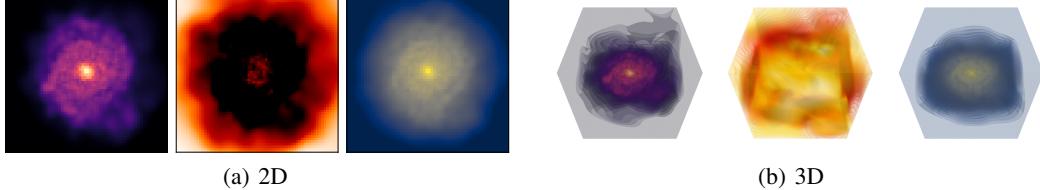


Figure 1: Each row showcases a sample galaxy represented in three stellar maps: metallicity, age, and mass from left to right. The pixel values have been normalized to the range of $[0, 1]$, where brighter pixels correspond to higher values. Note that in the stellar age maps, dark pixels represent young stars. The three-dimensional plots were generated using the Python package *plotly* [15].

2 Related Work

Hydrodynamical simulations of galaxy formation such as IllustrisTNG [1], EAGLE [2] CAMELS [3] or zoom-in models such as AURIGA [4], FIRE [5], VINTERGATAN [6] or NIHAO [7, 8] generate extensive, multidimensional datasets that include various physical properties of galaxies. Similarly, the GALAXY ZOO initiative [9] has collected a large set of real multicolor galaxy images that have been classified in a citizen science project. These datasets serve as valuable resources for rigorously and quantitatively studying galaxy morphology. Combined with modern ML techniques, they enable the development of interpretable generative models for galaxy morphology [e.g. 10].

For our initial benchmark with the **GAMMA** dataset, we employ PCA, which has been widely used in various fields for interpretable dimensionality reduction and feature extraction [11], including the decomposition of face images into "eigenfaces" [12] and its adaptation to galaxy classification, where the basis vectors of the reduced space are termed "eigengalaxies" [13]. In the context of galaxy morphology, [14] recently used PCA to transform a set of galaxy images into a space where "closeness" corresponds to visual similarity. Expanding upon these prior studies, we explore PCA's utility in jointly modeling the distribution of mass, metallicity, and stellar age in both two and three dimensions. This approach condenses the high-dimensional distribution of galaxies into a lower-dimensional space that retains essential morphological features and facilitates interpretable analysis.

3 GAMMA: A dataset for stellar mass, metallicity, and age maps

We extracted galaxy data from the publicly available IllustrisTNG-100 simulation suite³ [16, 1, 17] specifically from snapshot 99 (redshift $z = 0$) within the mass range $10^{9.5} M_{\odot}/h < M_{\star} < 10^{13} M_{\odot}/h$. We excluded galaxies with `SubhaloFlag = 0` and those without particles on the edges, producing a total of $N = 11727$ galaxies. Particles with negative `GFM_StellarFormationTime` are masked out, as they correspond to wind-phase gas cells and hence are not relevant for our analysis. For uniform rotation, each galaxy undergoes face-on rotation based on the eigenvectors of the moment of inertia tensor and a further tilt correction using PCA with two components to align the semi-major axis of the galaxy with the x-axis. Using the *SWIFTsimIO* library's render module [18] the 2D and 3D images were generated, properly reflecting the physical and spatial resolution by employing the smoothing length parameter. These images have 64^2 and 64^3 pixel resolutions and depict particle mass, mass-weighted metallicity and stellar age, spanning five stellar half-mass radii in all directions. Pixel values are normalized to $[0, 1]$ and adjusted with 25th percentile clipping to standardize intensities and reduce noise-induced variance. The final images for all fields are stored in HDF5 format, whereas the `SubhaloId` and stellar mass values for each galaxy are located in the `attributes` subgroup. In Fig. 1 we show a typical face-on spiral galaxy in the dataset.

4 Benchmark: Principal Component Analysis on Images and Data Cubes

We perform PCA in $\mathbb{R}^{N \times d}$ with $N = 11727$ and $d = 3 \cdot 64^2$ in 2D ($d = 3 \cdot 64^3$ in 3D) using the *scikit-learn* [19] Python package. PCA projects our high-dimensional data onto a lower-dimensional

³<https://www.tng-project.org>

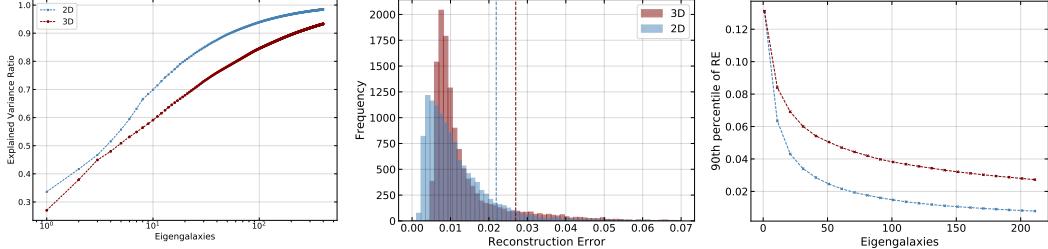


Figure 2: *Left:* Cumulative explained variance ratio (EVR) for up to 400 eigengalaxies: Achieving the same explained variance requires significantly more eigengalaxies in 3D compared to 2D. To surpass an 90% EVR, ~ 60 (215) eigengalaxies are needed in 2D (3D). *Middle:* Reconstruction error (RE) for fixed-dimensionality reduction on 60 (215) eigengalaxies in 2D (3D). The dashed line represents the 90% quantile. Impressively, 90% of all images exhibit a RE below 0.022 (0.027). *Right:* 90th percentile of RE as a function of the number of eigengalaxies. Reconstruction is a strong function of eigengalaxies, and already 15 (60) eigengalaxies lead to a RE better than 5% in 2D (3D).

subspace spanned by the top n eigengalaxies $\lambda_n \in \mathbb{R}^{N \times d}$, preserving most of the variance. We evaluate the effectiveness of the PCA model in preserving information through the Explained Variance Ratio (EVR). The cumulative sum of the EVR for up to 400 components is shown in Fig. 2. 60 (215) eigengalaxies account for $\sim 90\%$ of the total variance in 2D (3D), demonstrating substantial information retention despite a significant reduction in dimensionality by a factor of 205 (3641). With as few as 16 eigengalaxies, major features such as spiral arms are well reconstructed. Increasing this number results in more accurate and finer image details.

To assess the robustness of the model, we perform a 400-fold cross-validation with 60 components on a 75% random subsample of the data. The resulting EVR distribution has a mean of 0.9087 and a standard deviation of only 0.0003, demonstrating excellent robustness. This indicates that our dataset size is sufficient to encompass various morphologies and using 60 eigengalaxies robustly explains 90% of the variance between morphologies in 2D.

We define the reconstruction error (RE) as the fractional difference in pixel values between the PCA representation, $\hat{\mathbf{I}}$, and the original image, \mathbf{I} as

$$\text{RE} = \frac{\sum_{k=1}^d (I_k - \hat{I}_k)^2}{\sum_{k=1}^d I_k}$$

where \hat{I}_k and I_k represent the k -th elements (pixel) of the reconstructed and original vectors (images), respectively. Using 60 (215) eigengalaxies, we calculate the RE and show its distribution in the middle panel of Fig. 2 for the 2D (blue) and 3D (red) cases. Remarkably, 90% of all images have a RE below 0.022 in 2D (0.027 in 3D), highlighting the accuracy of our PCA model. The right panel of Fig. 2 shows how the RE scales with the number of eigengalaxies. In the 2D case, as few as ~ 15 eigengalaxies produce a RE below 0.05 for 90% of all images. In contrast, the 3D case requires ~ 60 eigengalaxies to achieve a similar RE. These results are consistent with the EVR findings and demonstrate that a small number of eigengalaxies accurately describe galaxy morphology.

One of the significant advantages of PCA over more complex ML methods such as neural networks is its inherent interpretability. PCA provides a linear decomposition of the original image space, where each individual component can be interpreted as an image representing a specific morphology. We have visualized how eigengalaxies contribute to different morphologies. Notably, vastly different eigengalaxies contribute to the reconstruction of spiral and triaxial galaxies, as expected due to their distinct morphologies. For spiral galaxies, eigengalaxies with pronounced power along diagonal edges are essential for reconstructing the spiral arm feature. Conversely, for triaxial galaxies, eigengalaxies with nearly point-symmetry or bar-like features play a more prominent role in the reconstruction. This qualitative analysis underscores the potential of PCA decomposition in describing the galaxy morphology. Fitting galaxies with our PCA model enables the study of general morphological trends across different galaxies through the PCA scores of contributing eigengalaxies. For instance, it becomes straightforward to quantify and assess whether a galaxy exhibits bar-like, spiral-like, or centrally concentrated mass, age, or metallicity distributions. This interpretability is valuable for galaxy classification and similarity searches.

To assess whether the lower-dimensional image space encodes meaningful morphological information,

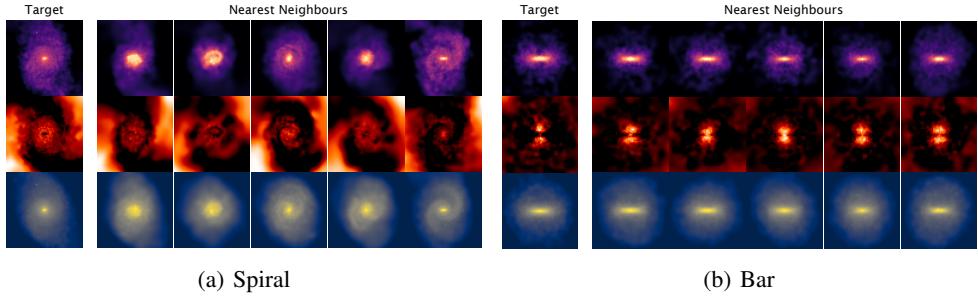


Figure 3: Nearest Neighbors in eigenspace: given a sample galaxy we find the five nearest neighbors in euclidean distance in the lower dimensional space (right) for a spiral (a) and bar shaped (b) galaxy.

we perform a morphological similarity search. This involves selecting a sample galaxy and searching for its nearest neighbors in the PCA eigenspace using the Euclidean distance metric: $d(\hat{s}, s_i) = \|\hat{s} - s_i\|_2$. Here, \hat{s} and s_i represent the scores of the sample and the remaining galaxies, respectively. In Fig. 3, we show the five nearest neighbors for two different sample galaxies, which demonstrates the relationship between Euclidean proximity in the PCA eigenspace and morphologically similar characteristics. For the spiral galaxy (left panels), the five nearest neighbors exhibit pronounced two-arm spiral structures. For the triaxial galaxy (right panels), all nearest neighbors are also triaxial and exhibit a prominent bar-like feature in mass and metallicity maps. Additionally, PCA scores can be used for various other analyses, including clustering data using methods such as Gaussian mixture models [20] or k-means [21], allowing unsupervised galaxy type classification. They can also be used for outlier detection, potentially identifying galaxies with unique and interesting morphological characteristics.

5 Conclusions and Limitations

We present the **GAMMA** dataset, a well-curated collection of 2D and 3D galaxy images that capture stellar age, metallicity, and mass. This dataset uniquely combines well-understood physical relations in the data with explicit and controlled image generation processes such as noise level or point-spread function. As such, this dataset paves the way for a deeper understanding of galactic structures and the public release of **GAMMA** will facilitate the community to go beyond the exploratory benchmarks presented in this work. So far, we have harnessed PCA to analyze galaxy images in both 2D and 3D, simultaneously modeling the distributions of mass, metallicity, and age. Our exploration has yielded several significant findings and insights into the potential applications of PCA for characterizing galaxy morphology and conducting morphological analysis:

- 1. Effective Reconstruction:** We have demonstrated that PCA can effectively reconstruct galaxy images with a relatively small number of eigengalaxies. The accuracy of these reconstructions was quantified using the reconstruction error metric, which measures the fractional difference in pixel values between the original and PCA-reconstructed images. Our results indicate that even with a modest number of 60 (215) eigengalaxies in 2D (3D), the PCA model achieves highly accurate reconstructions. Approximately 90% of the images have reconstruction errors below 0.022 (**0.027**), underscoring the potential of PCA to efficiently represent the intricate morphology of the galaxy.
- 2. Interpretability:** PCA offers some unique advantage over more complex ML methods. PCA allows for a straightforward interpretation of its components, with each eigengalaxy representing a specific morphological feature. Visualizing the top contributing eigengalaxies for different types of galaxies reveals their distinctive contributions and demonstrates the potential of PCA for classifying galaxies based on their morphology.
- 3. Applications in Morphological Analysis:** We present the utility of PCA to conduct morphological similarity searches. By calculating Euclidean distances in the PCA eigenspace, we successfully identified galaxies that were morphologically similar to a given sample galaxy. This suggests that PCA captures meaningful features of galaxy morphology, enabling efficient similarity analysis and clustering.

4. Limitations: PCA is a linear decomposition. Each galaxy is reconstructed as a linear combination of eigengalaxies, so the initial unitary rotation of the dataset is important for PCA to find meaningful features. PCA assumes that the data can be understood in terms of orthogonal axes of maximum variance, which is why certain morphological features, such as rotational positions of spiral arms, may not be accurately fitted.

In conclusion, our study underscores the potential of PCA as a powerful yet simple tool for analyzing and characterizing galaxy morphology. Its efficiency in image representation, interpretability of components, and facilitation of morphological similarity searches make it a valuable approach in the field of astrophysics. Future work could involve further refinement of the PCA model, such as exploring nonlinear PCA, investigating its applications in other astronomical datasets, and developing hybrid approaches that combine PCA with other ML techniques such as variational autoencoders, GANs, or diffusion models for even more comprehensive analyses of galaxy images. It could also be valuable to explore benefits from including simulation data from other cosmological simulations such as EAGLE, CAMELS or NIHAO, to further reduce potential biases in the data.

Broader Impact statement

The authors are unaware of any immediate ethical or societal implications of this work. This work purely aims to aid in the design of scientific algorithms and datasets. Looking more broadly, the **GAMMA** dataset and its connection to fundamental physics might help to benchmark new algorithms for image recognition. This may contribute to a broader application and development of interpretable and robust algorithms.

References

- [1] Annalisa Pillepich, Dylan Nelson, Lars Hernquist, Volker Springel, Rüdiger Pakmor, Paul Torrey, Rainer Weinberger, Shy Genel, Jill P Naiman, Federico Marinacci, and Mark Vogelsberger. First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. *Monthly Notices of the Royal Astronomical Society*, 475(1):648–675, dec 2017. doi: 10.1093/mnras/stx3112. URL <https://doi.org/10.1093/mnras/stx3112>.
- [2] S. McAlpine, J.C. Helly, M. Schaller, J.W. Trayford, Y. Qu, M. Furlong, R.G. Bower, R.A. Crain, J. Schaye, T. Theuns, C. Dalla Vecchia, C.S. Frenk, I.G. McCarthy, A. Jenkins, Y. Rosas-Guevara, S.D.M. White, M. Baes, P. Camps, and G. Lemson. The eagle simulations of galaxy formation: Public release of halo and galaxy catalogues. *Astronomy and Computing*, 15:72–89, apr 2016. doi: 10.1016/j.ascom.2016.02.004. URL <https://doi.org/10.1016/j.ascom.2016.02.004>.
- [3] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. , 915(1):71, July 2021. doi: 10.3847/1538-4357/abf7ba.
- [4] Robert J. J. Grand, Facundo A. Gómez, Federico Marinacci, Rüdiger Pakmor, Volker Springel, David J. R. Campbell, Carlos S. Frenk, Adrian Jenkins, and Simon D. M. White. The Auriga Project: the properties and formation mechanisms of disc galaxies across cosmic time. *Monthly Notices of the Royal Astronomical Society*, 467(1):179–207, 01 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx071. URL <https://doi.org/10.1093/mnras/stx071>.
- [5] Philip F Hopkins, Andrew Wetzel, Dušan Kereš, Claude-André Faucher-Giguère, Eliot Quataert, Michael Boylan-Kolchin, Norman Murray, Christopher C Hayward, Shea Garrison-Kimmel, Cameron Hummels, Robert Feldmann, Paul Torrey, Xiangcheng Ma, Daniel Anglés-Alcázar, Kung-Yi Su, Matthew Orr, Denise Schmitz, Ivanna Escala, Robyn Sanderson, Michael Y Grudić, Zachary Hafen, Ji-Hoon Kim, Alex Fitts, James S Bullock, Coral Wheeler, T K Chan, Oliver D Elbert, and Desika Narayanan. FIRE-2 simulations: physics versus numerics in galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 480(1):800–863, jun 2018. doi: 10.1093/mnras/sty1690. URL <https://doi.org/10.1093/mnras/sty1690>.
- [6] Oscar Agertz, Florent Renaud, Sofia Feltzing, Justin I Read, Nils Ryde, Eric P Andersson, Martin P Rey, Thomas Bensby, and Diane K Feuillet. VINTERGATAN – i. the origins of chemically, kinematically, and structurally distinct discs in a simulated milky way-mass galaxy. *Monthly Notices of the Royal Astronomical Society*, 475(1):103–118, 01 2018. doi: 10.1093/mnras/sty260. URL <https://doi.org/10.1093/mnras/sty260>.

- Society*, 503(4):5826–5845, apr 2021. doi: 10.1093/mnras/stab322. URL <https://doi.org/10.1093/2Fmnras%2Fstab322>.
- [7] Liang Wang, Aaron A. Dutton, Gregory S. Stinson, Andrea V. Macciò, Camilla Penzo, Xi Kang, Ben W. Keller, and James Wadsley. NIHAO project - I. Reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 454(1):83–94, November 2015. doi: 10.1093/mnras/stv1937.
 - [8] Tobias Buck, Aura Obreja, Andrea V. Macciò, Ivan Minchev, Aaron A. Dutton, and Jeremiah P. Ostriker. NIHAO-UHD: the properties of MW-like stellar discs in high-resolution cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 491(3):3461–3478, January 2020. doi: 10.1093/mnras/stz3241.
 - [9] Chris Lintott, Kevin Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C. Nichol, M. Jordan Raddick, Alex Szalay, Dan Andreeescu, Phil Murray, and Jan Vandenberg. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178, January 2011. doi: 10.1111/j.1365-2966.2010.17432.x.
 - [10] François Lanusse, Rachel Mandelbaum, Siamak Ravnbakhsh, Chun-Liang Li, Peter Freeman, and Barnabás Póczos. Deep generative models for galaxy image simulations. *Monthly Notices of the Royal Astronomical Society*, 504(4):5543–5555, July 2021. doi: 10.1093/mnras/stab1214.
 - [11] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, April 2016. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2015.0202. URL <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>.
 - [12] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 01 1991. ISSN 0898-929X. doi: 10.1162/jocn.1991.3.1.71. URL <https://doi.org/10.1162/jocn.1991.3.1.71>.
 - [13] Jorge De La Calleja and Olac Fuentes. Machine learning and image analysis for morphological galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 349(1):87–93, 03 2004. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2004.07442.x. URL <https://doi.org/10.1111/j.1365-2966.2004.07442.x>.
 - [14] Emir Uzeirbegovic, James E Geach, and Sugata Kaviraj. Eigengalaxies: describing galaxy morphology using principal components in image space. *Monthly Notices of the Royal Astronomical Society*, 498(3):4021–4032, sep 2020. doi: 10.1093/mnras/staa2651. URL <https://doi.org/10.1093/2Fmnras%2Fstaa2651>.
 - [15] Plotly Technologies Inc. Collaborative data science, 2015. URL <https://plot.ly>.
 - [16] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Rüdiger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: the galaxy colour bimodality. *Monthly Notices of the Royal Astronomical Society*, 475(1):624–647, nov 2017. doi: 10.1093/mnras/stx3040. URL <https://doi.org/10.1093/2Fmnras%2Fstx3040>.
 - [17] Volker Springel, Rüdiger Pakmor, Annalisa Pillepich, Rainer Weinberger, Dylan Nelson, Lars Hernquist, Mark Vogelsberger, Shy Genel, Paul Torrey, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: matter and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 475(1):676–698, dec 2017. doi: 10.1093/mnras/stx3304. URL <https://doi.org/10.1093/2Fmnras%2Fstx3304>.
 - [18] Josh Borrow and Alexei Borrisov. swiftsimio: A python library for reading swift data. *Journal of Open Source Software*, 5(52):2430, 2020. doi: 10.21105/joss.02430. URL <https://doi.org/10.21105/joss.02430>.
 - [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
 - [20] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009. ISBN 978-0-387-73003-5. doi: 10.1007/978-0-387-73003-5_196. URL https://doi.org/10.1007/978-0-387-73003-5_196.
 - [21] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998. doi: 10.1162/089976698300017467.