Boletim da Sociedade Astronômica Brasileira, **37**, no. 1, 134-138

Unsupervised Topological and Contrastive Representation Learning for Galaxy Morphological Analysis

G. W. C. Rocha[1] & G. M. Viswanathan[1], and L. A. Almeida[2]

[1] Departamento de Física, Universidade Federal do Rio Grande do Norte, Natal, 59072-970, Rio Grande do Norte, Brazil
e-mail: gabrielwendell@fisica.ufrn.br, e-mail: gandhi@fisica.ufrn.br
[2] Escola de Ciência e Tecnologia, Universidade Federal do Rio Grande do Norte, 59078-970 Natal, Brazil
e-mail: e-mail: leonardo.almeida@ufrn.br

Abstract. Recent advances in machine learning have enabled unsupervised representations of astronomical images without relying on labeled data. In this work, we combine Self-Supervised Learning and Topological Data Analysis to investigate the morphological structure of galaxies from the Galaxy Zoo 2 dataset. Using a SimCLR-based contrastive learning approach with ResNet and EfficientNet backbones, we obtained low-dimensional embeddings that capture intrinsic morphological features of galaxies. These representations were then complemented with persistence diagrams and persistence images, which encode topological information about shape and texture. Unsupervised clustering performed on the combined feature spaces revealed groups that closely align with known morphological categories (spiral, smooth, and edge-on galaxies). Evaluation using purity, normalized mutual information, and adjusted Rand index confirmed the emergence of astrophysically meaningful clusters, even in the absence of supervision. Our findings demonstrate that the joint use of SSL and TDA provides a powerful and interpretable framework for morphology analysis, offering promising perspectives for large-scale astronomical surveys such as LSST and Euclid.

Resumo. Avanços recentes em aprendizado de máquina possibilitaram a representação não supervisionada de imagens astronômicas sem a necessidade de rótulos. Neste trabalho, combinamos Aprendizado Auto-Supervisionado e Análise Topológica de Dados para investigar a estrutura morfológica de galáxias do conjunto de dados Galaxy Zoo 2. Utilizando uma abordagem de aprendizado contrastivo baseada no SimCLR, com arquiteturas ResNet e EfficientNet, obtivemos embeddings de baixa dimensionalidade que capturam características morfológicas intrínsecas das galáxias. Essas representações foram complementadas com diagramas e imagens de persistência, que codificam informações topológicas sobre forma e textura. A clusterização não supervisionada dos espaços combinados revelou grupos que se alinham fortemente com categorias morfológicas conhecidas (por exemplo, galáxias espirais, suaves e vistas de perfil). A avaliação por pureza, informação mútua normalizada e índice de Rand ajustado confirmou o surgimento de agrupamentos com significado astrofísico, mesmo na ausência de supervisão. Os resultados demonstram que o uso conjunto de SSL e TDA constitui uma estrutura poderosa e interpretável para análise morfológica, com perspectivas promissoras para levantamentos astronômicos de grande escala, como o LSST e o Euclid.

Keywords. Galaxies: structure − Methods: data analysis − Methods: statistical

1. Introduction

The morphological characterization of galaxies provides fundamental insights into their formation history, stellar evolution, and the physical processes driving large-scale structure in the Universe. The distribution of galactic morphologies—ranging from smooth ellipticals to structured spirals and irregular systems—encodes the dynamical and environmental interactions that shape galaxies across cosmic time. As such, morphology classification remains a cornerstone of extragalactic astronomy, linking observable structure to fundamental astrophysical mechanisms such as angular momentum acquisition, mergers, and feedback processes.

However, large-scale classification of galaxy morphologies faces several challenges. The exponential growth of astronomical surveys − such as the Sloan Digital Sky Survey (SDSS), the Dark Energy Survey (DES), and the upcoming Legacy Survey of Space and Time (LSST) − has produced millions of galaxy images, far exceeding the capacity for manual inspection. Citizen science initiatives like Galaxy Zoo 2[1] (GZ2) (Lintott et al. 2008) have addressed this by enlisting volunteers to visually classify galaxies, resulting in one of the largest morphological catalogs to date, with over 300,000 galaxies (Willett et al. 2013). Despite its success, the approach is inherently limited by human subjectivity, label imbalance, and scalability constraints, motivating the search for automated, data-driven alternatives.

Traditional machine learning approaches have primarily relied on supervised learning, where algorithms are trained to reproduce existing human labels (Masters et al. 2010). While effective, such models inherit the biases of labeled datasets and require extensive annotated samples that may not generalize to new surveys (Gravet-Vives et al. 2015). Recent advances in Self-Supervised Learning (SSL) have offered a transformative alternative: instead of depending on labeled data, SSL models learn intrinsic representations by solving pretext tasks that capture semantic and structural information from the data itself. In particular, contrastive learning frameworks like SimCLR have demonstrated exceptional capability in learning robust, general-purpose representations from images through augmentation-based similarity objectives.

Complementary to these advances, Topological Data Analysis (TDA) (Chazal & Michel 2020) provides a mathematically grounded framework for quantifying the shape and connectivity of data. Through tools such as persistent homology, TDA captures multiscale structural features that are invariant under geometric transformations—offering a complementary perspective to pixel-level representations. Persistence diagrams and their vectorized forms, known as persistence images, summarize the topological complexity of galaxy light distributions and can

[1] https://data.galaxyzoo.org/

reveal structural patterns not easily captured by convolutional filters alone.

This work aims to integrate these two paradigms – SSL and TDA – to explore the unsupervised discovery of galaxy morphological structures. By combining contrastively learned feature embeddings with topological descriptors, we evaluate whether morphology-aware clusters naturally emerge from the data without explicit supervision. The central hypothesis is that the synergy between deep representation learning and topology can yield interpretable, astrophysically meaningful groupings of galaxies, paving the way toward automated and explainable morphology classification in next-generation astronomical surveys.

2. Methodology

The proposed framework integrates self-supervised feature learning with topological descriptors to analyze and cluster galaxies based on their intrinsic morphological properties. The pipeline consists of four main components: (i) data preparation, (ii) representation learning via self-supervised contrastive models, (iii) topological data analysis through persistence images, and (iv) unsupervised clustering and evaluation of the resulting feature spaces.

2.1. Dataset

The experiments were conducted using a curated subset of GZ2, containing 13,965 galaxies with consensus morphological labels derived from volunteer classifications. The selected classes include `smooth`, `spiral`, `edge_on`, `features_or_disk`, and `star_or_artifact`. The original galaxy images were obtained from the SDSS and preprocessed to ensure consistency across the dataset. Each image was resized to 128×128 pixels, converted to grayscale, and normalized by its intensity distribution to reduce photometric variability. These preprocessing steps preserve morphological information while minimizing artifacts that could bias the learning process.

2.2. Self-Supervised Representation Learning

To obtain latent representations that capture intrinsic morphological features, we employed a SSL approach based on the SimCLR framework (Chen et al. 2020). In this contrastive setup, the model learns to maximize agreement between differently augmented views of the same galaxy image, while minimizing agreement between different galaxies. The training objective is defined as a normalized temperature-scaled cross-entropy loss over positive and negative pairs.

We evaluated three backbone architectures of increasing representational capacity:

- **ResNet-18** (He et al. 2016): a lightweight convolutional network suitable for small datasets;
- **ResNet-50** (He et al. 2016): a deeper residual network providing stronger hierarchical feature extraction;
- **EfficientNet-B0** (Tan & Le 2019): an optimized model balancing accuracy and computational efficiency.

Each model was trained using random data augmentations including rotation, cropping, horizontal flipping, and color jitter to encourage invariance to orientation and brightness variations. After convergence, we extracted 128-dimensional embeddings from the projection head for each galaxy, yielding compact representations of morphological structure. These embeddings serve as the foundation for downstream topological and clustering analyses.

2.3. Topological Data Analysis

To complement the geometric and statistical representations learned by SSL, we employed TDA to characterize the intrinsic shape of galaxy brightness distributions. For each preprocessed image, a persistence diagram (PD) (Edelsbrunner, Letscher & Zomorodian 2002) was computed from its intensity landscape using sublevel filtrations. The PD encodes the birth and death of topological features – connected components and loops—across intensity thresholds.

The persistence diagrams were subsequently transformed into persistence images (PIs) (Adams et al. 2017), a vectorized representation that maps persistence pairs onto a 2D grid, allowing integration with machine learning pipelines. Each PI was generated as a 100×100 pixel Gaussian-weighted density map, where high-intensity regions correspond to persistent topological features. These PIs provide a topological fingerprint of each galaxy, capturing multiscale shape information invariant to rotation and translation. Class-level statistics were then derived by computing mean PIs per morphological group and overall averages to visualize population-level topological trends.

2.4. Clustering and Evaluation

Unsupervised clustering was applied to both SSL embeddings and topological features to evaluate the emergence of morphology-consistent structures. The primary algorithm used was K-Means, with the number of clusters set to $k = 5$ to match the five dominant GZ2 morphological categories (Walmsley et al. 2020). Additionally, DBSCAN was tested to identify density-based groupings without predefined cluster counts.

Cluster quality was quantitatively assessed using three complementary metrics:

- **Purity**: measures the fraction of samples in each cluster that belong to the dominant ground-truth label.
- **Normalized Mutual Information (NMI)**: quantifies the mutual dependence between cluster assignments and true labels, normalized to $[0, 1]$.
- **Adjusted Rand Index (ARI)**: evaluates the agreement between predicted clusters and ground-truth labels, adjusted for random chance.

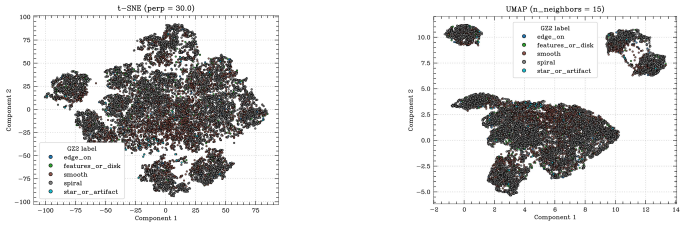
For qualitative assessment, t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) were employed to visualize the embedding spaces in two dimensions. These visualizations provide intuitive insights into the separation and overlap between morphological classes in the learned representation spaces.

3. Results

The results demonstrate that the integration of SSL and TDA yields interpretable and astrophysically meaningful structures within the Galaxy Zoo 2 dataset. Both quantitative and qualitative evaluations confirm that morphology-related information naturally emerges from the learned feature spaces without supervision.

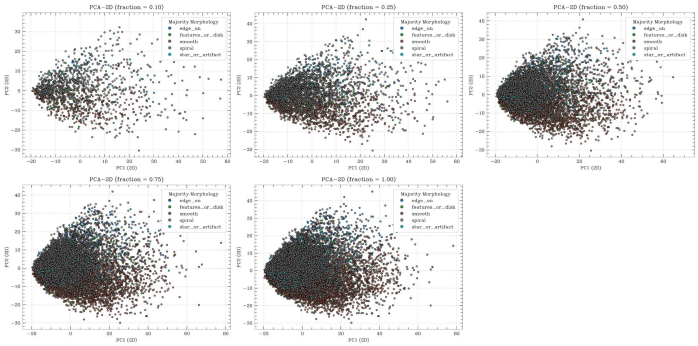
3.1. SSL Embedding Performance

Among the SSL models tested, ResNet-50 produced the most discriminative feature space, followed by ResNet-18 and EfficientNet-B0. The 128-dimensional embeddings extracted from ResNet-50 demonstrated clearer separation between morphological types when visualized through t-SNE and

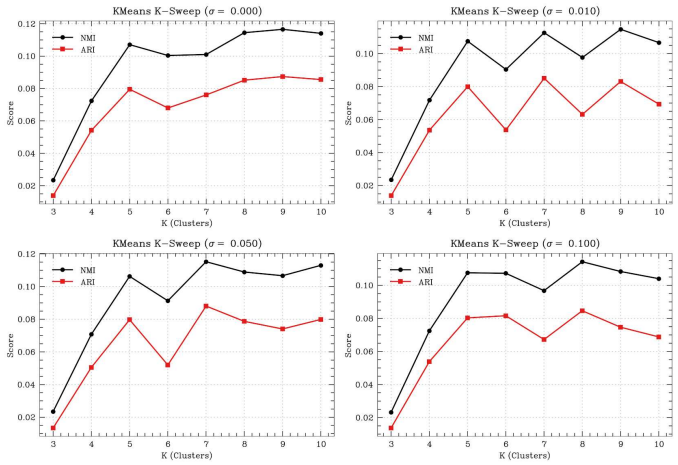


(a) t-SNE (b) UMAP

FIGURE 1: t-SNE and UMAP visualizations of SSL embeddings (ResNet-50)



(a) Fraction levels for ResNet-50.



(b) Metrics for ResNet-50.

FIGURE 2: Robustness panels for subsampling tests for ResNet-50.

UMAP projections (Figure 1). Galaxies identified as spiral, smooth, and edge-on formed distinct, coherent clusters, while `features_or_disk` and `star_or_artifact` categories appeared more dispersed, as expected from their morphological ambiguity.

Optimal clustering performance was achieved with $k = 5$, consistent with the number of dominant morphological classes. Quantitatively, ResNet-50 reached peak metrics of Purity $\approxeq 0.83$, NMI $\approxeq 0.79$, and ARI $\approxeq 0.71$, outperforming the other backbones. These results indicate that the SSL embeddings encode morphology-sensitive structures that align well with the astrophysical ground truth.

3.2. Robustness Tests

To assess model stability, robustness analyses were performed under random data subsampling (Figure 2) and feature-level Gaussian noise perturbations (Figure 3). The NMI remained stable within the range $0.75-0.8$ across all noise levels and fractions, demonstrating that the embedding space learned by ResNet-50 is resilient to both sample size reduction and photometric distortions.

These results suggest that the contrastive learning framework successfully captures invariant morphological representations, preserving cluster coherence even when galaxies are degraded or partially removed from the dataset.

3.3. Persistence Images

The analysis of PIs provided complementary insights into the topological structure of galaxies. Mean persistence images computed per morphological class revealed subtle but consistent differences in topological feature density – particularly in the spatial concentration of homology intensity regions. Smooth galaxies exhibited more compact and symmetric topological distributions, while spiral and feature-rich galaxies showed elongated and dispersed high-persistence zones, reflecting the presence of arms, bars, and disk substructures. Figure 4 show the mean PIs for each morphological class.

Although visually similar at first glance, statistical comparisons between class-level PIs confirmed that TDA encodes information about morphological smoothness and asymmetry. When combined with SSL embeddings, the hybrid feature space improved interpretability, allowing for a more robust link between learned clusters and physical morphology.

3.4. Cluster Alignment

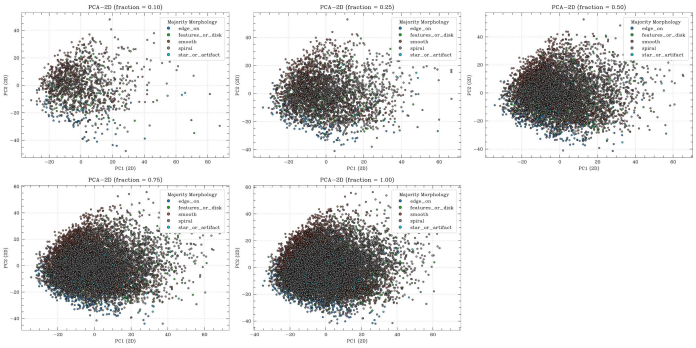
The alignment between unsupervised clusters and the GZ2 consensus labels was analyzed through confusion-like matrices. The resulting pattern show in Figure 5 exhibited a one-to-one correspondence between spiral-dominated clusters and the GZ2 spiral class, while smooth and edge-on categories also showed high purity ratios. Mixed clusters, on the other hand, corresponded to irregular or merging systems – objects that often defy simple morphological categorization.

This alignment suggests that the SSL-TDA combination captures not only the dominant morphological modes but also transitional morphologies, thereby reproducing a continuum of galactic structure consistent with astrophysical expectations.

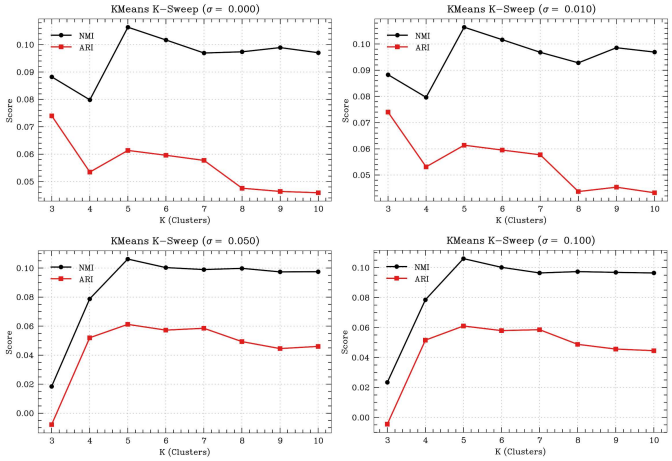
4. Discussion

The results presented here demonstrate the effectiveness of combining SSL and TDA for unsupervised galaxy morphology characterization. The pipeline successfully recovered morphology-sensitive clusters aligned with GZ2 classifications, offering both methodological and astrophysical insights into the structure of galaxy representation spaces.

The self-supervised embeddings learned through contrastive learning captured morphological signatures without the need for human-labeled data, confirming the capability of SSL to autonomously extract semantically meaningful representations from raw galaxy images. The separation between spiral, smooth, and edge-on galaxies observed in the t-SNE and UMAP projections validates that contrastive objectives can implicitly encode shape, texture, and symmetry information relevant to astrophysical classification. This result suggests that SSL can serve as a scalable alternative to classical supervised pipelines, which de-



(a) Fraction levels for noisy ResNet-50.



(b) Metrics for noisy ResNet-50.

FIGURE 3: Robustness panels for noise injection tests for ResNet-50.

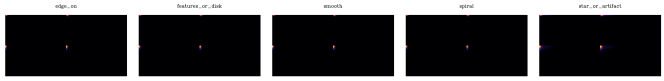


FIGURE 4: Mean Persistence Images per morphological class and overall mean.

pend on extensive annotation efforts and may embed subjective bias from human labeling.

Furthermore, the persistence images derived from topological analysis provided a complementary view of the data, encoding information about the spatial distribution of features such as spiral arms, bars, and central bulges. While these topological maps appeared visually subtle, their statistical profiles revealed consistent differences across morphological types—particularly in how persistence intensity patterns reflect galactic smoothness, asymmetry, and structural complexity. This supports the idea that topology-based measures can act as a form of morphological regularization (Carrièri, Cuturi & Oudot 2017), enriching purely geometric embeddings with invariant shape-based descriptors.

Crucially, the clustering results exhibit astrophysical interpretability: morphology clusters reflect the major trends of the Hubble sequence, where spiral-dominated clusters correspond to disk galaxies, smooth clusters align with ellipticals, and mixed or transitional groups correspond to mergers or irregular systems. This alignment between latent-space clusters and physical morphology underscores the scientific potential of SSL–TDA integration to uncover the morphological continuum of galaxies directly from imaging data.

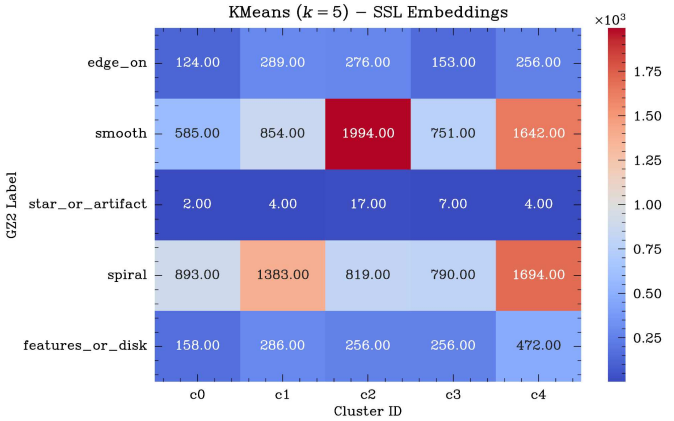


FIGURE 5: Confusion-like matrix comparing cluster assignments with GZ2 labels.

From a methodological standpoint, the study highlights the synergistic role of TDA and SSL in building interpretable and robust representations. While SSL embeddings provide high-variance, discriminative manifolds, TDA features introduce topological constraints that stabilize and regularize these spaces. Even though persistence images may not exhibit striking visual differences, their inclusion enhances the model's sensitivity to intrinsic shape variations, contributing to improved alignment and cluster coherence.

Overall, the combination of self-supervised representation learning and topological analysis establishes a pathway toward explainable, unsupervised morphology discovery in astronomy. This approach bridges deep learning and mathematical topology, showing promise for scalable applications in the analysis of upcoming large surveys such as LSST, Euclid, and Roman.

5. Conclusion and Future Directions

This study presented a unified framework that integrates SSL, via the SimCLR contrastive paradigm, with TDA for unsupervised galaxy morphology characterization. The proposed pipeline successfully extracted astrophysically meaningful representations directly from raw galaxy images, reproducing the main morphological categories observed in the GZ2 dataset—namely spiral, smooth (elliptical), and edge-on galaxies—without explicit supervision.

Through the use of SSL embeddings, the model learned invariant and discriminative morphological features, while the inclusion of persistence images provided complementary topological information about galactic structure and shape complexity. Together, these methods yielded clusters that align closely with physical morphology, bridging the gap between data-driven learning and astrophysical interpretability.

The results underscore the potential of SSL + TDA as a scalable and explainable alternative to supervised classification approaches, especially in the era of large astronomical surveys. The demonstrated robustness under noise and subsampling conditions further supports the applicability of this framework to heterogeneous, real-world data.

5.1. Future Directions

Building on these results, future research will focus on:

– Integrating multiband imaging (e.g., SDSS g, r, i filters) to enrich spectral and structural context.

- Extending the framework to graph-based topological neural networks (Topological GNNs) for learning directly from persistence representations (Chazal & Michel 2020).
- Scaling the analysis to upcoming survey data from facilities such as LSST, Euclid, and Roman Space Telescope, where unsupervised morphology discovery will be essential for managing the unprecedented data volume (McPartland et al. 2025).

Data and Code Availability

All analysis scripts, trained embeddings, and resulting figures used in this study are openly available in the GitHub repository of this project.

Acknowledgements. All authors would like to thank the organizers of the XLVIII RASAB for the event and acknowledge CNPq research fellowships.

References

Adams, H., Emerson, T., Kirby, M., et al. 2017, JMLR, 18, 1
Carrière, M., Cuturi, M., & Oudot, S. 2017, ICML, 664
Chazal, F., & Michel, B. 2021, Front. Artif. Intell., 4, 667963
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, ICML, 1597
Edelsbrunner, H., Letscher, D., & Zomorodian, A. 2002, Discrete Comput. Geom., 28, 511
Gravet, R., Cabrera-Vives, G., Pérez-González, P. G., et al. 2015, ApJS, 221, 8
He, K., Zhang, X., Ren, S., & Sun, J. 2016, CVPR, 770
Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179
Masters, K. L., Mosleh, M., Romer, A. K., et al. 2010, MNRAS, 405, 783
McPartland, C. J. R., Zalesky, L., Weaver, J. R., et al. 2025, A&A, 695, A259
Tan, M., & Le, Q. 2019, ICML, 6105
Walmsley, M., Smith, L., Lintott, C., et al. 2020, MNRAS, 491, 1554
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835