



HAL
open science

On the Nature of Disks at High Redshift Seen by JWST/CEERS with Contrastive Learning and Cosmological Simulations

Jesús Vega-Ferrero, Marc Huertas-Company, Luca Costantin, Pablo G. Pérez-González, Regina Sarmiento, Jeyhan S. Kartaltepe, Annalisa Pillepich, Micaela B. Bagley, Steven L. Finkelstein, Elizabeth J. Mcgrath, et al.

► To cite this version:

Jesús Vega-Ferrero, Marc Huertas-Company, Luca Costantin, Pablo G. Pérez-González, Regina Sarmiento, et al.. On the Nature of Disks at High Redshift Seen by JWST/CEERS with Contrastive Learning and Cosmological Simulations. *The Astrophysical Journal*, 2024, 961, 10.3847/1538-4357/ad05bb . insu-04822496

HAL Id: insu-04822496

<https://insu.hal.science/insu-04822496v1>

Submitted on 6 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



On the Nature of Disks at High Redshift Seen by JWST/CEERS with Contrastive Learning and Cosmological Simulations

Jesús Vega-Ferrero^{1,2,3} , Marc Huertas-Company^{1,2,4} , Luca Costantin⁵ , Pablo G. Pérez-González⁵ , Regina Sarmiento^{1,2} , Jeyhan S. Kartaltepe⁶ , Annalisa Pillepich⁷ , Micaela B. Bagley⁸ , Steven L. Finkelstein⁸ , Elizabeth J. McGrath⁹ , Johan H. Knapen^{1,2} , Pablo Arrabal Haro¹⁰ , Eric F. Bell¹¹ , Fernando Buitrago^{3,12} , Antonello Calabrò¹³ , Avishai Dekel^{14,15} , Mark Dickinson¹⁰ , Helena Domínguez Sánchez¹⁶ , David Elbaz¹⁷ , Henry C. Ferguson¹⁸ , Mauro Giavalisco¹⁹ , Benne W. Holwerda²⁰ , Dale D. Kocevski⁹ , Anton M. Koekemoer¹⁸ , Viraj Pandya^{21,25} , Casey Papovich^{22,23} , Nor Pirzkal¹⁸ , Joel Primack¹⁵ , and L. Y. Aaron Yung²⁴

¹Instituto de Astrofísica de Canarias, E-38200, La Laguna, Tenerife, Spain; astrovega@gmail.com

²Departamento de Astrofísica, Universidad de La Laguna, E-38205, La Laguna, Tenerife, Spain

³Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, E-47011 Valladolid, Spain

⁴LERMA, Observatoire de Paris, CNRS, PSL, Université de Paris, France

⁵Centro de Astrobiología (CAB), CSIC-INTA, Ctra. de Ajalvir km 4, Torrejón de Ardoz, E-28850 Madrid, Spain

⁶Laboratory for Multiwavelength Astrophysics, School of Physics and Astronomy, Rochester Institute of Technology, 84 Lomb Memorial Drive, Rochester, NY 14623, USA

⁷Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

⁸Department of Astronomy, The University of Texas at Austin, Austin, TX, USA

⁹Department of Physics and Astronomy, Colby College, Waterville, ME 04901, USA

¹⁰NSF's National Optical-Infrared Astronomy Research Laboratory, 950 North Cherry Avenue, Tucson, AZ 85719, USA

¹¹Department of Astronomy, University of Michigan, 1085 S. University Avenue, Ann Arbor, MI 48109-1107, USA

¹²Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, PT1349-018 Lisbon, Portugal

¹³Osservatorio Astronomico di Roma, via Frascati 33, Monte Porzio Catone, Italy

¹⁴Centre for Astrophysics and Planetary Science, Racah Institute of Physics, The Hebrew University, Jerusalem, 91904, Israel

¹⁵Santa Cruz Institute for Particle Physics, University of California, Santa Cruz, CA 95064, USA

¹⁶Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza de San Juan, 1, E-44001 Teruel, Spain

¹⁷Laboratoire AIM-Paris-Saclay, CEA/DRF/Irfu—CNRS - Université Paris Cité, CEA-Saclay, pt courrier 131, F-91191 Gif-sur-Yvette, France

¹⁸Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

¹⁹University of Massachusetts Amherst, 710 North Pleasant Street, Amherst, MA 01003-9305, USA

²⁰Physics & Astronomy Department, University of Louisville, Louisville, KY 40292, USA

²¹Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, New York, NY 10027, USA

²²Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242 USA

²³George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242 USA

²⁴Astrophysics Science Division, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA

Received 2023 February 14; revised 2023 September 19; accepted 2023 October 20; published 2024 January 12

Abstract

Visual inspections of the first optical rest-frame images from JWST have indicated a surprisingly high fraction of disk galaxies at high redshifts. Here, we alternatively apply self-supervised machine learning to explore the morphological diversity at $z \geq 3$. Our proposed data-driven representation scheme of galaxy morphologies, calibrated on mock images from the TNG50 simulation, is shown to be robust to noise and to correlate well with the physical properties of the simulated galaxies, including their 3D structure. We apply the method simultaneously to F200W and F356W galaxy images of a mass-complete sample ($M_*/M_\odot > 10^9$) at $3 \leq z \leq 6$ from the first JWST/NIRCam CEERS data release. We find that the simulated and observed galaxies do not exactly populate the same manifold in the representation space from contrastive learning. We also find that half the galaxies classified as disks—either convolutional neural network-based or visually—populate a similar region of the representation space as TNG50 galaxies with low stellar specific angular momentum and nonoblate structure. Although our data-driven study does not allow us to firmly conclude on the true nature of these galaxies, it suggests that the disk fraction at $z \geq 3$ remains uncertain and possibly overestimated by traditional supervised classifications. Deeper imaging and spectroscopic follow-ups as well as comparisons with other simulations will help to unambiguously determine the true nature of these galaxies, and establish more robust constraints on the emergence of disks at very high redshift.

Unified Astronomy Thesaurus concepts: [Galaxy formation \(595\)](#); [Galaxy evolution \(594\)](#); [High-redshift galaxies \(734\)](#); [Neural networks \(1933\)](#)

1. Introduction

Understanding how galaxy diversity emerges across cosmic time is one of the main goals of galaxy formation. How and when do stellar disks form? What are the main drivers of bulge growth? How and when did galaxy morphology and star formation get connected? Despite significant progress in the

²⁵ Hubble Fellow.



past years, thanks in particular to deep surveys undertaken with the Hubble Space Telescope (HST; e.g., Scoville et al. 2007; Grogin et al. 2011; Koekemoer et al. 2011), these questions remain largely unanswered. The general picture is that massive star-forming galaxies in the past were more irregular in their stellar structure (e.g., Abraham et al. 1996; Conselice 2003) than today’s disks even if observed in the optical rest frame (Buitrago et al. 2013; Huertas-Company et al. 2015). Galaxies above $z \sim 1$ also show the presence of giant star-forming clumps (e.g., Guo et al. 2015, 2018; Huertas-Company et al. 2020; Ginzburg et al. 2021), which might indicate a turbulent and unstable interstellar medium (e.g., Ceverino et al. 2010; Bournaud et al. 2014). Although the gas shows signatures of rotation at $z \sim 2$ (e.g., Wisnioski et al. 2015), the settling of disks seems to be a process happening at least from $z \sim 2$ (e.g., Kassin et al. 2012; Buitrago et al. 2014; Simons et al. 2017; Costantin et al. 2022) coincident with the decrease of gas fractions in massive galaxies (e.g., Genzel et al. 2010; Freundlich et al. 2019). Another important result of the past years is that the presence of bulges in galaxies is strongly anticorrelated with the star formation activity at all redshifts probed (e.g., van der Wel et al. 2014b; Barro et al. 2017; Costantin et al. 2020, 2021; Dimauro et al. 2022). This suggests that bulge formation and quenching are tightly connected physical processes (e.g., Chen et al. 2020b).

With its unprecedented sensitivity, spatial resolution, and infrared coverage, the James Webb Space Telescope (JWST) is offering a new window to the stellar structure of galaxies in the first epochs of cosmic history (Gardner et al. 2006). For the first time, we are able to explore the stellar morphologies of the first galaxies formed in the universe, which should enable new constraints on the physical processes governing galaxy assembly at early times and hopefully a better understanding of the physical processes leading to the formation of the first stellar disks and bulges. Some very recent works have already started this exploration by performing visual classifications (Ferreira et al. 2022, 2023; Kartaltepe et al. 2023), by applying supervised machine learning trained on HST images (Robertson et al. 2023) of galaxies observed in the first JWST deep fields or by using convolutional neural networks (CNNs) trained on HST/WFC3 labeled images and domain-adapted to JWST/NIRCam (Huertas-Company et al. 2023a). One of the main results of these early works is that JWST seems to be detecting star-forming disks even at $z > 3$, which would push the time of disk formation to very early epochs. Two questions naturally arise from these first works:

1. Are the galaxies seen by JWST true disks, i.e., flat, rotating systems? The aforementioned results are based primarily on qualitative morphological classifications, with quantitative tracers of morphology (e.g., Sèrsic fits) incorporated to further inform differences between the visually defined classes. However, galaxies might look morphologically disk-like but have significantly different stellar kinematics than local disk galaxies. Distinguishing edge-on flat disks from more prolate systems is also a very challenging task that could bias the results (e.g., van der Wel et al. 2014a; Zhang et al. 2019).
2. Do modern cosmological simulations reproduce the observed galaxy diversity at $z > 3$? Although some preliminary comparisons exist, a fair comparison in the observational plane is required to fully address this

question (e.g., Huertas-Company et al. 2019; Rodríguez-Gómez et al. 2019; Zanisi et al. 2021).

In this work, we attempt to provide new insights into these two main questions. To that purpose, we apply a novel data-driven approach based on contrastive learning (Hayat et al. 2021; Sarmiento et al. 2021) to a mass-complete sample of JWST galaxies at $z \geq 3$ observed within the Cosmic Evolution Early Release Science (CEERS; Finkelstein et al. 2017, 2022, 2023) survey. By calibrating the method with mock galaxies (Costantin et al. 2023) from the TNG50 cosmological simulation (Nelson et al. 2019a, 2019b; Pillepich et al. 2019) and by choosing the proper augmentations (i.e., transformations applied to the images such as rotations, flux normalizations, noise, etc.), we are able to build a morphological description, which is more robust to noise and galaxy orientation than more traditional approaches. Our morphological representation can then be correlated with the physical properties of galaxies from the simulation to provide new insights about the physical nature of disk-like galaxies, and to explore the agreements and disagreements between observations and simulations.

The paper proceeds as follows: in Section 2, we describe the galaxy data sets used in this work; Section 3 describes the contrastive learning setting used to derive unsupervised representations of galaxy morphologies; Section 4 explores the properties of the obtained representations on observed JWST/CEERS galaxies; a comparison of the self-supervised representations for simulated and observed galaxies is presented in Section 5; the results and implications are discussed in Section 6; finally, a summary and the final conclusions are presented in Section 7.

2. Data

2.1. CEERS

We use JWST imaging data from NIRCam obtained within CEERS (Finkelstein et al. 2017, 2022, 2023). This consists of short- and long-wavelength images in both NIRCam A and B modules, taken over ten pointings. Each pointing was observed with seven filters: F115W, F150W, and F200W on the short-wavelength side, and F277W, F356W, F410M, and F444W on the long-wavelength side. Here, we only use the F200W and F356W filters. A full description of this public data release²⁶ and the data reduction can be found in Bagley et al. (2023), Finkelstein et al. (2022).

In addition to the galaxy images, we use two different catalogs with physical properties of galaxies:

1. CEERS catalog (CEERS) is a photometric catalog (Finkelstein et al. 2022; Bagley et al. 2023) with derived stellar masses and photometric redshifts (z_{phot}) obtained through spectral energy distribution (SED) fitting of the latest data reduction photometry (Pablo G. Pérez-González private communication). For a fair comparison with the simulated TNG50 data set (see Section 2.2), we select 1,664 galaxies with $3 \leq z \leq 6$, stellar masses $M_* \geq 10^9 M_\odot$ and $F200W [AB] < 27$ mag. The magnitude cut ensures a large enough signal-to-noise ratio (S/N) to enable reliable morphological classification (see Kartaltepe et al. 2023). Obvious stars are removed using the

²⁶ <https://ceers.github.io/>

same procedure as in Huertas-Company et al. (2023a). Also in Huertas-Company et al. (2023a), the completeness limit is estimated at roughly $M_* \sim 10^{8.5} M_\odot$ over the $0 < z < 6$ redshift range. The cut in mass imposed of $M_* \geq 10^9 M_\odot$ is well above this completeness limit, and therefore, the main conclusions presented hereafter in this study should not be affected by incompleteness. For the galaxies in this data set, we also use the CNN-based morphological classifications that split them into four classes: spheroids (Sph), bulge + disk, disk, and disturbed (Irr). See Huertas-Company et al. (2023a), for more details.

- Visual classification catalog (VISUAL) is a redshift-selected $z \geq 3$ morphological catalog presented in Kartaltepe et al. (2023) containing 850 galaxies in common between CANDELS (Grogin et al. 2011; Koekemoer et al. 2011) and CEERS observations. This is intended to directly compare our morphological description to the visual classification of Kartaltepe et al. (2023). Redshifts and stellar masses are extracted from CANDELS v2 for the HST F160W-selected galaxies in the Extended Groth Strip field (see Kodra et al. 2023, for full details on the photometric redshift measurements and resulting catalogs). The visual classifications presented in Kartaltepe et al. (2023) of each galaxy are performed by three people. A given classification is assigned if two out of three people select that option. Galaxies classified in this way are broken down into the following morphological groups: disk only, disk + Sph, disk+irregular, disk+Sph+irregular, Sph only, Sph+irregular, and irregular only. See Kartaltepe et al. (2023) for more details on the different classification tasks and morphological groups. After selecting those galaxies with $M_* \geq 10^9 M_\odot$, $3 < z < 6$, and reliable visual classifications, we end up with a data set of 545 galaxies.

Both catalogs also include morphological measurements of Sèrsic index (n_e), semimajor axis (a), and axis-ratio (b/a) derived with `galfit` (Peng et al. 2010). More information about the fits can also be found in Kartaltepe et al. (2023).

The distributions of stellar masses of the galaxies in the CEERS and the VISUAL data sets are shown in Figure 1. The number of galaxies at the different redshifts analyzed is shown in Table 1.

2.2. Mock JWST Images of TNG50 Galaxies

We use the TNG50-1²⁷ suite of simulation (hereafter TNG50; Nelson et al. 2019a, 2019b; Pillepich et al. 2019) and their mock NIRCcam observations at $z > 3$ galaxies following the observational strategy of CEERS. The mock images²⁸ were produced by modeling the gas cells and star particles in the simulation as presented in Costantin et al. (2023). We consider four snapshots of the TNG50 simulation corresponding to $z = (3, 4, 5, 6)$ and galaxies with stellar masses $M_* \geq 10^9 M_\odot$. In total, the original data set consists of 1,326 galaxies (see Table 1). Each selected galaxy is then observed along 20 different line-of-sight orientations to increase the statistics, which produces a data set of 26,520 galaxy images that we consider as independent objects for the

purpose of this work. As described in Costantin et al. (2023), parametric and nonparametric morphological parameters for this data set are derived using the standard configuration of `statmorph`²⁹(Rodriguez-Gomez et al. 2019).

For this study, we use the noiseless images in the F200W and F356W bands from the Costantin et al. (2023) data set with a pixel scale of $0''.031$ and $0''.063 \text{ pix}^{-1}$, respectively. We decided to use these two filters simultaneously since they probe the UV, optical, and near-IR rest frame at $z > 3$ (see Figure 2 in Ferreira et al. 2022), allowing to probe simultaneously the distribution of young and old stars and offering a complete view of galaxy morphology for our data-driven approach. Another reason for including the F200W is its spatial resolution. The F200W filter is the filter with the longest wavelength at the $0''.031 \text{ pix}^{-1}$ resolution of the NIRCcam short-wavelength channels (F115W and F150W). Although resampled into the $0''.031 \text{ pix}^{-1}$ resolution after drizzling, the NIRCcam long-wavelength channels (F277W, F356W, F410W, and F444W) have a worse spatial resolution (originally $0''.063 \text{ pix}^{-1}$ resolution) than the NIRCcam short-wavelength channels.

The original field of view of each image is twice the total half-mass-radius (i.e., dark matter, gas, and star particles included) of the corresponding galaxy. This roughly corresponds to a field of view 10 times larger than the stellar half-mass-radius (i.e., only star particles included). However, our image classification scheme requires a fixed image size. Therefore, we select galaxy images with a field of view larger than 64×64 and 32×32 pixels in the F200W and F356W bands, respectively, and generate cutouts of those sizes. Then, to match both observations of the same galaxy, the images in the F356W band are resampled to the same pixel scale as the F200W images (see Section 3.2, for more details). Given the original field of view of each galaxy image, after cutting them to the input fixed size of our network (64×64 pixels), the cutouts will certainly include all the luminous matter in the stamps.

According to these criteria, $\lesssim 7\%$ of the galaxies (most of them at $z = 3$) are dropped out from our initial sample. The total number of galaxies considered is finally 1,238 distributed within $z = 3-6$ (see Table 1), which translates into 24,760 projections. Although the number of objects we remove is small, we check in Figure 2 if a specific population is systematically excluded. The figure shows the size-mass relation of the selected TNG50 data set along with the excluded galaxies based on the size of the field of view. The excluded galaxies are not necessarily the most compact and/or less massive galaxies in the data set. However, a fraction of them with lower-than-average stellar extent is indeed removed based on our selection and would reach otherwise sizes of a few hundred parsecs.

For comparison, we show in Figure 1 the distribution of the stellar masses in the simulated TNG50 data set, and the observed CEERS and VISUAL data sets. Note the good agreement between the TNG50 and the CEERS samples, even if we are comparing here the stellar masses directly extracted from the TNG50 simulations and those obtained through SED fitting of the latest JWST data. Also remarkable is the agreement, despite the selection effects, between the CEERS, VISUAL, and TNG50 data sets.

²⁷ <https://www.tng-project.org/>

²⁸ Data publicly released at <https://www.tng-project.org/costantin22>.

²⁹ `statmorph` is available at <https://statmorph.readthedocs.io>.

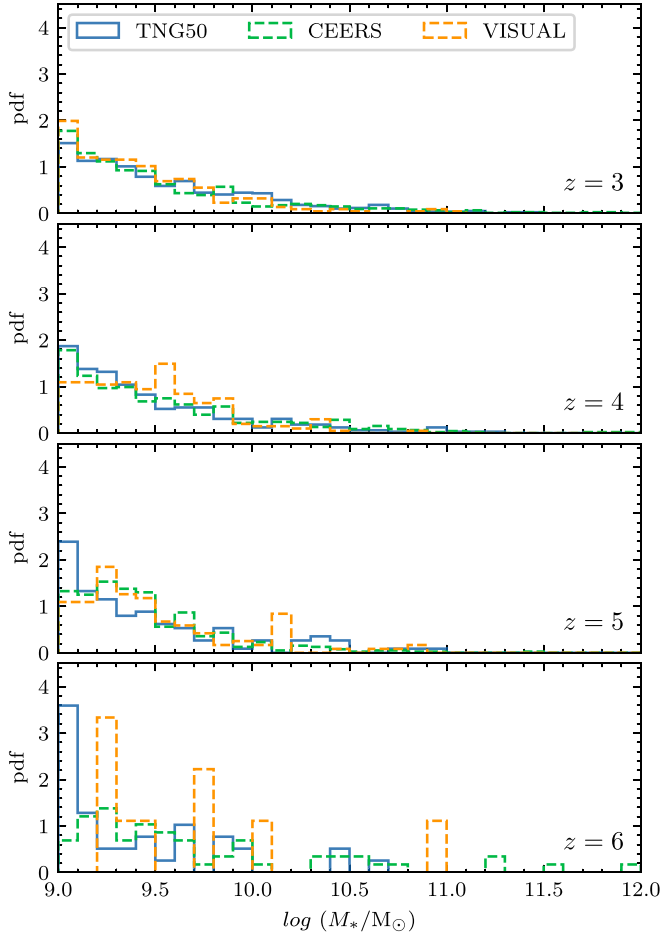


Figure 1. Probability density function of the logarithm of the stellar mass of the simulated galaxies in the TNG50 data set (blue histogram), and the observed galaxies in the CEERS and VISUAL data sets (green and orange dashed histograms, respectively). Different panels correspond to the redshifts analyzed, $z = (3, 4, 5, 6)$.

3. Self-supervised Learning Representation of Mock JWST Images of Simulated TNG50 Galaxies

In this section, we describe the main methodology we developed to obtain a data-driven morphological description of galaxies that is robust to noise and other nuisance parameters.

3.1. Contrastive Learning Framework

Our approach is based on an adaptation of the Simple framework for Contrastive Learning of visual Representations (SimCLR; Chen et al. 2020a). Very briefly, the idea behind the SimCLR framework is to obtain robust representations of images without labels by applying random augmentations as explained below. See Huertas-Company et al. (2023b), for a recent review of this technique applied to astrophysics.

Given an image, random transformations are applied to it to generate a pair of two augmented images, (x_i, x_j) . Each image in the pair is passed through a CNN to compress the images into a set of vectors, (h_i, h_j) . Then, a nonlinear fully connected layer (i.e., projection head) is placed to get the representations (z_i, z_j) . The representations are learned iteratively by maximizing agreement between the augmented views of the same image example (z_i, z_j) and minimizing agreement between all other pairs considered as negatives. This is achieved via a so-called

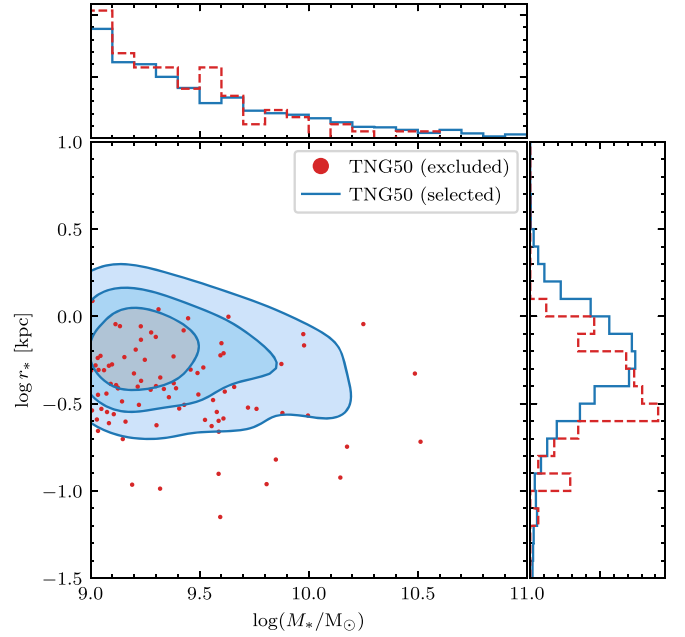


Figure 2. Logarithm of the physical size (stellar half-mass-radius, r_* , in kiloparsecs) vs. the stellar mass (M_* , in M_\odot) of the TNG50 galaxies. Blue-filled contours show the 25%, 50%, and 75% probabilities of the TNG50 selected galaxies. Red data points correspond to the excluded galaxies in terms of the size of the field of view. See Table 1.

Table 1
Summary of the Sample of TNG50 Simulated Galaxies and CEERS Observed Galaxies with $M_* \geq 10^9 M_\odot$

z	TNG50 (All)	TNG50 (Selected)	CEERS	VISUAL
3–6	1326	1238	1664	545
3	829	760	741	216
4	343	326	463	201
5	115	113	398	119
6	39	39	62	9

Note. The first column indicates the redshift (z); the second column shows the total number of galaxies in the simulated TNG50 data set; the third column refers to the number of selected galaxies according to image size limitations (i.e., 64×64 and 32×32 pixels in the F200W and F356W bands, respectively); the fourth column shows the number of galaxies in the CEERS data set; the fifth column indicates the number of galaxies in the VISUAL data set. Note that for the CEERS and VISUAL data sets galaxies are split into the following redshift bins: $z = 3$ for $3.0 \leq z < 3.5$, $z = 4$ for $3.5 \leq z < 4.5$, $z = 5$ for $4.5 \leq z < 5.5$, and $z = 6$ for $5.5 \leq z < 6.0$.

contrastive loss in the latent space:

$$l_{i,j} = -\log \frac{\exp(\langle z_i, z_j \rangle / h)}{\sum_{k=1, k \neq i}^{2N} \exp(\langle z_i, z_k \rangle / h)}, \quad (1)$$

where $\langle \mathbf{u}, \mathbf{v} \rangle$ denotes the dot product between L^2 -normalized \mathbf{u} and \mathbf{v} , and h denotes the temperature parameter that regulates the distribution of the output representations (see Hinton et al. 2015; Wu et al. 2018, for more details). The final loss is computed in batches of size N across all positive pairs, both (i, j) and (j, i) , while the rest of the augmented examples are treated as negative examples, which are denoted by k .

For this study, we follow the implementation from Sarmiento et al. (2021), which was successfully applied to

astronomical data. The CNN encoder consists of four convolutional layers with kernel sizes 5, 3, 3, 3, and 128, 256, 512, and 1024 filters per layer, respectively. Max-pooling layers and exponential linear unit activation functions are placed after each convolutional layer. Therefore, the representations before the projection head— (h_i, h_j) —for each galaxy image are encoded into 1024 features. Subsequently, the projection head (composed of three fully connected layers of 512, 128, and 64 neurons per layer) transforms the galaxy representations to a latent space— (z_i, z_j) —where the contrastive loss is computed.

3.2. Data Augmentation and Network Training

The choice of data augmentations is a key element in contrastive learning training (Chen et al. 2020a) as it allows us to turn the representations independent of some nuisance effects. In the context of this study, our goal is to obtain a morphological representation that is robust to S/N, rotation, and size, and does not depend on color. To reach this objective, we calibrate our algorithm on the mock TNG50 data set (Section 2.2) since it allows us to access noiseless versions of the images and, therefore, marginalize the noise.

For each simulated TNG50 galaxy image, we produce two augmented images (x_i, x_j) from the noiseless version. One of the images from the pair—the one used to produce the noise-added image—is then convolved by the corresponding point-spread function (PSF) in each filter (extracted from the observations from Finkelstein et al. 2022 in each band), while the other image—the one used to produce the noiseless image—is not convolved by the PSF to keep as much as spatial information as possible. Both images are rescaled in flux (as described below) independently. Then, we add the source Poisson noise only to the image used for the noise-added version. To match the same pixel scale in the two filters, we then rebin the images in the F356W filter to the same pixel scale as the images in the F200W filter. Finally, for the noise-added image, we include realistic noise by adding random patches of the sky extracted from the 10 CEERS pointings. Below, we described more precisely the augmentations applied to the images:

1. *Rotation.* We first apply a random flipping (horizontal or vertical, but not both) and a random rotation with 100% chance independently to the TNG50 noiseless image and the patch of the sky extracted from the CEERS pointings. Also, the noiseless and the noise-added version are rotated and flipped differently. This augmentation is intended to ensure the model is invariant to the galaxy orientation.
2. *Flux.* We randomly apply flux scaling to the noise-added version of the images after the PSF convolution, but before noise is added. The flux factor applied to the noise-added images in the two filters is randomly sampled from the flux distribution of the TNG50 data set in the F356W band. The same flux factor is applied to both the F200W and F356W filters. This augmentation is intended to stress the robustness to S/N. It may also help—as we will show in the following—to make the representations independent of galaxy size since the regions above the noise level will vary with the flux variations.

3. *Noise.* As described in Section 2, for each galaxy image in the F200W and F356W bands, we have a noiseless version that does not include any instrumental effects or noise. Using the available CEERS data, we construct mock CEERS galaxy images as a combination of the TNG50 noiseless images and random patches of the ten observed CEERS pointings. For the noise-added version, we first add the source Poisson noise, and then, we add real-time realistic noise (that may also include other sources and/or interlopers) to each of the 64×64 noiseless galaxy images by summing up one randomly chosen patch (different in each augmentation) from the CEERS pointing of the same size. From the contrastive learning point of view, these images with a real background are considered as an augmented copy of their noiseless analogs during the training process. These augmentations should enforce the representations to be robust to S/N as well as to background and foreground companions.
4. *Color.* Finally, in order to prevent the network from learning color information and/or the intrinsic brightness of the galaxies directly from the images, we apply two additional augmentations to both the noiseless and the noise-added images. First, each band is normalized individually after the augmentations are applied. Second, the noiseless and noise-added images are normalized independently and individually for each galaxy. Consequently, the maximum pixel value in every galaxy image (both noiseless and noise-added) is equal to one for each band.

We note that we decided not to apply direct size augmentation, (i.e., such as zoom-in or zoom-out), as that would force us to up-sample or down-sample the images with less or more than 64×64 pixels size, respectively, which creates some artifacts that the network is able to learn. The choice of the cutouts' size when producing the galaxy images is always complicated. One might prefer to make the stamps proportional to the size of the galaxy (see Vega-Ferrero et al. 2021, for an example). However, that requires reliable measurements of the galaxy sizes and also the resampling of the cutouts to the same size in pixels. Contrarily, it is possible to produce all the cutouts with the same size in pixels, independently of the galaxy size. By doing so, it is not needed to resample the cutouts since they already have the same dimensions. We decide to use a fixed size for the cutout to avoid the following: first, artifacts and noise correlations originating from the resampling phase that could mislead the contrastive learning representations; and second, losing spatial resolution of galaxies with large sizes after the resampling phase.

In Figure 3, we show several projections of a galaxy at $z = 3$, with $M_* \approx 6 \times 10^9 M_\odot$ extracted from the TNG50 simulation. In some of the augmented versions (along with PSF convolution, realistic noise, random rotations, flux variation, etc.), it is possible to distinguish several companions within the stamps. This is the result of adding randomly chosen patches of the CEERS pointings to the TNG50 noiseless images. These examples come from an extended bright galaxy that appears significantly dimmer in some of the augmented images due to the flux variations applied in the augmentations. In summary, the contrastive model should be able to extract a meaningful representation (h_i, h_j) that minimizes the distance between the

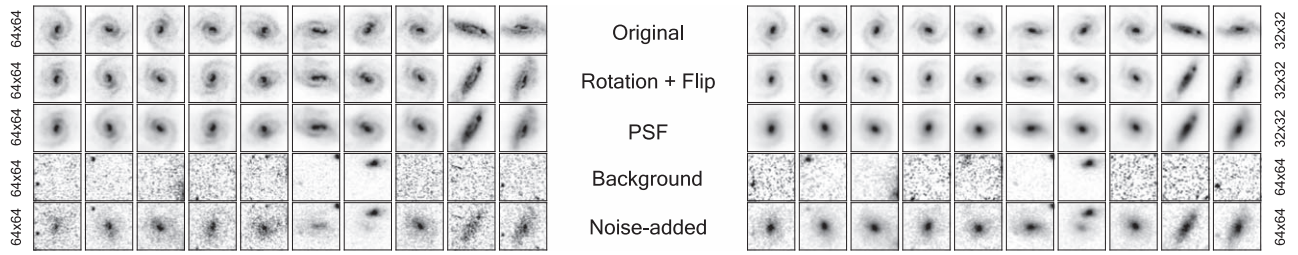


Figure 3. Images and augmentations of a TNG50 galaxy at $z = 3$ with $M_* \approx 6 \times 10^9 M_\odot$ in the F200W (left-hand panel) and F356W (right-hand panel) filters. The different rows (from top to bottom) correspond to: the original and/or noiseless image; the original image after rotation and flipping (vertical or horizontal); the image after PSF convolution; the background CEERS patch of the sky; the noise-added image, i.e., the sum of the PSF convolved image (flux variation and source Poisson noise applied) and the background patch of the sky. All the panels have a 64×64 pixels size, except the first three rows for the F356W filter that have the original size of 32×32 pixels before rebinning to 64×64 pixels. The pixel values have been *asinh* transformed with a 0.5% clipping.

representations (z_i, z_j) of the two images (original and/or noiseless and noise-added) of the same galaxy, even if they appear as different as shown in some panels in Figure 3.

Our contrastive SimCLR model is trained and tested with the mock JWST images for 24,760 different projections of 1,238 galaxies within $3 \leq z \leq 6$ and with stellar masses $M_* \geq 10^9 M_\odot$ in the two observed bands (F200W and F356W) with a temperature parameter $h = 0.5$ (that controls the strength of penalties on hard negative pairs). We randomly split our data set into a training and a test sample consisting of 1,100, and 138 galaxies, respectively. This translates into a training and a test data set of 22,000 and 2760 galaxy images, respectively. None of the projections of the galaxies in the test set has ever passed through the network during training. Additionally, we ensure that only one projection of the same galaxy enters each batch during training and that all the galaxy images are passed through the network at every epoch. We do so to avoid the algorithm learning the orientation of the same galaxy as seen from different line-of-sight projections since some of the projections are just a simple rotation of the galaxy in the sky.

The input tensors in our contrastive model have, therefore, a dimension of $(N, 64, 64, 2)$, with N being the batch size, 64 and 64 being the dimensions of the input images, and 2 being the number of channels or filters (i.e., the F200W and the F356W images). We train our algorithm with a batch size of $N = 550$ (i.e., half the number of galaxies in the training set) for 1,500 epochs in a GPU NVIDIA T4 Tensor Core with 16 GB of RAM. Random data augmentation is applied every 50 epochs, and therefore, we produce 30 (1500/50) different augmentations of the whole data set to increase the variability during the training process. To reduce the dynamic range and to be sensitive to both the center and outskirts of the galaxy, before training, we apply a *asinh* (inverse hyperbolic sin) transformation and a minimum–maximum normalization to each galaxy pair in each band.

In order to reduce the impact of possible discrepancies between the galaxies in the simulated TNG50 and observed CEERS data sets, when applying our model to data, we fine-tuned the model trained previously with a mixed data set of simulated TNG50 and observed CEERS galaxies. By doing so, we expect the model to learn important features from the observed CEERS galaxy images that were not present in the training set consisting only of simulated TNG50 galaxy images and, therefore, mitigate possible domain drift-related effects. We fine-tune the model with a training set consisting of 1,500 images randomly extracted from the previous training set (of noiseless and noise-added images) and 1500 images randomly selected from the CEERS data set of 1,664 galaxy images. Note

that for the observed data set we do not have noiseless images, so we fed the model with pairs of CEERS galaxy images to which different augmentations (only flip and rotation) are applied. We fine-tune the model up to 600 epochs (enough to converge) with a batch size of $N = 500$ galaxy images. In each epoch, random augmentations are applied to the observed CEERS pairs of images to increase the variability. For the simulated TNG50 images, the set of 1,500 also varies from epoch to epoch by selecting different galaxies and augmentations for each galaxy in the training TNG50 data set, but not from the reserved test set, which is always kept apart from the training.

3.3. Visualization of the Representation Space

We can hence analyze the properties of the representation space learned by the SimCLR framework presented in the previous section.

We start by visualizing how the TNG50 galaxy images, both with and without noise, are distributed in the representation space. It should be noted that, hereafter, the difference between the noiseless and the noise-added version of the TNG50 galaxy images is only the added patch of the CEERS paintings. Apart from the realistic noise added, no other augmentations are applied in this phase (only in the training phase described in the previous section). Since the representation space for each galaxy image is encoded into 1024 features, we perform a dimensionality reduction from the 1024 features to a 2D space to facilitate the visualization and interpretation of this representation. For that purpose, we use the uniform manifold approximation and projection (UMAP; McInnes et al. 2018) method with standard initial parameters (metric=euclidean, $n_neighbors = 15$, and $min_dist = 0.1$). The UMAP algorithm seeks to learn the manifold structure of the input data and find a low-dimensional embedding that preserves the essential topological structure of that manifold. It is therefore a way to visualize in 2D the representations learned by the self-supervised network. Before applying the UMAP technique, we assume the same distance metric in the representation space as the one used to calculate the contrastive loss in the head projection space, and therefore, we normalize the representations with an L^2 -norm such that the Euclidean and cosine distances between representations are equivalent. The two coordinate axes in the UMAP representation do not have any precise physical meaning and are a combination of the 1024 dimensions extracted by the contrastive learning setting.

It is important to emphasize that the contrastive approach is not intended for dimensionality reduction but for obtaining a robust representation of galaxy morphologies in a different

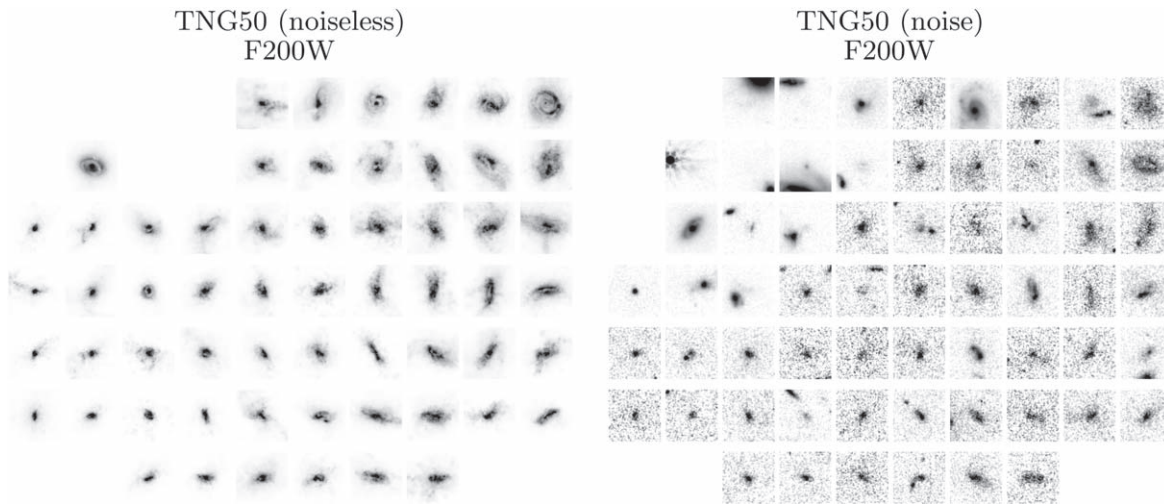


Figure 4. Randomly chosen images of the simulated TNG50 galaxies in the UMAP visualization. The UMAP space is binned, and one galaxy image per bin is shown. The left-hand panel shows the noiseless versions of the training galaxies, while the reduced versions (TNG50 + random CEERS patch) are shown in the right-hand panel. Extended and/or disk galaxies lie in the upper right, compact and/or rounder galaxies are located in the bottom left, and bright companions (only in the right-hand panel) concentrate toward the upper left section of the UMAP plane. Note that there is not a one-to-one correspondence between the galaxies shown in the two panels. Both panels correspond to galaxy images in the F200W filter.

space than images, which explains the high dimensionality of the representation space. Several works have shown indeed that the performance of contrastive learning increases with representations of higher dimension (e.g., Chen et al. 2020a).

In Figure 4, we show random examples of galaxies in the F200W filter (both with and without noise) in the UMAP 2D space. Note that some stamps on the right-hand panel (along with the addition of observed noise) show one (or more) foreground and/or background source(s) in the field of view. The figure clearly shows that galaxies are not randomly organized in the plane, indicating that the network has learned some relevant morphological features. The distributions are also similar for galaxies with and without noise. Galaxies with extended light distributions and with clear signs of a disk component—or interactions—tend indeed to appear on the right and upper right parts of the UMAP space, while more compact galaxies with smoother and concentrated light distributions tend to be placed on the left section of the plane. We can also see that galaxies showing more elongated shapes are found toward the bottom right section of the representation space. Also interesting to notice is how several galaxies with bright companions (off-center sources) tend to be placed on the upper left of the right-half panel (see Section 3.4 below for a more detailed discussion on this point).

3.4. A Morphological Description of Galaxies Robust to Noise and Background and/or Foreground Contaminants

We now examine in more detail the differences between the representations of noise-added and noiseless TNG50 galaxy images, and how the different augmentations of the same galaxy are represented by our contrastive model. As described in Section 3.1, one of the reasons for using the SimCLR framework is to obtain a data-driven representation that is robust to noise and other observational effects such as foreground and background companions.

On one hand, we quantify the effect of noise by computing the distance in the UMAP representation between the noiseless and the noise-added images of each galaxy, denoted as δ . On the other hand, we check how our contrastive model behaves

when more than one source (i.e., companions) is present in the stamp. To do so, we measure the total flux in the noise-added galaxy images (TNG50 + random CEERS patch) and in the noiseless TNG50 stamps (therefore, the intrinsic flux of the central galaxy after the flux scaling is applied) for the F356W filter. Then, we compute the ratio of the two, denoted as δ_{F356W} , as a proxy for the presence (or not) of companions and, if present, how bright they are with respect to the central galaxy.

In Figure 5, we show the UMAP plane for TNG50 galaxy images in our data set color-coded by δ and δ_{F356W} . For a reference of the UMAP axis ranges, the horizontal axis (UMAP 1) spans within (0.9, 12.1), and the vertical axis (UMAP 2) spans within (−1.7, 5.5). The total area covered by the data points in the UMAP plane is approximately 60 (in the arbitrary UMAP units). It is interesting to note how the main yellow clump in the upper left section of the UMAP plane where the stamps with bright companions (more than 3 times the flux than the flux of the central galaxy, i.e., $\delta_{F356W} \gtrsim 4$) tend to concentrate, and also their correlation with large values $\delta \gtrsim 2.5$. Some of these cases can be seen in the right-hand panel in Figure 4. For instance, there are several examples within these regions of $\delta_{F356W} \gtrsim 4$ that correspond to TNG50 images for which the companion is so bright (such as a star) that the central galaxy cannot be even identified in the stamp. In these cases with bright companions around the central galaxy, the model detects the brightest component (thus, the companion) instead of the central galaxy and tends to represent it in a particular region of the UMAP plane. Additionally, in the right-hand panel, it is also clear a yellow clump with large values of δ_{F356W} and moderate values of δ . These cases correspond to noise-added images with a bright companion that, contrarily, are still close to their noiseless counterpart (i.e., $\delta \lesssim 2$). We inspect in detail several of these cases and find that their noiseless counterparts tend to be located on the bottom right section of the UMAP plane. In any case, a galaxy image represented close to these regions of $\delta \gtrsim 2$ might be not well-represented and should be (at least) treated carefully or excluded from the subsequent analysis. It is also interesting to note the clump with large values of δ in the left edge of the representation space

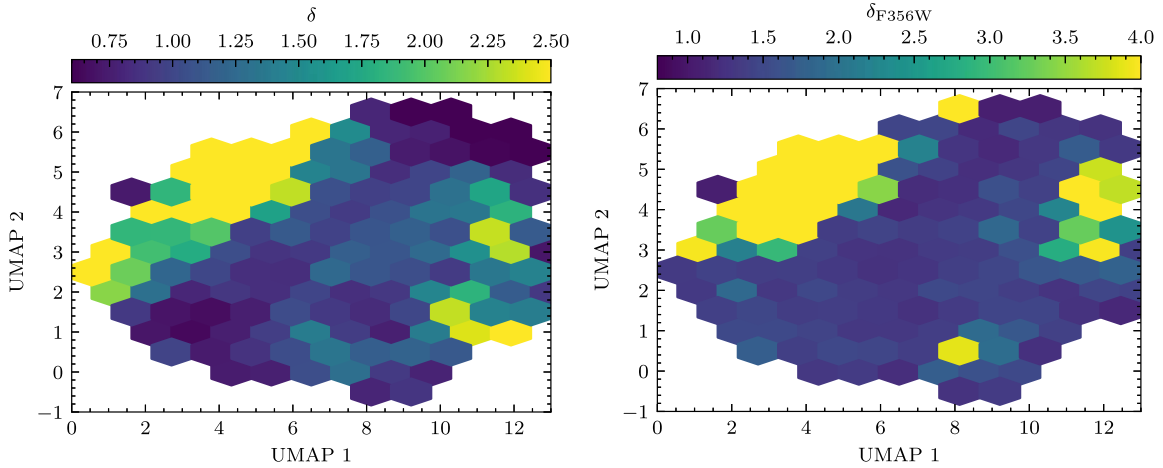


Figure 5. Left-hand panel: UMAP visualization for all the TNG50 galaxy images in our data set color-coded by the mean value of the distance in the UMAP representation between the noiseless and the reduced images of each galaxy (denoted as δ). Right-hand panel: UMAP visualization for all the TNG50 galaxy images in our data set color-coded by the mean value of the ratio of the flux measured in the TNG50 stamps and the flux derived from the noiseless TNG50 stamps for the F356W filter (denoted as δ_{F356W}). Large values of δ_{F356W} indicate the presence of a secondary (or even more) source. The larger δ_{F356W} is, the brighter the companions are with respect to the central galaxy.

(i.e., $0.0 < \text{UMAP 1} < 1.5$ and $2.0 < \text{UMAP 2} < 3.5$) with no clear correlation with large values of δ_{F356W} . Although this clump does not include a large number of galaxies, we find that those objects have their noiseless counterparts displaced up in the UMAP plane. These are very compact galaxies that look even more compact in their noiseless versions.

By excluding objects falling within the yellow clump shown in Figure 5, it is possible to clean our data set (or a JWST/CEERS sample of observed galaxies) of bright companions and/or contaminants that could bias our contrastive learning representations. As an alternative, we build up a data set of simulated TNG50 galaxy images that include all the possible augmentations described before but keeping a value of $\delta_{F356W} < 2.5$. By doing so, we ensure our data set does not include companions 1.5 brighter (in flux) than the central galaxy. Nevertheless, these extremely bright contaminants are much rare in real CEERS observations than in our augmented set of galaxy images because, if the contaminant is so bright that it outshines the central galaxy, the galaxy would not be detected.

In Figure 6, we show the cumulative distribution function of δ for the galaxy images in the training and the test sets, both including and excluding galaxy images with $\delta_{F356W} > 2.5$. We find that 75% and 90% of the projections in the training data set show values of $\delta \lesssim 1.5$ and $\delta \lesssim 3.0$ in the UMAP space, respectively. For the test set, 75% and 90% of the projections show values of $\delta \lesssim 1.4$ and $\delta \lesssim 2.7$ in the UMAP space, respectively. If we now exclude galaxy images with $\delta_{F356W} > 2.5$, for the training data set, 75% and 90% of the projections show values of $\delta \lesssim 1.2$ and $\delta \lesssim 2.4$ in the UMAP space, respectively; while, for the test data set, 75% and 90% of the projections show values of $\delta \lesssim 1.3$ and $\delta \lesssim 2.3$ in the UMAP space, respectively. A displacement of δ is analogous to saying that the noise and noiseless representations of the same galaxy pair are located within a circle of radius δ . Converted into an area, this means $\approx 8\%$ displacement in the UMAP plane for 75% of the galaxy images (i.e., $\delta \lesssim 1.2$) in the training set despite the level of noise, contamination, and augmentations applied to the input galaxy images, as can be seen in Figures 3 and 4. It should be noted that excluding cases with

$\delta_{F356W} > 2.5$ does not remove all the cases with companions since there are still cases of companions with fluxes up to 1.5 times the flux of the central galaxy. Even when these cases are included, the model performs satisfactorily well for a large fraction of the galaxy images presented in this study. Besides, the distributions of δ for the training and test samples are very similar. Therefore, we emphasize the model is not suffering from overfitting since none of the galaxy images in the test set has been shown previously to the network.

To further illustrate the effect of companions and noise on the representation space, we show some examples of the most extreme cases ($\delta > 5$ and $\delta_{F356W} > 10$) in Figure 7. The majority of images with the largest values of δ are mainly due to the presence of bright companions in the galaxy images (or artifacts). It is possible to identify compact bright companions (such as a star in case 3 and a compact galaxy in case 1), and more extended companions (galaxies in cases 2, 4, and 5).

Therefore, training our contrastive model with a combination of noiseless and noise-added TNG50 images leads to a robust representation of TNG50 images even in the case of the presence of companions in the image (at least, for those cases in which the companion is not extremely bright compared to the central galaxy). For the cases in which the companion is much brighter than the central galaxy, their locations in the UMAP may certainly help to find them in observed images and to treat them carefully in subsequent analysis.

Hereafter, we show results for the representation of this *clean* data set (i.e., only galaxies with $\delta_{F356W} < 2.5$ are considered) for which the representations obtained are not affected by extremely bright contaminants in the galaxy images.

3.5. Dependence on Physical and Photometric Parameters

An advantage of calibrating the neural network model with simulations is that we have access to a large number of physical properties of the galaxies. An additional test for our classification scheme is, therefore, to examine how the representation space is correlated with physical quantities as well as with other (more standard) morphological measurements.

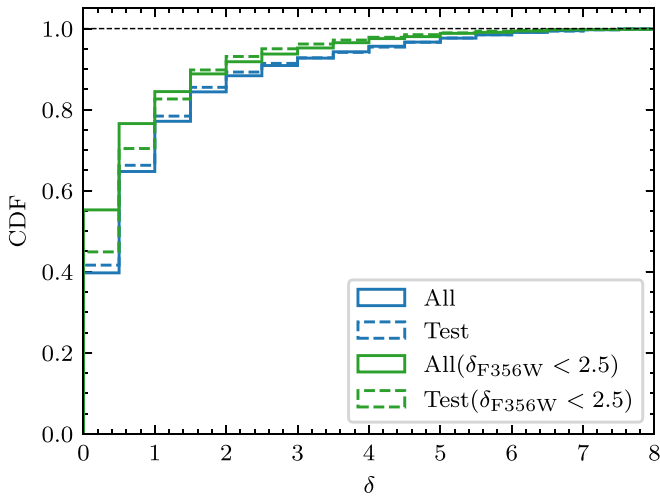


Figure 6. Cumulative distribution function of the distance in the UMAP plane between pairs of the same galaxy images, denoted as δ . Blue histograms correspond to the distribution of δ for the training (solid) and the test (dashed) data sets. Green histograms correspond to the distribution of δ for the training (solid) and the test (dashed) data sets when only galaxy images with $\delta_{F356W} < 2.5$ are considered.

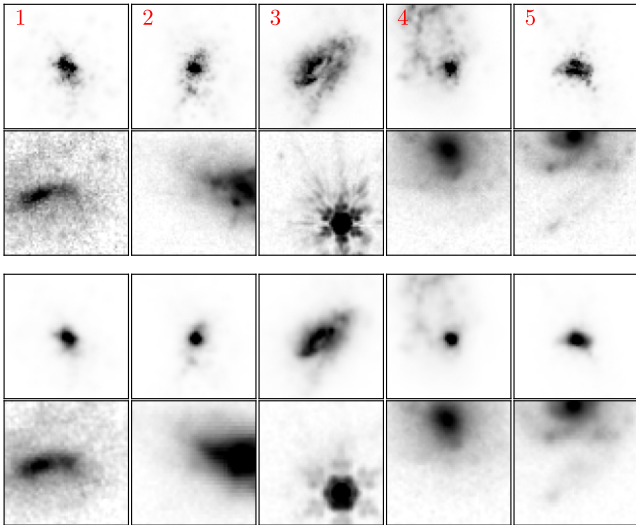


Figure 7. Randomly chosen examples of galaxy images with $\delta > 5$ and $\delta_{F356W} > 10$ (see Figure 5). For each of the examples, we show noiseless and noise-added images in the F200W (top rows) and F356W (bottom rows) filters.

3.5.1. Correlation with Physical Properties

In this section, we discuss how some physical properties extracted from the TNG50 simulation correlate with the representation in the UMAP plane. In Figure 8, we show the dependence in the UMAP plane with the total stellar mass ($M_*[M_\odot]$), the specific angular momentum of stars ($j_*[\text{kpc km s}^{-1}]$), the mass fraction in nonrotating stars (f_{nr}), and the flatness ($1 - f$) of the galaxy. The mass fraction in stars that have no net angular momentum around the z -axis is defined using the circularity parameter $\epsilon = J_z/J(E)$, as in Marinacci et al. (2014), for every star particle. It measures the maximum specific angular momentum possible at the specific binding energy E of the star. The mass fraction in nonrotating stars mass (denoted as f_{nr}) is then defined as the fractional mass of stars with $\epsilon < 0$ multiplied by 2. The flatness of the galaxy is computed as follows: $f = c/\sqrt{ba}$, where $c < b < a$ denote the

principal axes obtained as the eigenvalues of the mass tensor of the stellar mass inside $2r_*$. The larger $1 - f$ is, the flatter the system is in 3D. Here, and throughout the paper, we refer to the definitions and measurements of Pillepich et al. (2019). See also Section 6 for a more detailed discussion of the 3D shapes of the TNG50 galaxies.

Figure 8 shows remarkable correlations between the position of galaxies in the UMAP and their average physical properties. Overall, galaxies with larger specific angular momentum and a flatter stellar distribution tend to populate the upper right region of the UMAP. These galaxies are also the most massive ones although the correlation is less clear. Moreover, the galaxies with larger masses not only occupy the upper right section of the UMAP plane but also extend along the upper edge toward the left corner of the UMAP plane. On the contrary, low-mass galaxies populate predominantly the bottom left section of the UMAP representation. The left section of the UMAP plane is populated by rounder objects with lower specific angular momentum. It is also interesting to see that the transition between the variation of the physical properties is smooth, translating a continuum of galaxy morphology and/or structure.

Figure 8 only shows the median values of the physical properties in different regions of the UMAP. In order to quantify how constraining are these correlations, it is also important to measure the scatter of the different properties. This is shown in the Appendix A (Figure A1). In most cases, the scatter represents less than $\sim 20\%$ of the dynamical range, indicating that the distributions are overall relatively narrow, and therefore, the correlations with physical properties are informative.

We conclude that the representation space for images—in addition to being robust to observational and instrumental effects—carries information about the kinematics and intrinsic shapes of galaxies.

3.5.2. Connection to Standard Morphological Measurements

In Figure 9, we show the dependence on several photometric parameters estimated by Costantin et al. (2023) in the F200W filter: the effective radius (r_e), the Sèrsic index (n_e), the ellipticity from the Sèrsic fit ($1 - b/a$), the concentration parameter (C), the asymmetry ($|A|$), and the smoothness (S). There is a remarkable correlation between the position in the UMAP plane and n_e , C , A , and S . Gradually, n_e , C , and S grow from right to left in the UMAP space, while the A does it from left to right. Therefore, galaxy images with smoother, symmetric, and concentrated light distributions are found toward the left section of the UMAP plane. Also important is the correlation with the ellipticity ($1 - b/a$), with more elongated galaxies lying on the right (bottom right) section of the UMAP plane. To illustrate again the spread of these representations, we show in the Appendix A (Figure A2) the scatter of the parameters shown in Figure 9.

It is important to notice the existing correlation with the physical effective radius, r_e , with the largest galaxies populating the right section of the UMAP plane. This correlation with the physical size reflects the known correlation between morphological appearance and physical size (e.g., van der Wel et al. 2014b).

Based on the previous maps calibrated with the TNG50 simulation, asymmetric, more extended, flatter, and rotationally supported galaxies tend to populate the right and upper right

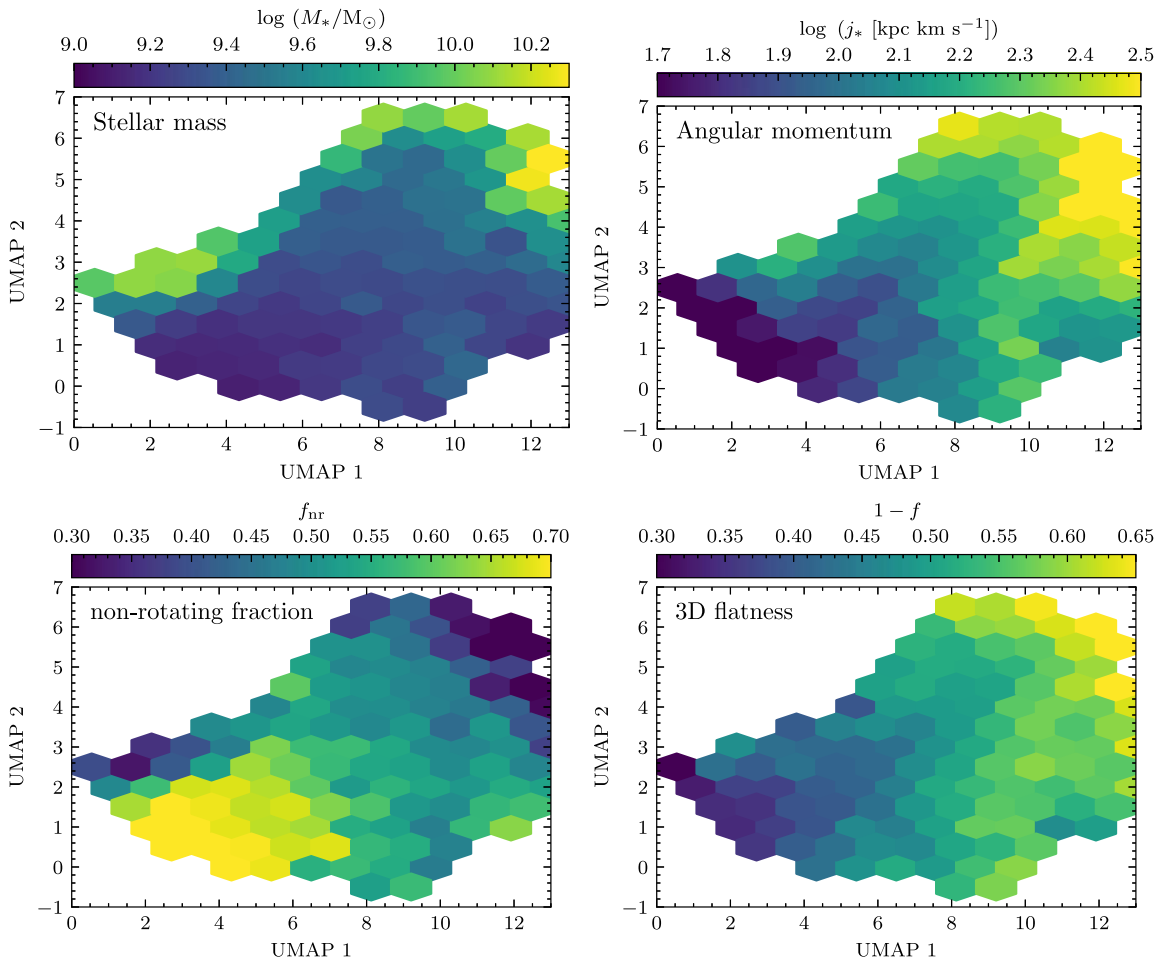


Figure 8. UMAP visualization for all the TNG50 galaxy images in our data set color-coded by the distribution of several physical properties extracted from the TNG50 simulation. Color code corresponds to the median values in each hexagonal bin in the UMAP plane. From left to right and top to bottom, the different panels show: the logarithm of the total stellar mass ($\log M_*/M_\odot$), the logarithm of the specific angular momentum of the stars ($\log j_* [\text{kpc km s}^{-1}]$), the mass fraction in nonrotating stars (f_{nr}), and the galaxy flatness ($1 - f$). The scatter maps of these parameters are presented in Appendix A.

sections of the UMAP representation. In more detail, the more to the right in the UMAP plane a galaxy is, the more elongated it appears. Smoother, more compact, rounder, and nonrotating galaxies are located toward the left section of the UMAP representation. Also, less massive galaxies can be found predominantly toward the bottom and bottom left sections of the UMAP plane. Although not shown, the results presented here are consistent (despite small variations) for the same morphological parameters measured in the F365W filter.

4. Self-supervised Learning Representation of JWST Galaxy Images

In this section, we apply the methodology described before to the two data sets of observed galaxies with JWST described in Section 2.1.

4.1. Representations of CEERS Galaxy Images

We feed the 1,664 observed CEERS galaxies to our contrastive model to retrieve their corresponding representations in the 1024 dimensions space. Then, we normalize the derived features and transform them into a 2D vector using the same UMAP embedding obtained for the features of the TNG50 galaxy images.

In the top row of Figure 10, we show the UMAP representation space for the observed CEERS data set. Interestingly, the observed galaxies tend to populate the complete UMAP plane, which indicates that both samples share similar morphological diversity. The UMAP visualization is however a projection of a higher-dimension space, which is not appropriate for outlier detection. Even if observed galaxies would not reside in the same manifold as simulated objects, the UMAP representation would tend to show them toward the edges of the plane but not outside. This is the behavior seen for observed CEERS galaxies, which tend to be concentrated in the edges of the UMAP cloud (toward the bottom and bottom left sections) independently of the source redshift. Given that the mass and flux distributions of both data sets are consistent—even though we have not performed a careful one-to-one match between simulations and observations—the differences in the distributions of points of both data sets are likely to originate in intrinsic differences in the morphological properties. Combining the distributions of points in Figure 10 with the information provided by Figures 8 and 9, we conclude that observed CEERS galaxies occupy more frequently than the simulated TNG50 galaxies the regions in the representation space where galaxies are more compact and with less specific angular

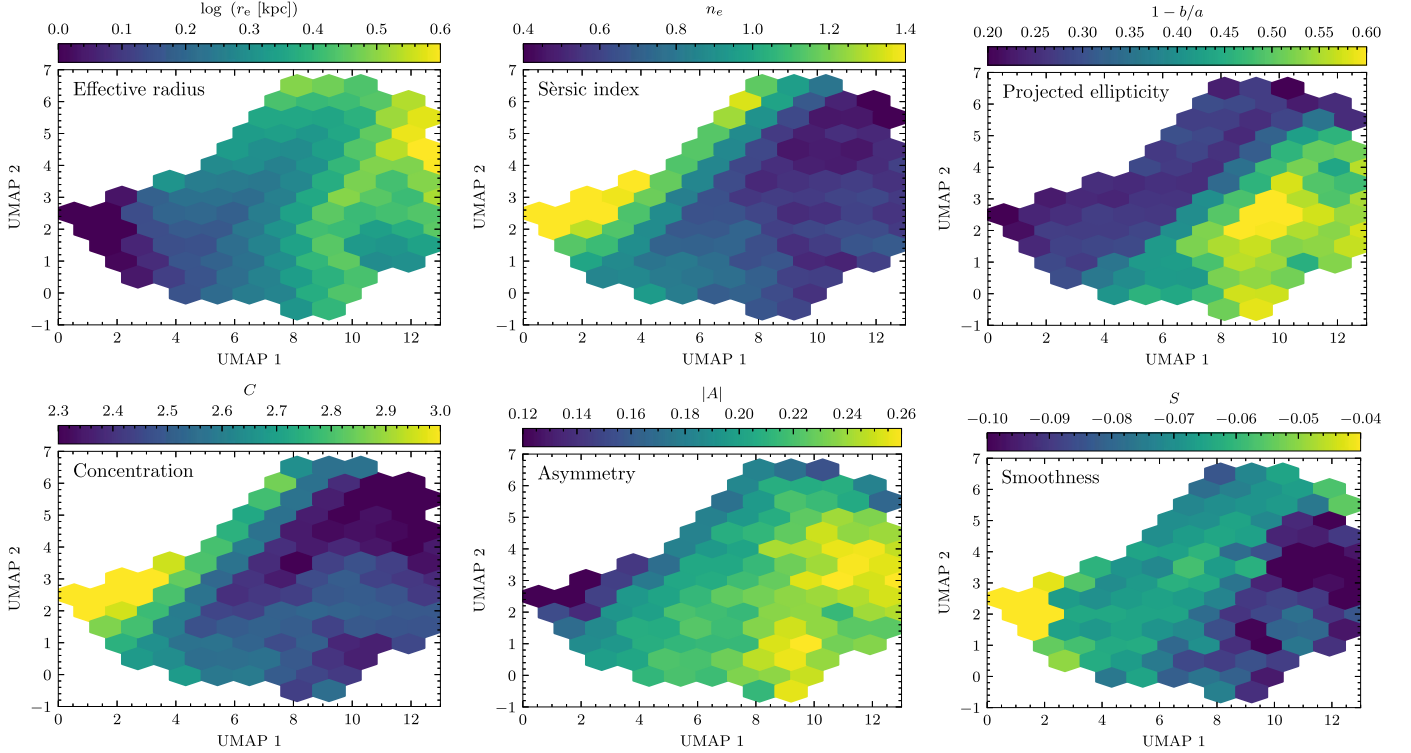


Figure 9. UMAP visualization for all the TNG50 galaxy images in our data set color-coded by the distribution of several morphological and photometric parameters. Color code corresponds to the median values in each hexagonal bin in the UMAP plane. From left to right and top to bottom, the different panels show: the logarithm of the effective radius (r_e [kpc]), in kiloparsecs, the Sersic index (n_e), the ellipticity based on Sersic fit ($1 - b/a$), the concentration (C), the asymmetry (A), and the smoothness (S). The scatter maps of these parameters are presented in Appendix A.

momentum. We investigate these differences in more detail in Section 5.

Also interesting is the presence of galaxy images with signs of interactions, multiple clumps, and gas accretion processes in the upper right section of the UMAP in Figure 10 (more clear in the F200W filter because of its better spatial resolution compared to the F356W filter). Moreover, the galaxy images with double nuclei (or even multiple clumps) closer in projection tend to appear in the bottom right section of the UMAP plane. These systems are apparently more elongated and, therefore, lay into the region of the UMAP plane where the projected ellipticities are on average larger (as shown in Figure 9), but less flat than the systems located on the upper right section of the UMAP plane (as shown in Figure 8).

Following Section 3.4, hereafter, we only include galaxies located within the black contours shown in Figure 10, for which it is unlikely to find bright companions or artifacts that could bias their representations. We find that $\sim 90\%$ (1 481) of the galaxies fulfill this criterion, while the remaining $\sim 10\%$ are excluded from the subsequent analysis. In this data set and according to the morphologies based on the F356W filter, 121 galaxies ($\sim 8\%$) are classified as Sph, 297 galaxies ($\sim 20\%$) are classified as disk, 96 galaxies ($\sim 6\%$) are classified as bulge + disk, and 967 galaxies ($\sim 65\%$) are classified as Irr. If we focus on the morphologies obtained from the F200W filter, the fraction of Irr galaxies raises up to $\sim 82\%$, and the fraction of disk galaxies decreases down to $\sim 6\%$.

4.2. Representations of VISUAL Galaxy Images

We also present a comparison of the representation obtained after applying our contrastive model to the VISUAL data set

for which visual morphological classifications are provided (Kartaltepe et al. 2023). After selecting those galaxies with $M_* \geq 10^9 M_\odot$, $3 < z < 6$, and reliable visual classifications, we end up with a data set of 545 galaxies. To avoid including in the analysis galaxy images with contaminants or artifacts, hereafter, we only consider those galaxies located within the black contours shown in Figure 10. In this case, we are confident about the representations obtained for $\sim 90\%$ (483) of the galaxy images, for which we find the following: 118 ($\sim 24\%$) disk galaxies, 102 ($\sim 21\%$) disk+Irr galaxies, 24 ($\sim 5\%$) disk+Sph+Irr galaxies, 56 ($\sim 11\%$) disk+Sph galaxies, 71 ($\sim 14\%$) Sph galaxies, 22 ($\sim 4\%$) Sph+Irr galaxies, 81 ($\sim 16\%$) Irr galaxies, and only 2 and 7 as point sources and unclassifiable galaxies, respectively.

In the bottom row of Figure 10, we show the representation of the galaxy in the VISUAL data set in the UMAP plane for the various morphological groups based on the provided visual classifications. Although the mass and redshift selection of the galaxies is based on different estimators (JWST photometry for the CEERS data set and CANDELS photometry for the VISUAL data set), the distribution of the representations in the UMAP plane for the VISUAL galaxies is similar to the CEERS representations (i.e., a significant fraction of galaxies occupy the bottom and bottom left section of the UMAP plane). The figure reveals some expected correlations with the traditional visual morphology. It is reassuring that disk+Sph and Sph groups from the VISUAL catalog populate the left corner of the UMAP plane, where compact, nonrotating galaxies with low angular momentum (according to TNG50 properties) are expected to be. However, we notice that galaxies classified as disk, Irr and/or disk+Irr in VISUAL are distributed throughout

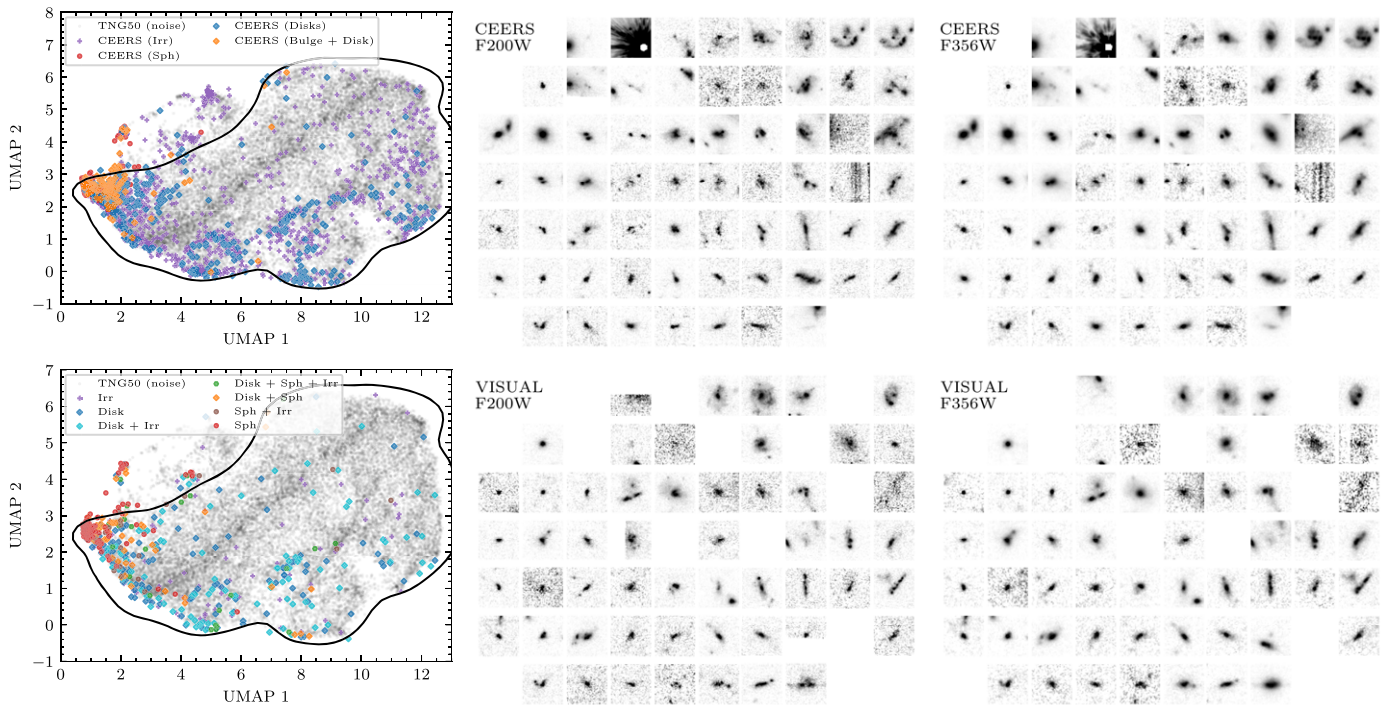


Figure 10. Comparison of distributions of observed and simulated galaxies in the representation space. The top row shows the CEERS mass-complete sample, and the bottom row shows the VISUAL sample. Top left-hand panel: UMAP visualization for the observed CEERS galaxy images selected in mass and redshift (color-coded by the CNN-based morphological classes derived in Huertas-Company et al. 2023a for the F356W filter) overlapped with the representation of noise-added TNG50 galaxy images. Black contour indicates the region that is not affected by extremely bright companions. Top middle panel: randomly chosen observed CEERS galaxy images in the UMAP visualization in the F200W filter. Top right-hand panel: randomly chosen observed CEERS galaxy images in the UMAP visualization in the F356W filter. Bottom left-hand panel: UMAP visualization for the observed VISUAL galaxy images selected in mass and redshift. Points are colored according to the visual classifications into several classes. Black contour indicates the region that is not affected by extremely bright companions. Bottom middle panel: randomly chosen observed VISUAL galaxy images in the UMAP visualization in the F200W filter. Bottom right-hand panel: randomly chosen observed VISUAL galaxy images in the UMAP visualization in the F356W filter.

the plane even toward the left section of the UMAP, very close to where Sphs lie. As shown in Figures 8 and 9, the left region of the UMAP where, according to VISUAL, disk-like morphologies are located corresponds to galaxies in TNG50 with physical and photometric properties typically shared by spheroidal systems, such as low specific angular momentum, large mass fractions in a nonrotating component, low flatness, and larger Sèrsic indexes. This raises interesting questions about the true nature of these disks that we discuss in Section 6.

5. A Comparison between Simulated and Observed Self-supervised Morphologies

In this section, we examine in more detail the differences found in previous sections between the simulated TNG50 and the observed JWST galaxy images.

5.1. Distribution of Self-supervised Representations

The representations of the simulated TNG50 and the observed CEERS galaxy images inferred by our contrastive model seem to be distributed differently (Sections 4.1 and 4.2). Observed CEERS galaxies tend to concentrate in the left and bottom left sections of the UMAP plane, while simulated TNG50 galaxies expand over the whole UMAP range with similar number densities. As previously mentioned, the UMAP representation is not well suited for the detection of outliers. Therefore, even if observed galaxies seem to (overall) lie in the same region as simulated ones, they can still live in different manifolds in the higher-dimensionality representation space.

To further quantify this distribution shift, we first derive the distance—in the 1024 dimensionality space—to the 10th closest TNG50 neighbor for each galaxy in the VISUAL and CEERS data sets (denoted as δ_{10}). In order to have a fair reference distribution, second, for each observed galaxy, we find the closest simulated TNG50 neighbor in the representation space and compute the distance to its 10th closest TNG50 neighbor (also denoted as δ_{10}). In other words, the former corresponds to the distance between the observed galaxy and its 10th closest TNG50 neighbor, while the latter corresponds to the distance to the 10th closest TNG50 neighbor of the closest TNG50 galaxy to each observed galaxy. If both data sets—observed and simulated—are distributed likewise in the same manifold, the distribution of distances should be similar. If, on the contrary, observed galaxies occupy differently the parameter space, their representations should be disconnected, and therefore, we should measure larger values of δ_{10} . The distributions of δ_{10} are shown in Figure 11. It can be clearly seen that the distributions for observed galaxies are shifted toward larger values compared to the reference distribution. This indicates that a significant fraction of observed galaxies are located in regions of the UMAP representation space with lower number densities (i.e., along the edges, not in the central regions) than the average of the TNG50 data set. This separation could be interpreted as an additional indication that the representations obtained for the TNG50 and the JWST observations do not exactly live in the same manifold. Nevertheless, we find a small fraction of simulated TNG50 galaxies with values of $\delta_{10} \gtrsim 0.9$, meaning that the separation

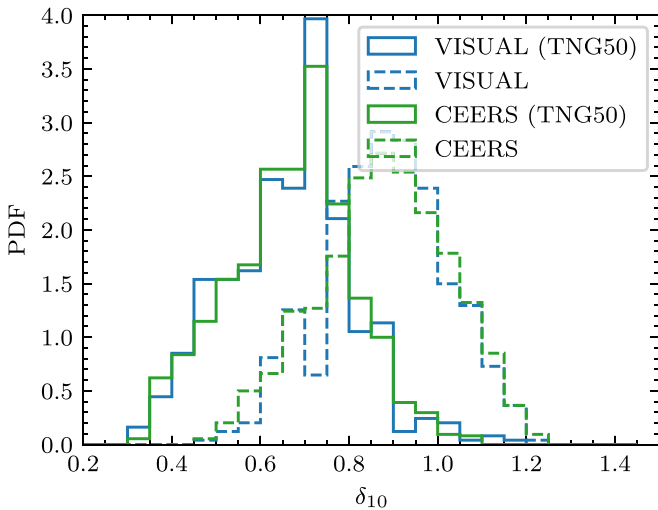


Figure 11. Probability density functions of the distances to the 10th closest neighbor in the 1024 dimensions of the representation space, denoted as δ_{10} . Solid histograms correspond to the distance to the 10th closest neighbor in TNG50 of the closest TNG50 neighbor of each galaxy in the VISUAL (in blue) and the CEERS (in green) data sets. Dashed histograms correspond to the distance to the 10th closest neighbor in TNG50 of each galaxy in the VISUAL (in blue) and the CEERS (in green) data sets.

for the most extreme cases is also present for some galaxies in the TNG50 simulation.

To better understand these measured discrepancies, we quantify the differences between observed and simulated galaxies in terms of more standard morphological properties in Figure 12. We show the distributions of observed and simulated galaxies in the $\log M_* - \log r_e$, $\log M_* - \log n_e$, and $\log M_* - b/a$ planes in four redshift bins. To divide in redshift, for the simulated data set, we take all galaxies in a given snapshot, while, for the observations, we include all galaxies that are associated with the closest snapshot based on their photometric redshifts. It should be kept in mind that this figure (as the previous ones) does not include all TNG50 galaxies, but those for which the JWST mocks are available for a field of view larger than 64×64 pixels (see Section 2.2 and Figure 2 for more details). Given the small number of galaxies removed, we do not expect the distributions to change significantly, though.

It is manifest, first, that the TNG50 simulated galaxies overlap with CEERS observed ones in the parameter space of Figure 12. This is per se, again, a nonnegligible confirmation of the zeroth-order good functioning of the underlying TNG50 model. However, it is also apparent, differently than what could be deduced from the representation space distributions, that the TNG50 galaxies studied here actually exhibit less galaxy-to-galaxy variation in sizes, Sèrsic indices, and shapes than CEERS observed galaxies, at fixed stellar mass and redshift. Furthermore, the TNG50 simulation predicts galaxies with larger sizes (at $z = 3-4$, but not $z = 5-6$), with smaller values of n_e (at all $z = 3-6$) and that are rounder in projection (more so the higher the redshift) than what is measured in the observed CEERS galaxies. These differences at least partly explain the different distributions in the representation space of contrastive learning and also go in the expected direction of observed galaxies mainly populating the left and bottom left corner of the UMAP.

These reported differences could originate from a resolution-induced effect (see, e.g., Zanisi et al. 2021) or could be an

indication of more fundamental physical differences. Resolution is certainly an important concern since galaxies at these redshifts are generally small. We recall that, although the TNG50 has a softening length for stellar particles of ~ 300 pc, it does not mean that galaxies’ stellar disks cannot be thinner than that softening length. The interplay of the various numerical resolution choices (such as gravitational softening of the different matter components, hydrodynamical smoothing length of the gas out of which stars form, mass resolution, etc.) manifests itself in very complex manners in the final structures of the simulated galaxies (see, e.g., Section 2.3 of Pillepich et al. 2019; or Section 2.4 of Pillepich et al. 2023). As shown in Pillepich et al. (2019; Figures 4 and B2), the half-light or half-mass heights of TNG50 galaxies can be smaller than ~ 300 pc depending on mass and redshift. Similarly, the stellar minor-to-major axis ratios of the stellar mass distributions of TNG50 galaxies can be smaller than $b/a \sim 0.3$, as shown in, e.g., Figure 8 (top panels) of Pillepich et al. (2019), again depending on mass and redshift. In fact, in Figure 12, it can be clearly seen how the axis ratios extend down to $b/a \sim 0.2$ (mainly at $z = 3-4$). Moreover, as shown in Appendix B2 of Pillepich et al. (2019), TNG50 disk heights can be considered converged to better than 20%–40% across all studied masses and redshifts, when compared to the same galaxies simulated at worse numerical resolution.

We note as well that the stellar masses reported for the TNG50 simulation correspond to the 3D stellar mass, while those obtained for the CEERS data set are based on the SED fitting to the JWST photometry. Also, the Sèrsic parameters for the TNG50 and the CEERS galaxies are derived using different methodologies: for the TNG50, the morphological parameters are obtained with `statmorph` (as described in Costantin et al. 2023); while, for the CEERS data set, they are derived with `galfit`.

More in-depth comparisons of simulated and observed data—likely beyond images—are required to reach a final conclusion.

5.2. Self-supervised Clustering in TNG50 and CEERS

As an additional way to quantify the differences between TNG50 galaxies and observed CEERS galaxies, we compare the abundances of TNG50 and CEERS galaxies retrieved from the separation into different classes using a clustering technique.

In particular, we apply the k –means algorithm to cluster data in the representation space by trying to separate samples in k groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (WCSS). We find $k = 5$ as the optimal number of clusters based on the elbow curve. The elbow method is a graphical representation of finding the optimal k in a k –means clustering. It works by finding WCSS, i.e., the sum of the square distance between points in a cluster and the cluster centroid. This result is also confirmed using an alternative method based on the silhouette score. We implement the elbow and silhouette methods using the Yellowbrick package in PYTHON (see Bengfort et al. 2018, for more details).

We label the different clusters according to the properties (photometric and morphological) of the galaxies belonging to each of them. In Figure 13, we show the properties (and correlations between them) of the different classes. Therefore, we define the following classes: *Extremely Compact* (EC), *Compact* (Cm), *Intermediate* (In), *Elongated* (El), and

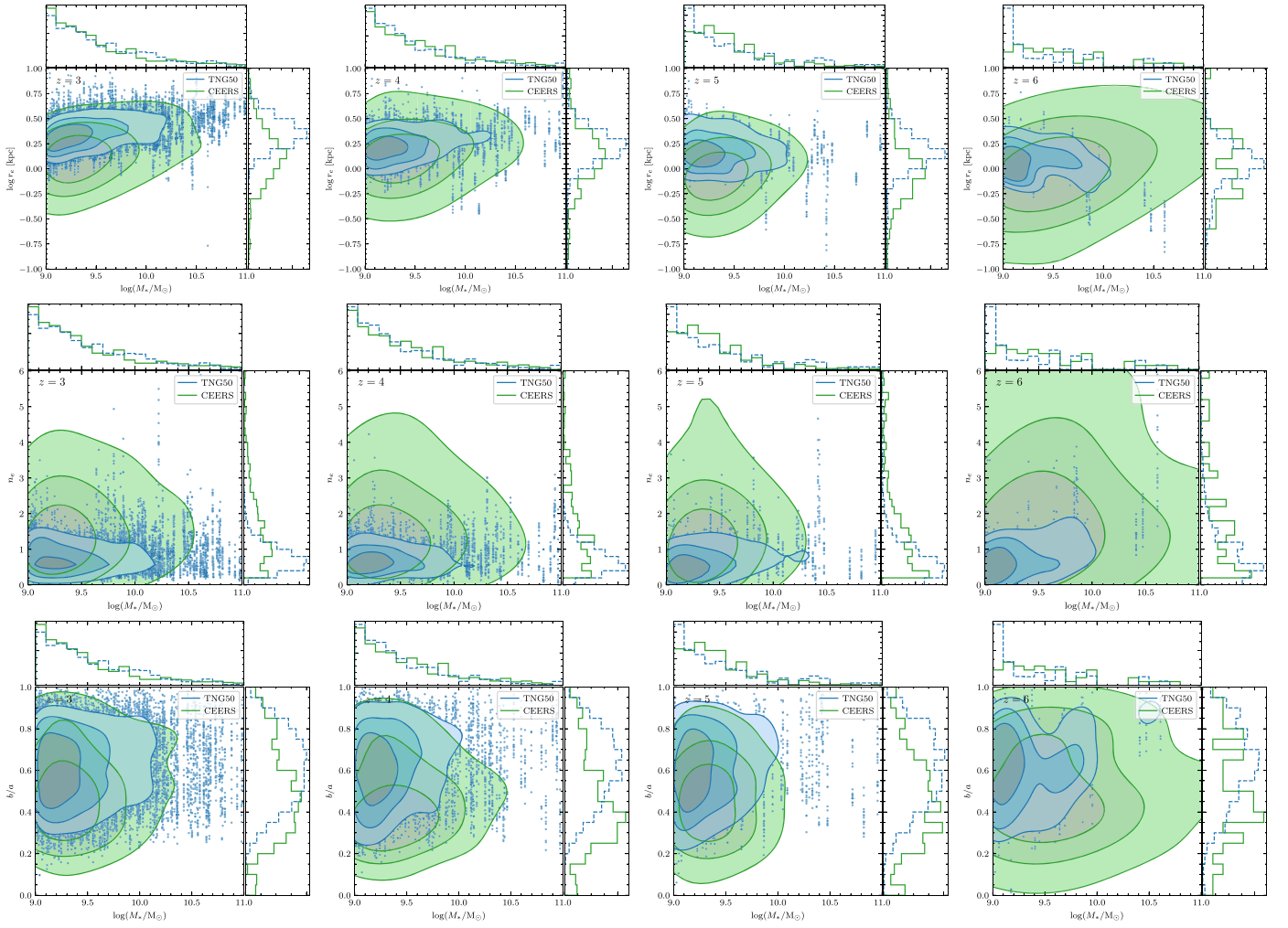


Figure 12. Distribution of the logarithm of the effective radius ($\log r_e$ in kiloparsecs, top row), the Sèrsic index (n_s , middle row), and the axis ratio (b/a , bottom row) as a function of the logarithm of the stellar mass ($\log M_*$) for the TNG50 (in blue) and the CEERS (in green) data sets. From left to right, the panels show the distributions at $z = 3, 4, 5, 6$ for the TNG50 data set. For the CEERS data set, galaxies are included in the closest redshift value. Contour levels enclose 25%, 50%, and 75% of the data. Blue points correspond to galaxies outside the 75% contour for the TNG50 data set. Blue dashed and green solid histograms in the horizontal and vertical axes show the PDF of the TNG50 and CEERS data sets, respectively. The photometric parameters shown are measured in the F200W filter.

Extended (EX). It is clear how EC and Cm galaxies show low-mass, low angular momentum, large fractions of nonrotating stars, and low flatness. They are also smaller in size with larger Sèrsic index values, rounder in projection, and more compact than the rest of the classes. As going from the In class to the El and EX classes, the masses and sizes of the galaxies progressively increase along with the angular momentum of the stars and the flatness. These classes also exhibit smaller Sèrsic indexes and are less concentrated and more asymmetric than the Cm classes. Also interesting is the separation in the projected ellipticity of the El class, showing extremely large values of the $1 - b/a$ compared to the rest of the classes.

For the simulated TNG50 galaxies, we find the following: $\sim 12\%$ of EC galaxies, $\sim 26\%$ of Cm galaxies, $\sim 19\%$ of In galaxies, $\sim 22\%$ of El galaxies, and $\sim 20\%$ of EX galaxies. While for the observed CEERS data set, we find $\sim 55\%$ of EC galaxies, $\sim 17\%$ of Cm galaxies, $\sim 2\%$ of In galaxies, $\sim 25\%$ of El galaxies, and $\sim 2\%$ of EX galaxies. Therefore, and also clear from Figure 14, there is a systematic lack of observed CEERS galaxies in the rest of the classes beyond the EC class, with the exception of the El class, for which the fractions of observed galaxies are slightly larger than for the simulated ones. In

particular, the fractions of observed CEERS galaxies in the In and EX classes are significantly smaller than those for the simulated TNG50 ones.

In Figure 14, we show the UMAP visualization color-coded by the five classes for the simulated TNG50 and the observed CEERS data sets. Note that galaxies with artifacts or bright companions are not included in the derivation of the different class fractions. Given the division and the correlations shown in Figures 8 and 9, we denote the galaxies located in the left section of the UMAP that belong to the cluster in red as EC galaxies, while the remaining galaxies (i.e., those not assigned to the red EC cluster) are considered as noncompact (NC) galaxies. We find that on average $\sim 12\%$ of the TNG50 galaxies belong to the EC class. For the CEERS data set, we find that $\sim 55\%$ of the CEERS galaxies belong to the EC class. In order to mitigate the possible effects of resolution in the simulation, we additionally impose a minimum size threshold for CEERS galaxies and only include CEERS galaxies with $r_e > 1\text{kpc}$, measured in the F200W filter. This excludes the low-radius tail of 5% of the TNG50 galaxies and $\sim 50\%$ of the CEERS data set (mainly Sphs since they are typically smaller in size). We find that the fraction of galaxies in the EC class for

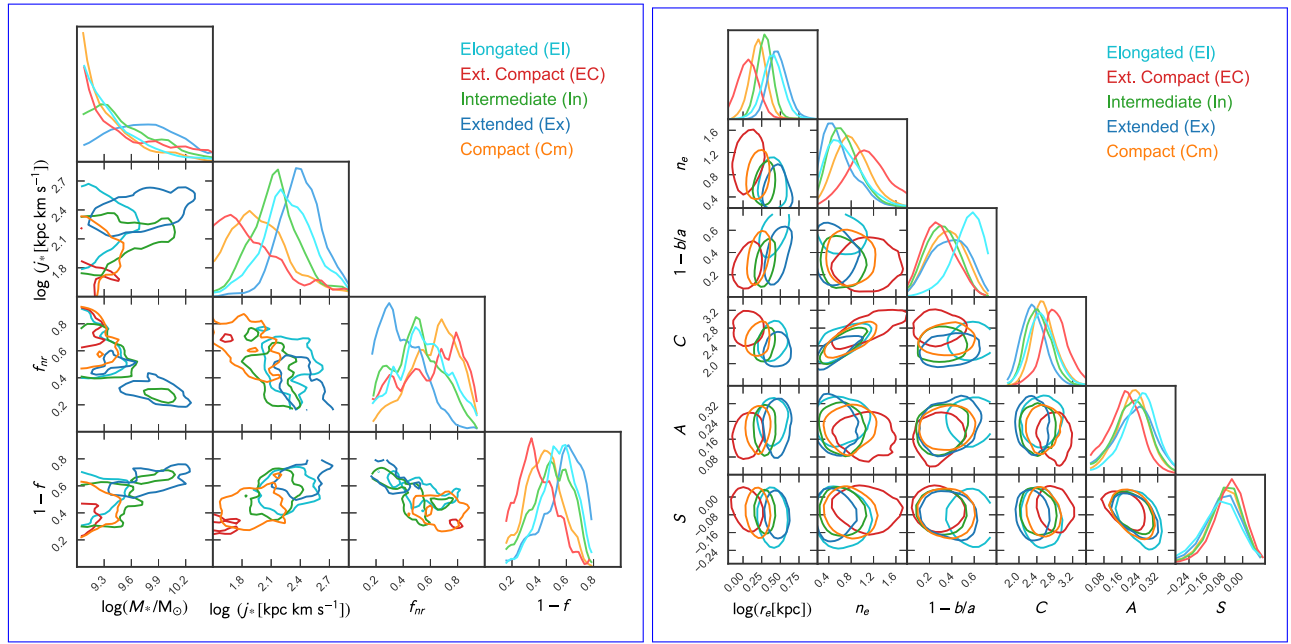


Figure 13. Triangle plot with the photometric (left-hand panel) and morphological (right-hand panel) properties of the morphological clusters shown in the left-hand panel in Figure 14. For clarity, only the 68% contour levels are shown.

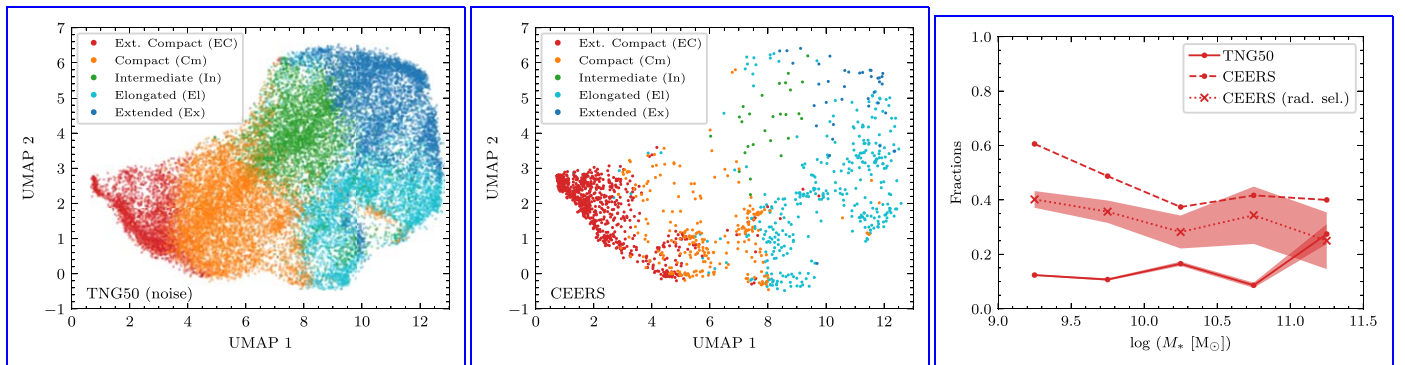


Figure 14. Left-hand panel: UMAP visualization of noise-added TNG50 galaxy images color-coded by classes according to the k -means method for five clusters. Morphological classes are labeled as: extremely compact (EC, in red), compact (Cm, in orange), intermediate (In, in green), elongated (EI, in cyan), and extended (EX, in blue). Middle panel: same as the left-hand panel but for the CEERS galaxy images. Right-hand panel: fractions extremely compact (EC) galaxies (i.e., those belonging to the EC cluster in red) in TNG50 (solid lines), CEERS (dashed lines), and CEERS with $\log r_e [\text{kpc}] > 0$ (dotted lines) in 5 logarithmic mass bins of width 0.5 dex in the range $9 < \leq \log(M/M_\odot) < 11.5$. The shaded regions correspond to the fraction errors considering Poisson errors in the number of selected galaxies and the total number of galaxies in each mass bin. Note that for clarity we do not show the shaded region for the CEERS data set without applying any cut in size.

the CEERS data set is reduced to $\sim 37\%$, but the discrepancy between observations and simulations is still significant.

Our results tend to confirm that the TNG50 model systematically underpredicts the abundance of EX galaxies from a purely data-driven perspective, similar to the findings of, e.g., Flores-Freitas et al. (2022). The effect does not seem a pure consequence of resolution.

6. A Comparison between Self-supervised and Supervised Morphologies

We now compare in more detail the CNN-based and visual classifications for the CEERS and VISUAL data sets with the self-supervised classifications—which have been shown to correlate with physical properties—and speculate about the abundance of disks at $z > 3$. As shown in Section 4.1, visually classified disks are spread all over the representation space, which suggests that, according to the self-supervised

representations, they represent a heterogeneous group of galaxies with different physical properties.

6.1. Self-supervised Clustering versus Morphological Classifications

In Figures 15 and 16, we show a comparison between the CNN-based and visual classifications with the two broad contrastive learning clusters containing EC and NC galaxies, respectively.

The figures show that, for both the CNN and visual classifications, almost all galaxies classified as Sphs or with a bulge component belong to the EC cluster. This is expected as the EC cluster lies in a region of the representation space dominated by round and compact galaxies. However, the trend for disks and irregular galaxies is not so clear. The figures show that both the EC and NC clusters contain a large fraction of disks and irregular galaxies, which reflect the fact that disks and irregulars are distributed over all the representation space.

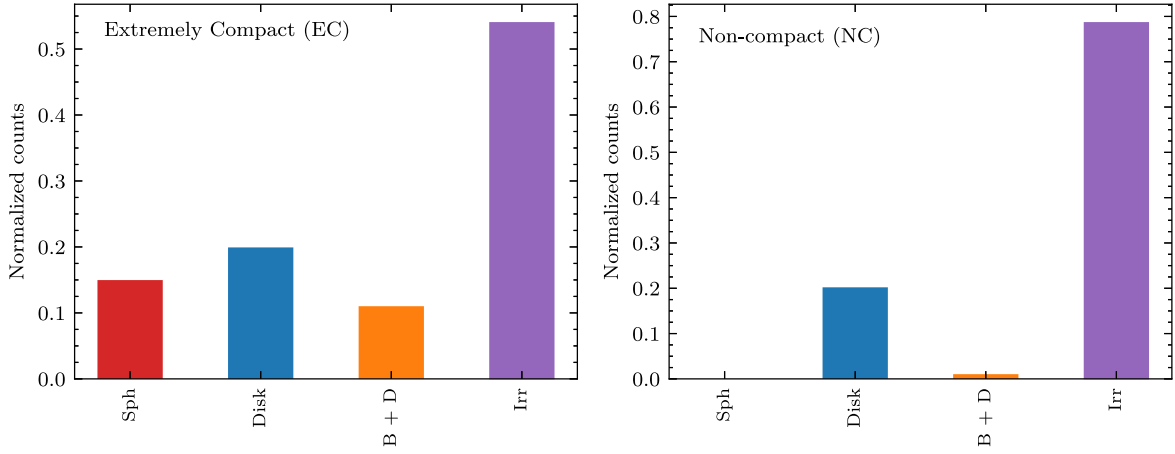


Figure 15. Comparison of the classification into extremely compact (EC, left-hand panel) and noncompact (NC, right-hand panel) galaxies in the CEERS data set and the CNN-based classifications derived by Huertas-Company et al. (2023a). Histograms are color-coded according to the CNN-based morphological classes into: Sph in red, disk in blue, bulge + disk (B+D) in orange, and disturbed (Irr) in purple.

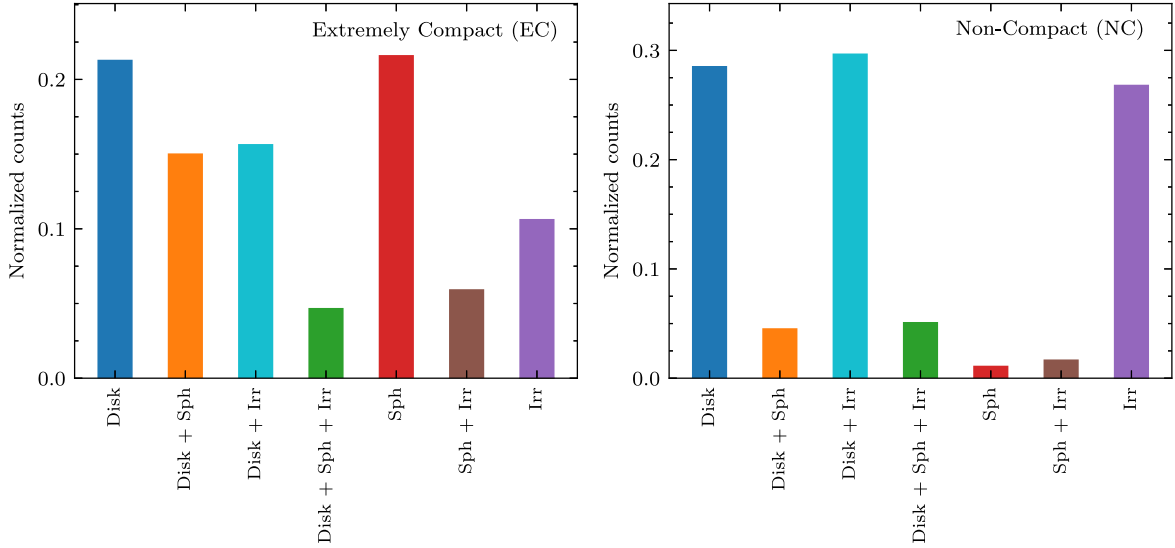


Figure 16. Comparison of the classification into extremely compact (EC, left-hand panel) and noncompact (NC, right-hand panel) galaxies in the VISUAL data set and the visual classifications derived by Kartaltepe et al. (2023). Histograms are color-coded according to the visual morphologies as in Figure 10.

In fact, we measure that each cluster contains roughly $\sim 50\%$ of the disk and Irr galaxies for both the CEERS and the VISUAL data sets. This discrepancy between the self-supervised and the traditional classes is somehow surprising and suggests that the population of visually classified disks and irregulars present a large spread of morphological properties that the contrastive learning representation is capturing. We quantify this further in the following. We emphasize that this discrepancy between supervised and self-supervised classifications is independent of the degree of agreement between simulations and observations discussed in Section 5.

6.2. Two Populations of Visually Classified Disks?

Based on the positions of the visual disks in the contrastive learning representation space, we identify two different populations of visually classified disks, which we call EC disks and NC disks, for EC and NC disks, respectively.

In order to get more insights about why the self-supervised learning algorithm tends to locate them in different clusters, we then examine the properties of the EC disks and NC disks using parametric morphologies obtained via Sèrsic fitting in the

F356W filter (see Section 2 for more details). In Figure 17, we show the distributions of axis-ratios (b/a), semimajor axes (a), and Sèrsic indices (n_e) for the two disk classes, and for the CEERS and VISUAL data sets. We also show, for reference, the same distributions for galaxies visually classified as Sphs.

The figure clearly shows different distributions. As expected, EC disks have smaller effective radii. The NC disks exhibit a distribution of $\log a$ that peaks at $\log a \sim 0.25$, while for the EC disks it peaks at smaller values of $\log a \sim -0.10$. The distribution of n_e is also different and reflects that EC disk are more concentrated. For NC disks, it peaks at $n_e \lesssim 1.2$, characteristic of an exponential profile. However, for EC disks, the distribution is skewed toward larger values of the Sèrsic index, although smaller than for Sphs. Regarding the b/a distribution, the NC disks tend to be more EI, and, as expected, Sphs are rounder on average. The EC disk candidates are in between, with a peak at $b/a \sim 0.5$.

In view of these differences and the correlations between structural properties and the positions in the representation space highlighted in previous sections, it is expected that the NC and EC disks fall in different regions of the parameter

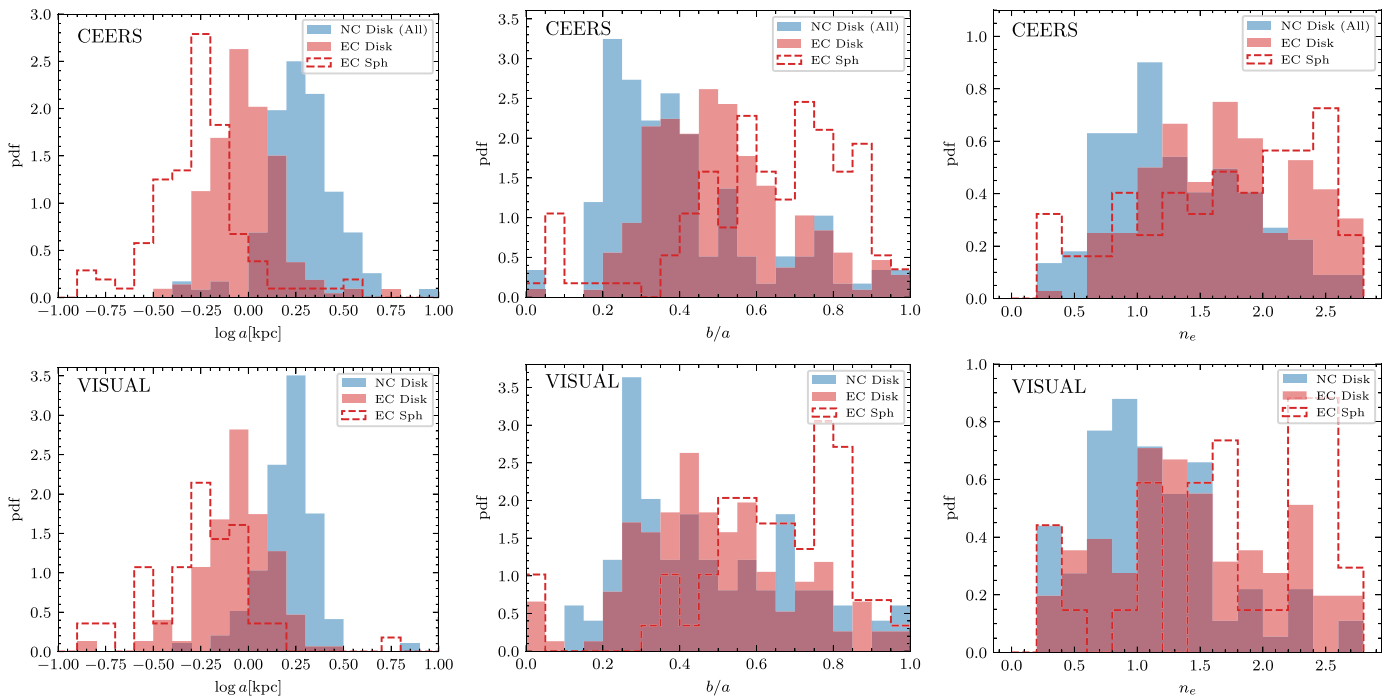


Figure 17. Morphological and photometric properties of the two populations of disk-like galaxies in the CEERS (top panels) and the VISUAL (bottom panels) data sets. From left to right, the different panels show the probability distribution function of the size of the galaxy ($\log a$ [kpc]), the projected ellipticity (b/a), and the Sèrsic index (n_e), respectively. The different histograms correspond to EC disk candidates (red shaded histograms), NC disk candidates (blue shaded histograms), and EC Sph candidates (red dashed histograms). For the CEERS data set, we consider as disk candidates those classified as disk and bulge+disk, while, for the VISUAL data set, we consider as disk candidates the four classes of visually classified disks (disk, disk + Sph, disk + Irr, and disk + Sph + Irr). All quantities are derived from the Sèrsic fits in the F356W filter.

space, even if they share the same visual classification. It also makes sense that these EC disks are generally classified as disks by expert classifiers given the proposed classification scheme, in the sense that they do not show strong signs of irregularities and are more elongated and less concentrated than pure Sphs. Therefore, by default, they end up in the disk class. In Appendix B, we show examples of several NC disks and EC disks candidates (including some cases with $b/a < 0.3$) along with various EC Sph candidates for comparison.

The difference between the completely supervised and self-supervised classifications illustrates how a purely data-driven approach can offer an alternative description of the information content of the data. It is also important to highlight that the difference between the two types of classifications is not only driven by differences in size. The distributions of $\log a$ plotted in Figure 17 overlap significantly. In addition, if we only include in the analysis galaxies with $r_e > 1$ kpc, there is still a significant fraction ($\sim 40\%$) of visually classified disks in the EC cluster.

An interesting question, therefore, is whether NC and EC disks are all *true* disks—if by disk, we understand a flat system predominantly supported by the rotation of the gas and stars—as this has important implications on the frequency of disk formation at these very early cosmic epochs. It is certainly impossible to unambiguously answer this question with the data at hand. However, we can speculate based on the information inferred from the TNG50 simulation. As shown in the previous subsections, there is a strong correlation between the contrastive learning representations and the stellar specific angular momentum. Interestingly, the location of EC disks is predominately populated by low angular momentum

galaxies, which would suggest that these systems are preferentially velocity dispersion supported.

As an additional attempt to shed some light on the physical properties of EC and NC disks, we explore in more detail how the contrastive learning representation space distributes galaxies with different 3D stellar structures. We characterize the shape of galaxies in the TNG50 sample as done and studied in Pillepich et al. (2019), i.e., with an ellipsoid with three semiaxes $c < b < a$ and use the axial ratios $q = b/a$ (intermediate-to-major), and $s = c/a$ (minor-to-major) to define three main 3D shape classes: oblate, Sph, and prolate following the definitions of van der Wel et al. (2014a), Zhang et al. (2019). The axial ratios are derived after diagonalizing the stellar mass tensor in an iterative way while keeping the major axis length fixed to $2r_*$. We consider oblate or disk galaxies those with $a \sim b > c$, El, or prolate objects those with $a > b \sim c$, and spheroidal systems those with similar values for the three semiaxes. Note that, by definition, the 20 projections of the TNG50 galaxies have the same 3D shape. At these redshifts, we find that $\sim 67\%$ of the galaxies in the simulation have a spheroidal shape according to this definition. Only 18% and 15% of the galaxies have oblate and prolate shapes, respectively. The fact that, at high redshift and low stellar masses, galaxies tend to present a more prolate structure has been found both in observations (van der Wel et al. 2014a; Zhang et al. 2019) and simulations (Tomassetti et al. 2016; Pillepich et al. 2019).

We show in Figure 18 the distribution of the mock images of oblate galaxies in the UMAP projection of the representation space obtained with contrastive learning. For clarity, we do not show the distributions of prolate and spheroidal galaxies. Interestingly, the bottom right corner, where the EC disks lie,

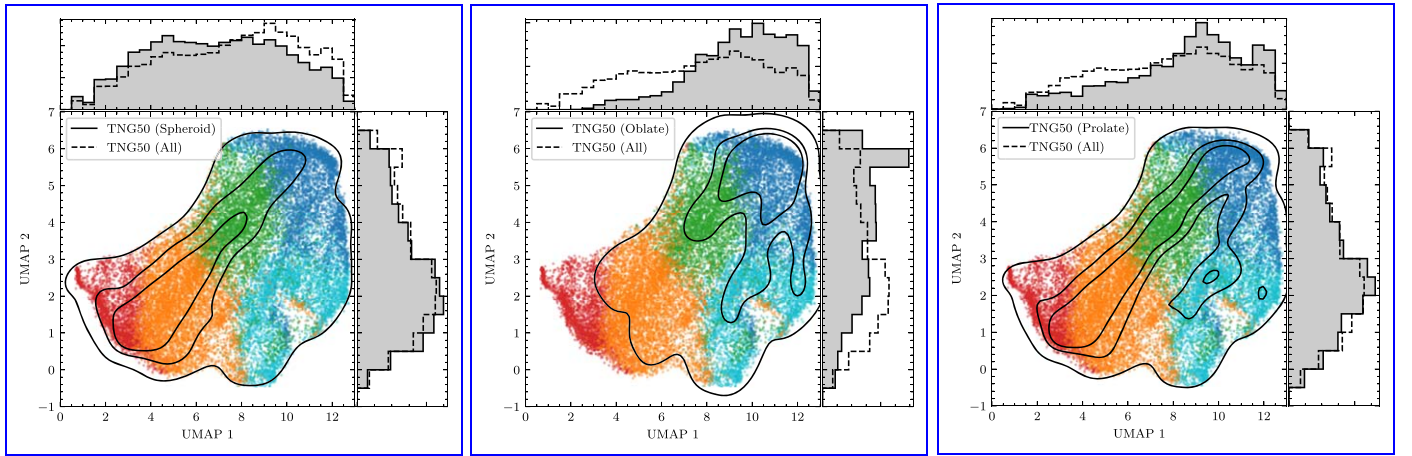


Figure 18. Location in the UMAP plane of TNG50 galaxies according to the 3D shape inferred from the stellar particles: spheroid (left-hand panel), oblate (middle panel), and prolate (right-hand panel). Points are color-coded according to the morphological classes described in Section 5.2. The contour levels indicate the 25%, 50%, and 95% probabilities. Unfilled black dashed and filled solid histograms in the horizontal and vertical axes show the PDF of the whole data set and the shape-selected galaxies of the TNG50 data set, respectively. Morphologically disk (i.e., oblate) galaxies tend to populate the right and upper right regions of the UMAP plane, while it is unlikely to find them in the left section of the UMAP plane.

does not contain almost any oblate system. The result would suggest that EC disks are very unlikely to be true disks (i.e., flat rotating systems) even if they appear as EI in the images. Despite the small number of TNG50 galaxies with $b/a \lesssim 0.4$, we find that EC TNG50 galaxies with $b/a \lesssim 0.4$ are more likely to be prolate than Sph in shape, as shown in Figure 19.

However, a big caveat is that the previous statements rely of course on the calibration with the TNG50 simulation, which, as shown in the previous sections, does not properly reproduce the observed morphological diversity. Therefore, no firm conclusion can be established without more observations and comparisons with other state-of-the-art simulations. As a matter of fact, an alternative explanation for the different representations of EC disks and NC disks could be that the simulation cannot generate compact disks because of resolution issues, as already acknowledged. This would imply that the location of the EC disks in the representation space cannot be interpreted in terms of physical properties, as these systems simply do not exist in the simulation. We expect the fine-tuning of the model presented in Section 3.1 should mitigate this effect, but cannot be guaranteed.

Nevertheless, our data-driven analysis suggests that robustly establishing the fraction of disk galaxies at $z > 3$ remains an open issue.

7. Summary and Conclusions

This work presents a novel data-driven method based on contrastive learning to infer the morphological properties of galaxies observed with JWST. The method is calibrated on mock JWST galaxy images extracted from the TNG50 cosmological simulation that, thanks to its large number of qualitatively realistic galaxies, allows us to produce a morphological description—without any assumption on galaxy classes—robust to background noise, S/N, color, and orientation. In addition, we show that the obtained representations of galaxies based on their images correlate well with some other physical properties inferred from the simulation (such as the specific angular momentum of stars, j_* , and the intrinsic 3D shape) along with some measured photometric and structural properties (such as Sèrsic index and the projected ellipticity).

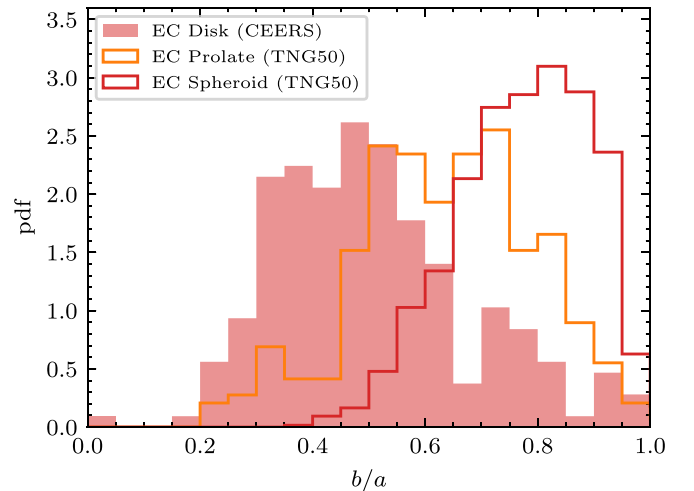


Figure 19. Distributions of projected ellipticity (b/a) for EC disk candidates in CEERS (red shaded histogram), and EC prolate and spheroid TNG50 candidates (orange and red empty histograms, respectively). Elongated galaxies in projection (i.e., $b/a \lesssim 0.4$) are compatible with being prolate in shape according to the TNG50 EC prolate candidates.

We have applied the method to JWST images from the CEERS survey in the F200W and F356W bands of the following: (1) a mass-complete sample ($M_* \geq 10^9 M_\odot$) of galaxies at $3 < z < 6$ in the CEERS survey for which CNN-based classifications are available; and (2) a mass- and a redshift-selected sample of CEERS galaxies at $3 < z < 6$ with $M_* \geq 10^9 M_\odot$ for which visual morphological classifications are available.

Our main results are as follows:

1. Simulated galaxies from the TNG50 cosmological simulation seem to cover well the observed morphological diversity at $z > 3$. However, the morphological distributions of CEERS and simulated galaxies are measured to be different. When compared at the pixel level, simulated and observed galaxies seem to populate in different proportions in the different regions of the TNG50-trained manifolds. We show that these differences can be at least partly explained because observed

galaxies can be more compact and more elongated than simulated ones. In fact, the galaxy-to-galaxy variation in sizes, Sèrsic indices, and shapes at fixed stellar mass and redshift are larger in the observed CEERS population than in TNG50 simulated ones. These differences might be partly explained by the limited resolution of the TNG simulation, but not completely since the discrepancies are not erased when only large galaxies are considered.

- Our morphological description also suggests that CNN-based and visually classified disks comprise two different populations: one made of EC disks and another of NC disks. Half of the galaxies that are classified as disks are indeed located in the representation space very close to EC Sphs and, therefore, are more consistent with not being pure disk-like galaxies (i.e., having a prolate or spheroidal stellar structure). Although some of these conclusions might be affected by the calibration with the TNG50 model, our study suggests that some EC disk candidates can be misclassified as disks, as they appear (on average) more elongated in the images than EC Sph candidates. The coexistence of prolate and oblate systems at high redshift is in qualitative agreement with the predictions of other models (e.g., zoom-in simulations), which also found that low-mass galaxies at high- z tend to present a prolate shape (Ceverino et al. 2015; Tomassetti et al. 2016). More in-depth follow-up of these two populations of galaxies, possibly with spectroscopy and additional comparisons with other simulations, is required to further constrain their true nature.

Acknowledgments

The authors acknowledge the referee for the comprehensive and constructive comments that played a crucial role in enhancing the accuracy, clarity, and exhaustiveness of the paper. M.H.-C. thanks Shy Genel, David Koo, and Sandy Faber for insightful discussions. J.V.-F., M.H.-C., R.S., and J.H.K. acknowledge financial support from the State Research Agency (AEIMCINN) of the Spanish Ministry of Science and Innovation under the grant “Galaxy Evolution with Artificial Intelligence” with reference PGC2018-100852-A-I00 and under the grant “The structure and evolution of galaxies and their central regions” with reference

PID2019-105602GB-I00/10.13039/501100011033, from the ACIISI, Consejería de Economía, Conocimiento y Empleo del Gobierno de Canarias and the European Regional Development Fund (ERDF) under grants with reference PROID2020010057 and PROID2021010044, and from Instituto de Astrofísica de Canarias (IAC) projects P/300724 and P/301802, financed by the Ministry of Science and Innovation, through the State Budget and by the Canary Islands Department of Economy, Knowledge and Employment, through the Regional Budget of the Autonomous Community. J.V.-F. and F.B. also acknowledge support from the grant “Galactic Edges and Euclid in the Low Surface Brightness Era (GEELSBE)” with reference PID2020-116188GA-I00 financed by the Spanish Ministry of Science and Innovation. L. C. acknowledges financial support from Comunidad de Madrid under Atracción de Talento grant 2018-T2/TIC-11612 and Spanish Ministerio de Ciencia e Innovación MCIN/AEI/10.13039/501100011033 through grant PGC2018-093499-B-I00. This work was supported by the Amazon Web Services (AWS) Cloud Credits for Research Program.

Appendix A

Scatter of Physical and Photometric Parameters in the UMAP Visualization

In Figures A1 and A2, we show the variability in the physical and photometric parameters, respectively, in the UMAP visualization shown in Figure 9 in Section 3.5. The scatter is quantified as a normalized median absolute deviation, denoted here as NMAD. The median absolute deviation (MAD), defined as $MAD(y) = median(|y - median(y)|)$, is a robust measure of the variability of a univariate sample of quantitative data. The MAD is less affected by outliers and nongaussianity than the typical variance and standard deviation. To facilitate the comparison between different variables, we normalized the MAD by the dynamical range of the data, defined as the percentile range containing 98% of the data. The resulting normalized MAD, denoted as NMAD, is an indicator of the variability of the data that, in this case, shows how informative the correlation with the different parameters shown in Figures 8 and 9 are. The values of NMAD $\lesssim 0.2$ are indicative of a low variability of the data.

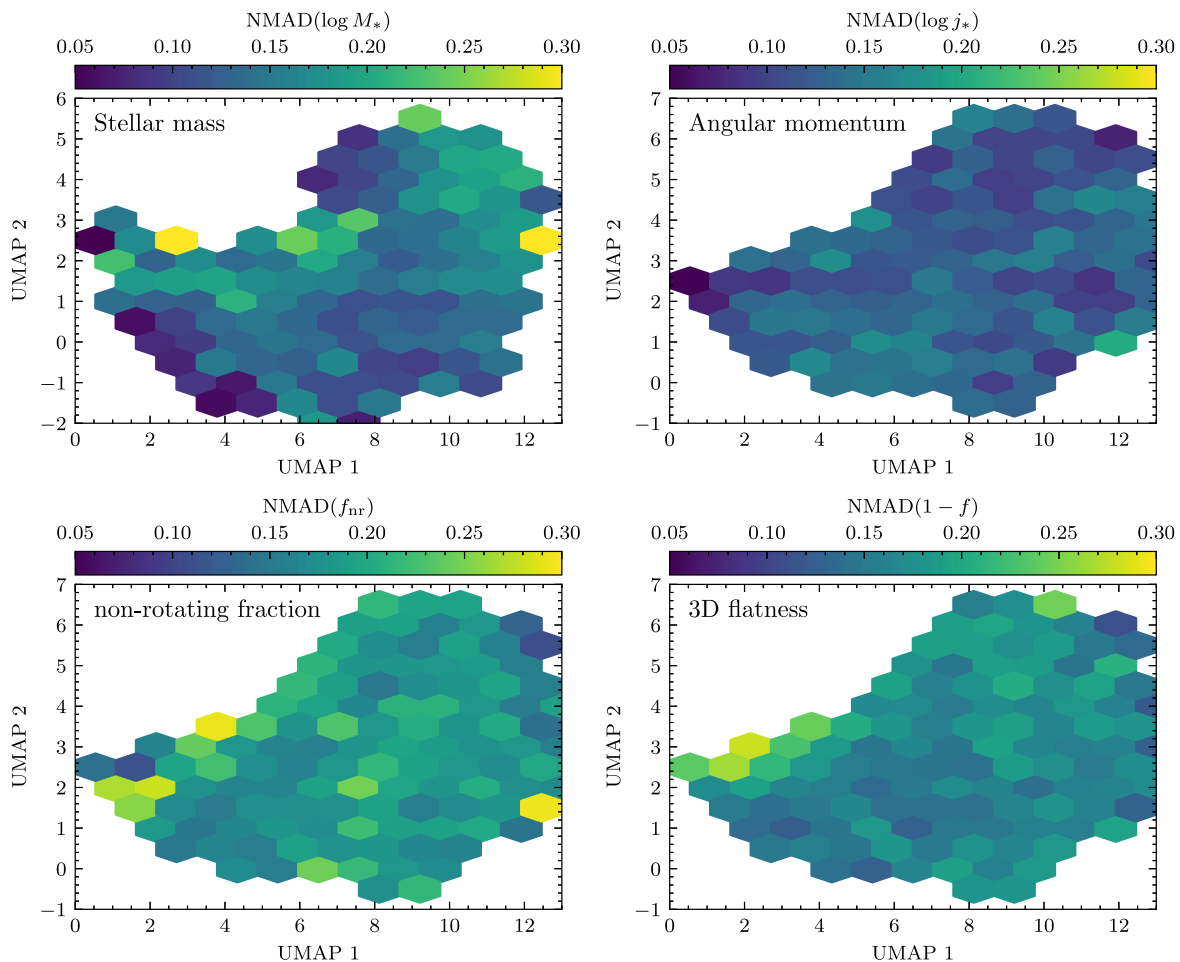


Figure A1. UMAP visualization for all the TNG50 galaxy images in our data set color-coded by the distribution of several physical properties extracted from the TNG50 simulation. From left to right and top to bottom, the different panels show the UMAP plane color-coded by the NMAD of: the logarithm of the total stellar mass ($\log M_*$ [M_\odot]), the logarithm of the specific angular momentum of the stars ($\log j_*$ [kpc km s^{-1}]), the mass fraction in nonrotating stars (f_{nr}), and the galaxy flatness ($1 - f$).

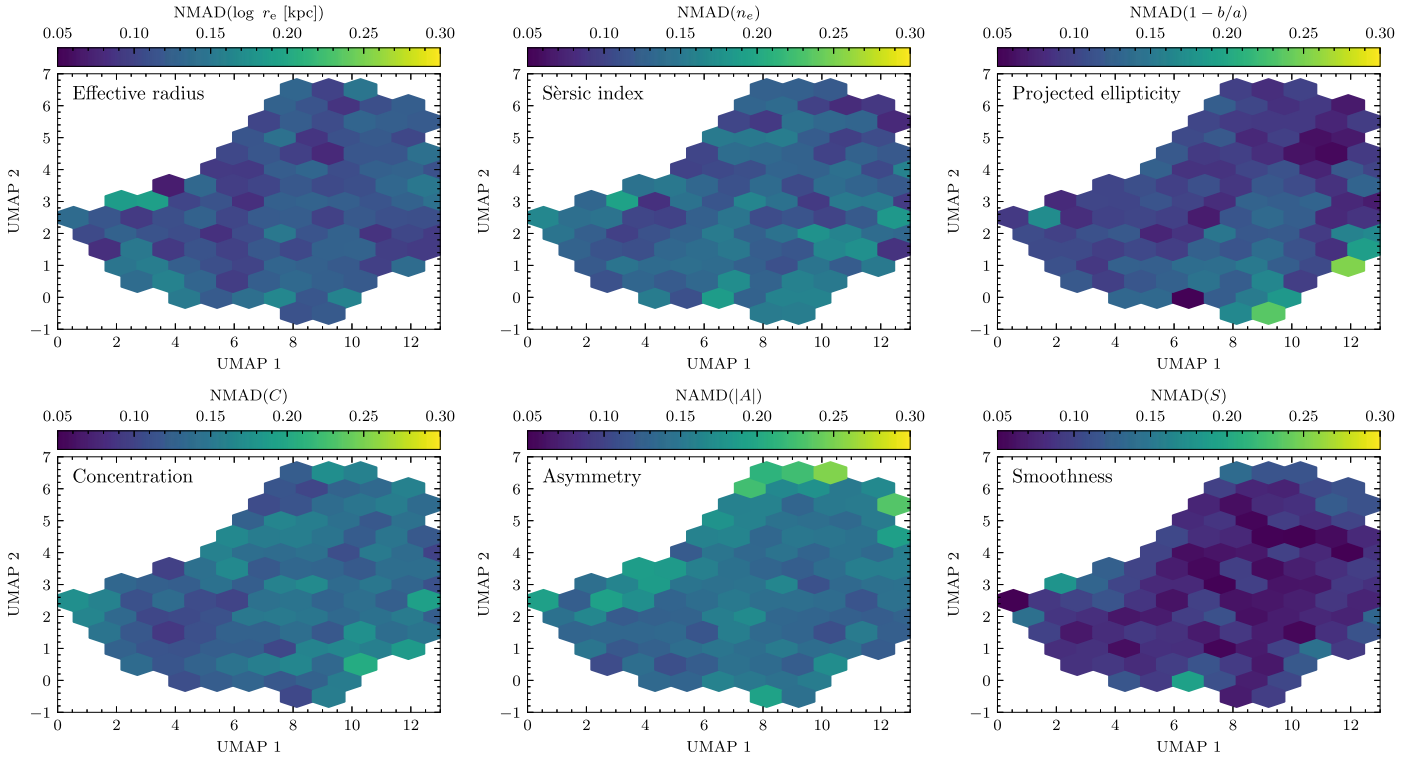


Figure A2. UMAP visualization for all the TNG50 galaxy images in our data set color-coded by the distribution of several morphological and photometric parameters. From left to right and top to bottom, the different panels show the UMAP plane color-coded by the NMAD of: the logarithm of the effective radius (r_e [kpc]), the Sérsic index (n_e), the ellipticity based on Sérsic fit ($1 - b/a$), the concentration (C), the asymmetry (A), and the smoothness (S).

Appendix B Examples of Observed JWST Galaxy Images

In Figure B1, we show examples of galaxies in the CEERS data set that are classified into EC and NC, according to the criterion described in Section 6.1. For comparison, we show examples of disk and Sph galaxies following the CNN-based morphological classification presented in Huertas-Company et al. (2023a). It is clear the differences between the NC disks and the EC disks, with more EX and EI (in projection) light distributions for the NC disks candidates than for the EC disks candidates. We also include examples of NC disks and EC

disks candidates that contribute to the low end of the b/a distribution in Figure 17. In fact, all these examples have values of $b/a < 0.3$. For the NC disks candidates with $b/a < 0.3$, they show in almost all the examples signs of multiple clumps and/or in interaction with the central galaxy. For several cases of the EC disks candidates with $b/a < 0.3$, there are also signs of multiple clumps that could lead to an underestimation of the true b/a values.

Moreover, it is difficult to distinguish between some EC disks and EC Sph candidates, as they exhibit Cm and round light distributions, although the EC disk galaxies appear slightly more EI (on average) than the EC Sph ones.

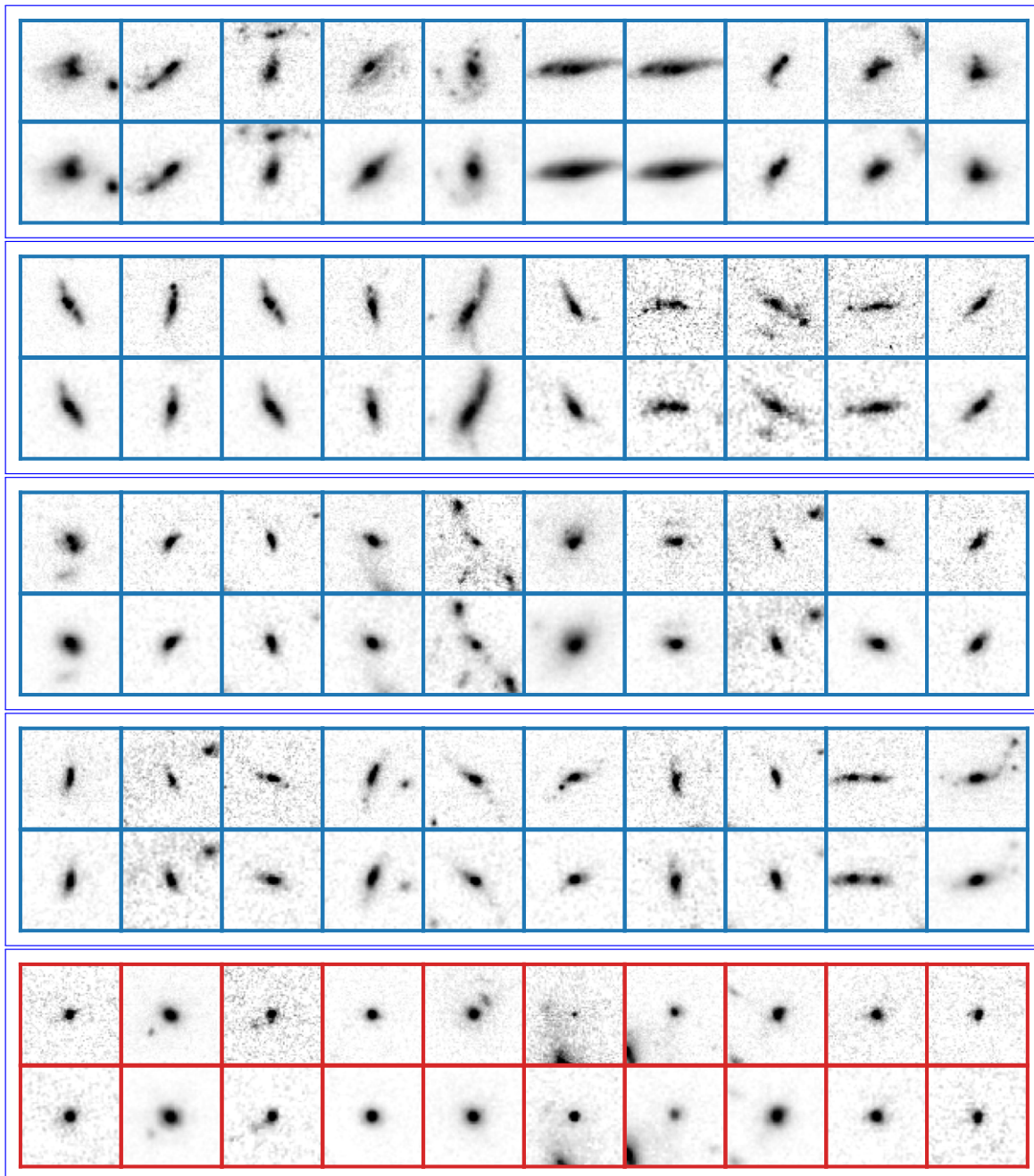































Figure B1. Examples of galaxy images in the CEERS data set considered as: NC disks (first row), NC disks with $b/a < 0.3$ (second row), EC disks (third row), EC disks with $b/a < 0.3$ (fourth row), and EC Sph (fifth row) candidates. Images are framed in color according to the CNN-based classification: blue for disk and red for Sph candidates. Each galaxy image is shown in the two F200W (top row) and the F356W (bottom row) filters.

ORCID iDs

Jesús Vega-Ferrero  <https://orcid.org/0000-0003-2338-5567>
 Marc Huertas-Company  <https://orcid.org/0000-0002-1416-8483>
 Luca Costantin  <https://orcid.org/0000-0001-6820-0015>
 Pablo G. Pérez-González  <https://orcid.org/0000-0003-4528-5639>
 Regina Sarmiento  <https://orcid.org/0000-0002-3803-6952>
 Jeyhan S. Kartaltepe  <https://orcid.org/0000-0001-9187-3605>
 Annalisa Pillepich  <https://orcid.org/0000-0003-1065-9274>
 Micaela B. Bagley  <https://orcid.org/0000-0002-9921-9218>
 Steven L. Finkelstein  <https://orcid.org/0000-0001-8519-1130>

Elizabeth J. McGrath  <https://orcid.org/0000-0001-8688-2443>
 Johan H. Knapen  <https://orcid.org/0000-0003-1643-0024>
 Pablo Arrabal Haro  <https://orcid.org/0000-0002-7959-8783>
 Eric F. Bell  <https://orcid.org/0000-0002-5564-9873>
 Fernando Buitrago  <https://orcid.org/0000-0002-2861-9812>
 Antonello Calabrò  <https://orcid.org/0000-0003-2536-1614>
 Avishai Dekel  <https://orcid.org/0000-0003-4174-0374>
 Mark Dickinson  <https://orcid.org/0000-0001-5414-5131>
 Helena Domínguez Sánchez  <https://orcid.org/0000-0002-9013-1316>
 David Elbaz  <https://orcid.org/0000-0002-7631-647X>
 Henry C. Ferguson  <https://orcid.org/0000-0001-7113-2738>

Mauro Giavalisco  <https://orcid.org/0000-0002-7831-8751>
 Benne W. Holwerda  <https://orcid.org/0000-0002-4884-6756>
 Dale D. Kocevski  <https://orcid.org/0000-0002-8360-3880>
 Anton M. Koekemoer  <https://orcid.org/0000-0002-6610-2048>
 Viraj Pandya  <https://orcid.org/0000-0002-2499-9205>
 Casey Papovich  <https://orcid.org/0000-0001-7503-8482>
 Nor Pirzkal  <https://orcid.org/0000-0003-3382-5941>
 Joel Primack  <https://orcid.org/0000-0001-5091-5098>
 L. Y. Aaron Yung  <https://orcid.org/0000-0003-3466-035X>

References

- Abraham, R. G., van den Bergh, S., Glazebrook, K., et al. 1996, *ApJS*, 107, 1
 Bagley, M. B., Finkelstein, S. L., Koekemoer, A. M., et al. 2023, *ApJL*, 946, L12
 Barro, G., Faber, S. M., Koo, D. C., et al. 2017, *ApJ*, 840, 47
 Bengfort, B., Danielsen, N., Bilbro, R., et al. 2018, Yellowbrick v0.6, v0.6, Zenodo, doi:10.5281/zenodo.1206264
 Bournaud, F., Perret, V., Renaud, F., et al. 2014, *ApJ*, 780, 57
 Buitrago, F., Conselice, C. J., Epinat, B., et al. 2014, *MNRAS*, 439, 1494
 Buitrago, F., Trujillo, I., Conselice, C. J., & Häußler, B. 2013, *MNRAS*, 428, 1460
 Ceverino, D., Dekel, A., & Bournaud, F. 2010, *MNRAS*, 404, 2151
 Ceverino, D., Primack, J., & Dekel, A. 2015, *MNRAS*, 453, 408
 Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020a, arXiv:2002.05709
 Chen, Z., Faber, S. M., Koo, D. C., et al. 2020b, *ApJ*, 897, 102
 Conselice, C. J. 2003, *ApJS*, 147, 1
 Costantin, L., Méndez-Abreu, J., Corsini, E. M., et al. 2020, *ApJL*, 889, L3
 Costantin, L., Pérez-González, P. G., Méndez-Abreu, J., et al. 2021, *ApJ*, 913, 125
 Costantin, L., Pérez-González, P. G., Méndez-Abreu, J., et al. 2022, *ApJ*, 929, 121
 Costantin, L., Pérez-González, P. G., Vega-Ferrero, J., et al. 2023, *ApJ*, 946, 71
 Dimauro, P., Daddi, E., Shankar, F., et al. 2022, *MNRAS*, 513, 256
 Ferreira, L., Adams, N., Conselice, C. J., et al. 2022, *ApJL*, 938, L2
 Ferreira, L., Conselice, C., Sazonova, E., et al. 2023, *ApJ*, 955, 94
 Finkelstein, S. L., Bagley, M. B., Ferguson, H. C., et al. 2023, *ApJL*, 946, L13
 Finkelstein, S. L., Bagley, M. B., Haro, P. A., et al. 2022, *ApJL*, 940, L55
 Finkelstein, S. L., Dickinson, M., Ferguson, H. C., et al. 2017, The Cosmic Evolution Early Release Science (CEERS) Survey, JWST Proposal ID 1345. Cycle 0 Early Release Science
 Flores-Freitas, R., Chies-Santos, A. L., Furlanetto, C., et al. 2022, *MNRAS*, 512, 245
 Freundlich, J., Combes, F., Tacconi, L. J., et al. 2019, *A&A*, 622, A105
 Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, *SSRv*, 123, 485
 Genzel, R., Tacconi, L. J., Gracia-Carpio, J., et al. 2010, *MNRAS*, 407, 2091
 Ginzburg, O., Huertas-Company, M., Dekel, A., et al. 2021, *MNRAS*, 501, 730
 Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, *ApJS*, 197, 35
 Guo, Y., Ferguson, H. C., Bell, E. F., et al. 2015, *ApJ*, 800, 39
 Guo, Y., Rafelski, M., Bell, E. F., et al. 2018, *ApJ*, 853, 108
 Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, *ApJL*, 911, L33
 Hinton, G., Vinyals, O., & Dean, J. 2015, arXiv:1503.02531
 Huertas-Company, M., Guo, Y., Ginzburg, O., et al. 2020, *MNRAS*, 499, 814
 Huertas-Company, M., Iyer, K. G., Angeloudi, E., et al. 2023a, arXiv:2305.02478
 Huertas-Company, M., Pérez-González, P. G., Mei, S., et al. 2015, *ApJ*, 809, 95
 Huertas-Company, M., Rodríguez-Gomez, V., Nelson, D., et al. 2019, *MNRAS*, 489, 1859
 Huertas-Company, M., Sarmiento, R., & Knapen, J. H. 2023b, *RASTI*, 2, 441
 Kartaltepe, J. S., Rose, C., Vanderhoof, B. N., et al. 2023, *ApJL*, 946, L15
 Kassin, S. A., Weiner, B. J., Faber, S. M., et al. 2012, *ApJ*, 758, 106
 Kodra, D., Andrews, B. H., Newman, J. A., et al. 2023, *ApJ*, 942, 36
 Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, *ApJS*, 197, 36
 Marinacci, F., Pakmor, R., & Springel, V. 2014, *MNRAS*, 437, 1750
 McInnes, L., Healy, J., & Melville, J. 2018, arXiv:1802.03426
 Nelson, D., Pillepich, A., Springel, V., et al. 2019a, *MNRAS*, 490, 3234
 Nelson, D., Springel, V., Pillepich, A., et al. 2019b, *ComAC*, 6, 2
 Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2010, *AJ*, 139, 2097
 Pillepich, A., Nelson, D., Springel, V., et al. 2019, *MNRAS*, 490, 3196
 Pillepich, A., Sotillo-Ramos, D., Ramesh, R., et al. 2023, arXiv:2303.16217
 Robertson, B. E., Tacchella, S., Johnson, B. D., et al. 2023, *ApJL*, 942, L42
 Rodríguez-Gomez, V., Snyder, G. F., Lotz, J. M., et al. 2019, *MNRAS*, 483, 4140
 Sarmiento, R., Huertas-Company, M., Knapen, J. H., et al. 2021, *ApJ*, 921, 177
 Scoville, N., Aussel, H., Brusa, M., et al. 2007, *ApJS*, 172, 1
 Simons, R. C., Kassin, S. A., Weiner, B. J., et al. 2017, *ApJ*, 843, 46
 Tomassetti, M., Dekel, A., Mandelker, N., et al. 2016, *MNRAS*, 458, 4477
 van der Wel, A., Chang, Y.-Y., Bell, E. F., et al. 2014a, *ApJL*, 792, L6
 van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014b, *ApJ*, 788, 28
 Vega-Ferrero, J., Domínguez Sánchez, H., Bernardi, M., et al. 2021, *MNRAS*, 506, 1927
 Wisnioski, E., Förster Schreiber, N. M., Wuyts, S., et al. 2015, *ApJ*, 799, 209
 Wu, Z., Xiong, Y., Yu, S., & Lin, D. 2018, in 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 3733
 Zanisi, L., Huertas-Company, M., Lanusse, F., et al. 2021, *MNRAS*, 501, 4359
 Zhang, H., Primack, J. R., Faber, S. M., et al. 2019, *MNRAS*, 484, 5170