Dissertation
submitted to the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of Heidelberg University, Germany
for the degree of
Doctor of Natural Sciences

Put forward by
Lukas, Eisert
born in: Künzelsau
Oral examination: 18.07.2024

INFERRING THE ASSEMBLY AND MERGER HISTORIES OF GALAXIES WITH THE ILLUSTRISTNG SIMULATIONS AND MACHINE LEARNING

LUKAS EISERT

May 2024

Referees:
Dr. Annalisa Pillepich
Apl. Prof. Andreas Koch-Hansen

ABSTRACT

This thesis presents an investigation into galaxy formation and evolution, utilizing cutting-edge cosmological simulations and machine learning methodologies. Galaxy data from the full cosmological simulations TNG50 and TNG100 within the IllustrisTNG project are employed, and machine learning techniques are trained to extract the assembly and merging history of these simulated galaxies. These machine-learning models can then be applied to observational data, offering a novel method to connect simulations and observations. Initially, this is achieved by using scalars representing integrated observable galaxy features, from which we are able to infer scalars describing the assembly and merging history accurately. However, scalars encompass only a fraction of the complete observational data (images, spectra, IFUs, etc.), and accurately reconstructing scalars from simulations with the same observable bias as observations is not trivial. To address this, contrastive learning is employed to perform representation learning on both survey-realistic mocks of TNG and observed Hyper Suprime-Cam (HSC) image data (in r, g, and i bands), ensuring comparability between simulated and observed images. Remarkably, our findings demonstrate sufficient similarity between the simulated and observed images to justify the idea of simulation-based inference with images. Subsequently, an inference model trained on TNG data is successfully applied to HSC data, allowing the retrieval of information regarding the ex-situ fraction and the time and mass of the last major merger undergone by a galaxy. This interdisciplinary approach merges the domains of cosmological simulations, observational astronomy, and machine learning, offering a new perspective on galaxy formation and evolution. The developed methodologies not only enhance our ability to interpret observational data but also enable the assessment of the realism of cosmological simulations.

ZUSAMMENFASSUNG

Diese Arbeit präsentiert eine Untersuchung zur Bildung und Evolution von Galaxien unter Verwendung hochmoderner kosmologischer Simulationen und Methoden des maschinellen Lernens. Dabei werden Galaxiendaten aus den vollständigen kosmologischen Simulationen TNG50 und TNG100 des IllustrisTNG-Projekts genutzt, um maschinelles Lernen darauf zu trainieren, die Entstehungs- und Verschmelzungsgeschichte dieser simulierten Galaxien zu extrahieren. Diese maschinellen Lernmodelle können dann auf Beobachtungsdaten angewendet werden und bieten damit eine neue Methode, um Simulationen und Beobachtungen zu verbinden. Zunächst wird dies durch die Verwendung von skalaren Werten erreicht, die beobachtbare Galaxienmerkmale repräsentieren, aus denen wir dann in der Lage sind, Skalare abzuleiten, welche die Entstehungs- und Verschmelzungsgeschichte genau beschreiben. Skalare umfassen jedoch nur einen Bruchteil der vollständigen Beobachtungsdaten (Bilder, Spektren, IFUs usw.), und die genaue Rekonstruktion von Skalaren aus Simulationen mit der gleichen beobachtbaren Verzerrung wie Beobachtungen ist nicht trivial. Um dies zu adressieren, wird kontrastives Lernen verwendet, um Repräsentationslernen sowohl auf realistischen Umfragemockups von TNG als auch auf beobachteten Hyper Suprime-Cam (HSC)-Bilddaten (in r, g und i Bändern) durchzuführen, um die Vergleichbarkeit zwischen simulierten und beobachteten Bildern sicherzustellen. Bemerkenswerterweise zeigen unsere Ergebnisse ausreichende Ähnlichkeit zwischen den simulierten und beobachteten Bildern, um die Idee der simulationsbasierten Inferenz mit Bildern zu rechtfertigen. Anschließend wird ein Inferenzmodell, das auf TNG-Daten trainiert ist, erfolgreich auf HSC-Daten angewendet, was die Rückgewinnung von Informationen über den Ex-situ-Fraktion und die Zeit und Masse der letzten großen Verschmelzung ermöglicht, die eine Galaxie durchlaufen hat. Dieser interdisziplinäre Ansatz vereint die Bereiche kosmologischer Simulationen, beobachtende Astronomie und maschinelles Lernen und bietet einen neuen Blickwinkel auf die Bildung und Evolution von Galaxien. Die entwickelten Methoden verbessern nicht nur unsere Fähigkeit, Beobachtungsdaten zu interpretieren, sondern ermöglichen auch die Bewertung des Realismus kosmologischer Simulationen.

PUBLICATIONS

During the duration of the PhD and up to the time of submission, two papers have been published in peer-reviewed journals and are part of this thesis:

[1] Lukas Eisert, Connor Bottrell, Annalisa Pillepich, Rhythm Shimakawa, Vicente Rodriguez-Gomez, Dylan Nelson, Eirini Angeloudi, and Marc Huertas-Company. "ERGO-ML: comparing IllustrisTNG and HSC galaxy images via contrastive learning." In: *MNRAS* 528.4 (Mar. 2024), pp. 7411–7439. DOI: 10.1093/mnras/stae481. arXiv: 2310.19904 [astro-ph.GA].

[2] Lukas Eisert, Annalisa Pillepich, Dylan Nelson, Ralf S. Klessen, Marc Huertas-Company, and Vicente Rodriguez-Gomez. "ERGO-ML I: inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks." In: *MNRAS* 519.2 (Feb. 2023), pp. 2199–2223. DOI: 10.1093/mnras/stac3295. arXiv: 2202.06967 [astro-ph.GA].

In addition, the following co-authored papers are relevant to the topic of this thesis:

[1] Eirini Angeloudi, Jesús Falcón-Barroso, Marc Huertas-Company, Regina Sarmiento, Annalisa Pillepich, Daniel Walo-Martín, and Lukas Eisert. "ERGO-ML: towards a robust machine learning model for inferring the fraction of accreted stars in galaxies from integral-field spectroscopic maps." In: *MNRAS* 523.4 (Aug. 2023), pp. 5408–5429. DOI: 10.1093/mnras/stad1669. arXiv: 2306.01056 [astro-ph.GA].

[2] Connor Bottrell et al. "IllustrisTNG in the HSC-SSP: image data release and the major role of mini mergers as drivers of asymmetry and star formation." In: *arXiv e-prints*, arXiv:2308.14793 (Aug. 2023), arXiv:2308.14793. DOI: 10.48550/arXiv.2308.14793. arXiv: 2308.14793 [astro-ph.GA].

[3] Diego Sotillo-Ramos, Annalisa Pillepich, Martina Donnari, Dylan Nelson, Lukas Eisert, Vicente Rodriguez-Gomez, Gandhali Joshi, Mark Vogelsberger, and Lars Hernquist. "The merger and assembly histories of Milky Way- and M31-like galaxies with TNG50: disc survival through mergers." In: *MNRAS* 516.4 (Nov. 2022), pp. 5404–5427. DOI: 10.1093/mnras/stac2586. arXiv: 2211.00036 [astro-ph.GA].

ACKNOWLEDGMENTS

I want to express my sincere gratitude to the following individuals without this Thesis would not exist:

- Annalisa Pillepich, my supervisor and the main reason I pursued a PhD. Since 2019, her influence has been profound, shaping my academic and personal life like no other.
- Prof. Andreas Koch-Hansen, who volunteered at short notice to referee this thesis.
- Prof. Dr. Luca Amendola and Prof. Dr. Carsten Rother, who volunteered to participate as examiners in my final PhD Exam.
- Connor Bottrell, my close collaborator, whose HSC mocks heavily contributed to this thesis and who also encouraged me to pursue a Postdoc.
- Dylan Nelson and Ralf. S. Klessen, who served as second and third PhD supervisors.
- Rhythm Shimakawa, Vicente Rodriguez-Gomez, Eirini Angeloudi, and Marc Huertas-Company, who provided fruitful inputs as co-authors for my papers, now part of this thesis.
- Regina Sarmiento and Victor Ksoll, who gave valuable input on contrastive learning and Invertible Neural Networks.
- The Team of the GC Theory Group, who always provided me with good advice and thoughts about my results.
- Eric Rohr, Nils Hoyer, Noa Hoffmann, Mila Chadayammuri, and Selina Nitschai, who gave me good advice in making the thesis as readable as possible.
- Sofía Rojas, who helped me with applications and also recommended the LaTeX style for the thesis.
- André Miede, who actually created this LaTeX thesis style (Classic Thesis Style v4.2).
- Christian Fendt and Nina Zhao, who helped in administrative affairs many times.
- All Outreach Fellows, doing all those public tours with you was a welcome change to the daily PhD routine.
- My office mates and friends, Jan and Taavet, who provided motivation and made my time at MPIA unforgettable. Naturally, this also includes all the other PhD students, postdocs, and undergrad students.

- A special thanks to the group of interns and master students 2022/2023 who gave me an additional motivation boost during their stay at MPIA: Valeria, Anastasia, Chandana, Simona, Valentine, Clara, and Sam.
- My parents, who supported my lengthy studies financially and morally.

And finally, my cat, who passed away during the duration of the PhD.

CONTENTS

I MOTIVATION AND OVERVIEW 1

1 MOTIVATION 3

2 OVERVIEW 5

II INTRODUCTION 7

3 GALAXY EVOLUTION IN ΛCDM COSMOLOGY 9
- 3.1 ΛCDM Cosmology 9
- 3.2 Galaxies and Galaxy Evolution 13
- 3.3 Hierarchical Growth of Structure and Galaxy mergers 14

4 GALAXY EVOLUTION WITH HPC SIMULATIONS 19
- 4.1 Cosmological Simulations of Galaxies 19
- 4.2 The IllustrisTNG Project 20
- 4.3 Galaxy Evolution with IllustrisTNG 22

5 CONNECTING SIMULATED AND OBSERVED GALAXIES 29
- 5.1 Overview of recent and future Galaxy Surveys 29
- 5.2 Observational Mocks of Simulated Galaxies 30
- 5.3 Existing comparisons of simulated and observed galaxy populations 31

6 MACHINE LEARNING 35
- 6.1 Short Introduction to Machine Learning and Deep Learning 35
- 6.2 Representation Learning and Contrastive Learning 37
- 6.3 Simulation-Based-Inference via conditional Invertible Neural Networks (cINN) 38

III INFERRING THE ASSEMBLY HISTORIES OF ILLUSTRISTNG GALAXIES FROM INTEGRAL OBSERVABLE PROPERTIES VIA INVERTIBLE NEURAL NETWORKS 41

7 INTRODUCTION 43

8 SETUP, SIMULATION DATA AND METHODS 47
- 8.1 Rationale 47
- 8.2 The IllustrisTNG simulations for galaxy formation and evolution 47
- 8.3 Galaxy identification and past history 48
- 8.4 Galaxy sample 49
- 8.5 Galaxy observables and unobservables 49
- 8.6 Inference method: Conditional Invertible Neural Networks (cINN) 54

9 NETWORK TRAINING 55
- 9.1 Preprocessing of the sample 55
  - 9.1.1 Feature Scaling 55
  - 9.1.2 Sample split, in the context of galaxy populations across cosmic epochs 55
- 9.2 Network configurations and training 56
  - 9.2.1 cINN 57
  - 9.2.2 Ensemble Training 58

10 RESULTS: RECOVERING THE UNOBSERVABLE PROPERTIES OF GALAXIES' PAST 61
10.1 Testing procedure 61
10.2 Performance of the cINN 62
10.2.1 Posterior distributions from the cINNs 62
10.2.2 cINN predictions vs. ground truth 64
10.2.3 Interpretation of the uncertainties of the cINN predictions 67
10.2.4 Recovery of secondary dependencies and cross correlations 68
10.3 Sensitivity analysis: what are the most informative input galaxy properties? 71
11 DISCUSSION AND OUTLOOK 75
11.1 Understanding the cINN results 75
11.1.1 The case of the mean merger mass ratio 75
11.1.2 Stellar mass and time of the last major merger 78
11.1.3 Dependence on the shape of the TNG100 prior 82
11.2 For which galaxies can we get the best predictions on their past history? 83
11.3 Towards the application to real observations 86
12 SUMMARY AND CONCLUSIONS 89

IV COMPARING ILLUSTRISTNG AND HSC GALAXY IMAGES VIA CONTRASTIVE LEARNING 91
13 INTRODUCTION 93
14 GALAXY IMAGE DATA 97
14.1 Subaru Hyper Suprime-Cam images 97
14.2 TNG simulated galaxies and their HSC mock images 98
14.2.1 Forward modeling of simulated galaxies into HSC observations 98
14.2.2 Simulated galaxy sample used in this work 99
14.3 Data preparation and homogenization 99
14.3.1 FOV selection 100
14.3.2 Galaxy-population matching 101
14.3.3 Band selection and stretching 103
14.3.4 Splitting of the sample(s) 103
15 THE MACHINE LEARNING METHODS 105
15.1 E(2) equivariant Residual Networks 105
15.2 Contrastive Learning: NNCLR 105
15.3 Training of the Residual Network via NNCLR 106
16 RESULTS 111
16.1 Interpreting the representation space with UMAPs 111
16.2 How similar are the mocks from the TNG simulations to the observed HSC images? 112
16.2.1 Comparing the representation space with UMAPs 113
16.2.2 Quantifying the similarity via the normalized 8th-nearest-neighbour distance and the OOD Score 119
16.2.3 Connecting the OOD Score to galaxy properties 123
17 FIRST APPLICATIONS, INSIGHTS, DISCUSSION AND OUTLOOK 127

17.1 Identifying counterparts between TNG and HSC galaxies 127
17.2 A preview: are we able to infer the redshift, Petrosian radii and i-band-magnitude from the HSC images? 128
17.3 Discussion 133
17.3.1 Are IllustrisTNG galaxies sufficiently realistic for simulation-based inference? 133
17.3.2 The relation between OOD Score and the degree of realism 134
17.3.3 Implications for simulation-based inference 136
18 SUMMARY AND CONCLUSIONS 137

V THE ASSEMBLY HISTORIES OF HSC GALAXY IMAGES VIA INVERTIBLE NEURAL NETWORKS, CONTRASTIVE LEARNING, AND COSMOLOGICAL SIMULATIONS 139
19 INTRODUCTION 141
20 GALAXY DATA 143
20.1 Merger/Assembly Statistics 143
20.2 Galaxy Image Data 143
20.2.1 Observations: Subaru Hyper Suprime-Cam images 143
20.2.2 Simulations: TNG50 and TNG100 145
20.3 Data Preparation 145
21 FROM IMAGES TO POSTERIORS: MACHINE LEARNING METHODS 147
21.1 ResNet and Contrastive Learning 147
21.2 cINN 148
21.3 Training Procedure 148
22 RESULTS 151
22.1 Generating Posteriors and MAPs 151
22.2 Validation of inference results 153
22.3 Are the results transferable to the Domain of observed HSC images? 153
22.4 Recovering the Assembly History of HSC Galaxies via Images 156
23 DISCUSSION 161
23.1 What can be done to further improve the inference results? 161
23.2 Can we trust the inference? 162
24 SUMMARY AND CONCLUSIONS 163

VI DISCUSSION AND SUMMARY 165
25 DISCUSSION 167
25.1 Learning the posterior and the role of the prior 168
25.2 The representation space: challenges and opportunities 169
25.3 The Out of Domain score and the curse of dimensionality 170
25.4 Information Content of the data 170
25.5 Can we trust the resulting posteriors? 171
26 SUMMARY AND OUTLOOK 173

VII APPENDIX 177
A CALIBRATION ERRORS OF THE CINN MODEL 179
B THE BASELINE MODEL: A MULTILAYER PERCEPTRON (MLP) 181

B.1 Training of the MLP 181
B.2 Comparison between MLP and cINN: do they yield the same results? 183
C HYPERPARAMETERSEARCH WITH OPTUNA 187
D DEPENDENCE OF THE OUT OF DOMAIN SCORE ON THE MODEL AND PARAMETERS 191
E EXAMPLE FOR FALSE FRIENDS BECAUSE OF A FAILING FOV DETERMINATION 193

BIBLIOGRAPHY 195

Figure 1 **Visualization of the observed and simulated cosmic web.** On large scales, namely $\approx$ 100Mpcs, galaxies form a web-like formation that is the largest structure in the universe we know up to this date. This holds for observations like from SDSS (top plot) and cosmological simulations like the Millennium simulation (bottom plot). Note, that the observations show observed galaxies, while the simulation is based on the dark matter density in the simulation. Plots and results from [139]. 12

Figure 2 **The color-magnitude diagram of galaxies within the SDSS legacy survey.** The plots show the number density of galaxies in the color (u-r) - magnitude (r-band) plane. In the left plot, there is a clear bi-modality visible between blue/star-forming galaxies (at the bottom) and red/quenched galaxies at the top. In the right plot, this is visualized by plotting blue and red populations separately. The bi-modality is an indicator that the quenching of a galaxy happens in general on a significantly shorter timescale than the time spent in the star-forming or quenched population. It should also be noted how both galaxy populations adhere to clear scaling relations in the color-magnitude plane. The relationship between star-forming galaxies and their star formation rate is often referred to as the galaxy main sequence. Plots and results from [13]. 14

Figure 3 **Examples for nearby interacting galaxies.** From top to bottom: M51 or 'Whirlpool-galaxy', a spiral galaxy in Canes Venatici that undergoes a merger with a small companion galaxy. M63 or 'Sunflower-galaxy', a spiral galaxy in Canes Venatici with prominent stellar streams visible in its stellar halo. Those stellar streams are theorized to originate from accreted satellite galaxies. Markarian's Chain, a fraction of galaxies within the Virgo Cluster consisting of over 1000 galaxies. The Virgo Cluster is the center of the Virgo Supercluster which also contains the Local Group and therefore our own Galaxy. NGC3718, a spiral galaxy in Ursa Major that is heavily distorted due to gravitational interaction with the smaller companies visible at the bottom of the image. Images from [62]. All galaxies have been observed from the area around Heidelberg. 17

Figure 4 **Comparison of Cosmological Simulations from [121]** The computational cost scales on the one hand with the number of galaxies at the x-axis (and in the case of full cosmological simulations therefore with the volume of the simulation box) and with the resolution of the simulation on the y-axis (here the baryon mass resolution). The zoom regime is given as diamonds, with in average higher resolutions but lower numbers in galaxies. The regime of full cosmological simulations, as circles, are less well resolved but provide magnitudes of larger samples. 21

Figure 5 **Color Bi-modality in TNG and the diversity of evolution Tracks in TNG100.** Overview map of $z = 0$ galaxies in TNG100 in the color-stellar mass plane. The main bulk of galaxies are in 2 clusters: one blue/low mass cluster and one red/massive cluster. Each line represents an example galaxy moving across the plane from $z = 2$ to $z = 0$. Although in general, galaxies move from the blue population to the red population, the evolutionary pathways of individual galaxies are very diverse. Plot and results from [123]. 24

Figure 6 **Stellar Masses and Halo Masses in TNG100 and TNG300** Left: Number density of stellar masses in TNG100, TNG300 and Illustris at $z = 0$. TNG overall reproduces the trends in the number-densities observed (grey). Note how the availability of galaxies at a given stellar mass depends on the resolution and volume of a cosmological simulation: While the resolution of TNG300 is too low to resolve low-mass galaxies, the box of TNG100 (or, here not shown, TNG50) is too small to yield a significant number of galaxies with stellar masses above $10^{12} M_{\odot}$. Right: Stellar-mass to halo-mass (SMHM) relation. How many stars does a DM halo of a given mass contain? Interestingly, the relation between stellar mass and hallo mass is rather tight with a maximum at $10^{12} M_{\odot}$. Towards the lower mass end, Star Formation Feedback, and to the higher mass end, AGN Feedback is decreasing the ability of halos to form stars. Another interesting feature is the overall lower ability of the TNG model to produce stellar mass, compared to the original Illustris simulation lacking magnetic fields. Plots and results from [134]. 25

Figure 7 **Accretion of Stars vs Formation of Stars.** Left: Fraction of accreted, or ex-situ, stars in TNG300 at redshift $z = 0$ as a function of stellar mass or halo mass. Color-coded according to the aperture of the measurement. Ex-situ stars are all stellar particles that have not formed in the respective halo during their lifetime. Within 30kpc to the galaxy center, the fraction of ex-situ stars is strongly increasing with mass while the stellar halos outside of 100kpc are almost entirely made out of ex-situ stars. Right: Quenched Fraction at $z = 0$ as a function of stellar mass for TNG and several other cosmological simulations. Overall the fraction of quenched galaxies (in this case galaxies with a specific Star Formation Rate $sSFR < -11$) is increasing with stellar mass. The onset of quenching and the gradient with stellar mass differ. For TNG, quenching starts around stellar masses of $10^{10} M_\odot$ and the quenched fraction reaches almost 1 between $10^{10.5-11} M_\odot$. Beyond this point, galaxies' stellar inventory grows mostly due to merging. Plots and results from [134] and [56]. 26

Figure 8 **AGN activity can interrupt star formation.** Star Formation efficiency (star formation rate divided by the gas mass) vs the black hole mass and color-coded by the ratio of stellar mass and black hole mass. At the onset of AGN Feedback, a steep decrease in star formation efficiency can be seen. Plots and results from [195]. 27

Figure 9 **The connection between the stellar ex-situ fraction with the merger history of a galaxy.** Ex-situ stellar fraction vs stellar mass for three merger summary statistics: mean merger gas fraction, mean merger lookback time, and mean merger mass ratio. Especially the mass ratio and the lookback time is correlating with the ex-situ fraction. Plots and results from [147]. 27

Figure 10 **Example for Idealized and Realistic HSC Mocks** At the top we show idealized HSC mocks of TNG50 images without observational realism. At the bottom, we show the same images with full observational realism as they are used in [26] and [61]. Images adapted from [176]. 32

Figure 11 **Visual scheme of the setup and goals of this paper, ERGO-ML I.** We can predict the observable properties of thousands of galaxies given their individual histories and evolution using a cosmological hydrodynamical galaxy simulation, here TNG100. The goal of this paper is to model the inverse of this, i.e. the mapping from the observables to the unobservable stellar mass assembly and merger history (see also Table 1). Because of the possible loss of information in the forward step – that is, the observables are not sufficient to fully constrain the history –, we are especially looking for the posterior distribution, which is the probability that a certain merger history x is compatible with a set of observations c. In this work, we focus on utilizing scalar integral properties, i.e. galaxy features, and test this scheme on simulated galaxies. However, in subsequent work, we will go beyond features and use maps of galaxies and of their stellar haloes as inputs, and we will also apply the models developed here to actual observational data. 46

Figure 12 **Overview of the observable statistics used in this work, as a function of galaxy stellar mass and redshift, and predicted by TNG100**: in each row we plot the whole simulated sample (182'625 galaxies in total) introduced in Section 8.4 as 2D histograms between one of the observable quantities and the stellar mass (within $2 \times r^{\star}_{1/2}$). Curves denote the medians as function of the stellar mass at redshift $z < 0.2$ (solid), $0.2 < z < 0.5$ (dashed) and $0.5 < z < 1$ (dot-dashed). Note that there is in fact a bi-modality in the data between disc-blue and elliptical-red galaxies, which is not well represented by the median lines. 53

Figure 13 **Overview of the unobservable statistics used in this work and predicted by TNG100**: in each row we plot the whole simulated sample introduced in Section 8.4 as 2D histograms between one of the unobservable quantities and the stellar mass (within $2 \times r^{\star}_{1/2}$). Curves denote the medians as function of the stellar mass for redshift $z < 0.2$ (solid), $0.2 < z < 0.5$ (dashed) and $0.5 < z < 1$ (dot-dashed). Here, for the mass and the lookback time of the last major merger, we plot the galaxies without any major merger in their history as a bar at the top. Especially for the stellar ex-situ fraction and the mass of the last major merger, the simulations predict a strong positive correlation with the stellar mass of the respective galaxy; on the other hand, these relationships depend only weakly on redshift. This is naturally different for the unobservables that are defined relative to the redshift of the galaxy at hand, namely the mean merger lookback time and the time of the last major merger. 53

Figure 14 **Posterior distributions obtained with our cINN** (Sections 8.6 and 9.2.1) for 15 randomly-chosen galaxies from the TNG100 simulation test set (out of in total 17982 galaxies). Each row corresponds to one of the 15 galaxies (we show the redshift and the (snapshot-wise) unique SUBFIND ID on the left) whereas each column shows results for the five inferred unobservable statistics of their assembly and merger history. We show the prior distributions of the overall sample of test galaxies (grey), the predicted posterior by our cINN approach (blue), the MAP prediction (Maximum A-Posteriori estimation, yellow), and the ground truth (red) – i.e. the actual outcome of the TNG100 simulation. The distributions are normalized such that the maximum value is equal to 1. For the last major merger, we introduce an additional bin to show the fraction of posterior samples that fall outside the valid regime, i.e. the probability that the galaxy has never had a major merger in its merging history. This bin is chosen as the MAP for this galaxy only if the fraction of posterior samples falling into this bin is > 0.5 i.e. if the predicted probability for a major merger to have occurred is larger than 50 per cent. 59

Figure 15 **Prediction performance of our cINN based on training, validating and testing against the outcome of the TNG100 cosmological simulation.** The columns show results for the five inferred unobservable statistics of the assembly and merger history of galaxies, for 17'982 TNG100 test galaxies . **Top:** 2D histograms of the cINN posteriors (y-axis) over simulation ground truth (x-axis) for each of the 5 output quantities. Namely, for each test galaxy, we draw 100 random posterior samples and stack them across all test galaxies. The measured median is given as solid black line while the area included within the black dashed lines contain 80 per cent of the datapoints. For comparison, we also show the ideal result (i.e. the case whereby MAP prediction and ground truth concur) as dotted black diagonal. **Middle:** 2D histograms of the cINN MAPs (y-axis) vs. simulation ground truth (x-axis) for each of the 5 output quantities. Meaning of the lines and annotations as in the top panels. **Bottom:** 2D histograms of the cINN MAP errors, i.e the MAP - Ground Truth (y-axis) vs. simulation ground truth (x-axis) for each of the 5 output quantities. The solid black line shows the median, the area within the dashed and dash-dotted lines contain 68% (1σ) and 95% (2σ) of the datapoints in bins of ground truth values. We show for reference the ideal (zero error) baseline as red dotted line. For the time and mass of the last major merger (third and fourth panels from the left), we also show the galaxies that have no major merger (NMM) in their main progenitor branch according to the cINN (top), the ground truth (right) and both (upper right corner). The percentage denotes the overall fraction of galaxies in each of those 3 bins. Note that for the bottom plot, the right bin is empty as the error is undefined if the galaxy had no major merger according to the ground truth. 63

Figure 16 **Uncertainties of the cINN predictions. Top:** 2D histograms of the standard deviation of the cINN posterior distributions vs. the absolute value of the MAP prediction error for all TNG100 test galaxies. In bins of posterior standard deviation, we plot where the 68th and 95th percentiles of the test galaxies lie: thick black curves. The dashed lines correspond to the expected "ideal" result, assuming that the MAP prediction error is normal distributed. **Bottom:** 2D histogram of the standard deviation of the cINN posterior distribution vs. the simulation ground truth. The black curves denote the median relationships, whereas 80 per cent of the test galaxies are contained within the area enclosed by dashed lines, in bins of predicted properties (x-axis). For the time and mass of the last major merger, we plot the galaxies that have never had a major merger according to the ground truth (i.e. the TNG100 simulation) in an extra bin (NMM) and show the fraction of galaxies that fall into this bin: in these cases, the standard deviations are only calculated from the fraction of posterior samples that fall into the physical range (i.e. lookback time $\leqslant$ 13.7 Gyr and stellar mass $\geqslant 10^7\,M_\odot$). 66

Figure 17 **Recovery of the cross-correlations among the statistics of the stellar assembly and merger history.** We show how test galaxies occupy the parameter space among pairs of the cINN-predicted properties with contours at fixed number density: distributions of the ground truth test sample (ground truth priors, red), the posterior distribution predicted by the cINN trained and tested on the TNG100 galaxies (blue), and the MAP estimates based on these posteriors (green). For the posterior distribution, we include 10 randomly drawn posterior samples per galaxy. Here we focus on galaxies with stellar mass in the $10^{10-10.7}\,M_\odot$ range. The prior distributions are in general well reconstructed by the cINN although a certain level of smoothing is visible especially in the less sampled higher-mass bin: see next Figure. The MAPs are in general heavily biased towards the peaks of the prior distributions; however, also the "point-predictions" from the cINN recover well the underlying true distributions (green vs. red contours) in the case of the stellar ex-situ fraction and the mean merger lookback time. 69

Figure 18 As in Fig. 17 but for high-mass galaxies, with stellar mass in the $10^{10.7-12}\,M_\odot$ range. 70

Figure 19 **What inputs are the most informative?** Sensitivity analysis from the MLP model. Each field and color denote the change in the mean prediction error (in absolute value) for the target properties (left) if an observable input (top) is randomly shuffled across the test sample. As the shuffling breaks any information given by the respective input, a shuffled input is considered to be important if the error between the original prediction and the shuffled prediction rises accordingly. As the change in error is evaluated target-wise, it is only comparable for each output property separately but not across targets (i.e. not across rows). We see for example a strong dependence of the ex-situ fraction on the stellar mass and on the morphological/structural parameters such as the disk-to-total ratio and the half light radius. Very prominent is also the influence of the (luminosity-weighted) stellar age on the prediction of most merger statistics. 72

Figure 20 **UMAP representation of TNG100 galaxies at** $z = 0$, that embeds the six-dimensional space of the observable statistics (excluding the galaxies' redshift) into a 2D representation space. We deliberately omit axes to the panels as they have no immediate physical meaning. One dot corresponds to one galaxy. The colors denote, in each panel, the ground truth values from the simulation of the observable inputs (top) and the unobservable statistics of the stellar assembly and merger history (bottom). From top-left to bottom-right: stellar mass, stellar size, fraction of disc stars, g-r color, stellar age, ex-situ fraction, mean merger lookback time and mean mass ratios, and the lookback time and mass of the last major merger. For the last two quantities, galaxies with no major merger in their history are plotted with fixed (unphysical) masses and lookbacktimes according to Section 8.5. To obtain the UMAP representation, here we have used the whole TNG100 sample, including train, validation and test sets. In this representation space, the sample clearly decomposes into star-forming (right grouping) and quenched (left grouping) galaxies. It is remarkable how smoothly and coherently the ex-situ fraction is represented across this 2D space, even if only observable features have been used to create this representation. 77

Figure 21 **Capability of the cINN to deal with galaxies with and without a major merger.** We investigate how the fraction of cINN samples that fall into the NMM (No Major Merger) bin, i.e. the probability, relates to the ground truth from the TNG100 simulation. We plot the histograms of these NMM probabilities for test galaxies with no major merger in their ground truth (red) and with at least one major merger in their ground truth (blue). A cINN probability of 0.5 (vertical dotted line) has been chosen as the boundary above which galaxies are *classified* by the cINN to have had no major merger, i.e. we set the MAP estimate to the NMM bin (unphysical time or mass of the major merger) only if more than 50 per cent of the posterior samples fall into this bin. Our expectation is that the fraction of cINN posterior samples in the NMM bin is a good proxy for the fraction of ground-truth galaxies that had no major merger. This is the case: the red thick curve denotes the number of ground-truth NMM galaxies divided by the total number of galaxies in bins of cINN probability. We see that this quantity (red curve) indeed follows the fraction of cINN NMM samples (red dotted line). 79

Figure 22 **Which galaxies (in terms of observable properties) have the least ambiguous prediction in relation to having had or not a major merger?** We plot how the TNG100 test galaxies populate the space of the observables when divided in three classes, based on the probability assigned by the cINN to be a No-Major-Merger (NMM) case: < 20 (red), $20 - 50$ (green), and > 50 (blue) per cent, respectively. The galaxies with higher probability to be classified as NMM (blue curves) are typically lower-mass, bluer and diskier. Galaxies with masses $\gtrsim 10^{11}\,M_\odot$ are in general assigned by the cINN a very low probability that there was never a major merger in their history (red curves). However, the galaxies with ambiguous cINN probability (green curves) share similar physical properties with those with higher ones. The redshift and the stellar metallicity of galaxies do not exhibit significant biases across these three classes of galaxies, and so they are omitted. 81

Figure 23 **Prediction performance as a function of galaxy mass.** We investigate how the MAP error of our cINN model (MAP - Ground Truth) depends on galaxy stellar mass, by using all galaxies in our TNG100 test set in 5 uniformly-spaced mass bins. **Left**: Median relative errors (i.e. normalized by the ground truth values on a galaxy-by-galaxy basis) for the five inferred unobservable statistics of the assembly and merger history of galaxies. For the time and mass of the last major merger, we neglect galaxies without a major merger in their history. **Right**: Errors (MAP - Ground Truth) for the ex-situ fraction as a function of galaxy stellar mass. The median error is drawn as solid black line while the area containing 68% of the galaxies (1σ) is included within dashed lines. For comparison, we also plot the median of the posterior standard deviations given by our cINN model as a solid red line. To quantify how our results improve upon the bare relationship between galaxy mass and ex-situ fraction that could be extracted directly from the TNG100 simulation, we plot as grey shaded area the 68 per cent confidence interval of the prior distribution in the ex-situ vs. mass plane (i.e. the 16th-84th percentiles in bins of stellar mass of Fig. 13, rightmost panel, having subtracted the median value). 83

Figure 24 **Which galaxies (in terms of observable properties) have the least ambiguous assembly and merger history?** We answer this question by electing the ex-situ fraction as example. We separate the TNG100 test set in two subsets (stacking all redshifts): galaxies with a predicted ex-situ cINN posterior standard deviation that is below (blue) and above (red) the median of the whole test sample. The blue subset hence represents the galaxies whose predictions are less uncertain than those in the red subset. We show how these two sets populate the space of stellar mass vs. (from left to right) g-r colour, fraction of disc stars, half-light radius and average stellar age. In the plot edges, we show the projections of the two distributions on the x and y axis. We see that, for the ex-situ fraction, the cINN returns especially small error bars in the regime of low-mass, blue, and disc-like galaxies. 85

Figure 25 **Overview over the method proposed in this paper to compare observed and simulated galaxies at the image level.** We use a set of Hyper Suprime Cam (HSC) galaxy images and survey-realistic mocks of TNG50 and TNG100 simulated galaxies. We then use them together to train the contrastive learning method NNCLR, in order to map all images into the same representation space. This allows us in the following to compare the samples either in the representation space itself, via an Out-Of-Domain score and its distributions, or visually in a further dimensionally-reduced 2D UMAP representation. 95

Figure 26 **Matching the observed HSC dataset to the two simulated TNG50 and TNG100 samples of galaxy images, respectively.** We show the distribution of all three sets in photometric redshift, apparent I-band magnitude and petrosian radius before (top) and after (bottom) the matching (refer to Table 4 for definitions). We split the set of matched HSC galaxies into two subsets: one that is matched to TNG100 and one that is matched to TNG50. This is necessary as the prior distributions between the two simulated image samples are different, because of the different volume and resolution of the simulations. Note that we apply a larger smoothing kernel for the redshift: this is necessary as the simulated TNG galaxies are given only for a set of discrete snapshots. 102

Figure 27 **Are the representations related to observable features?** Each panel shows a 2D hexbin histogram of TNG100 galaxy images (left panels) and TNG50 galaxy images (right panels) in the 2D UMAP parameter space. The UMAP mapping itself was trained using the 256-dimensional representations of TNG50/100 and HSC test galaxies. Note that because of the smaller sample size of TNG50, we choose a larger bin size in the right panels. The bins are coloured according to the median image/galaxy property in each bin. We show from top to bottom, and in the continuation figures: total stellar mass, fraction of disk stars, integrated galaxy colour index (g-r), stellar half-mass radius, Sérsic half-light radius, Sérsic index, Sérsic ellipticity, point asymmetry of the light distribution, concentration of the light, smoothness of the light distribution, Gini-M20 bulge parameter and Gini-M20 merger parameter. The UMAP, and therefore the representations upon which it is based, are clearly related to observable properties of the TNG50/100 galaxies. 108

Figure 28 Continuation of Figure 27. 109

Figure 29 Continuation of Figure 27. 110

Figure 30 **How well do the representations of the observed and simulated galaxy images align to each other?** We compare the distributions of TNG50/100 and HSC images in the corresponding 2D-UMAP mapping of the 256-dimensional representations. In the uppermost panel we show KDE density plots of TNG100 (TNG50) images in blue (orange). Additionally the HSC sets that are matched against TNG100/TNG50, respectively, are shown in light blue/orange in two additional plots. We see that there is a large overlap among the three sets but also a slight offset and differences in point density. In the lower two panels we further investigate the offset in a visual way: we show the 2D hexbins of the UMAPs coloured bin-wise according to the median value of the three common properties used for the galaxy-sample matching: from left to right, redshift, i-band magnitude, Petrosian radius. In each row we show one of the datasets: from top to bottom, TNG100, HSC matched to TNG100, TNG50, HSC matched to TNG50. We also include the contours covering 80 per cent of the datasets from the uppermost panel: TNG100 in blue, TNG50 in orange and in light orange/light blue the subsets of HSC matched to TNG50/TNG100. 115

Figure 31 **What images underlie the UMAP representations?** We align TNG100 galaxy images to their positions in the UMAP representations from Figure 30. For each node on a rectangular grid in 2D UMAP space, we choose the closest data point (i.e. image representation) and plot its corresponding image. As a guideline, we report the contours as in Figure 30 covering 80 per cent of the datasets: TNG100 in blue, TNG50 in orange, and the corresponding HSC matched samples in light blue and light orange. With this visual inspection, we see that the galaxy structures are indeed well grouped together. 116

Figure 32 Same as Figure 31 but for TNG50 galaxies. Note that there are more empty regions due to the smaller sample size. 117

Figure 33 Same as Figure 31 but for HSC galaxies, specifically the HSC sample matched to TNG100. For brevity, we omit showing the analog figure for the HSC sample matched to TNG50, which would appear very similar. 118

Figure 34 **Our measure to quantify the discrepancy between the individual galaxies and a target dataset: the Out of Domain Score (OOD).** Here we show the distributions for the measures that we have introduced and that aim at quantifying the overlap between the domains of the observed (HSC) and simulated (TNG50 and TNG100) datasets. Left: Distributions of the 'raw' normalized 8th nearest neighbour distance. We measure this quantity for various scenarios: between pairs of the three data sets and also between two random splits of each dataset to investigate the intrinsic scatter that persists even if two distributions are drawn from the same population. For convenience we only show the HSC self-distance as the self-distances are qualitatively similar for all three sets. Note that there is a difference from which 'side' this quantity is measured: E.g. TNG100-HSC is the distribution of the distances for the TNG100 galaxies relative to the HSC galaxies. HSC-TNG100 is the distribution of the distances for the HSC galaxies relative to the TNG100 galaxies. However, for convenience we only show the distances in one direction as the distances are also qualitatively similar for both directions. Right: The measure that we use in the rest of the paper: the Out of Domain Score (OOD). It is overall equal to the left panel but normalized by the 95th percentile of the self distances. In doing so we define galaxies with scores larger than 1 as out-of-domain while scores smaller than 1 as in-domain. With this definition we measure that 33 per cent of the TNG100 and 30 per cent of the TNG50 galaxies are out-of-domain relative to the HSC images. Overall we measure that the comparison in 256 dimensions yield similar results as the ones in the 2D UMAP representations in Figure 30. There is a large overlap between all three datasets. However, it is also noteworthy that there are fewer extreme outliers between TNG50 and TNG100 than between the simulations and HSC. This is a hint for a general offset between TNG and reality. 121

Figure 35 **The Out of Domain Score as measurement for the offset between the TNG and HSC datasets.** Top-Left: UMAP of the representations of TNG100 galaxies. Bottom-Left: UMAP of the representations of HSC galaxies matched to TNG100. Top-Right: UMAP of the representations of TNG50 galaxies. Bottom-Right: UMAP of the representations of HSC galaxies matched to TNG50. In contours we show the 80 per-cent coverings of TNG50 (orange), TNG100 (blue), HSC matched to TNG50 (light orange) and HSC matched to TNG100 (light blue) galaxies. The UMAPs are coloured according to the bin-wise median of the Out of Domain Score assigned to the galaxies relative to the matched HSC galaxies (for the TNG sets) and relative to the TNG50/100 galaxies (for the HSC sets). As a reminder, a OOD Score larger than 1 denotes discrepancy. 123

Figure 36 **What galaxy observables drive the OOD Scores of the TNG images relative to the HSC images?** We plot 'in-principle' observable galaxy properties against the OOD Score; these are: total stellar mass, fraction of disk stars, integrated galaxy colour index (g-r), half mass radius, Sérsic half light radius, Sérsic index, Sérsic ellipticity, asymmetry of the light distribution, concentration of the light, smoothness of the light distribution, Gini-M20 bulge parameter and Gini-M20 merger parameter (these are the same as in Section 16.1). We use these distributions to investigate which types of TNG galaxies fall into the domain of HSC galaxies and which types do not. We perform this study on TNG100 galaxies (blue) and TNG50 galaxies (orange), relative to their respective matched HSC samples. The dashed lines show the distributions while the solid lines show the median in 10 fixed bins of OOD Score. We define all galaxies with OOD Scores smaller than 1 as within the domain, while all galaxies larger than 1 outside of the domain. Note that the TNG100 and TNG50 samples shown here represent distinct galaxy image populations: e.g. the stellar mass of the TNG50 sample is significantly lower than the one of TNG100. 125

Figure 37 **Identified TNG50/100 counterparts of HSC galaxy images.** We show TNG50/100 counterparts of randomly selected HSC galaxies which have been identified as the closest neighbour in the representation space (using the sum-norm). Additionally we show the OOD Score for each HSC galaxy. We see overall a good visual match among the galaxies. 129

Figure 38 **Identified HSC counterparts of TNG50/100 galaxy images.** Same as Figure 37 but with HSC counterparts for randomly selected TNG50/TNG100 galaxies. 130

Figure 39 Same as Figure 37 but with randomly selected HSC galaxies with high OOD Scores. We do not see any significant differences, which instead would have been expected given the high OOD Scores. So, based on this visual comparison, it appears that, in fact, the TNG sets seem to extrapolate well beyond their own domains. 131

Figure 40 Same as Figure 38 but with very high OOD scores. Here we clearly see that the identification of out-of-domain galaxies works. In particular, we see galaxies that are in a very crowded environment and/or for which the Petrosian fitting did not work i.e. captured not only the light profile of the central galaxy but a whole cluster leading therefore to a very large FOV. We also find cases of very large neighbouring galaxies and galaxies with a very clumpy appearance. These galaxies could be excluded successfully for scientific applications by using the OOD Score. 131

Figure 41 **How well can we reproduce galaxy properties of HSC by interpolating from TNG100 galaxies in representation space?** For this exercise we assign to each HSC galaxy the properties of the closest TNG100 neighbour in the 256-dimensional representation space and we focus on redshift, i-band magnitude and Petrosian radius. In the top panel we show the 2D histogram of the results for each property: on the x axis is the HSC ground truth while the y axis shows the quantities derived from the interpolation from TNG100. In the lower panel, we show the dependence of the prediction error (HSC Ground Truth - TNG100-based prediction) on the Out of Domain Score. Additional to the histogram we show in each plot the 'ideal' result as black dotted line and outline the distribution of the histogram also with red dotted lines. We see that with our model we can nicely reproduce the i-band magnitude and the Petrosian radius while the quality of the prediction is also dependent on the OOD Score. Note that this result is fully unsupervised – at no point was any active training performed to reproduce these quantities. The large scatter in the prediction error is especially caused by the coarse interpolation to the nearest neighbor in a high dimensional – and sparsely populated - space: the results of this figure represent only a lower estimation for the overall capabilities of an inference model which is based on these representations. 132

Figure 42 **Posterior distributions of the three assembly history statistics for 15 randomly selected galaxies from the TNG100 simulation test set.** Each row represents a single galaxy, with its redshift and unique SUBFIND ID listed on the left, while each column displays the inferred unobservable statistics of its assembly and merger history. The distributions are depicted for the entire test galaxy sample, i.e. the prior (in grey), our cINN model predictions (in blue), Maximum A-Posteriori (MAP) estimates (in yellow), and the ground truth from the TNG simulations (in red). These distributions are normalized so that the maximum value equals 1. Additionally, we include an extra bin for the last major merger quantities, indicating the fraction of posterior samples falling outside the valid regime, representing the probability that the galaxy has not experienced a major merger in its history. The MAP estimate for this bin is utilized only if the fraction of posterior samples exceeds 50 per cent, indicating a greater than 50 per cent likelihood that no major merger occurred during the lifetime of the respective galaxy. 152

Figure 43 **Validation of our cINN's predictive accuracy based on testing against ground truth data from the TNG50 and TNG100 simulations.** The columns display results for three inferred but unobservable statistics concerning the assembly and merger history of galaxies, assessed using TNG50 and TNG100 test galaxies. **Top:** Two-dimensional histograms between cINN Maximum A-Posteriori (MAP) predictions (y-axis) and the simulation ground truth (x-axis) for each of the three output quantities. The solid black line represents the median, while the area encompassed by the black dashed lines encompasses 80 per cent of the data points. The ideal scenario, where MAP prediction and ground truth align perfectly, is depicted by a dotted black diagonal line. **Bottom:** Similar to the top panel, but instead of MAPs, we use the predicted value of the posterior peak closest to the ground truth. This adjustment allows for the visualization of inferred values for misclassified non-major merger galaxies of the mass and lookback time of the last major merger. Additionally, for the time and mass of the last major merger (2 panels on the left), we show galaxies identified by the cINN as having no major merger (NMM) (top), or according to the ground truth (right), and both (upper right corner). The percentage denotes the overall fraction of galaxies in each of these three categories. 154

Figure 44 **Uncertainties of the cINN Predictions. Top:** Two-dimensional histograms of the standard deviation of cINN posterior distributions (y-axis) versus the absolute value of MAP prediction errors (|MAP predictions - ground truth|, x-axis) across all TNG50 and TNG100 test galaxies. Thick black curves represent the 68th and 95th percentiles of test galaxies within bins of posterior standard deviation. Dashed lines illustrate the expected outcome under the assumption of normal distributed posteriors. **Bottom:** Two-dimensional histogram of the standard deviation of cINN posterior distributions (y-axis) versus the simulation ground truth (x-axis). The black curves indicate the median, with 80 per cent of test galaxies falling within the area enclosed by dashed lines. For the time and mass of the last major merger, galaxies identified by the ground truth as never having experienced a major merger (NMM) are represented in an additional bin. The fraction of galaxies falling into this bin is denoted. In such cases, standard deviations are calculated solely from posterior samples within the physical range (i.e., lookback time $\leqslant$ 13.7 Gyr and stellar mass $\geqslant 10^{8.5}, M_\odot$). 155

Figure 45 **The Distribution in Out of Domain Score (OOD) after retraining the pre-trained ResNet with a cINN head.** We show the overlap between the domains of the observed (HSC) and simulated (TNG50 and TNG100) datasets after retraining the ResNet of [61] with the cINN head to infer merger properties. This exercise is important to ensure that the inference has not lead to a strong domain shift between simulations and observations. The plot shall therefore be compared to Figure 10 in [61]. We evaluate the Out of Domain Score (OOD) for various scenarios, including comparisons between pairs of the three datasets and between two random splits of each dataset. This investigation aims to understand the inherent scatter that persists even when two distributions are drawn from the same population. For simplicity, we present only the self-distance for the HSC dataset, as the self-distances exhibit qualitatively similar trends for all three sets. It's important to note that there is a distinction in which 'side' this quantity is measured: For instance, TNG100-HSC represents the distribution of distances for TNG100 galaxies relative to HSC galaxies, while HSC-TNG100 represents the distribution of distances for HSC galaxies relative to TNG100 galaxies. 157

Figure 46 **For which HSC galaxies can we trustfully infer merger properties from the TNG simulations?** We plot the inferred merger parameters of HSC galaxies introduced in Table 5 against the OOD score of HSC galaxies. With this, we can identify for which subset of HSC galaxies we can perform a trusted inference based on the TNG simulations. We perform this study on TNG100 galaxies (blue) and TNG50 galaxies (orange), relative to their respective matched HSC samples. The dashed lines show the distributions while the solid lines show the median in 10 fixed bins of properties. 158

Figure 47 **Merger History of HSC galaxies based on galaxy images:** We show the distribution of MAP predictions for the tree merger quantities (Lookback time and Mass of the Last Major Merger, Ex-situ Fraction) for the galaxies in our HSC test image set (grey line). We also show the distribution for HSC galaxies that are well within the domain of TNG images by choosing an OOD cutoff $\leqslant 1$ (dashed grey line). Here the OOD is calculated for all HSC galaxies relative to the combined set of TNG50 and TNG100 galaxies. For comparison, we also show the distribution for TNG50 and TNG100 galaxies. The fraction of HSC galaxies that are predicted to have –most likely– no major merger is ≈ 80 per cent. 159

Figure 48 **Calibration error of the cINN model as function of the confidence level for each of the 5 target quantities.** The calibration error is the deviation between the fraction of test galaxy ground truths contained within a certain posterior confidence interval and the expected fraction according to the respective confidence level. A positive calibration error therefore denotes an under-confident posterior while a negative value denotes an over-confident posterior. We see that especially the ex-situ fraction posterior are well within $\pm 2\%$, whereas the mean merger mass ratio and the lookback time of the last major merger tends to be slightly over-confident. We exclude galaxies with no major merger in their history for the calculation of the last major merger calibration error. 180

Figure 49 **Comparison of the MLP with the cINN MAP predictions**. On the x-axis, we plot the difference between the prediction errors from the two methods: |MLP − Ground Truth| − |MAP − Ground Truth|, where here MLP represents the MLP point-prediction value and MAP is the cINN Maximum A-Posteriori estimation. Therefore, a positive value of this quantity means that the cINN performs better in the predition than the MLP. **Top:** ground truth vs. difference between the errors of the MLP and cINN methods. We observe that the distributions are in general symmetrical around 0, but for the case of the mean merger mass ratio (second column from the left). The performance of the two models is therefore similar in terms of their point predictions, with a slightly better performance for the cINN. **Bottom:** standard deviation of the cINN posterior vs. difference between the errors of the MLP and cINN methods. The galaxies where the MLP and cINN MAPs disagree more substantially are also predicted to have a broader posterior distribution. This is especially pronounced for the ex-situ fraction. We speculate that this is an effect caused by the intrinsic uncertainty in the prediction of the selected unobservables. In these plots, for the time and mass of the last major mergers, we exclude galaxies with no major merger in their history. 184

Figure 50 We show the influence of the model to model variation on the OOD Score defined and calculated in Section 16.2.2. To do so we recalculate the OOD Scores for the 15 best Optuna runs and compare them at the galaxy-population level (left plot) and for individual galaxies (right plot). On the left we reproduce Figure 34 but showing the curves for all 15 runs. We see that the run to run scatter is too small to actually influence the plot qualitatively. On the right we show the model to model variance in OOD Score for each individual galaxy. The most galaxies have a variance below 0.2, which is small relative to the overall range in OOD values. 192

Figure 51 We show the influence of the choice in nearest neighbour on the calculation of the Normalized Nearest Neighbour Distance introduced in Section 16.2.2. We show the same plot for (from top-left to bottom-right) 4 choices in nearest neighbour: 8 (our fiducial choice), 4, 16 and 32. While the variation has an influence on the curve, the differences are too marginal to influence the claims made in this work. 192

Figure 52 Examples for "False Friends" because of a failing FOV determination. The depicted HSC galaxy on the left is given a very low score as it got closely grouped together with other galaxies (from HSC and TNG) whose FOV is significantly larger than the average. It is possible to identify such failing cases either by visual inspection or by comparing the known galaxy properties for the TNG mock images. 194

LIST OF TABLES

Table 1 **Observable and unobservable galaxy properties used and connected in this work.** At the top, we list and define the observable integral properties of galaxies (i.e. features) that are are used in this work to infer unobservable properties of galaxies' past assembly and merger history, bottom. 51

Table 2 **Layer configuration used for the** *conditioning network* **of the conditional Invertible Neural Network adopted in this work** (cINN, Sections 8.6). From left to right, we give the layer type, the output shape of the respective layer, and the number of trainable parameters per layer. 57

Table 3 **Layer configuration used for the cINN** *sub-networks* (see Section 9.2.1). Columns as in Table 2. 57

Table 4 Overview over the fundamental galaxy properties used to match the observed and simulated galaxy samples and for validation tests. 100

Table 5 **Unobservable galaxy properties used in this work and predicted for observed HSC galaxies.** We list and define the unobservable properties of galaxies' past assembly and merger history that are used in this work and that are obtained by applying a ResNet and cINN to HSC stellar-light galaxy images. Quantities are adapted from [63], note however the changed definition of a major merger, which now includes a mass cut at 8.5 $\log M_\odot$ of the secondary. 144

Table 6 **Layer configuration used for the Multilayer Perceptron neural network adopted in this work** (MLP, Sections B and B.1). We give the layer type (left column), the output shape of the respective layer (middle column), and the number of trainable parameters per layer (right column). Within our physical problem at hand, the 7 input parameters of the MLP are the 7 observable galaxy properties, whereas the target/output quantities are 5 summary statistics of the galaxies' assembly and merger history; both chosen and defined in Section 8.5. In total, this model has 202'757 trainable parameters. 182

Table 7 **Overview over the model and training parameters that have been optimized by using the hyperoptimization tool Optuna.** 188

ACRONYMS

CNN Convolutional Neural Network

cINN Conditional Invertible Neural Network

ERGO-ML Extracting Reality from Galaxy Observables with Machine Learning

FOV Field of View

HSC Hyper Suprime-Cam

HSC-SSP Hyper Suprime-Cam Subaru Strategic Program

MAP Maximum A Posteriori (Estimate)

ML Machine Learning

MLP Multilayer Perceptron

NMM No Major Merger

NNCLR Nearest-Neighbor Contrastive Learning of Visual Representations

OOD Out of Domain (Score)

ResNet Residual Neural Network

SDSS Sloan Digital Sky Survey

TNGxxx Referring to Specific Runs of the IllustrisTNG Project

UMAP Uniform Manifold Approximation and Projection for Dimension Reduction

Part I

MOTIVATION AND OVERVIEW

1 MOTIVATION

Galaxies are the fundamental structure of baryonic matter in the universe. For decades astronomers have been trying to understand how they form and evolve from the Big Bang until today. However, many observations and an unfeasible amount of telescope time would be required to detail all the physics of galaxies across cosmic epochs. Even then, information is lost due to the limited information the observations contain, either due to projection effects, observational biases, or the fact that we can only observe each galaxy at a single point in time. This especially hampers the study of galaxy mergers, which represent an important building block in the formation of (massive) galaxies. On the other hand, we can infer galaxy properties from simulations based on the physics we understand. Currently, the field of galaxy formation studies is being revolutionized by state-of-the-art cosmological simulations such as IllustrisTNG [106, 116, 121, 122, 134, 136, 171], EAGLE [44, 155], SIMBA [47], and HorizonAGN [91], which grant a profound insight into the history and evolution of galaxies and the entire Universe. While earlier cosmological simulations failed to replicate some of the straightforward relationships observed in the real cosmos, for example, the bi-modality in galaxy-color [188], recent advancements in simulation technology and understanding of galaxy physics have brought current simulations closer to producing genuinely realistic galaxy populations and morphologies [61, 199]. Such an increased realism –in-principle– allows us to learn physical parameters of interest from cosmological simulations and then infer them for observations. In this thesis, we especially use machine learning techniques for this complex task. By this, we can for example deduce the unobservable merger history from observations. Expanding on this idea, every parameter that can be extracted from the simulations could then be transferred to observations. However, in order to perform the inference, it is essential to ensure that the simulations are sufficiently realistic, and this necessitates a thorough comparison between simulations and observations. The conventional methods of comparing simulations and observations using mainly relations of single summary statistics, like stellar mass, are no longer adequate. This is especially true when considering the authenticity of mock data generated from recent simulations [26, 117, 197]. On the one hand, summary statistics are hard to define, especially consistently between simulations and observations. Furthermore, they are always connected to a potential loss of information compared to the original data. This thesis is centered around two focal points:

1. **Reassessing the Means of Comparison:** How can we move beyond evaluating simulations and observations solely through scalar properties, but instead transition to a direct comparison of the outputs from the observations? Which types of galaxies can the simulations reproduce, and which ones can they not?

2. **Knowledge Transfer Between Simulations and Observations:** Is it feasible to train Machine Learning models on simulations to infer unobservable prop-

erties from galaxy images captured by observational surveys? Can we bridge the gap between these two domains and transfer insights gained from simulations to enhance our understanding of the real universe?

2

OVERVIEW

This thesis aims to develop a robust machine-learning framework for simulation-based inferences of unobservable galaxy properties, in particular the assembly and merger history of galaxies from observational visual image data. Starting with a short introduction to the relevant fields (Cosmology and Galaxy Evolution, Cosmological Simulations, Comparisons between simulations and observations and Machine Learning), this thesis consists of 2 published papers: 'ERGO-ML I: Inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks' [63] (Part iii) and 'ERGO-ML: Comparing IllustrisTNG and HSC galaxy images via contrastive learning' [61] (Part iv). Furthermore, in Part v results are shown which are yet unpublished:

1. **ERGO-ML I: Inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks:** This paper on the ERGO-ML (Extracting Reality from Galaxy Observables with Machine Learning) series demonstrates the feasibility of inferring the otherwise unobservable stellar assembly and merger history of central galaxies using machine learning techniques. Using the TNG100 simulation from the IllustrisTNG project, the study uses a set of seven observable scalar integral properties of galaxies to infer unobservable parameters including the stellar ex-situ fraction, average merger lookback times and mass ratios, and the time and mass of the last major merger. We use a conditional Invertible Neural Network (cINN) to not only point predictions but posterior distributions: by this, we can model the uncertainty distribution which is caused by the inverse nature of the problem. This work successfully demonstrates the feasibility of using observable galaxy properties to constrain the history of galaxy evolution using machine learning and cosmological simulations.

2. **ERGO-ML: Comparing IllustrisTNG and HSC galaxy images via contrastive learning:** This sequel to the ERGO-ML series utilizes contrastive learning to directly assess the realism of simulated galaxies against observed counterparts, using a large sample of galaxies spanning $z = 0.1$ to $z = 0.4$. This dataset contains real Hyper Suprime-Cam (HSC) observations, in the form of r,g,i-band images, and a set of survey-realistic galaxy mock images, representing HSC observations, from the cosmological simulations TNG50 and TNG100. Through self-supervised contrastive learning, a comprehensive 256-dimensional representation space is constructed, allowing the identification of simulated galaxies closely resembling real ones and simulated galaxies that look dissimilar to the observed galaxies. Notably, the majority of simulated galaxies align well with observed HSC images, albeit with identifiable discrepancies in certain morphological attributes. We furthermore introduce a metric to measure and quantify these discrepancies and similarities.

3. The forthcoming work pioneers the inference of merger and assembly-related galaxy properties from observational HSC images, a task previously unfeasible with classical methods. Starting from the self-supervised contrastive learning framework from [61] and the cINN from [63], the study allows the modeling of the posterior distributions for merger/assembly statistics based on observed HSC images.

Summarizing, these contributions allow for a deeper understanding of galaxy evolution by bridging the gap between observational data and theoretical simulations through machine learning methodologies.

Part II

INTRODUCTION

3

GALAXY EVOLUTION IN ΛCDM COSMOLOGY

Humanity's view of the heavens has gradually evolved over time. Initially, light in the nighttime sky, with its appearance, arrangements, motion, and changeability, was subject to legends. The Earth was considered to be at the center of the universe, with the sky following its own rules separate from the laws of physics on Earth. The invention of the telescope allowed for deeper exploration, enabling the observation of planetary surfaces, their moons, and even detailed observations of nebula-like objects among the fix-stars. It took a few hundred years until it was proven that some of these nebulae are much farther away than the stars surrounding Earth. We now know these objects as galaxies: gigantic gravitational bound masses of stars and gas, on a par with our own Galaxy. In this thesis we are especially interested in the evolutionary history of those objects: For this, astronomers have to understand a broad field of physics, magnitudes of scales and time: from quantum physics, which is e.g. relevant on small scales for the physics of stellar object, over to hydrodynamics, which is relevant for the gas dynamics in galaxies and gravity which dominates the motions of masses on large scales to the cosmological model which describes how spacetime itself was and is evolving. In this chapter, we will examine the basics of the most successful cosmological theory, ΛCDM. We will then have a short excursion into galaxies, the most important physical processes, and their evolution. Furthermore, we will introduce the key aspects of structure growth and galaxy mergers, not looking at galaxies as single identities but as objects that participate in interactions with each other and their surroundings during their lifetime.

3.1 ΛCDM COSMOLOGY

At the beginning of the 20th century, it was widely believed that the Universe consisted solely of the Milky Way, and galaxies were merely nebulas within it (e.g., the "Andromeda Nebula"). Using Cepheid variables (i.e., variable stars with a well-defined luminosity-period relation), Edwin Hubble demonstrated that these "nebulas" were galaxies in their own right, far beyond the Milky Way. He also showed that these galaxies exhibited increasing redshifts with distance from the observer. This increasing redshift is described by Hubble's law:

$$z \approx H_0 \cdot d$$

where z represents the redshift, d denotes the (luminosity) distance between the object and the observer, and H_0 is the Hubble constant.

To explain this redshift, the theory of general relativity is necessary. In brief, general relativity replaces the concept of a three-dimensional space, where masses are attracted by gravity, with a four-dimensional spacetime. The curvature of this spacetime controls the dynamics of matter, which in turn determines the curvature.

This interplay between the space-time-geometry and the matter within is described by the Einstein field equations:

$$G_{\mu\nu} = \kappa T_{\mu\nu}$$

where $G_{\mu\nu}$ represents the Einstein tensor, a measure of spacetime curvature, and $T_{\mu\nu}$ denotes the energy-momentum tensor in (cosmological) spacetime.

This equation can be solved for the whole universe under the assumption of the cosmological principle, which states that the universe is homogeneous and isotropic. While this assumption is necessary to construct a theory of the entire universe based on observations centered around Earth, it is not fulfilled on 'small' scales (planet systems, galaxies, or even clusters of galaxies). Nevertheless, observations (e.g. in Figure 1) uncovered a web-like large-scale structure, with matter accumulating in filaments, sheets, and knots, where the cosmological principle seems to be valid on scales above $\approx 100\mathrm{Mpc}$ [152]. On those scales, one can therefore theorize an analytical model of the universe, neglecting density perturbations on smaller scales. Utilizing this assumption and the Einstein field equations, all possible spacetimes can be described by the Friedmann – Lemaître – Robertson – Walker metric:

$$ds^2 = dt^2 - a(t)^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right)$$

where k represents the position/time-independent scalar curvature, and $a(t) = \frac{d(t)}{d(t_0)}$ describes the expansion or contraction of space. This expansion of space explains the redshift, with $a(t) = \frac{1}{1+z}$. In such a framework, there is no absolute definition of distance anymore; one can differentiate between the comoving distance d_c, which remains constant over time, and the proper distance $d_p(t)$, which varies between two points in space at a certain time. The connection between these distance measures is given by $d_p(t) = a(t) \cdot d_c$ when d_c is defined such that it equals d_p at $z = 0$.

But why is space expanding at all? To determine the evolution of a, one must speculate about the constituents of the Universe, i.e., determine $T_{\mu\nu}$. Similar to spacetime, this tensor is assumed to be homogeneous and isotropic and can be written as $T_{\mu\nu} = \mathrm{diag}(-\rho, p, p, p)$, with density ρ and pressure p values dependent on the composition and nature of the universe's constituents. With this tensor, the general solution of the Einstein equations is given by:

$$\begin{aligned} H^2 &= \frac{8\pi}{3}\rho - \frac{k}{a^2} \\ \dot{H} + H^2 &= -\frac{4\pi}{3}\left(\rho c^2 + 3p\right) \end{aligned}$$

These equations, known as the Friedman equations, return the evolution of spacetime from its ingredients. The Hubble parameter $H = \frac{\dot{a}}{a}$ measures the rate of expansion and can be viewed as the time-dependent extension of the Hubble constant H_0 measured today. All solutions of these equations share the property that as t

approaches 0, the scale parameter a approaches 0. This singularity is also referred to as the Big Bang. However, the Big Bang is neither a component nor a prediction of the theory but rather the point at which the theoretical model breaks down.

We will now trace the evolution of the Universe from shortly after the Big Bang. At this time, the Universe was filled with a hot, densely interacting plasma of elementary particles. As space continued to expand, the energy density decreased, allowing the formation of protons (i.e., ionized hydrogen) and basic nuclei such as helium, deuterium, and lithium. The temperature further decreased until the nuclei recombined with electrons to form neutral atoms. At this point, radiation decoupled from matter, and the universe became transparent. This event is known as the epoch of recombination $z(\approx 1100)$, and we observe this decoupled radiation today as the Cosmic Microwave Background (CMB). Fluctuations in temperature within this background, presumed to be caused by quantum fluctuations in the early Universe, can be studied to determine the parameters of our cosmological model. The matter then continued to gravitational collapse into the overdensities seeded by those fluctuations. This collapse turned the comparatively small fluctuations in the plasma over the last 13.8 Gyrs into the large density differences we can observe today.

Up to this point, we have used the term "ingredient" as a placeholder. But is there more to the Universe than just matter and radiation? While the concept of Dark Matter, which does not interact electromagnetically, was initially proposed to explain galactic rotation curves, it also turned out to be necessary to reproduce the observed CMB within the model. Additionally, observations of supernovae and then distant quasars revealed not only an expanding universe but also that this expansion is accelerating. Although the physical mechanism driving this acceleration remains largely unknown, it can be easily incorporated into the Friedman equations by introducing a constant vacuum density ρ_{vac} known as Dark Energy.

The model described above is known as ΛCDM cosmology. The term Λ is a nod to Einstein's cosmological constant, which is an additional degree of freedom in the field equations (this aspect was simplified in the introduction) and can be linked to the vacuum density ρ_{vac}. CDM stands for Cold Dark Matter and represents one specific type of proposed dark matter candidate that provides the best results in predicting the large-scale structure of our universe today. While CDM behaves similarly to normal matter, except for its lack of electromagnetic interaction, it leads to a bottom-up formation scenario in which smaller structures form first.

We want to emphasize that the ΛCDM model is just one possible cosmological model but, at this point by far, the most successful one. This said, there are still many problems with this theory: with dark matter, the model requires an invisible mass that stubbornly evades direct observations while the nature of dark energy is completely unknown. Considering this, the success of ΛCDM is even more fascinating. Nevertheless, there are yet open conflicts between theory and observations. One of the dominant ones is the Hubble Tension where measurements of the Hubble Constant based on Cepheids and based on the CMB are significantly differing from each other [15].

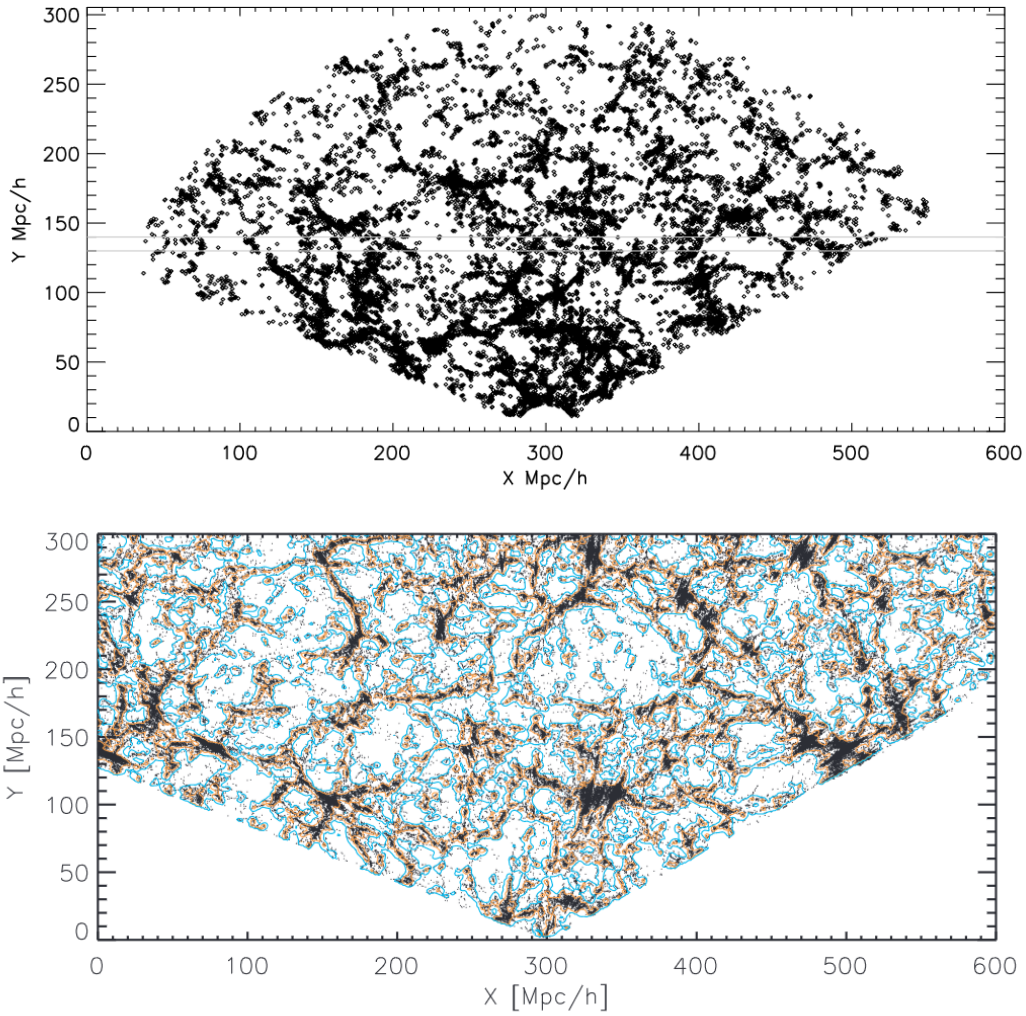


Figure 1: **Visualization of the observed and simulated cosmic web.** On large scales, namely ≈ 100Mpcs, galaxies form a web-like formation that is the largest structure in the universe we know up to this date. This holds for observations like from SDSS (top plot) and cosmological simulations like the Millennium simulation (bottom plot). Note, that the observations show observed galaxies, while the simulation is based on the dark matter density in the simulation. Plots and results from [139].

3.2 GALAXIES AND GALAXY EVOLUTION

We now provide basic inputs to the objects of interest in this thesis: galaxies. As the entire field of Galactic Physics employs a large number of scientists worldwide, one cannot expect to cover the entire field but only provide a brief overview of what galaxies are and the most basic mechanisms that dominate their appearance.

We want to first describe the observations and then gradually add theory to understand what we see and also comprehend their long-term evolution. When we visually observe galaxies, the dominant feature we see is a large number of gravitationally bound stars. These groups of stars can take on many shapes: at the extremes, they are perfectly aligned in a disk or diffuse in elliptical haloes. These shapes were first attempted to be classified in the so-called Hubble scheme. However, the Hubble scheme is very simplistic and not fully representative of the large variety of shapes and morphologies we can observe. Observations of many galaxies have revealed that there are many scaling relations that are valid, such as the Faber-Jackson Relation [65], the Fundamental Plane for ellipticals [54], and the Tully-Fisher Relation for disks [181].

Interestingly, when observing the rotation speed of disk galaxies, one can observe that the galaxies need additional mass for the observed stellar rotation to be explained by Newtonian gravity [16]. This need for an additional mass led to the proposal of Dark Matter (DM), which was then nicely incorporated into the cosmological ΛCDM model we previously introduced. From theory, it is assumed that each galaxy resides inside a much more massive DM halo and stars form out of gas that cools and concentrates at the bottom of the potential well of DM haloes.

In fact, another important constituent of galaxies is gas, the so-called Interstellar Medium (ISM), mostly composed of hydrogen with various amounts of metals (i.e., elements heavier than helium). Depending on the temperature and density, this gas can be ionized, neutral, or molecular hydrogen. Stars are formed out of the ISM, but the gas has to cool down enough to gravitationally collapse into a star, i.e., Molecular Hydrogen is needed. On the other hand, if a star is born, the radiative pressure will push out the gas again and also reheat the gas, which might prevent further star formation near a star. When a star dies, gas (including the metals formed due to fission inside the star) is ejected into the ISM.

In a closed system, the ISM in a galaxy will therefore be consumed by star formation until all gas is converted to stars, removed, or too hot to form more stars. With this knowledge, one can now, at first order, understand the two most dominant observed galaxy types: disks and ellipticals.

Gas that falls into a DM halo conserves its angular momentum. Due to collisions between the gas particles, the motion of the gas will align in a disk orthogonal to the total angular momentum vector. Within this gaseous disk, stars are formed and inherit the shape of a disk. The color of the galaxy is dominated by young, extremely bright blue O and B stars ($\approx 10000 L_\odot$). If star formation ultimately ceases (due to the lack of gas), those blue massive stars vanish much faster than their low-mass red counterparts (with lifetimes of a few million years compared to the 10 billion years of the sun), and the galaxy will turn red. Furthermore, the orbits of the stars will start to thermalize, and the shapes will transform towards ellip-

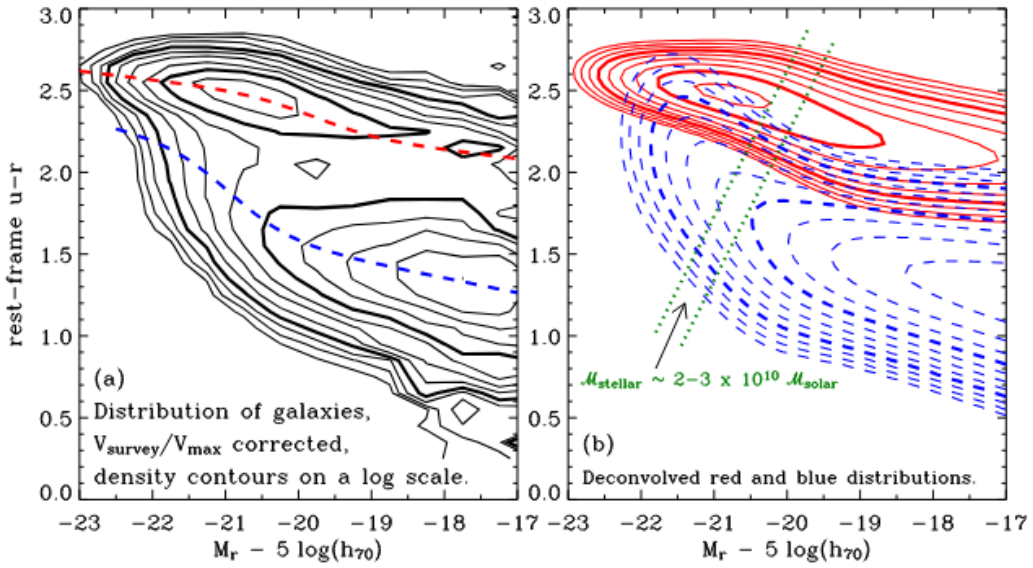


Figure 2: **The color-magnitude diagram of galaxies within the SDSS legacy survey.** The plots show the number density of galaxies in the color (u-r) - magnitude (r-band) plane. In the left plot, there is a clear bi-modality visible between blue/star-forming galaxies (at the bottom) and red/quenched galaxies at the top. In the right plot, this is visualized by plotting blue and red populations separately. The bi-modality is an indicator that the quenching of a galaxy happens in general on a significantly shorter timescale than the time spent in the star-forming or quenched population. It should also be noted how both galaxy populations adhere to clear scaling relations in the color-magnitude plane. The relationship between star-forming galaxies and their star formation rate is often referred to as the galaxy main sequence. Plots and results from [13].

tical galaxies. Because of this, one can observe the so-called color bimodality, for example, in a color-magnitude diagram like in Figure 2.

However, galaxies are not isolated systems. They are bound to other galaxies by gravity, which will eventually lead to interactions that are important for their evolution. As these interactions will be a special focus of this thesis, we will take a close look at this in the next section.

Another important mechanism is Active Galactic Nuclei (AGN), supermassive Black Holes in the central region of a galaxy with masses above $10^5 M_\odot$. Gas that falls into such a black hole will release huge amounts of energy, which will heat and/or eject gas from the galaxy and can therefore also favor the quenching of a galaxy.

3.3 HIERARCHICAL GROWTH OF STRUCTURE AND GALAXY MERGERS

To fully understand the formation of galaxies, we must not only examine isolated galaxies but also comprehend the entire evolution of, both individual and populations of, galaxies within a cosmological context and time frame.

How do galaxies come into existence? After recombination, the universe becomes filled with neutral gas that begins to gravitationally collapse, following dark matter

halos which are expected to form even faster due to the absence of repelling electromagnetic forces between solely gravitationally interacting dark matter. These halos are forming the cosmic web shown in Figure 1. Within these halos, infalling gas forms the first stars. These very massive Population III stars emit high-energy photons, which ionize the gas and render the universe transparent again. This happens around redshift $z \approx 10 - 6$ and is called the epoch of reionization. From this point onward, gas within dark matter halos accumulates, forming the stellar inventory observable today through star formation, which peaks at $z = 2$.

However, this evolution is not straightforward and is influenced by a large number of diverse processes. For gas to fall into a dark matter halo and form stars, it must cool down. This might explain why we observe significantly fewer baryons in galaxies today than expected from ΛCDM, with approximately 40 per cent 'missing baryons'. Another complexity arises from galaxies being congregated in large groups (e.g., the local group, including the Milky Way, Andromeda galaxy, Triangulum galaxy, and many smaller galaxies) and clusters (e.g., the Virgo cluster with thousands of galaxies). In such large clusters, we can observe parts of these missing baryons as hot X-ray-emitting intra-cluster medium (ICM), which failed to fully collapse and form stars. Furthermore, the ICM can deplete already collapsed gas in infalling galaxies due to ram pressure stripping, ultimately halting star formation in those galaxies (although it is theorized that increased gas density inside the galaxy could enhance star formation for a period of time).

Star formation can only advance galaxies so far in stellar mass, as the age of the universe is finite. How can galaxies like M87 in the center of the Virgo cluster form? Spending time with each other, galaxies yield another interesting phenomenon: dynamical friction. A mass (e.g., a galaxy) moving within a field of mass (a dark matter halo) causes the mass density to rise behind itself, inducing deceleration without physical contact. By this mechanism, galaxies moving close enough deposit some of their kinetic energy into their stellar and dark matter components, potentially leading to their merging into one. Observationally, (groups of) galaxies have been observed in the act, theorized to be related to such gravitational interaction. Importantly, merging is not solely responsible for building very massive galaxies but is a common process in galaxy evolution that gradually forms more and more massive galaxies. The signatures of mergers can also be observed a significant time after the merging event: The stellar components of merging galaxies usually end up on thermalized orbits around the final galaxy's main body, forming extended stellar halos that can be observed around galaxies, often exhibiting distinct features like shells and streams of stars. A small example of such interacting galaxies is shown in Figure 3.





Figure 3: **Examples for nearby interacting galaxies.** From top to bottom: M51 or 'Whirlpool-galaxy', a spiral galaxy in Canes Venatici that undergoes a merger with a small companion galaxy. M63 or 'Sunflower-galaxy', a spiral galaxy in Canes Venatici with prominent stellar streams visible in its stellar halo. Those stellar streams are theorized to originate from accreted satellite galaxies. Markarian's Chain, a fraction of galaxies within the Virgo Cluster consisting of over 1000 galaxies. The Virgo Cluster is the center of the Virgo Supercluster which also contains the Local Group and therefore our own Galaxy. NGC3718, a spiral galaxy in Ursa Major that is heavily distorted due to gravitational interaction with the smaller companies visible at the bottom of the image. Images from [62]. All galaxies have been observed from the area around Heidelberg.

4

GALAXY EVOLUTION WITH HPC SIMULATIONS

As we outlined in the previous chapter, galaxy evolution is a complex field, and analytic models can only go to a certain level. Because of this, modern astrophysics utilizes High Performance Computing (HPC) simulations of Galaxy Formation and Evolution. Such simulations can be, such called, zoom-in simulations (e.g. FIRE, AURIGA) which focus on the evolution of a single or a group/cluster of galaxies. While such models help to understand the complex interactions and processes in galaxy evolution they lack the cosmic variance of galaxies as they would be observed (i.e. by producing a realistic galaxy population). This can be circumvented by running a full cosmological which evolves a whole representative part of a model universe across cosmic time. In the following, we will introduce cosmological simulations in general and then focus on the details of the IllustrisTNG Simulation Project and its predictions.

4.1 COSMOLOGICAL SIMULATIONS OF GALAXIES

Our Galaxy has many billion stars spanning over 10th of kpc. How could one even simulate a single galaxy or even a whole universe? Early theorists used N-body simulations (gravitation only) with Dark Matter particles only to study the large-scale structure of the universe which would arise in ΛCDM from the initial fluctuation observed at the cosmological microwave background. A famous example is the Millennium simulation [166] which incorporated 10 billion dark matter particles (still fewer particles than stars in our own galaxy) in a cube spanning $\approx$ 750 Mpcs on each side and provided us with a better understanding how galaxy halos form, their masses and distribution in knots and filaments. The reasoning behind DM-only simulation is based on the fact that ΛCDM predicts that over 80 per-cent of the matter in the universe is dark and the structure formation might be therefore also dominated by DM.

But this kind of simulation does not yield any information about the material we can actually observe: baryonic matter like stars and gas (For completeness it has to be mentioned that there are semi-analytic models to calculate galaxy properties based on the DM-only simulation [45] [174]). Furthermore, simulations incorporating only Dark Matter lead to significant conflicts with observations. Famous problems are for example the 'missing satellites'-problem with ΛCDM simulations producing way more low mass satellite galaxies than observed. 'Cusp-Core'-problem: Dark Matter halos from observed rotation curves suggest a flat DM density in the center of halos while DM profiles from N-Body simulations suggest a 'Cuspy' -center with a density peak at the center. 'The Too Big To Fail' problem: satellites reside in too massive halos compared to observations. Those problems (and others) can be caused by two distinct sources: Is ΛCDM failing or is the assumption that Baryonic Feedback is negligible? Or maybe a mixture of both?

Because of this, over a decade after the initial release of the Millennium Simulation, we are now in the era of large-scale hydrodynamical cosmological simulations. This means that –in addition– to the simulation of only gravitational interacting DM Particles one adds gas and stars into the simulation. This however opens many new challenges to solve: hydrodynamical equations have to be solved for the gas spanning tremendous magnitudes in density between gas in star-forming regions, sparse Circum Galactic Medium (CGM), and even the less dense intergalactic medium (IGM). The formation of stars out of gas, the evolution of stars, and the feedback stars back in the ISM have to be dealt with. The evolution of central supermassive black holes (SMBHs), the AGN, and AGN Feedback has to be accounted for. And as we are interested in a statistically relevant sample this has to be achieved in a cosmological volume with many 10s of Mpc edge length. Because of that, resolutions of those simulations have to be adjusted accordingly: Stellar particles represent stellar populations with $\approx 10^5 - 10^7 M_\odot$ (IllustrisTNG Project) and gas cell sizes of $\approx 0.3 - 1.5$kpc (IllustrisTNG Project). Physics below this resolution limit has to be approximated by effective theories. Differences in those so-called –Sub Grid Models– but also numerical solvers lead in the recent past to various simulation projects: Illustris, Eagle [44, 155], Horizon-AGN [91], SIMBA [47]. Depending on the scientific question at hand, models, and resolutions can be tailored accordingly. With TNG50 of the IllustrisTNG project, for example, the barionic mass resolution almost reaches the realm of zoom simulation therefore allowing a detailed study of galaxy formation within a representative cosmological volume. TNG-Cluster [120] re-simulates specifically 352 massive galaxy clusters drawn from a very large volume of 1Gpc enabling especially the study of those massive objects. MilleniumTNG [127] on the other hand simulates the full Millenium simulation box ($\approx$ 750 Mpcs) but with a magnitude lower resolution than TNG-Cluster/TNG300 aiming for the study of baryons in the large scale structure. CAMELS [186] on the other hand is a suite of many 1000s of simulations, computed with various codes, subgrid models, and varying cosmological parameters. With this, a study of the influence of all those boundary conditions on the resulting cosmos is possible. An overview of past relevant simulation projects is also shown in Figure 4.

4.2 THE ILLUSTRISTNG PROJECT

IllustrisTNG [TNG hereafter, 106, 116, 121, 122, 134, 136, 171] is a suite of full cosmological simulations consisting of three main volumes: TNG50, TNG100, and TNG300, with respective comoving box sizes of 50 Mpc, 100 Mpc, and 300 Mpc. The volume is largest for TNG300 (therefore especially tailored for the study of massive galaxy clusters) and the resolution highest for TNG50 (baryonic mass resolution of $5.7 \times 10^4 M_\odot$ compared to $7.6 \times 10^6 M_\odot$ of TNG300) which almost reaches into the realm of zoom-in simulations. There are additional subruns with lower resolutions for resolution studies and DM-only runs for the study of the impact baryons have on galaxy formation. IllustrisTNG is based on the cosmological hydrodynamical simulation code AREPO, like its predecessor Illustris. In contrast to Illustris, it includes magnetic fields, thereby solving coupled Magnetohydrodynamic equations. The astrophysical model incorporates gas cooling/heating, star

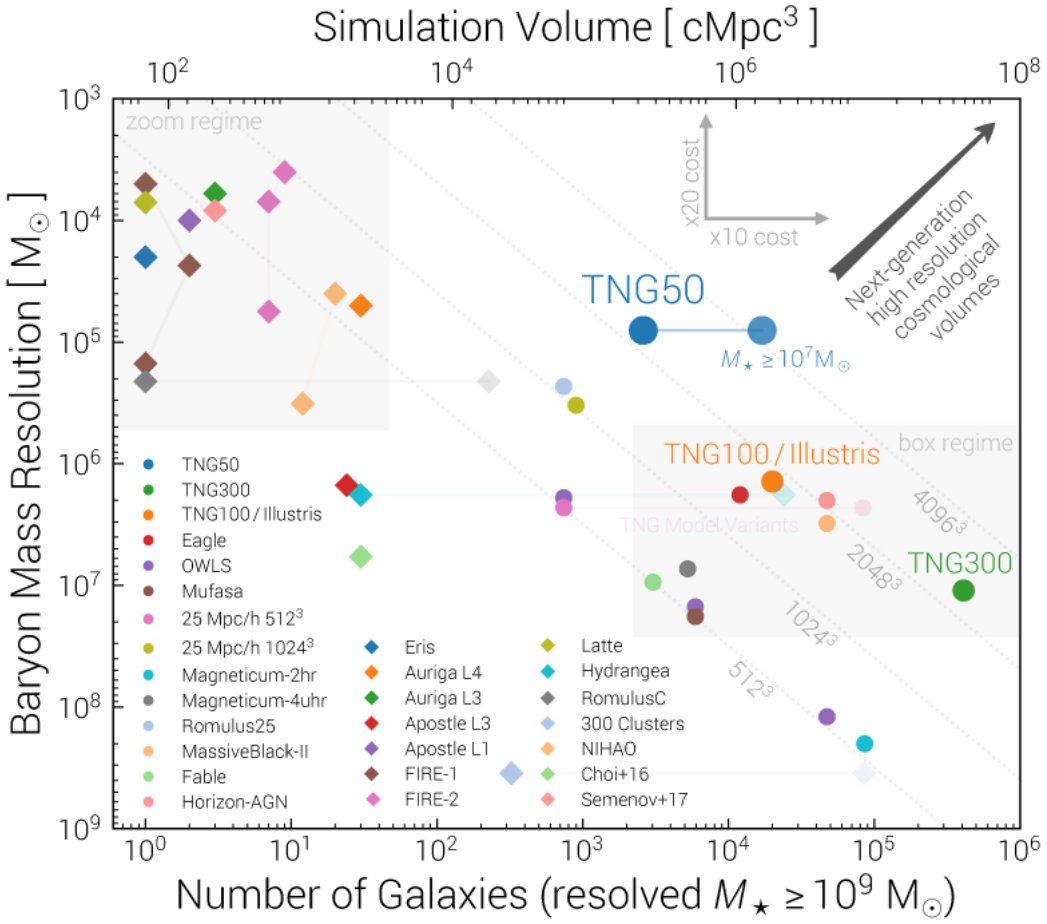


Figure 4: **Comparison of Cosmological Simulations from [121]** The computational cost scales on the one hand with the number of galaxies at the x-axis (and in the case of full cosmological simulations therefore with the volume of the simulation box) and with the resolution of the simulation on the y-axis (here the baryon mass resolution). The zoom regime is given as diamonds, with in average higher resolutions but lower numbers in galaxies. The regime of full cosmological simulations, as circles, are less well resolved but provide magnitudes of larger samples.

formation, stellar/AGN evolution, and feedback. The simulation boxes are initialized at $z = 127$ from a Zel'dovich approximation reflecting the measured cosmological microwave background as measured by the Planck satellite [138].

The resulting data products of the simulation runs are the positions and properties of star particles (representative of star populations), gas cells, dark matter particles, and black hole particles representing the supermassive black holes in each galaxy center, as well as wind particles, at 99 snapshots equally spaced in cosmic time.

Two codes have first been run on this particle data: Subfind [168], which identifies minima in the gravitational potential and assigns bound particles to those so-called SubFind Halos or Subhalos. These Subhalos are the simulated counterparts of actual galaxies.

The second code is a friends-of-friends (FoF) algorithm that links dark matter particles according to a linking length. By this, groups and clusters of galaxies can be identified. In these groups, the first subhalo is called the central subhalo, while the other subhalos in a FoF group are called satellites.

The results are publicly available either as direct particle data, integrated subhalo properties, or FoF group properties. In this thesis, we will also rely on the merger and assembly catalogs, which have been calculated from the subhalos catalogs by matching stellar particles across snapshots. By this, a subhalo's stellar inventory at a given snapshot can be identified with the stellar particles contained in one or more subhalos in the previous snapshot. The result is a merger tree containing all progenitors of a subhalo. From this merger tree, further properties used in this work can be achieved: average merger times and masses, masses and times of the last major merger, and also the ex-situ fraction: the fraction of stars that have accreted by mergers over the total stellar mass.

4.3 GALAXY EVOLUTION WITH ILLUSTRISTNG

The TNG Project allows the study of galaxy formation across various orders of magnitudes in size and resolution. This should be just a short overview where we focus on the most basic numbers and results relevant to the understanding of the following work. A success of TNG is the reconstruction of the observed galaxy color bi-modality in the universe [121]. In Figure 5 the bimodality in TNG100 can be seen between blue galaxies with stellar masses below $\approx 10^{10.5} M_\odot$ and red galaxies with stellar masses above. $\approx 10^{10.5}$. Because of the large sample sizes of TNG (in the order of 10^3 for TNG50 to 10^6 for TNG300 with stellar masses above $10^9 M_\odot$), various evolutional tracks can be observed spanning between the blue population towards the red population. This incorporates e.g. galaxies that rapidly quench towards the red population without any additional mass accumulation after that, galaxies that accumulate large amounts even after the star formation is quenched, and galaxies that undergo a second star formation era, probably fueled by an infalling gas-rich merger.

Figure 6 shows that IllustrisTNG produces realistic distributions in stellar mass compared to observational surveys in the local universe at $z \approx 0$ and even up to redshifts of $z = 4$ [134]. This said, all three TNG runs are needed to get a full understanding of galaxy evolution in the full mass and scale range. Extremely massive

galaxies with stellar masses above $10^{12}M_\odot$ or halo masses above 10^{14} are only sampled by the sufficiently large cosmological Volume of TNG300. On the other hand, only the increased resolution of TNG50 allows a throughout study of low-mass satellites below $10^8M_\odot$. By comparing the mass of stars to the overall halo mass (i.e. also the mass of the around a magnitude more massive dark matter component), the efficiency of halos to convert gas to their stellar inventory can be judged: this fraction peaks around a halo mass of $10^{12}M_\odot$ which then roughly translates to a stellar mass of $10^{11}M_\odot$. On the low mass end, feedback from star formation might lead to reduced efficiency in binding baryons to the DM halo. On the high mass end, AGN feedback is invoked reducing the star formation efficiency in the galaxy limiting ultimately the number of stars that can form in a galaxy [195] (Figure 8). According to Figure 7, the transition of quenching today at $z = 0$ is around stellar masses of $10^{10.5}M_\odot$ [56]. While the stellar inventory of low mass galaxies (stellar masses below $10^{10}M_\odot$) have almost exclusively formed in-situ (i.e. formed within the potential well of the respective halo), more massive galaxies' main contribution is ex-situ stars [134]. Stars that have not formed in the galaxy itself but are accreted by mergers. The fraction of accreted stars rises significantly with stellar mass with TNG300 galaxies at the higher mass end almost consisting of 90 per-cent exsitu stars. When looking at stars in the outskirts surrounding galaxies, the stellar halos, consistently large ex-situ fractions are observed across halo masses.

Lastly, the ex-situ fraction can be directly connected to the merger history of a galaxy. Recent and massive mergers have e.g. naturally a large contribution to raising the ex-situ fraction in a galaxy [147]. Examples of quantities describing the merger history of a galaxy and their relations with the ex-situ fraction are shown in Figure 9 and are also of interest to this thesis.

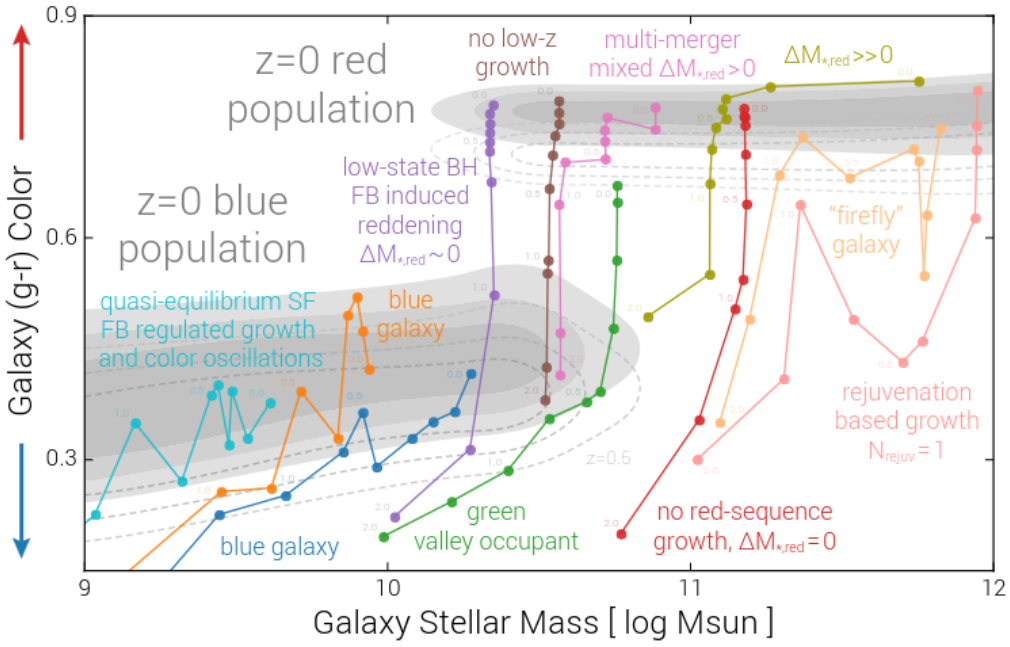


Figure 5: **Color Bi-modality in TNG and the diversity of evolution Tracks in TNG100.** Overview map of $z = 0$ galaxies in TNG100 in the color-stellar mass plane. The main bulk of galaxies are in 2 clusters: one blue/low mass cluster and one red/-massive cluster. Each line represents an example galaxy moving across the plane from $z = 2$ to $z = 0$. Although in general, galaxies move from the blue population to the red population, the evolutionary pathways of individual galaxies are very diverse. Plot and results from [123].

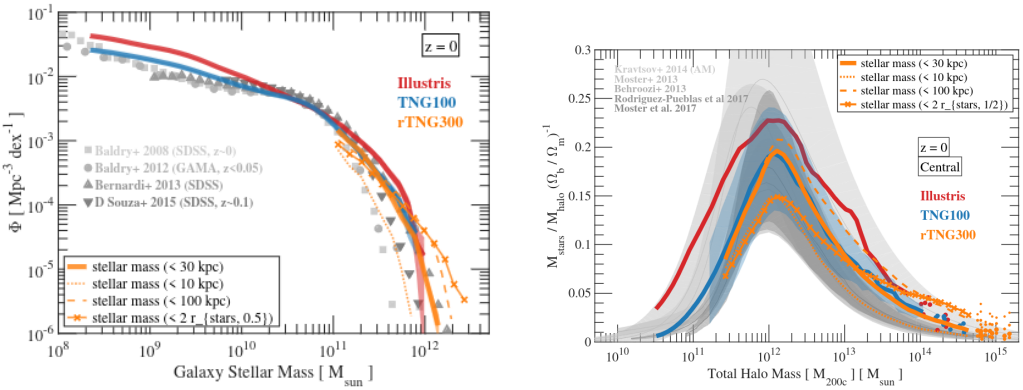


Figure 6: **Stellar Masses and Halo Masses in TNG100 and TNG300** Left: Number density of stellar masses in TNG100, TNG300 and Illustris at $z = 0$. TNG overall reproduces the trends in the number-densities observed (grey). Note how the availability of galaxies at a given stellar mass depends on the resolution and volume of a cosmological simulation: While the resolution of TNG300 is too low to resolve low-mass galaxies, the box of TNG100 (or, here not shown, TNG50) is too small to yield a significant number of galaxies with stellar masses above $10^{12}M_{\odot}$. Right: Stellar-mass to halo-mass (SMHM) relation. How many stars does a DM halo of a given mass contain? Interestingly, the relation between stellar mass and hallo mass is rather tight with a maximum at $10^{12}M_{\odot}$. Towards the lower mass end, Star Formation Feedback, and to the higher mass end, AGN Feedback is decreasing the ability of halos to form stars. Another interesting feature is the overall lower ability of the TNG model to produce stellar mass, compared to the original Illustris simulation lacking magnetic fields. Plots and results from [134].

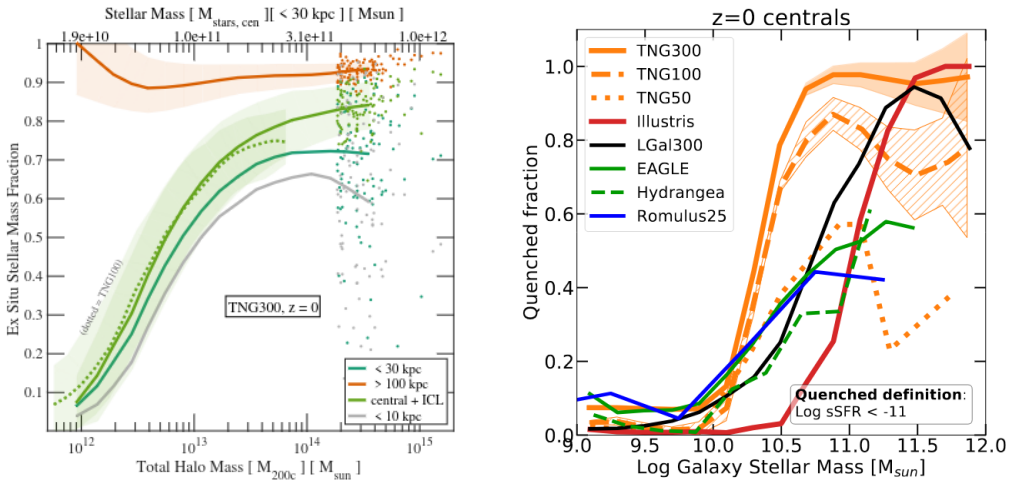


Figure 7: **Accretion of Stars vs Formation of Stars.** Left: Fraction of accreted, or ex-situ, stars in TNG300 at redshift $z = 0$ as a function of stellar mass or halo mass. Color-coded according to the aperture of the measurement. Ex-situ stars are all stellar particles that have not formed in the respective halo during their lifetime. Within 30kpc to the galaxy center, the fraction of ex-situ stars is strongly increasing with mass while the stellar halos outside of 100kpc are almost entirely made out of ex-situ stars. Right: Quenched Fraction at $z = 0$ as a function of stellar mass for TNG and several other cosmological simulations. Overall the fraction of quenched galaxies (in this case galaxies with a specific Star Formation Rate sSFR < -11) is increasing with stellar mass. The onset of quenching and the gradient with stellar mass differ. For TNG, quenching starts around stellar masses of $10^{10}M_{\odot}$ and the quenched fraction reaches almost 1 between $10^{10.5-11}M_{\odot}$. Beyond this point, galaxies' stellar inventory grows mostly due to merging. Plots and results from [134] and [56].

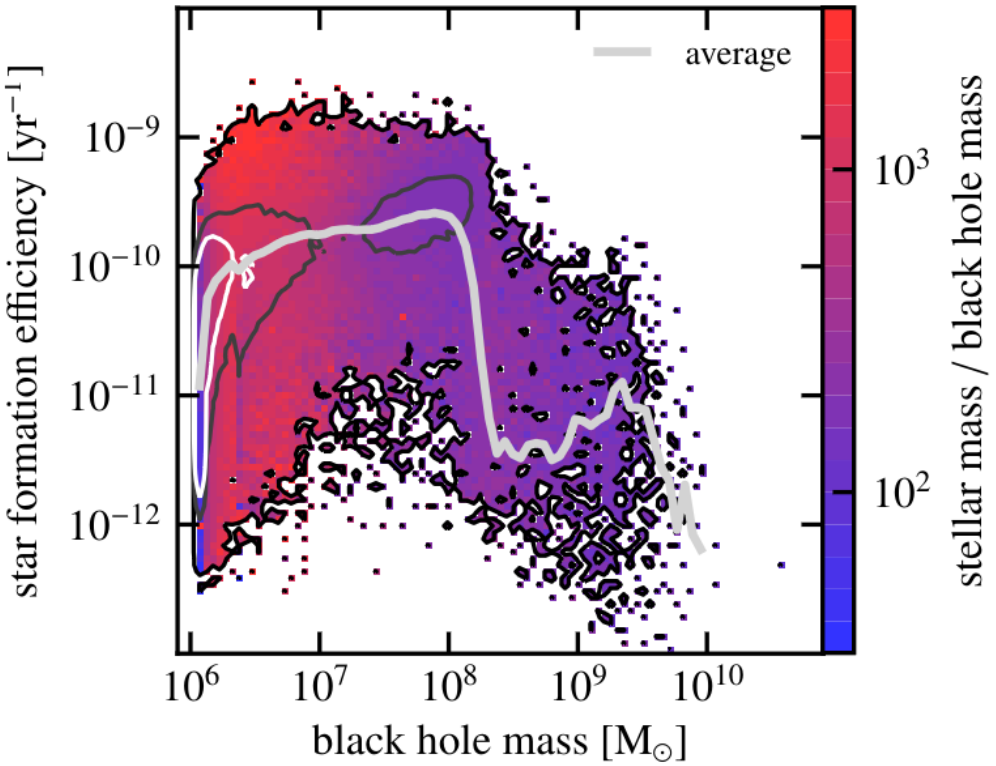


Figure 8: **AGN activity can interrupt star formation.** Star Formation efficiency (star formation rate divided by the gas mass) vs the black hole mass and color-coded by the ratio of stellar mass and black hole mass. At the onset of AGN Feedback, a steep decrease in star formation efficiency can be seen. Plots and results from [195].

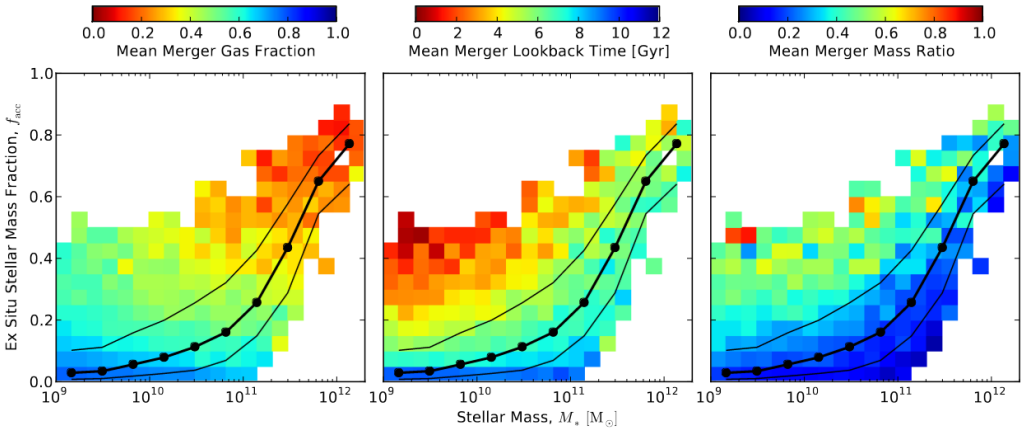


Figure 9: **The connection between the stellar ex-situ fraction with the merger history of a galaxy.** Ex-situ stellar fraction vs stellar mass for three merger summary statistics: mean merger gas fraction, mean merger lookback time, and mean merger mass ratio. Especially the mass ratio and the lookback time is correlating with the ex-situ fraction. Plots and results from [147].

5 CONNECTING SIMULATED AND OBSERVED GALAXIES

The results of simulations are just numbers on hard disks. To gain knowledge out of these numbers we have to put those results in contrast to observations: Are the simulations reflecting the reality and to what extent? And if not, how can we adjust the simulation model and/or the underlying theory to accommodate for the discrepancies?

For cosmological simulations, we deal with the problem that we do not simulate observed galaxies which we can compare one to one. But we have a (hopefully) large number of simulated galaxies and many observed galaxies and can compare them population-wise.

On a first order, we know that TNG is producing a realistic universe: for example observed two-point correlation functions [172] and the color-bi-modality of galaxies [123] are reproduced. In the following section, we take a deeper look into how those comparisons are performed, with a focus on comparisons performed on visual image mocks of stellar light with respect to large galaxy-imaging surveys.

5.1 OVERVIEW OF RECENT AND FUTURE GALAXY SURVEYS

To the public, the big telescopes are often the most fascinating: e.g. Very Large Telescope (VLT), Extremely Large Telescope (ELT), Thirty Meter Telescope (TMT), Giant Magellan Telescope (GMT), or the James Web Space Telescope (JWST). They all have in common a very high angular resolution power (10^{-3} arcseconds for VLT). An astronomical object can be studied in high detail with these kinds of telescopes which are achieved by large Apertures, Adaptive Optics, or by placing the instruments in space. This is not only of use for galactic and extragalactic astrophysics but also in other fields, especially Exoplanet research or the detailed study of nebulas.

For the field of galaxy evolution, especially when comparing them population-wise against full cosmological simulations, we are more interested in less detailed but deep images of many galaxies. These are delivered by large-scale galaxy surveys. Such survey telescopes are comparatively smaller, without the strict necessity for adaptive optics but a high focal ratio and Field of View (FOV). In the following, we briefly introduce a few datasets of utility for these terms.

5.1.0.1 *Recent galaxy surveys*

One of the largest galaxy surveys up to date is the Sloan Digital Sky Survey (SDSS) which started in the year of 2000 with a 2.5m telescope at the Apache Point Observatory. Over time the SDSS launched additional phases incorporating new instruments and telescopes. From the legacy survey, images for 930.000 galaxies are available [1] but are rather shallow compared to modern surveys (r band magnitude limit of 22.7). Of special interest for the topic of galaxy evolution is also the

MaNGA IFU survey which was part of the 4th phase of SDSS [30] and provided high-quality IFU maps for $\approx$ 10000 nearby galaxies. The current phase of SDSS is dominated by three new programs: Milky Way Mapper, Local Volume Mapper, and the Black Hole Mapper which will yield highly resolved stellar spectra in the order of 10^6 and galaxy spectra in the order of 10^5 [5]. Other examples for large galaxy surveys are: GAMA [58, 59], a multi-wavelength photometric and spectroscopic galaxy survey. The Dragonfly Survey [46], a deep ($1-\sigma$ of 31 $mag/arcsec^2$ in g-band) photometric (dwarf- and low surface brightness-) galaxy survey using comparatively small aperture camera lenses. Pan-STARRS [35], a synoptic imaging survey not only observing galaxies but especially also near-earth-objects (NEOS), variable stars, and supernovae.

5.1.0.2 *Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP)*

The survey of choice for this thesis is the Wide Layer survey of the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP [3] or short HSC). The HSC-SSP is a 330-night observing program utilizing the Subaru 8.2m telescope on Mauna Kea in Hawaii with the Wide Layer survey covering 1200 square degrees of sky with a median i-band seeing of 0.61 arcseconds and a 5σ point-source depth of 26.2 AB mag in five optical broad-band filters grizy. The combination of depth, angular resolution, and comparatively large survey area yields an ideal galaxy image dataset for the population-wise comparison to simulated galaxies.

5.1.0.3 *Other relevant upcoming galaxy surveys*

Upcoming surveys are for example the EUCLID Space Telescope which started in July 2023, the Roman Space Telescope which will serve as an Infrared Survey Counterpart to JWST, and the Legacy Survey of Space and Time (LSST) which will start in Early 2025 at the Rubin Observatory. Those surveys will reach unprecedented Depth and Field Sizes but also data volumes. LSST for example will reach a 5σ point-source depth of 26.9 AB mag covering large fractions of the southern sky (18,000 square degrees of sky) with the expected detection of $\approx 10^9$ galaxies [90]. LSST, EUCLID, and Roman will produce many TB per day and multiple PB across the duration of the surveys. This serves as another motivation for the increasing application of automated Machine Learning methods to process large amounts of image data.

5.2 OBSERVATIONAL MOCKS OF SIMULATED GALAXIES

When comparing the results of simulations with observed galaxies, the quantities that are compared must be in their definition as close to each other as possible. This itself is not a trivial task. E.g. when comparing the mass of observed and simulated galaxies, the observed mass can be derived by integrating the galaxy light in combination with a model of the stellar populations which then yields the combined mass. This however is affected by the limitation of the observers which might neglect mass which is very far out and below the detection limit of the respective survey. Comparing such a statistic therefore to the naively summed mass of simulated particles will result in a systematic error.

Because of this, it is advisable to bring the simulations in the same domain as the observations which are usually images, spectra, or IFU observations. Especially for the field of Machine Learning, which might be sensitive to subtle differences, such a mocking procedure must result in realistic mocks of the observations. This naturally includes unavoidable annoyances, that observers usually have to deal with, like the telescopes' PSF, photon noise, artifacts, and background and foreground objects.

The mocking is usually performed in the following steps: (1) deriving the emission of light from stellar and gas particles/cells. For this, a model is needed that maps the properties of the cells and particles to the Spectral Energy Distribution which is emitted by the resolution element. As stellar particles represent not single stars but populations of stars, the assumption of an Initial Mass function is needed which yields the actual number distribution of stellar masses for each stellar particle. Additionally, the light of young stars has to be corrected for dust that surrounds the birthplace of such stars. (2) Raytracing-based dust modeling can be applied to the emitted light. By this absorption, scattering, and reemission by dust are taken into account. The state-of-the-art algorithm to do so is called SKIRT [33] and is also utilized in this thesis. Lastly, it depends on the simulation model how the dust properties are derived: In the case of TNG, the dust is not explicitly modeled but has to be derived from the gas metallicity distribution. (3) The light can then be projected on a 2D image plane. To calculate the actual observed values of each pixel, the redshift, the filter transmission curve, and the survey resolution (arcsec per pixel) are needed. (4) Further survey-specific realism is added. This includes the convolution with the telescopes' Point Spread Function (PSF), insertion of photo noise, and the introduction of the resulting image into an observed background containing background sources.

With this, a galaxy originating from the computer can be translated into an image of a galaxy as it would be observed by a telescope. In this thesis, we will utilize images that have been achieved by such a mocking procedure from all galaxies contained in TNG50 and TNG100 above a certain mass limit. If one wants to mimic also a survey population-wise a so-called light cone is needed: this accounts for the fact that the volume of the local universe (small redshift) is much smaller than the volume of the distant universe (high redshift). The number distribution in redshift is therefore reproduced.

For TNG there are many mock data available. E.g. SDSS mocks [150] [86], MANGA IFU mocks [153] [118], LEGA-C/VIMOS mock spectra [197], JWST-CEERS mock images [42] and finally also the HSC mocks which have been co-developed for [26] and [61] (i.e. this thesis). Examples of those mocks are provided in Figure 10.

5.3 EXISTING COMPARISONS OF SIMULATED AND OBSERVED GALAXY POPULATIONS

There are many examples of how mocks of simulated galaxies have been used to extract scalar properties and summary statistics using methods similar to those employed by observers: Morphology parameters extracted from low redshift galaxy images, such as the Gini-M20 for Illustris [164], compared against HSC galaxies, and also IllustrisTNG [151], with a comparison against Pan-STARRS observations.

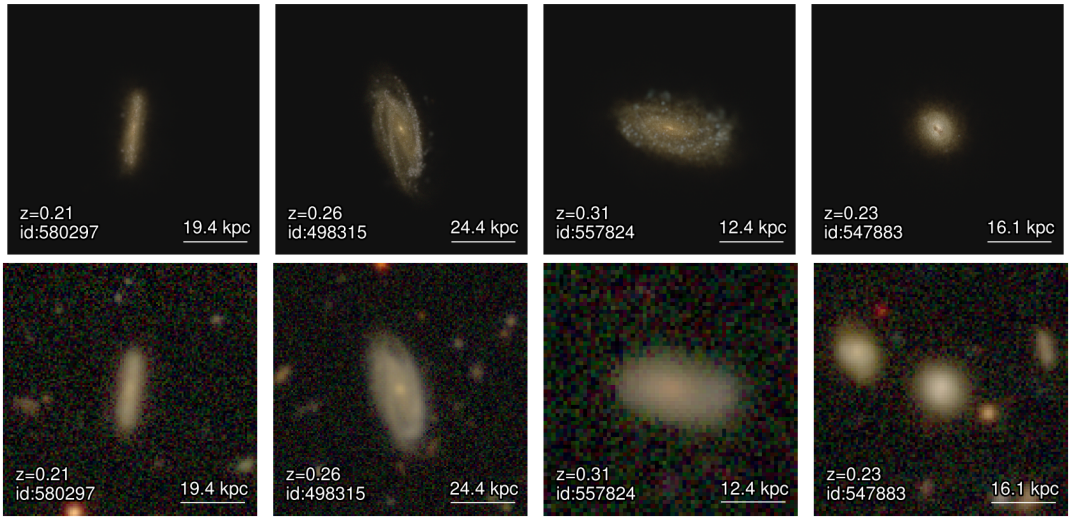


Figure 10: **Example for Idealized and Realistic HSC Mocks** At the top we show idealized HSC mocks of TNG50 images without observational realism. At the bottom, we show the same images with full observational realism as they are used in [26] and [61]. Images adapted from [176].

Comparison of TNG100 galaxies' stellar mass surface density profiles to Dragonfly Nearby Galaxies Survey galaxies [110]. Comparison of dark matter distribution based on shear measurements of TNG100/300 and observed DECaLS, SDSS, and HSC galaxies [6].

Despite these successes, the conventional approach to comparing simulated and observed galaxies remains constrained by the examination of a limited set of scalar properties or summary statistics. This approach risks overlooking crucial information content or cross-correlations present in the data. Moreover, comparing non-self-consistently derived properties, such as kinematically derived morphologies from simulation data versus photometry-derived morphologies of observed galaxies, further complicates the analysis.

To address these limitations, a direct comparison between simulations and observations at the map level is of advantage. This entails contrasting full images of real galaxies with survey-realistic mocks of simulated galaxies. Such an approach ensures the utilization of all available information contained in the images while bypassing the challenging task of selecting and constructing summary statistics for comparison.

Especially machine learning methods are well suited to tackle the complex task of evaluating and comparing simulated and observed images on an image-to-image level: One of the first studies comparing TNG galaxy mocks to observed SDSS galaxies used galaxy morphology classifications via machine learning [85]: overall the classifications found a good match between simulated and observed galaxies but interestingly the high mass end in TNG is dominated by massive disk galaxies while elliptical galaxies dominate in the observations. [199] was able to identify and quantify the improvement from Illustris to TNG50 galaxy mocks but also found that 'fine details of galaxy structure are still different between observed and simulated galaxies' of small/spheroidal/quenched galaxies speculating that these differences are caused by numerical resolution.

[131] classified visual galaxy merger between EAGLE mocks and SDSS, especially encountering problems that networks that are trained on one of both samples are performing badly on the other. Overall the field of galaxy merger identification is suffering from a degradation of performance when moving from the training set to a test set which is sampled from another simulation or observations [104]. Such a –domain shift– is always present if the underlying distributions of the dataset are not the same: either because the galaxy images are not exactly the same or the network is focusing on domain-specific features on the images (i.e. the domains are not overlapping) or also because the distributions are not matching: e.g. observations contain without corrections much fewer low mass galaxies than simulations (i.e. domains are overlapping but the galaxy images have a different distribution within). These domain shifts are not only present between simulated mocks and observed galaxies but also between cosmological simulations [185].

Therefore an important task in this thesis is to bridge the domains: the difficulty is to make the network insensitive to minor differences between the datasets but not to a degree that the comparison becomes pointless (a-priori we know that there are differences). Therefore bridging the domains means also to identify subsets of the overall data used to train a model. E.g. by limiting aperture and input quantities, domain convergence could be reached for the ex-situ determination between TNG and EAGLE [7].

Of particular interest for this thesis is the class of 'contrastive learning' methods, which are self-supervised algorithms capable of identifying features of interest from images without prior labeling of galaxies [84]. By doing so, these methods circumvent the need for manual identification and calculation of labels and marginalizing over small differences in domain. This will be heavily utilized in this thesis and deeper explained in the next Section.

6 MACHINE LEARNING

The talents of machines and humans are different from each other. While the machine is good at arithmetic, which is a reason why cosmological simulations are run by HPC and not by humans, it has much greater difficulties with tasks that are simple for us humans. One is for example the classification of galaxies based on images. Because of this, citizen-since projects like Galaxy Zoo [108] have been run with great success. Current and future surveys and simulations as well produce more and more data which makes manual processing by humans increasingly unsuitable. Furthermore, the complexity a human can comprehend is limited: while a simple classification based on a 2D image might be feasible, a quantitative recursion based on a high-dimensional dataset will be tough to handle. Because of this scientists are eager to teach machines to work like humans. In the following, we will have a short introduction to Machine Learning, and then focus on the two main methods used in this thesis: Contrastive Learning and Simulation-Based-Inference via conditional Invertible Neural Networks.

6.1 SHORT INTRODUCTION TO MACHINE LEARNING AND DEEP LEARNING

The Term Machine Learning (ML) is today often directly identified with the field of Neural Networks. However, it can be applied to each algorithm (i.e. the machine) which can adapt or learn from data and then apply to previously unseen data. A very traditional ML method is for example the Principal component analysis (PCA), which has been already proposed in 1901 [130] way before there was a machine to run it.

Usually, ML methods are grouped into two main categories: Supervised and Unsupervised methods. Supervised methods work on labeled data which serve as a ground truth that can be trained on and then transferred to previously unseen data. An unsupervised method on the other side works without such labels but learns automatically on the structure of the given data. The clustering of data points is e.g. an unsupervised method or the previously mentioned PCA.

In this thesis, we will use both classes: an unsupervised method (Contrastive Learning) which will be able to learn galaxy morphologies just by looking at the images but without getting any additional information provided. Building on this we then use a supervised method (Simulation-Based-Inference) which relies on the merger and assembly history ground truth provided by the TNG simulations to learn quantities of interest from the mythologies/galaxy images.

For a simple task, one should not underestimate the power of non-DL methods like Random Forest, Support Vector Machine, etc. The task at hand however is to construct arbitrarily shaped posteriors related to the merger and assembly history of a galaxy to a visual multi-band image. To do so, we use powerful methods out of the field of Deep Learning (DL) which use Deep Neural Networks as key building blocks: Simply phrased such neural networks consist out of layers

of linear functions that use the output of previous layers as input. Many layers (i.e. 'deep' networks) of such linear functions have been proved to be able to represent arbitrary functions between the input of the first layer and the output(s) of the last layer (Universal approximation theorem, for example, [165]). The function itself can be learned by the minimization of a predefined loss function. Example: A supervised problem to infer a variable y from the input x. The network can be described as a function $f(x, w) = y$ with the parameters of the linear functions w. To learn this function we can no define the Mean Squared Error as the loss function $L_w = \sqrt{\Sigma(y_{i,true} - f(x_i, w))^2}$ with a batch of data points i. The correct function f will trivially minimize the Error. Starting now from a random set of parameters w and therefore function f we will get by a high chance a not so optimal f. But as f consists of a chain of linear functions we can calculate the derivative of the loss function against the parameters dL/dw. With this, we can –back-propagate– the error through the network adjusting the parameters of the model such that the loss function gets minimized. This minimization is usually an iterative process, only following the gradient dL/dw according to a pre-defined learning rate and then recalculating dL/dw at the new position in the parameters space.

The differences in models there are in Deep Learning are mainly in the exact architecture of those layers of the network and the Loss Functions used in the training. The architectures we use in this Thesis are:

- Fully Connected Networks/Layers or Multi-Layer Perceptron (MLP): Simple Linear Layers that are connected to all previous and following layers. Width (number of nodes per layer) and Depth (number of layers) are the relevant properties describing the architecture. Usually combined with an activation function after the linear function which nullifies negative values.

- Convolution Neural Networks (CNN): As 2D image data is high dimensional and would lead to computational problems with fully connected layers (Each pixel is connected to all nodes in a layer) CNN uses fixed-size kernels in each layer which work on an image like a convolution (hence the name). A 3x3 kernel e.g. has 3x3 nodes with parameters, each layer can have multiple kernels that are sensitive to different image features. Chaining multiple such conventional layers can iteratively learn more complex features: while the first layer might be only sensitive to edges, the last will be able to identify the galaxy. Relevant properties for the CNN are the kernel size, the number of kernels per layer, and the number of layers (depth).

- Residual neural networks (ResNet) [82]: Deep CNNs have in theory the advantage of learning very complex features from the input data. However training them is a not simple task and might ultimately lead to worse results (This is also, but not only, connected to the so-called Vanishing gradient problem, i.e. dL/dw decays in very deep networks making training slow or even impossible). Because of this, ResNets have taken a dominating role in image recognition. Simply speaking, CNN layers are grouped into Residual Blocks where the input of the Block is added to the output. Such a skip connection leads to an identity mapping of the block: the residual block is not learning the transformation but the perturbation with respect to the input.

- E(2) Equivariant CNNs [193]: Standard CNNs are equivariant regarding translations, i.e. moving the position of an object in an input image will result in a translation of the output 'image'. The orientation of galaxies in the image plane on the other hand should be irrelevant to the information content (assuming the cosmological principle). E(2) equivariant convolutional layers are such a modification to the kernels of a standard CNN to ensure rotation equivariance. A rotation of an object will therefore lead to a rotation in the output. We use this for CNNs and the ResNets as well.

- Normalizing Flows and conditional Invertible Neural Networks (cINN): Normalizing flows are a general class of differentiable networks that can map arbitrary distribution to a Multivariate Normal Function. We will later use this to sample arbitrarily shaped posterior distribution. cINNs are one method to construct such a normalizing flow and we will describe those in more detail in Section 6.3.

Note that this is just a very basic overview, details are in the respective papers. In the next 2 chapters, we describe in detail the methods of training used in this thesis.

6.2 REPRESENTATION LEARNING AND CONTRASTIVE LEARNING

For machines, an image is a complex input. Even a small image with 256 pixels per side and 3 channels already spans a $3 \times 256 \times 256 = 196608$ dimensional space. Most of the points in that 196608 dimensional space relate to images that just consist of 'noise' or to images we are not interested in for the application at hand. The number of relevant images, galaxy images in this thesis, therefore only span a small but complex subspace in those 196608 dimensions.

In representation learning, the aim is to transform the high-dimensional raw image data, into a more compressed format that only captures meaningful features for the given task. This process involves mapping the input data into a representation space where similar images are clustered together, easing downstream tasks like classification or inference.

The ResNet architecture is one approach to mapping images into a representation space. However, the challenge lies in ensuring that these representations effectively encode the relevant features of the input images. To address this, self-supervised methods such as 'Nearest Neighbour Contrastive Learning' (NNCLR) have been developed [60].

NNCLR is inspired by SimCLR [36] which trains the ResNet to learn representations by contrasting transformed versions of the same image with those of other images. The transformations applied include reflections, rotations, zooms, affine translations, Gaussian blur, and Gaussian noise. These transformations encourage the model to focus on essential features while ignoring irrelevant variations.

By training the ResNet with NNCLR, the network automatically learns to distinguish between different types of galaxies by grouping similar ones together in the representation space. However, NNCLR takes an additional step beyond SimCLR. Instead of solely contrasting transformed images, NNCLR searches for the nearest neighbor of a transformed image within the representation space and uses it as the

transformed 'same' galaxy. This additional step enhances the model's generalization ability by interpolating between training galaxies.

To make the nearest neighbor search computationally feasible, a queue-based approach is employed. At each training step, the latest batch of input images and their representations are added to the queue, while the oldest batch is removed. This maintains a manageable set of representative images for matching nearest neighbors.

Moreover, to facilitate contrastive learning, a 'Projection Network' is incorporated. This network, consisting of fully connected layers, projects the representation space into a lower-dimensional space where the contrastive loss is calculated and the network is trained. However, during testing or application, this projection network is removed, and only the learned representation space is utilized.

For completeness it is mentioned that there are also other methods in the field of self-supervised image representation learning, e.g. Variational Auto Encoder [183] or other contrastive learning methods like BYOL [77].

6.3 SIMULATION-BASED-INFERENCE VIA CONDITIONAL INVERTIBLE NEURAL NETWORKS (CINN)

What is the probability distribution of a galaxy with given observable features c to have unobservable properties x? This probability is described by the conditional probability distribution or posterior distribution $p(x|c)$. In Bayesian statistics, the posterior distribution can be written as

$$p(x|c) = \frac{p(c|x)}{p(c)} p(x)$$

with $p(c)$ as the evidence or marginal distribution, i.e. the distribution in the space of the observables. The prior $p(x)$ is the a-priori known distribution in unobservable properties without having any information about the appearance of the galaxy. The likelihood $p(c|x)$ describes the probability that a galaxy with given unobservable properties x has observable features c. For simple analytical models, one can directly calculate the likelihood and, based on this, receive the posterior distribution. For complex scenarios, like the connection between galaxy morphologies and the merger history, this likelihood is not traceable. Because of this, we utilize methods out of the field of likelihood-free inference or simulation-based inference. By this, we train a machine learning method to directly model the posterior $p(x|c)$ based on data made available from full cosmological simulations. It is important to emphasize, that posteriors learned by such a method are always constructed relative to the prior and marginal distribution of the underlying training data.

Traditional regression models, such as a Multilayer Perceptron (MLP), can only perform a regression to the statistical moments of the posterior distribution. When training with a Mean Square Loss this e.g. corresponds to the mean of the distribution. This is only feasible if the connection between galaxy features and unobservables is bijective or if the uncertainty in unobservable inference from the observables is explainable by simple distributions like normal distributions.

In reality, galaxies with very similar observable parameters can have distinct assembly and merger histories. On the other hand, the same galaxy might look very

different from different sides. Consequently, for a full understanding, we should not only be interested in point predictions but rather the characterization of the posterior distribution $p(x|c)$—the conditional probability of merger statistics x given galaxy observables/images c.

Conditional Invertible Neural Networks (cINNs) are a tool for addressing the complexities inherent in modeling conditioned arbitrarily shaped posteriors [10]. They are building upon Invertible Neural Networks (INNs) [9] and the invertible architecture proposed by [53] within those and constitute a type of normalizing flow.

This normalizing flow architecture is a differentiable and invertible function: $f(x, c) = z$ with x as the merger/assembly statistics of interest, c observable quantities or images, and z as the latent variable. The trick is now to train this invertible function such that z constitutes a normal distribution (hence normalizing flow). By using the invertibility of f the original posterior distribution, learned from the simulation can then be sampled from the well-known normal distribution $f^{-1}(z, c) = x$.

By training the network with a negative log-likelihood loss it is achieved that the latent variables z indeed approximate a multivariate Gaussian. Mathematically, the negative log-likelihood loss is defined as:

$$L_{NLL} = \left| \frac{|z|^2}{2} - \log(|J|) \right|,$$

where J denotes the Jacobian of the invertible function f.

For the implementation of cINNs we use the FrEIA framework [11], a toolkit built via PyTorch [129]. Notably, cINNs have already demonstrated usability in analogous astronomical problems [e.g., 99], underscoring their potential in modern astrophysical data analysis.

Part III

INFERRING THE ASSEMBLY HISTORIES OF ILLUSTRISTNG GALAXIES FROM INTEGRAL OBSERVABLE PROPERTIES VIA INVERTIBLE NEURAL NETWORKS

The following chapter is based on the work 'ERGO-ML I: Inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks' which was published in MNRAS [63] as part of the PhD. A fundamental prediction of the LambdaCDM cosmology is the hierarchical build-up of structure and therefore the successive merging of galaxies into more massive ones. As one can only observe galaxies at one specific time in cosmic history, this merger history remains in principle unobservable. By using the TNG100 simulation of the IllustrisTNG project, we show that it is possible to infer the unobservable stellar assembly and merger history of central galaxies from their observable properties by using machine learning techniques. In particular, in this first paper of ERGO-ML (Extracting Reality from Galaxy Observables with Machine Learning), we choose a set of 7 observable integral properties of galaxies (i.e. total stellar mass, redshift, color, stellar size, morphology, metallicity, and age) to infer, from those, the stellar ex-situ fraction, the average merger lookback times and mass ratios, and the lookback time and stellar mass of the last major merger. To do so, we use and compare a Multilayer Perceptron Neural Network and a conditional Invertible Neural Network (cINN): thanks to the latter we are also able to infer the posterior distribution for these parameters and hence estimate the uncertainties in the predictions. We find that the stellar ex-situ fraction and the time of the last major merger are well determined by the selected set of observables, that the mass-weighted merger mass ratio is unconstrained, and that, beyond stellar mass, stellar morphology and stellar age are the most informative properties. Finally, we show that the cINN recovers the remaining unexplained scatter and secondary cross-correlations. Our tools can be applied to large galaxy surveys in order to infer unobservable properties of galaxies' past, enabling empirical studies of galaxy evolution enriched by cosmological simulations.

7

INTRODUCTION

One of the main predictions of the ΛCDM paradigm is the hierarchical formation of structure whereby galaxies merge into progressively more massive systems [e.g. 72, 100, 166].

This hierarchical growth is a fundamental physical principle for the evolution of galaxies, both within populations and for individual objects. For example, it determines that more massive and luminous galaxies, which are thought to reside in more massive dark matter haloes [192], are rarer than less massive and luminous ones [19, 157]. Moreover, already a few decades ago, it had been suggested that the (dry) mergers of massive disc galaxies produce galaxies with elliptical morphologies [e.g. 144], i.e. with most of their stars in hot orbits, such that massive elliptical galaxies have often been associated to a recent violent merger history.

Over the last years, cosmological simulations have been able to quantitatively follow the evolution and interactions of representative populations of dark-matter haloes [e.g. 27, 89, 142] and galaxies [124, 156, 187] within ever-increasing volumes. They have hence demonstrated that mergers between galaxies are more frequent at higher redshifts [66, 146] and that the number of massive mergers is a steep function of galaxy and halo mass [146]. Moreover, it has been shown, via cosmological hydrodynamical galaxy simulations, that at high stellar masses ($> 10^{10-11}\,M_\odot$) mergers become an important source for the stellar inventory of a galaxy in addition to in-situ star formation. As a result, massive galaxies are made of larger fractions of stellar material that is accreted from mergers and orbiting satellites [e.g. 134, 147].

Astronomical observations provide snapshots of the state of galaxies at distinct points in time, such that their actual merger and accretion histories remain in principle unobservable. Whereas it is not uncommon to observationally capture galaxies in the act of merging with one another [80] and whereas recent mergers can leave clear signatures in the stellar mass maps of observed galaxies, via e.g. stellar shells [182], tidal tails [107], and disturbed stellar morphologies [64], the inference of more ancient events and interactions is substantially more challenging.

Recent developments in both theory and observations have enabled progress in this direction, namely in inferring aspects of the past histories of galaxies given their properties at the time of inspection. For example, the combination of models with 6D + metallicity data for millions of individual stars from the *Gaia* mission [69] and other surveys [e.g. LAMOST, GALAH, and APOGEE: 51, 52, 103] have allowed the reconstruction of discrete merger events of the assembly history of our own Galaxy [17, 83]. It is now believed that the Milky Way has undergone its last major merger sometime at $z = 1 - 2$ [20], even if the merging galaxy, dubbed Gaia-Sausage-Enceladus and with estimated $5 - 6 \times 10^8 M_\odot$ in stars [114], has long been destroyed.

In a similar fashion, but going beyond the Local Volume, [201] and Zhu et al. (2022, submitted) have uncovered that two early-type galaxies in the Fornax cluster

each underwent a merger as early as 8 and 10 billion years ago, respectively, and quantified their stellar mass and merger time. To achieve this, they have combined population-orbit super-position methods [202], the results of the cosmological simulation TNG50 [121, 136], and the observed luminosity density, stellar kinematic, age and metallicity maps from the MUSE IFU instrument on the VLT [154].

Although spectacular, these results are destined to be limited to small samples of galaxies in the nearby Universe, mostly because of the scope of the required observational techniques. In this paper, we investigate whether, and to what degree, less-costly observational galaxy data can be queried to successfully infer aspects of the past stellar assembly and merger history of large samples galaxies. In particular, we aim to see whether scalar features that summarize galaxy integral properties, such as galaxy stellar masses, galaxy colors or labels of stellar morphology, possess sufficient information content to determine, e.g., the amount of accreted stellar mass or the mass and time of the last major merger. This possibility would open up a powerful new avenue to quantify the merger history of thousands of observed galaxies across cosmic epochs, because it would be achievable with data from large photometric surveys such as SDSS[1], GAMA[2], DES[3] and HSC-SSP[4], but also future ones such as LSST[5].

To do so, we use data from the IllustrisTNG project[6], which provides a series of cosmological, gravity + magnetohydrodynamics simulations of galaxies in representative portions of synthetic universes [124]. As such simulations predict the physical state of thousands of individual and realistic galaxies across cosmic time, they allow us to understand and quantify how unobservable dynamical processes of galaxy evolution, including the merger history, influence the observable appearance of galaxies. The "by hand" investigation of these connections has been addressed in the past [e.g. 147, 149], also using as observational inputs the properties of the stellar haloes [41, 110, 133, 140]. Here we take a leap forward and utilize modern machine learning (ML) methods, training from the IllustrisTNG galaxies to derive a continuous mapping function between their observable and their unobservable properties.

This is the first paper of a wider project (ERGO-ML) where we aim at Extracting Reality from Galaxy Observables with Machine Learning, across a wide range of physical properties of galaxies and utilizing diverse sets of observations, e.g. from photometry to spatially-resolved spectroscopy, from stellar light to gaseous signatures, from the innermost regions of galaxies to their dim stellar and gaseous haloes, from scalar features to maps and multi-dimensional data cubes. We will use and combine state-of-the-art cosmological simulations of galaxies to test methods for direct application to observational data in order to "train on simulations and apply to observations". All this is now possible thanks to the breadth of scope of current cosmological galaxy simulations [189], the large numbers of the galaxies simulated therein [119, 124] and their increasing realism, which since a few years has also being quantified with ML methods [e.g. 86, 200].

1 Sloan Digital Sky Survey, https://www.sdss.org/
2 Galaxy And Mass Assembly, http://www.gama-survey.org/
3 Dark Energy Survey, https://www.darkenergysurvey.org/
4 Hyper Suprime-Cam Subaru Strategic Program, https://hsc.mtk.nao.ac.jp/ssp/
5 Large Synoptic Survey Telescope/Vera C. Rubin Observatory, https://www.lsst.org/
6 www.tng-project.org

The potential of combining ML and galaxy (simulation) data for similar scientific goals has been shown over the past couple of years in a wide array of works. A large fraction of these have focused on determining the properties of the underlying dark-matter haloes [e.g. 50, 190] and on identifying merging galaxies or merger remnants from images [e.g. 22, 23, 38, 67]. Very recently, [159] have also used data from IllustrisTNG to determine, via a Random Forest, the accreted stellar mass fraction.

In this paper, we build further by focusing on summary statistics of the entire merger and assembly history of galaxies. We also go beyond the methodologies adopted thus far in the field by obtaining not only point predictions, but also by quantifying the full posterior distributions, and hence uncertainties, of the model predictions.

Our paper is organized as follows. In Section 8, we provide the details for the rationale of this work and for the galaxy simulations and the ML algorithms adopted throughout. The architecture and training of the latter are described in Section 9. We show that the unobservable past history of galaxies can be decoded from a handful of galaxy features in Section 10, where we quantify the results from a Conditional Invertible Neural Network. We comment on the technical aspects of the inference and on the scientific implications of our findings in Section 11 and conclude and summarize in Section 12.

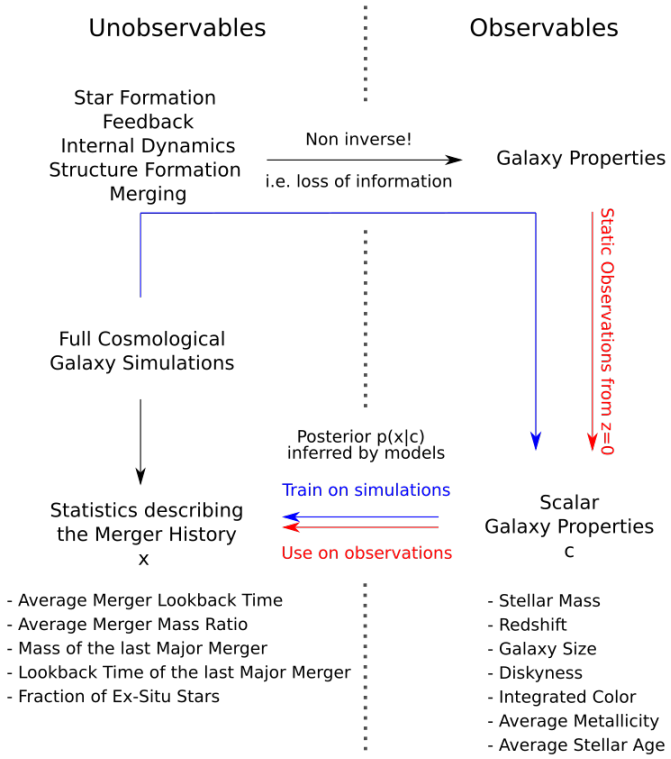


Figure 11: **Visual scheme of the setup and goals of this paper, ERGO-ML I.** We can predict the observable properties of thousands of galaxies given their individual histories and evolution using a cosmological hydrodynamical galaxy simulation, here TNG100. The goal of this paper is to model the inverse of this, i.e. the mapping from the observables to the unobservable stellar mass assembly and merger history (see also Table 1). Because of the possible loss of information in the forward step – that is, the observables are not sufficient to fully constrain the history –, we are especially looking for the posterior distribution, which is the probability that a certain merger history x is compatible with a set of observations c. In this work, we focus on utilizing scalar integral properties, i.e. galaxy features, and test this scheme on simulated galaxies. However, in subsequent work, we will go beyond features and use maps of galaxies and of their stellar haloes as inputs, and we will also apply the models developed here to actual observational data.

8

SETUP, SIMULATION DATA AND METHODS

8.1 RATIONALE

We aim to uncover and investigate the connections between observable (input or feature) and unobservable (output or desiderata or target) properties of galaxies, with a focus on their unobservable past merger history. To do so, we use data from the IllustrisTNG simulations to train a machine learning algorithm: a Conditional Invertible Neural Network (cINN).

The known parameters are 7 galaxy integral galaxy properties (i.e. features) that describe the state of the galaxy at a certain point in time. We focus on galaxy properties that are in principle directly observable or derivable from observations, namely: galaxy stellar mass, galaxy redshift i.e. lookback time, stellar half-light radius, galaxy stellar morphology (i.e. the discyness, measured via the disc-to-total mass ratio), integrated galaxy color, average stellar metallicity, average stellar age.

The output target quantities we elect to investigate are 5 summary statistics of the galaxies' assembly and merger histories that are directly predicted by the cosmological simulations and are extractable from the simulation output data. These are the total stellar ex-situ mass fraction, the lookback time and stellar mass of the last merger with a stellar mass ratio $\geqslant \frac{1}{4}$, the mass-weighted average merger mass ratio, and the mass-weighted average merger lookback time.

The rationale of our study is summarized in Fig. 11 and a succinct overview of the observable and unobservable galaxy properties we focus on is in Table 1. In the following, we briefly describe the simulation data we use to uncover the connections between observables and past history of galaxies, the galaxy properties, and the functioning of the cINN neural network and the way we use it.

8.2 THE ILLUSTRISTNG SIMULATIONS FOR GALAXY FORMATION AND EVOLUTION

We utilize the publicly-available results [124] of the large-scale cosmological magneto-hydrodynamical simulation project IllustrisTNG [TNG hereafter, 106, 116, 121, 122, 134, 136, 171]. These consist of multiple simulation runs for galaxy formation and evolution based on the moving mesh code AREPO [167] and cover a broad range of resolutions, simulation volumes and additional dark matter only runs to trace the influence of baryons on the cosmological evolution.

The full-physics runs of TNG follow the evolution of not only dark-matter particles but also of gas cells, and stars and black holes particles (collectively referred to as resolution elements), from redshift $z = 127$ to $z = 0$. The initial conditions defined in periodic-boundary cubic domains are given by the Zeldovich approximation, via the N-GenIC code [170]. The underlying cosmology is given by Planck [137].

The TNG simulations are hereby the direct successor of the Illustris project [187], but introduce magneto-hydrodynamics and several updates to the physical model [see details in the method papers: 135, 194]. In this work, we will use the outcome of the run called TNG100, as further elaborated in Section 8.4.

8.3 GALAXY IDENTIFICATION AND PAST HISTORY

The TNG simulations return tens of thousands of well-resolved galaxies across 14 billion years of cosmic evolution, for which integral properties can be measured from the constituents resolution elements. To contain the amount of saved data to a manageable size, the resolution elements for each of the simulated ingredients have been stored only at 100 equally distributed points in (simulated) cosmic time, called snapshots. We have at our disposal 50 snapshots at $z \leqslant 1$ and 50 at $z > 1$ [see 124, for more details].

Within the simulation output, galaxies – i.e. sets of resolution elements that together constitute individual galaxies – are identified by the FRIEND-OF-FRIENDS [FoF, 55] and SUBFIND [169] algorithms. Based on these two methods, a hierarchical set of two catalogues are available:

1. The SUBFIND catalogue, which lists the subhaloes within each snapshot and provides a number of integral statistics for each object, such as e.g. the total mass of each subhalo for each particle type.

2. The FoF catalogue, which groups these subhaloes into FoF haloes.

Typically, the most massive subhalo in a FoF halo is called *central* or primary subhalo, while all other subhaloes in that group are *satellites*. Any subhalo with non-vanishing stellar mass is a galaxy, whether central or satellite.

As we want to investigate the formation history of the galaxies, we need the connection of identified subhaloes along their cosmological evolution. In this paper, we hence utilize merger trees created by SUBLINK [146]. This method constructs a tree-like structure out of the previously identified subhaloes, such that for each root subhalo there is a link to one or more progenitors in the previous snapshot. If there is more than one progenitor, the progenitor along the most massive branch is chosen as the main or first progenitor. Throughout this work, the time evolution of the galaxies of interest is given by such main progenitor branches, whereas the secondary progenitors constitute *merging galaxies*, in short: mergers.

In particular, in this paper we adopt the so-called SUBLINKGAL code, a version of SUBLINK whereby star particles and star-forming gas cells – instead of dark-matter particles – are tracked in time to construct the trees of galaxies. In comparison to the default functioning of the SUBLINKGAL code, we apply additional measures to obtain a cleaner catalog of merger events. Namely, we consider only galaxies with cosmological origin[1], with at least 50 stellar particles (galaxy stellar mass $\gtrsim 5 \times 10^6 M_\odot$), and whose main progenitor history is recorded for at least three snapshots.

1 Namely, throughout this analysis, we exclude so-called spurious subhaloes, i.e. gravitationally-bound sets of stars that formed out of gas that in turn fragmented in already-established, parent galaxies. To do so, we impose SubhaloFlag $\equiv 1$ [124].

8.4 GALAXY SAMPLE

The TNG suite consists of three flagship simulations of different resolutions and volumes: TNG50, TNG100 and TNG300. These evolve cubic volumes of approximately 50, 100, and 300 comoving Mpc a side, with the best resolution achieved in the smaller-volume TNG50 [121, 136]. However, a larger volume allows for more simulated galaxies, importantly especially at the high-mass end. With about ten thousand galaxies above $10^9\,M_\odot$ at every snapshot, TNG100 provides a galaxy sample that is sufficient for our ML approach and is therefore our choice for the following study.

From TNG100, we consider only central galaxies (i.e. the first galaxy within a FoF group and hence the galaxy at the deepest minimum of the FoF potential) with cosmological origin [i.e. non-spurious subhaloes as per 124], and in the $z = 0-1$ range and with total stellar masses between 10^{10} and $10^{12}\,M_\odot$. We choose this limit as the highest mass end is not well sampled and would require a dedicated work.

The lower limit in stellar mass is set for a number of reasons:

- to ensure that the galaxy properties are well captured not only at the time of inspection, with about $\gtrsim 10^4$ stellar particle per galaxy, but also at earlier times when the inspected galaxies and their potential merging companions have lower mass and are hence less resolved;
- to focus on a mass regime with a richer merger history, given that the merger rates of galaxies decline towards low masses [146];
- to focus on a mass regime that is accessible by current large photometric surveys of galaxies, such as SDSS, that are complete down to $\gtrsim 5-7\times10^9\,M_\odot$.

With 49 discrete snapshots between $z = 0$ and $z = 1$, our TNG100 sample includes 182'625 galaxies with total stellar mass in the $10^{10-12}\,M_\odot$ range.

8.5 GALAXY OBSERVABLES AND UNOBSERVABLES

We characterize the simulated galaxies by the fundamental properties listed and described in the top panel of Table 1. These include the galaxy stellar mass and the redshift, as from previous studies [e.g. 146, 149], we know that the merger activity of galaxies is highly dependent on these (see Introduction).

Additionally, we also use as observable inputs information on the galaxy stellar extent and structure, such as the size of the galaxies and their morphology (i.e. the stellar disk-to-total mass ratio). While the stellar half-light radius gives information about the spatial distribution of the stars, the fraction of disc stars gives us a possibility to differentiate between disc-like and elliptical galaxies. Disc galaxies are thought to be mainly build by in-situ star formation, whereas a large fraction of non-circular orbits could be an indicator for past merger events. In observations, similar morphological estimators can be obtained via photometric morphological decompositions or by using integral field unit spectroscopy [e.g. ManGA, 29].

Finally, as further information that can be gained via photometric observations of galaxies, we also characterize the simulated galaxies by their integrated galaxy

Galaxy Observables

Name	Definition	Note
Stellar Mass	Stellar mass within twice the radius that contains half of the total galaxy mass i.e. within two times the stellar half-mass radius.	With this mass definition, which we use throughout, the minimum and maximum mass of our sample fall slightly below the nominal limits of 10^{10} and $10^{12}\,M_\odot$ used for the sample selection.
Lookback Time	Time at which the galaxy is considered.	For plotting and when utilizing ML methods, we use the lookback time rather than redshift, as it represents a linear measure of time and the TNG snapshots are also approximately uniformly distributed in time.
Half Light Radius	Stellar half-light radius in the SDSS-r band.	This radius is measured relative to the center of the galaxy, which is defined as the position of the resolution element with the lowest potential energy in the respective subhalo. We measure this quantity using the dust-attenuated SSDS magnitudes from [122].
Fraction of Disc Stars	Disc-to-total ratio, obtained via the fractional mass of disc stars: $[D/T] = M^\star_{circ}/M^\star_{tot}$, i.e. of stars on circular orbits.	To identify stars on circular orbits within the simulation, we use the star-wise epsilon parameter defined by e.g. [105] as $\epsilon = J_z/J(E)$, whereby J_z is the angular momentum around the symmetry axis and $J(E)$ is the maximum momentum J for a certain binding energy E. We define a stellar particle to be on a circular orbit when $\epsilon > 0.7$. As galaxy-wide statistics, we use the fraction of stellar particles that are identified as circular within an aperture of $10\times$ the stellar half-mass radius around the galaxy center, from [71].
$g-r$ Color	$g-r$ in SDSS bands	These are dust-attenuated and measured within twice the stellar half-mass radius; from [122].
Stellar Metallicity	SDSS-r-band luminosity-weighted average stellar metallicity.	The average is taken across gravitationally-bound stellar particles within twice the stellar half-mass radius. In the simulation, the stellar metallicity is the fraction of all elements heavier than Helium divided by the total stellar-particle mass.
Stellar Age	SDSS-r-band luminosity-weighted average stellar age.	The average is taken across gravitationally-bound stellar particles within twice the stellar half-mass radius.

Galaxy Unobservables		
Name	Description	Note
Mean Merger Time	Mass-weighted average lookback time of all past merger events in a galaxy history.	Adopted from [147]. The lookback time is equal to the last snapshot the galaxies were both identified as unique subhaloes by SublinkGal. "Mass weighted" in this context means that the mass of the secondary progenitor is used as weight in the averaging. This makes both this and the following statistics more sensitive to *massive* merger events.
Mean Merger Mass Ratio	Mass-weighted average stellar mass-ratios of all past merger events in a galaxy history	Adopted from [147]. The mass ratio is measured at t_{max} i.e. the time the secondary reached its maximum stellar mass.
Last Major Merger Mass	Total stellar mass of the last merger with a stellar mass ratio $\geqslant 1/4$.	The mass is defined as the maximum total stellar mass the secondary had during its lifetime. If there is no major merger, the parameter is set to the value of $10^6\,M_\odot$, i.e. below the resolution limit of TNG100.
Last Major Merger Time	Lookback time of the last merger with a stellar mass ratio $\geqslant 1/4$.	The lookback time is relative to the redshift of the galaxy under consideration. If there is no major merger, the parameter is set to the unphysical value of 15 Gyr.
Stellar Ex-Situ Fraction	Ratio between ex-situ to total stellar mass, where total = ex-situ + in-situ.	We can track individual stellar particles through cosmic time using their unique particle ID and call in-situ the stars of a galaxy that have formed within its main progenitor branch and call ex-situ those that formed in galaxies on secondary branches of the galaxy merger tree, following the definitions and results of [134, 147]. This mass ratio includes all stellar particles identified by Subfind to belong to the respective subhalo i.e. all gravitationally-bound stellar particles with no galactocentric distance cut.

Table 1: **Observable and unobservable galaxy properties used and connected in this work.** At the top, we list and define the observable integral properties of galaxies (i.e. features) that are are used in this work to infer unobservable properties of galaxies' past assembly and merger history, bottom.

color, average stellar metallicity and average stellar age. These may especially contain information about the formation history of the stellar populations.

We elect five summary statistics to characterize the past history of a galaxy: these are listed in Table 1, bottom, and we will further comment on our choices in the Discussion. In fact, finding a set of variables that fully represent the merger history of any individual galaxy is not a trivial task. From the simulations, we can construct this history from the corresponding SUBLINK tree of progenitor subhaloes/galaxies that have previously merged with the galaxy of question. In principle, it would be possible to use the whole list of progenitors as an output. Furthermore, recently, [125] have proposed to characterize the structure of merger trees using a single parameter: the tree entropy. However, here we opt to streamline our analysis by deriving a set of meaningful numbers for each galaxy that concatenate the information of the merger tree into a set of scalar statistics. These have been used broadly in the astronomical community, as we detail below, and thus represent a more straightforward and intuitive representation of the past history of a galaxy than more complex or more novel alternatives.

As anticipated in the Introduction, cosmological simulations have demonstrated that galaxies can assemble their stars via two channels: by in-situ star formation, whereby gas is converted into stars, and by the stripping and accretion of stars that formed in other galaxies that in turn became satellites of, or merged with, others [e.g. 126, 133]. The second stellar mass assembly channel is referred to as "ex-situ", and accreted or ex-situ stars can be found throughout galaxy bodies, from their innermost regions to the stellar haloes [e.g. 132, 147]. Furthermore, the fraction of stars formed ex-situ has been frequently used to distinguish the evolutionary pathways of galaxies of different sizes and morphologies, both in simulations [e.g. 49, 143, 148] as well as in the discussion of observations [e.g. 87]. It is therefore a natural choice to use the ex-situ stellar mass fraction as a primary indicator for summarizing the past merger history of a galaxy.

For any galaxy, we also record the mass-weighted average mean lookback time and the mass-weighted average stellar mass ratio of all past merger events; this is the simplest method to summarize the overall history of merger events in terms of timing and massiveness. Previous analyses of previous cosmological simulations have shown that there exist a strong correlation between these statistics and the stellar ex-situ fractions [147]: we will further investigate on these within the TNG framework in the next Sections.

Finally, as recent major mergers may have a strong impact on the properties of galaxies, we also consider the lookback time and the stellar mass of the last major merger a galaxy underwent [147]. As described in the Introduction, these parameters are in general of great interest for the Milky Way [e.g. 115] and for other galaxies [81, 201] as well. There is no universal definition of a major merger: in this work, we choose mergers with stellar mass fraction larger than $1/4$. However, not all galaxies undergo such massive mergers ever: for galaxies with no major merger in their lifetime, we set these mass and time to unphysical values, so that we can later on distinguish these no-major-merger galaxies from the remaining population.

We show an overview of all observable and unobservable quantities and how they relate to galaxy mass in Fig. 12 and Fig. 13.

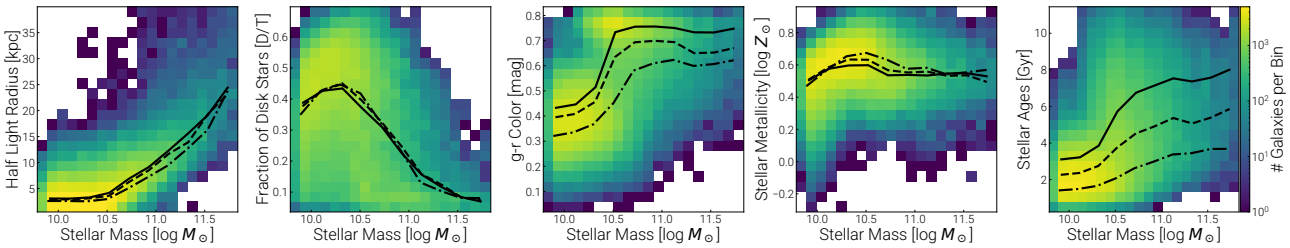


Figure 12: **Overview of the observable statistics used in this work, as a function of galaxy stellar mass and redshift, and predicted by TNG100**: in each row we plot the whole simulated sample (182'625 galaxies in total) introduced in Section 8.4 as 2D histograms between one of the observable quantities and the stellar mass (within $2 \times r^{\star}_{1/2}$). Curves denote the medians as function of the stellar mass at redshift $z < 0.2$ (solid), $0.2 < z < 0.5$ (dashed) and $0.5 < z < 1$ (dot-dashed). Note that there is in fact a bi-modality in the data between disc-blue and elliptical-red galaxies, which is not well represented by the median lines.

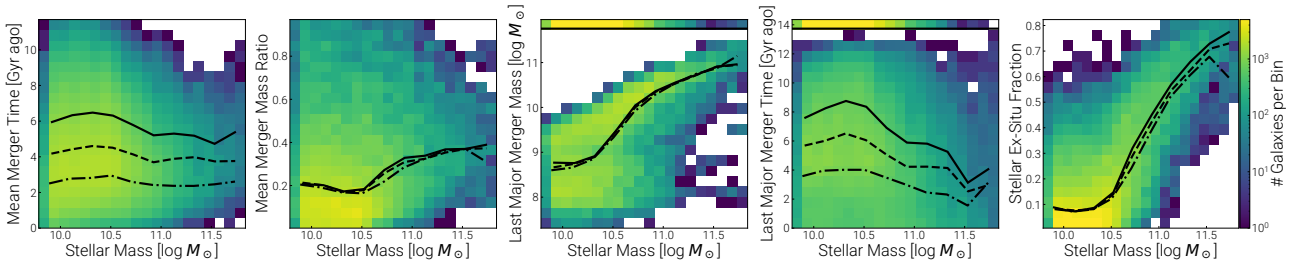


Figure 13: **Overview of the unobservable statistics used in this work and predicted by TNG100**: in each row we plot the whole simulated sample introduced in Section 8.4 as 2D histograms between one of the unobservable quantities and the stellar mass (within $2 \times r^{\star}_{1/2}$). Curves denote the medians as function of the stellar mass for redshift $z < 0.2$ (solid), $0.2 < z < 0.5$ (dashed) and $0.5 < z < 1$ (dot-dashed). Here, for the mass and the lookback time of the last major merger, we plot the galaxies without any major merger in their history as a bar at the top. Especially for the stellar ex-situ fraction and the mass of the last major merger, the simulations predict a strong positive correlation with the stellar mass of the respective galaxy; on the other hand, these relationships depend only weakly on redshift. This is naturally different for the unobservables that are defined relative to the redshift of the galaxy at hand, namely the mean merger lookback time and the time of the last major merger.

8.6 INFERENCE METHOD: CONDITIONAL INVERTIBLE NEURAL NETWORKS (CINN)

A regression model like e.g. a Multilayer Perceptron (for comparison we introduce, train and test such a model in Appendix B) can converge to an exact solution only if the underlying connection between the 7 observables and the 5 merger statistics is a bijective function and is therefore unambiguous. However, it is possible that the information contained in the chosen set of observables is insufficient and that galaxies with (nearly) identical observable parameters may actually have distinct assembly and merger histories. It is also possible that the observable properties of galaxies may not be uniquely determined solely by their past stellar assembly and merger history, but also by other physical processes and conditions, as it is certainly the case. It is hence important to not only perform a simple scalar regression with one single set of point predictions, but to also know the posterior distribution $p(x|y)$: this is the conditional probability that a set of merger statistics x is compatible with the condition that a set of galaxy observables y is given.

We achieve this by using the Conditional Invertible Neural Networks (cINNs) developed by [10], which are a generalization of their Invertible Neural Networks [9]. The latter utilize as building blocks the invertible architecture proposed by [53] and an additional set of latent variables z, that store the information that would be otherwise lost. This network is therefore describable as an invertible function $f(x) = [y, z]$ from the merger-history statics x, to a tuple of observables y, and a set of latent variables z. It was shown by [9] that by training this network as proposed, the inverse $f^{-1}(y, z)$ will converge against the posterior $p(x|y)$ when sampling over the gaussian latent space z. However, a simple INN demands for the forward process $f(x) = [y, z]$ to be well defined, this not being the case for our physical problem. The cINNs circumvent this issue by allowing for the set of observables y, to be introduced in the network as conditions c, which are then concatenated into the conditional affine coupling layers. The resulting neural network is therefore equivalent to an invertible function $f(x, c) = z$ with latent variables z whose dimension is equal to the dimension of x. By applying a negative log likelihood loss

$$L_{NLL} = \left| \frac{|z|^2}{2} - \log(|J|) \right|, \tag{1}$$

with the Jacobian J of the invertible function f, the distribution of z is enforced to take the form of a multivariate gaussian.

In practice, the cINN provides a mapping between the posterior $p(x|c)$ and a multivariate gaussian distribution conditioned by the galaxy observables c; by sampling the gaussian latent space z, one can therefore recover the posterior $f^{-1}(z, c) = p(x|c)$.

Conditional Invertible Neural Networks can be easily implemented by using the FrEIA framework [Framework for Easily Invertible Architectures, 11], which is based on PyTorch [129]. They have been already successfully applied to similar problems in astronomy [e.g. 99].

9 NETWORK TRAINING

9.1 PREPROCESSING OF THE SAMPLE

Before tackling the training of the cINN, a couple of interventions need to be executed on the data.

9.1.1 *Feature Scaling*

As standard in ML applications, we scale all quantities to a comparable range of values, for the following reasons:

1. the loss functions we are using are not necessarily invariant with the magnitude of the unobservable quantities. A feature with large absolute variations would be thus more important in terms of loss than a feature with small absolute variations;
2. the gradient descent algorithms that are used to optimize the neural networks are sensitive to the scale of the input – the change in weights is proportional to the feature values;
3. many default parameters of ML algorithms are given relative to standardized inputs.

Therefore, we scale each of the quantities introduced in Section 8.5 by subtracting the mean and dividing by the standard deviation of each variable distribution. We hence get a sample of inputs and outputs in scaled units, whereby the mean is 0 and the standard deviation is 1 for each quantity across the whole sample. During the whole training and testing, we only use these scaled quantities and scale them back to physical units only when enunciating and visualizing the final results.

9.1.2 *Sample split, in the context of galaxy populations across cosmic epochs*

We randomly split the TNG100 galaxy sample into a training, a validation and a test subset, in fractions of 80, 10, and 10 per cent, respectively. We fit the ML models only to the training set, separately, whereas the validation set is used during the fitting process to evaluate the performance of the models. The test set is needed to test the generalization performance of the final models and consists of galaxies that are never seen before by the networks.

Because cosmological simulations follow the same galaxy populations across cosmic epochs, progenitors and descendants of the same galaxy (i.e. of the same main branch) may appear over multiple, consecutive snapshots. This special tree-like structure of the sample under consideration may require special precaution, as some information may unintentionally slip between the training and validation/test subsamples. For example, a galaxy in the validation/test set may have

progenitor(s) and/or descendant(s) in the training sample with possibly similar observable and unobservable parameters: training a network under these circumstances may lead to a good optimization along the merger tree of each galaxy, but possibly to a bad generalization performance across galaxies throughout the TNG100 galaxy catalogues.

The issue is more severe for more closely spaced snapshots: in the case of TNG, the time span between consecutive snapshots is about 150 Myr; moreover, this issue is progressively more problematic for smaller training sets. In the real Universe, the properties of the galaxy populations at different redshifts certainly reflect the underlying evolution of individual galaxies and their properties across time. However, in early tests, we observed that training our network(s) without reducing the merger tree-like information between training and validation/test sets causes overfitting and therefore can eventually lead to a bad performance on real observational data. To be more specific: without paying attention to the tree-like structure of the data, our ML model(s) performed worse when trained (and tested) on larger galaxy samples (i.e. available from the TNG300) in contrast to the expectation that a larger training sample should yield more information that a network could learn from. We believe that this is caused by the fact that a larger sample (where galaxies are in general closer to each other in parameter-space) counteracts the overfitting caused by the information that is present in the tree-like structure.

To overcome this problem we split the TNG100 galaxy sample into training, validation and test subsets according to the merger trees, such that all progenitors and descendants of each galaxy are contained in the same subset.

In practice, we perform the following steps:

1. get the root descendant for each galaxy (at each snapshot); i.e. the ID of the galaxy it will eventually result in according to SublinkGal;
2. split the list of these root descendants in training, validation and test sets;
3. sort each galaxy according to the associated root descendant into the 3 subsets.

With this procedure it is ensured that the progenitors and descendants of each test and validation galaxy are not contained in the training set. We have verified that, after splitting according to root descendants, training and validation galaxies indeed lie much further apart in parameter space than with the default random splitting.

9.2 NETWORK CONFIGURATIONS AND TRAINING

We want to find the optimal architecture and optimal set of training parameters. To do so, we perform for each network type a parameter search i.e. we train each network multiple times on the same set of TNG100 training galaxies but alternating the architecture and training parameters at each step. Each training is stopped as soon as the validation loss (i.e. the model loss when applied to the validation galaxy subsample) does not decrease anymore. From all resulting models, we choose then the network configuration that yields the lowest validation loss. In the following,

Layer	Output Shape	Parameters
Input	7	0
Dense	128	1024
ReLU Activation	128	0
Dense	128	16512
ReLU Activation	128	0
Dense	128	16512
ReLU Activation	128	0
Dense	128	16512

Table 2: **Layer configuration used for the *conditioning network* of the conditional Invertible Neural Network adopted in this work** (cINN, Sections 8.6). From left to right, we give the layer type, the output shape of the respective layer, and the number of trainable parameters per layer.

Layer	Output Shape	Parameters
Input	130	0
Dense	128	16768
ReLU Activation	128	0
Dense	128	16512
ReLU Activation	128	0
Dense	128	16512
ReLU Activation	128	0
Dense	4-6	516-774

Table 3: **Layer configuration used for the cINN** *sub-networks* (see Section 9.2.1). Columns as in Table 2.

we describe the architectures and training parameters we have found to suit best for the particular problem at hand.

9.2.1 *cINN*

For the cINN, we rely on conditional affine coupling blocks [in so-called GLOW configuration, 97] that utilize two standard (and therefore non-invertible) MLP sub-networks. Additionally, a further "conditioning" MLP network is introduced between the conditions (i.e. the 7 galaxy observables of Section 8.5) and the coupling layers. In practice, this conditioning network works like the point-prediction model described in Appendix B, whereas the coupling layers provide a model for the posterior around the point predictions.

For the conditioning network we use the architecture shown in Table 2; the sub-networks are summarized in Table 3. For the cINN itself, we use in total 12 coupling layers with a random perturbation layer between each. With this configuration, we have in total 1'262'584 parameters to train.

As already described in Section 8.6, this network is trained by propagating the obervables and unobservables of the training subset through the network and then adjust the layer weights of the model in such a way that the negative log likelihood loss L_{NLL} is minimized.

During the training, we add a small amount of gaussian noise with $\sigma = 0.025$ to all inputs and outputs. This is done to lower the risk of overfitting, especially in the case of the parameters based on lookback time, which enter the network in a discrete fashion as the simulation only provides us with 50 discrete snapshots. Again, we use ADAM as optimizer and start with a learning rate of 2×10^{-3}, which is then gradually reduced (i.e. multiplied by 0.7) after each 5 runs through the whole training sample. We train the cINN in batches of 2048 training galaxies. To avoid overfitting, we calculate the loss for the validation set and terminate the training if the validation loss has not decreased for 5 following training epochs.

These training parameters (and model architecture) were determined by a parameter search: on the one hand we choose the parameters such that the validation loss is minimized; on the other hand, we also check that the latent space z indeed becomes a gaussian and we also test if the prior distribution of the validation set is successfully recovered (both utilizing a Maximum Mean Discrepancy between the two distributions).

9.2.2 *Ensemble Training*

As the network's initial weights as well as the gaussian noise augmentation are randomly chosen, the training of deep learning models is a stochastic process. Therefore, even when using the same data, the same network architecture and the same training parameters, the results of a network may differ, albeit slightly, from training to training: some will perform better or worse than others for specific galaxies. To take these network-to-network variations into account, we train multiple networks and average their results. In this way, we ensure a globally-optimal result over the whole sample while reducing the risk of overfitting as we can keep the single networks simpler and profit from the combined prediction capability. Furthermore, we ensure that our results (Section 10) are less dependent on the random choice of initial parameters.

For the cINN, we also train 7 networks with random initial weights. However, in this case we cannot simply take medians of scalar point predictions. Instead, when sampling the posterior, the posterior samples are drawn alternately from the 7 single networks. The resulting ensemble posterior for each galaxy is therefore the normalized sum of the 7 individual network posteriors.

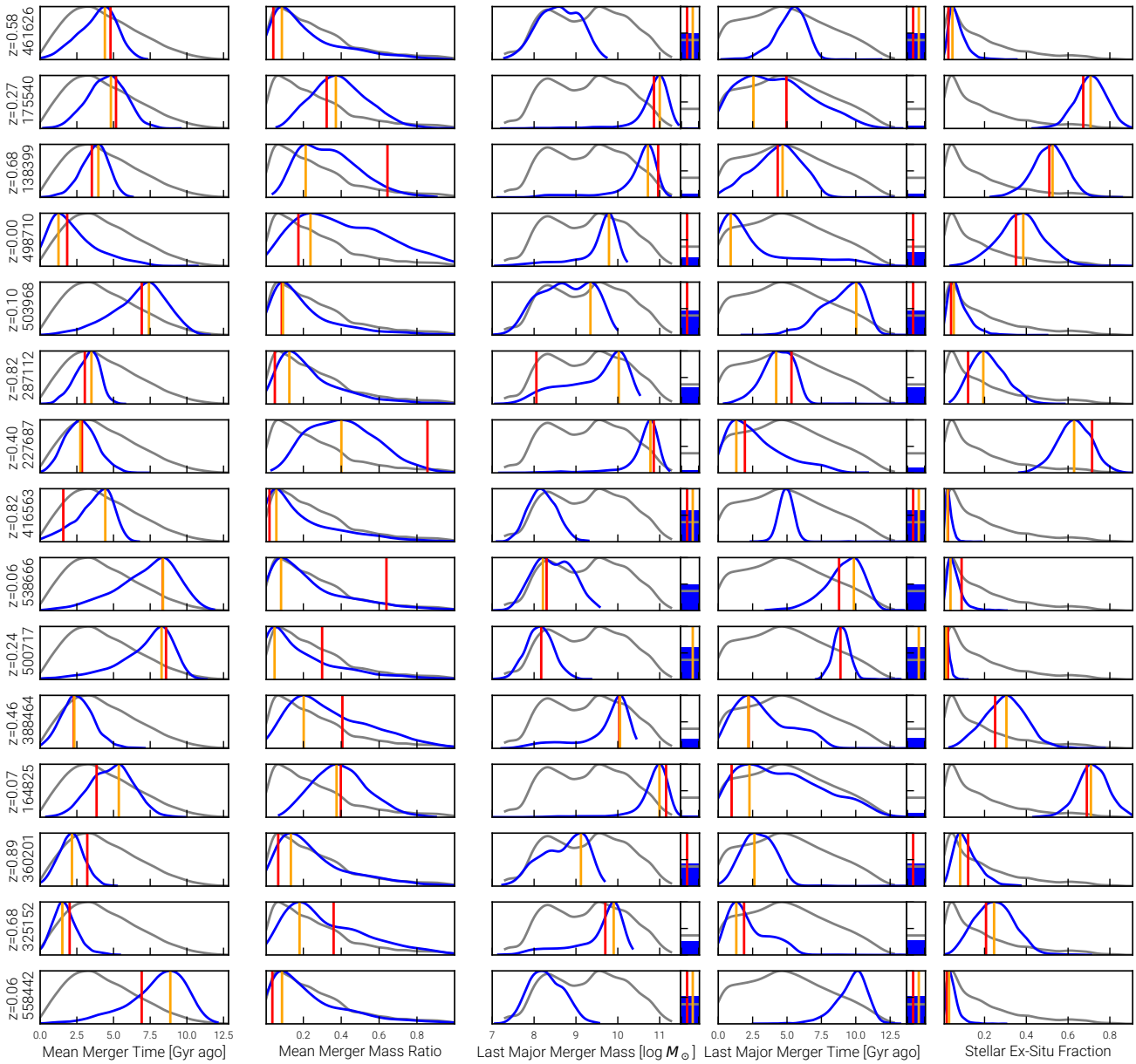


Figure 14: **Posterior distributions obtained with our cINN** (Sections 8.6 and 9.2.1) for 15 randomly-chosen galaxies from the TNG100 simulation test set (out of in total 17982 galaxies). Each row corresponds to one of the 15 galaxies (we show the redshift and the (snapshot-wise) unique SUBFIND ID on the left) whereas each column shows results for the five inferred unobservable statistics of their assembly and merger history. We show the prior distributions of the overall sample of test galaxies (grey), the predicted posterior by our cINN approach (blue), the MAP prediction (Maximum A-Posteriori estimation, yellow), and the ground truth (red) – i.e. the actual outcome of the TNG100 simulation. The distributions are normalized such that the maximum value is equal to 1. For the last major merger, we introduce an additional bin to show the fraction of posterior samples that fall outside the valid regime, i.e. the probability that the galaxy has never had a major merger in its merging history. This bin is chosen as the MAP for this galaxy only if the fraction of posterior samples falling into this bin is > 0.5 i.e. if the predicted probability for a major merger to have occurred is larger than 50 per cent.

10

RESULTS: RECOVERING THE UNOBSERVABLE PROPERTIES OF GALAXIES' PAST

In this Section, we show the main scientific results of the paper, i.e. that it is possible to recover aspects of the past assembly and merger history of galaxies from a set of integral properties at a given time. In practice, in what follows, we apply the previously introduced and trained networks to the unseen test data set from the TNG100 simulation. We first focus on the overall performance of the cINN and also show how the cINN recovers the cross-correlations among input and output galaxy properties. Furthermore we identify the most informative galaxy properties for the inference of the past history using a MLP point-recursion model as a simpler-to-use surrogate: the MLP is introduced and discussed in Appendix B.

10.1 TESTING PROCEDURE

To test the performance of the cINN and its prediction capabilities, we apply the trained cINN to a previously-unseen test sample. Our main measure for the goodness of fit is the prediction error for each galaxy: this is the difference between the predicted values, inferred by the cINN given the observables, and the ground-truth values given by the TNG100 simulation.

In the case of a simple recursion network (MLP), one gets point predictions for each galaxy that can be directly used for the error measurement. However, for the cINN the procedure is more articulate: we first sample the posterior for each galaxy (i.e. conditioned by the corresponding set of observables) by drawing 1000 random normal-distributed points in the latent space z and by propagating them backwards through the cINN. The resulting posterior is then represented by the point density of the sampled point cloud and can be further postprocessed. To be able to compare this point cloud to the point predictions of a MLP, we identify as "point prediction" of the cINN the Maximum A-Posteriori estimation (MAP) i.e. the position of the highest point density in unobservable space.

Technically, to obtain the MAP for each unobservable quantity, we perform a gaussian kernel-density estimation on the set of posterior samples and evaluate the estimated density on a uniform grid of 512 points along each dimension of the unobservable space and use the position of the highest density. Note that with this, we impose a strict upper limit on the prediction accuracy due to the discrete nature of this grid approach. However, with 512 grid points we are well below the level of noise that was added during the training.

For the lookback time and mass of the last major merger, we define an additional bin for the galaxies or predictions that have no major merger in their history (we call them for convenience NMMs). The cINN MAP is set to this bin if at least 50 per cent of the posterior sample points fall into this bin i.e. have a predicted lookback time larger than 13.7 Gyr – the ground truth for these galaxies is set to 15 Gyr;

see Section 8.5 – or a predicted stellar mass below $10^7\,M_\odot$, as the ground truth for these galaxies is set to $10^6\,M_\odot$.

To take also the additional information of the posterior distribution into account, we calculate the standard deviation of the posterior point cloud along each dimension i.e. we take the root of the diagonal elements of the covariance matrix. So, whereas the prediction errors give an assessment on the accuracy of the predictions, we investigate below if the standard deviation of the posteriors serves as a meaningful error bar (i.e. level of uncertainty or precision) for the MAP predictions.

10.2 PERFORMANCE OF THE CINN

10.2.1 *Posterior distributions from the cINNs*

A strength of cINN architectures is that they return a posterior distribution for each target quantity[1]. However, the performance evaluation is not as straightforward as e.g. the validation of a simple recursion method. To get a first impression of the performance of the cINN, we show in Fig. 14 a few examples of the cINN posteriors inferred for a random subset of TNG100 test galaxies. Each row shows results for each test galaxy, 15 in total; columns refer to the five summary statistics elected to summarize the assembly and merger history of galaxies (see Section 8.5): from left to right, mean merger lookback time, mean merger mass ratio, stellar mass and time of the last major merger, fraction of ex-situ stars. Gray curves denote the prior distributions, i.e. the distributions of values encompassed by the whole sample of test galaxies; blue curves show the posterior distribution predicted by the cINN for each test galaxy and unobservable property. In each panel, we mark the MAP (the Maximum A-Posteriori estimation, Section 10.1) in orange and the simulation ground truth in red. It is important to note that here the posteriors (blue curves) are a convolution of the "true" data-driven posterior and the uncertainty introduced by the network-to-network variations dealt with in Section 9.2.2.

Fig. 14 shows that the posteriors of the mean merger lookback time and particularly of the stellar ex-situ fraction are typically peaked and rather symmetrically distributed around this peak, denoting qualitatively a good prediction power of the cINN model in terms of point predictions. In contrast to that, the posterior of the mean merger mass ratio of the tested galaxies is often similar to the prior of this parameter. For the mass and time of the last major merger, we can also see broader and more complex posteriors, at times with multiple peaks e.g. fifth galaxy from the top. Note that the additional bin, the one for galaxies with no major merger in their history, can be seen as a posterior peak as well. It is important to emphasize, however, that everywhere where the posterior probability is > 0, a certain merger statistic configuration is in general predicted to be possible: i.e. we have to check if the inferred posteriors follow the distribution of the ground truth later on.

1 It would in principle be possible to obtain a posterior distribution also for the MLP predictions: however, for the MLP, a parameterized distribution and therefore prior assumptions about the shape of the posterior would be needed, which is not the case for the cINN

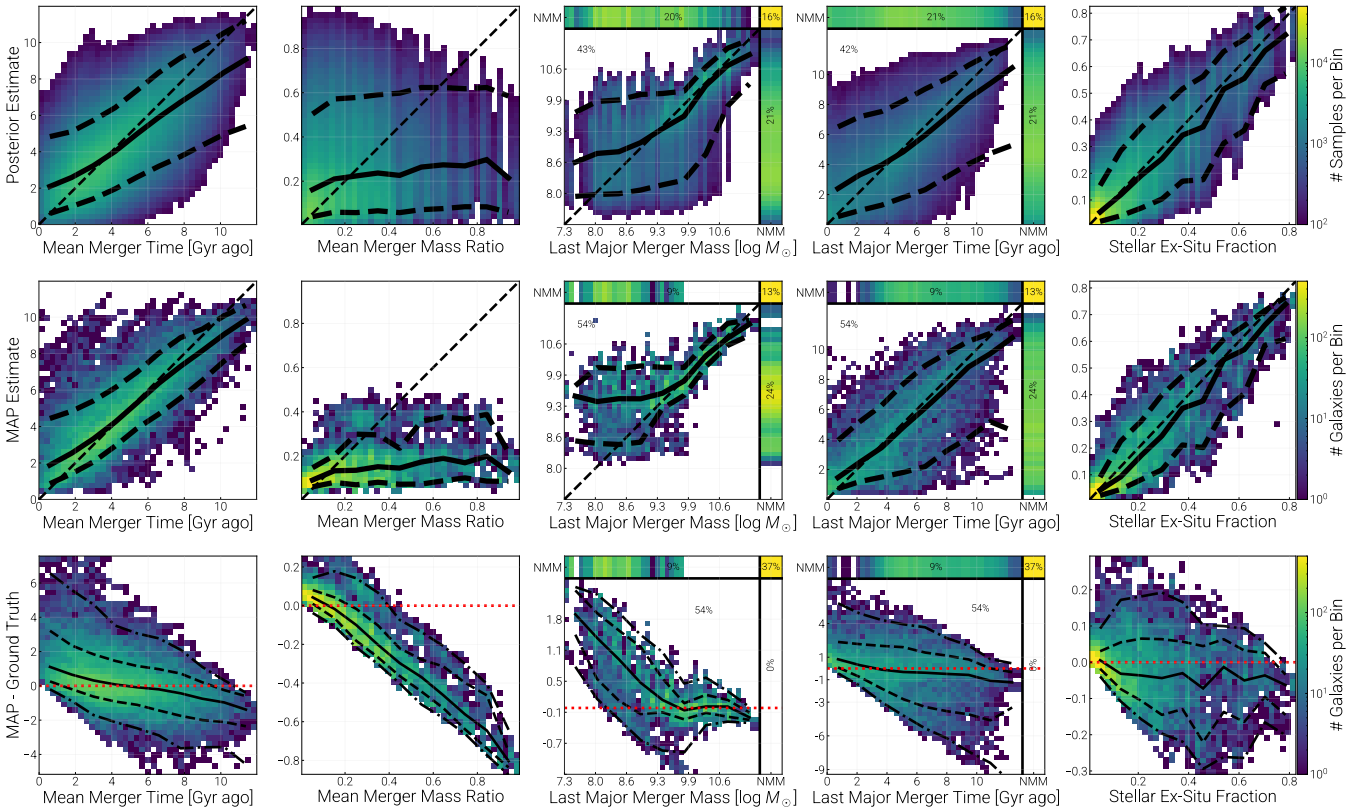


Figure 15: **Prediction performance of our cINN based on training, validating and testing against the outcome of the TNG100 cosmological simulation.** The columns show results for the five inferred unobservable statistics of the assembly and merger history of galaxies, for 17'982 TNG100 test galaxies . **Top:** 2D histograms of the cINN posteriors (y-axis) over simulation ground truth (x-axis) for each of the 5 output quantities. Namely, for each test galaxy, we draw 100 random posterior samples and stack them across all test galaxies. The measured median is given as solid black line while the area included within the black dashed lines contain 80 per cent of the datapoints. For comparison, we also show the ideal result (i.e. the case whereby MAP prediction and ground truth concur) as dotted black diagonal. **Middle:** 2D histograms of the cINN MAPs (y-axis) vs. simulation ground truth (x-axis) for each of the 5 output quantities. Meaning of the lines and annotations as in the top panels. **Bottom:** 2D histograms of the cINN MAP errors, i.e the MAP - Ground Truth (y-axis) vs. simulation ground truth (x-axis) for each of the 5 output quantities. The solid black line shows the median, the area within the dashed and dash-dotted lines contain 68% (1σ) and 95% (2σ) of the datapoints in bins of ground truth values. We show for reference the ideal (zero error) baseline as red dotted line. For the time and mass of the last major merger (third and fourth panels from the left), we also show the galaxies that have no major merger (NMM) in their main progenitor branch according to the cINN (top), the ground truth (right) and both (upper right corner). The percentage denotes the overall fraction of galaxies in each of those 3 bins. Note that for the bottom plot, the right bin is empty as the error is undefined if the galaxy had no major merger according to the ground truth.

10.2.2 *cINN predictions vs. ground truth*

In Fig. 15, we therefore systematically compare the cINN predictions with the ground truth from the TNG100 simulation. In particular, in the top panels of Fig. 15 we show the stacked posterior distributions predicted by the cINN for all galaxies of the test sample. In practice, we show the probability distribution functions of Fig. 14 for all test galaxies by stacking them in bins of true mean merger lookback time, mean meger mass ratio, mass and time of the last major merger, and fraction of ex-situ stars, from left to right, with colors denoting the normalized probability distribution function. In each panel, the black thin dashed line denotes the 1:1 relation; the solid black curve shows the median of the stacked posteriors and the dashed thick black curves encompass their 10th and 90th percentiles.

In the middle panels of Fig. 15, we plot the MAP estimates for each test galaxy vs. the simulation ground-truth values, again different column for the five merger statistics. As we aim to quantify the outcome for 17'982 test galaxies, the results are given as color-coded numbers of galaxies per image pixel. In each panel, the black thin dashed line denotes the 1:1 relation; the solid black curve shows the running median of the galaxy test sample in the depicted plane; and the dashed black curves indicate the location of the 10th and 90th percentiles of the test galaxy samples in bins of true merger statistics.

In the bottom panels of Fig. 15, we replace the MAP estimates of the middle panel with the error on the prediction, i.e. the difference between the MAP estimate and the simulation ground truth for each test galaxy, for each inferred target quantity. The median prediction error is given by the solid black curves whereas the dashed (dot-dashed) curves encompass 68 (95) per cent of the test galaxies.

As also seen in Fig. 14, Fig. 15 demonstrates that the cINN returns very good predictions for both the mean merger lookback time (leftmost panels) and the mass fraction of ex-situ stars (rightmost panels). The MAP estimates for the ex-situ fractions are on average slightly ($\approx$ 5 per cent) below the ground truth, this being the case also from the stacked overall posteriors: this could be caused by the unevenly-distributed prior distribution of the ex-situ parameter, which strongly favours lower ex-situ fractions. Nevertheless, for 68 per cent of the galaxies, the prediction error (bottom panels) for the ex-situ fraction does not exceed $\pm 10 - 15$ percentage points from the ground truth: the model is therefore well able to discriminate between low and high ex-situ galaxies. The MAP error is especially low in the low ex-situ-fraction regime (i.e. ex-situ mass fractions lower than 10 per cent) where the predicted ex-situ fraction does not exceed $\approx \pm 5$ percentage points for 95 per cent of the test galaxies. Similar promising results hold for the mass-weighted mean merger time, with errors on the MAP prediction smaller than ± 1.5 lookback billion years for $\approx$ 69 per cent of the test galaxies.

On the other hand, the cINN predictions for the time and mass of the last major merger, and especially for the mean merger mass ratio, are more complex. As noted before, the mean merger mass ratio posteriors are often similar to the prior of this parameter. We can see this also in terms of the MAPs, whose peaks are (independently from the ground truth) always close to the prior peak. In practice, our models cannot recover this particular summary statistics of the galaxies' history. We will further elaborate on this in Section 11.1.1.

For the inference of the mass and time of the last major merger, it should be noted that 37 per cent of the TNG100 test galaxies have never had a major merger in their history according to the simulation (i.e. ground truth). The cINN predicts in 24 per cent of the cases that galaxies have had a major merger even though they have experienced in truth no major merger (so, a wrong prediction). In contrast to that, 9 per cent of galaxies with ground-truth major merger have been predicted to have no major merger, while 13 per cent have been correctly identified to have no major merger in their history. These numbers are consistent for both the mass and time of the last major merger.

For the remaining galaxies that have undergone a major merger and that the cINN has correctly recognized as such ($\approx$ 54 per cent of the overall test sample), the cINN performs well, particularly in timing the merger, with a good alignment to the ideal line. For the median galaxy, the cINN predicts an MAP value for the time of the last major merger that is close to the true one by less than ± 1 billion years, also for very ancient mergers. Galaxies with a more recent major merger (within the last $2-3$ Gyr) have lower MAP errors throughout the galaxy population than ancient ones (and there is a much lower tendency to wrongly classify them as NMM). In contrast to that, there is a large fraction of galaxies with an early major merger (> 8 Gyr ago) that are wrongly predicted to have a much more recent major merger. For the stellar mass of the major merger, for the bulk (68 per cent) of the galaxy population, the cINN predicts the correct mass of the last major merger to within $\pm 0.1-0.2$ dex but only for mergers more massive than $\sim 10^{9.5}\,M_\odot$ in stars. On the other hand, the prediction becomes ambiguous at lower masses. These two trends are naturally related to each other: because of the monotonically mass growth of galaxies, massive mergers will also happen at later cosmological times. The lower ambiguity for recent major mergers translates therefore directly to the better performance for more massive mergers and vice versa. In conclusion, the network performs especially well in identifying recent and/or massive major mergers, while the predictions become less accurate if the last major merger was longer ago (or not present at all). This supports the initial expectation and foundation of the whole exercise, namely that recent mergers have a more pronounced impact on the appearance of a galaxy.

It should be noted that the trends in the prediction error for the mean merger lookback time, the lookback time of the last major merger and (partially) the ex-situ fraction are also caused by the physical limits that our model learns: e.g. there are no negative lookback times or ex-situ fractions. Additionally, it is important to emphasize that choosing the MAP estimate is only one of the possible ways to condense the information contained in the posterior distributions. We have also evaluated the cINN performance based on the median and the mean values of the posteriors: the latter yields slightly better results regarding the under-prediction of the ex-situ fraction and the mass ratio but at the cost of returning worse performances for the other target statistics.

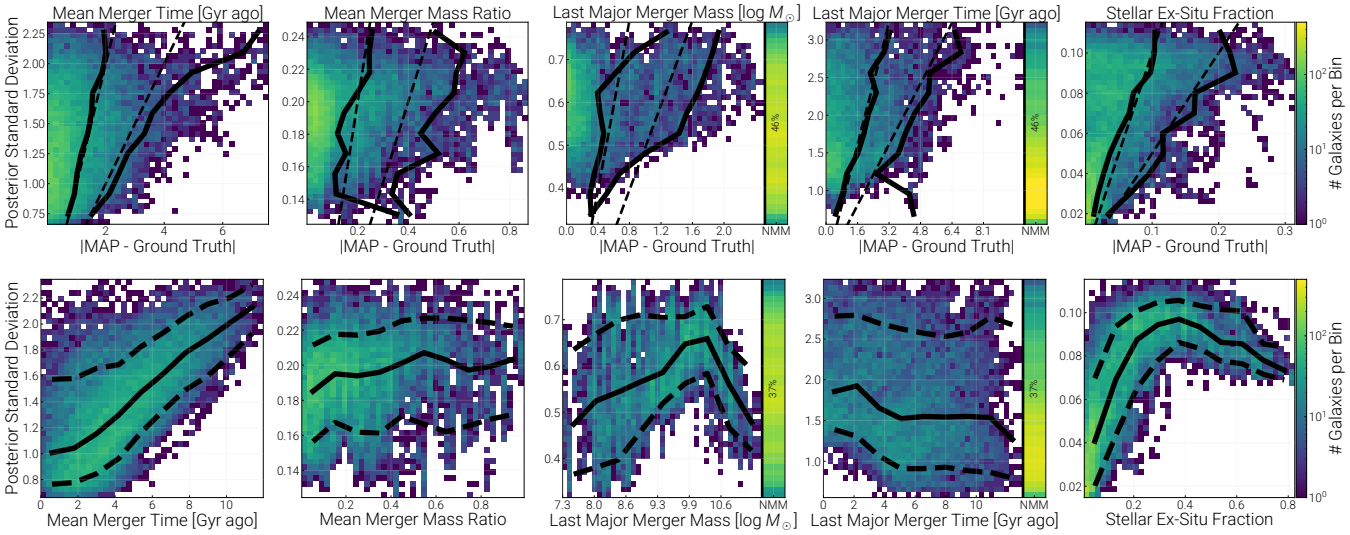


Figure 16: **Uncertainties of the cINN predictions. Top:** 2D histograms of the standard deviation of the cINN posterior distributions vs. the absolute value of the MAP prediction error for all TNG100 test galaxies. In bins of posterior standard deviation, we plot where the 68th and 95th percentiles of the test galaxies lie: thick black curves. The dashed lines correspond to the expected "ideal" result, assuming that the MAP prediction error is normal distributed. **Bottom:** 2D histogram of the standard deviation of the cINN posterior distribution vs. the simulation ground truth. The black curves denote the median relationships, whereas 80 per cent of the test galaxies are contained within the area enclosed by dashed lines, in bins of predicted properties (x-axis). For the time and mass of the last major merger, we plot the galaxies that have never had a major merger according to the ground truth (i.e. the TNG100 simulation) in an extra bin (NMM) and show the fraction of galaxies that fall into this bin: in these cases, the standard deviations are only calculated from the fraction of posterior samples that fall into the physical range (i.e. lookback time $\leqslant$ 13.7 Gyr and stellar mass $\geqslant 10^7\,\mathrm{M}_\odot$).

10.2.3 *Interpretation of the uncertainties of the cINN predictions*

In a next step, we want to estimate the errorbars on the cINN point predictions, i.e. on the chosen MAP estimates, by evaluating the meaning of the standard deviation of each posterior.

It should be emphasized that the cINN posteriors throughout this work include prediction uncertainties given by both the limited information given by the selected observables as well as by the network-to-network variations (see Section 9.2.2). We proceed under the ansatz that the former is dominant.

To use the standard deviation of the cINN posteriors as a meaningful measure for this uncertainty we need to answer the following questions:

- Are the posteriors "gaussian-like"? Namely, are the posteriors symmetric around one peak and hence with the well-known relations between the variance and the confidence intervals?
- Is the standard deviation of the posterior related to the prediction error of our test sample? In other words, is the width of the posteriors related to the difference between the MAP estimate and the ground truth?

For example, from Fig. 14, we already know that for the mass and time of the last major merger and the mean merger mass ratio the cINN often returns asymmetric posteriors. In such cases, the meaning of the posterior standard deviation is therefore in question.

To investigate these relations, in the top panels of Fig. 16 we plot the standard deviation of the posteriors vs. the prediction errors (here the absolute value of the difference between the MAP estimate and the ground truth) for the whole TNG100 test sample. To evaluate whether the error is indeed gaussian, we indicate with thin dashed lines the relationships we would expect if the cINN posteriors were normally distributed. Namely, in general terms, we would expect that 68 per cent of the test galaxies with a posterior standard deviation equal to e.g. 1 have a prediction error $\leqslant 1$; similarly, we would expect that 95 per cent of the test galaxies with a posterior standard deviation equal to e.g. 1 have a prediction error $\leqslant 2$; and so on. We can then compare such idealized gaussian expectations with the locus of the galaxies in the plane of posterior standard deviation vs. prediction error: namely, in bins of the posterior standard deviation, we measure where the 68 and 95 percentiles of the galaxies lie: solid thick lines.

For the ex-situ fraction we find that the empirical standard deviations (solid lines) of the cINN align well with the 'ideal' gaussian ones (dotted lines): we conclude that the ex-situ standard deviations can be well interpreted as a gaussian error. With caution, this can also be stated for the mean merger lookback time (especially at the 1-σ level) and for the time of the last major merger. For the other two target quantities, we can see that more attention needs to be placed on the shape of the posteriors (as expected from the examples in Fig. 14).

In the bottom panels Fig. 16, we also quantify the connection between the standard deviation of the posteriors and the ground truth values. There are two strong correlations to highlight and comment upon. Firstly, the standard deviation of the mean merger lookback time increases with the true mean merger lookback time. However, we find that this linear dependency is build up by galaxies at different

redshifts, with the galaxies at $z = 0$ populating the top right corner and those at $z = 1$ populating the lower left region.

Secondly, the standard deviation of the ex-situ fraction increases strongly with the true ex-situ fraction, reaches a peak at around fractions of $\approx 0.3 - 0.4$ and then slightly decreases again at higher ex-situ fractions. The low ex-situ fractions can therefore be estimated with small uncertainties (percentages of a few) for a large part of the overall galaxy sample. However, as the fraction of ex-situ stars increases, also the uncertainty increases up to $\approx \pm 10$ per cent. We interpret this as a result of the diverse and complex evolution history that leads to an increased ambiguity between the chosen set of observable inputs and the ex-situ fraction. The downturn at the high ex-situ end could be explained with the fact that this region of the parameter space is constituted again by a simpler population of galaxies, namely massive red galaxies whose stellar mass growth is dominated by the contribution from mergers.

We further quantify the degree to which the posterior distribution itself is related to the distribution of ground truth galaxies in Appendix A.

10.2.4 *Recovery of secondary dependencies and cross correlations*

In addition to the predictive power of the cINNs to return unobservable assembly and merger statistics from observed quantities, it is important also to check to what degree the cINN is capable to learn the cross correlations among the 5 statistics themselves. In fact, we know from the analyses of cosmological simulations that galaxies properties within both sets of observable and unobservable statistics are highly correlated with one another. So for example, at fixed galaxy stellar mass, galaxies with more massive mergers have larger ex-situ mass fractions [147]. In Figs. 17 and 18, we compare a) the prior distributions of the ground truth, i.e. how the test galaxies populate the parameter space of the 5 target statistics (red); b) the distributions from the cINN by integrating its posterior samples (blue): namely, for each galaxy in the test sample we use the ground truth observables and randomly sample 10 posterior points for the 5 statistics; the overall distribution is then given by the sum of all points for all test galaxies; finally c) the distributions of the MAP values from the cINN (green). The resulting distributions are shown in Fig. 17 for low-mass galaxies (stellar mass $< 10^{10.7}\,M_\odot$) and in Fig. 18 for high-mass ones ($> 10^{10.7}\,M_\odot$).

The cross correlations among the target quantities are well reproduced by the cINN posteriors (blue vs. red contours), although a certain amount of smoothing is visible, particularly in the less-sampled high-mass end population. For example, the cINN posteriors recover the following trends:

- The lookback time of the last major merger and the mean merger lookback time are tightly correlated, with only a small fraction of galaxies having a significant amount of minor mergers after their last major merger event (galaxies with no major mergers are excluded here).

- Smaller mean merger lookback times, larger mean mass ratios, larger masses of the last major merger and smaller times since the last major merger imply larger ex-situ stellar mass fractions at the time of inspection. These rela-

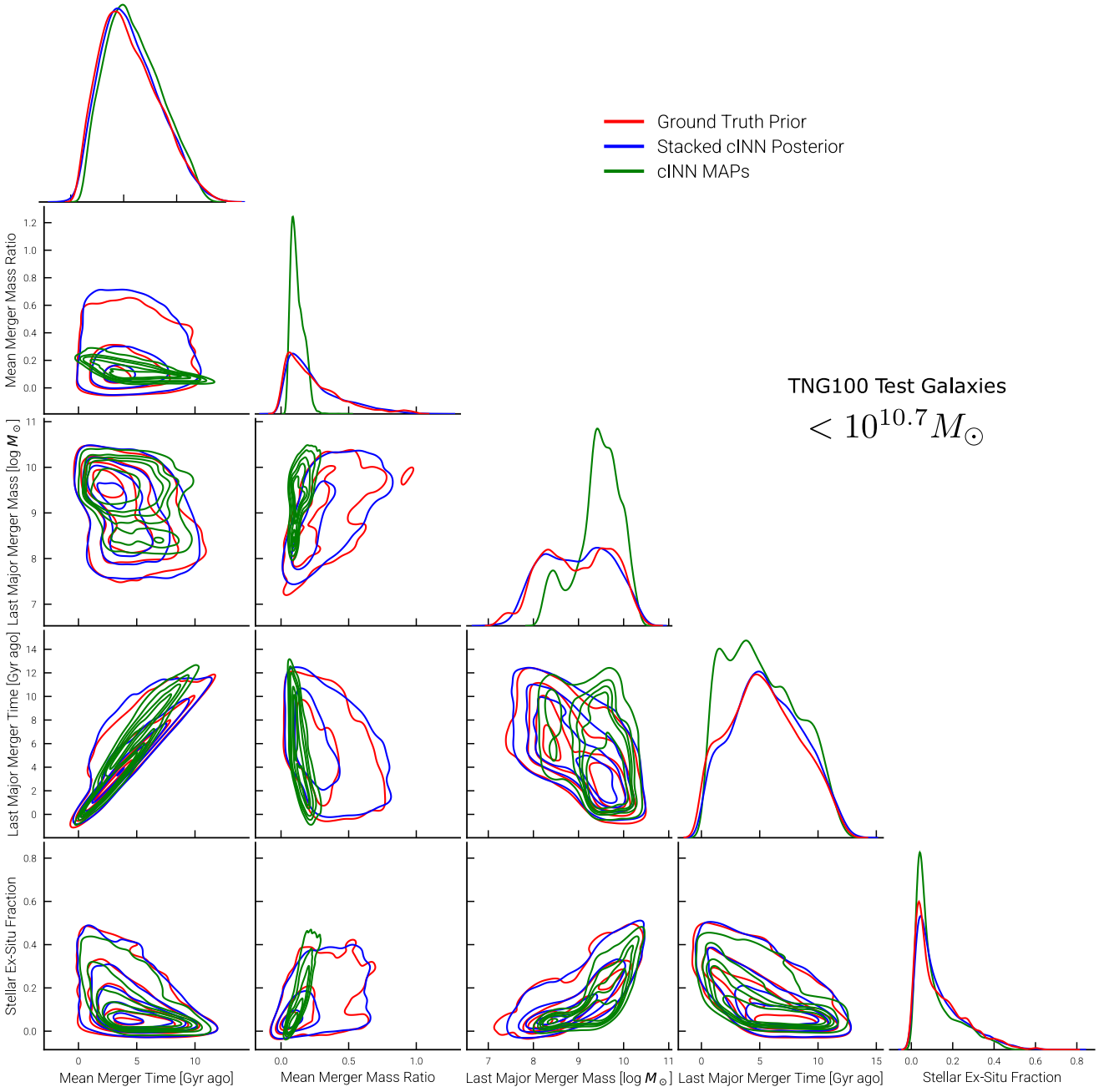


Figure 17: **Recovery of the cross-correlations among the statistics of the stellar assembly and merger history.** We show how test galaxies occupy the parameter space among pairs of the cINN-predicted properties with contours at fixed number density: distributions of the ground truth test sample (ground truth priors, red), the posterior distribution predicted by the cINN trained and tested on the TNG100 galaxies (blue), and the MAP estimates based on these posteriors (green). For the posterior distribution, we include 10 randomly drawn posterior samples per galaxy. Here we focus on galaxies with stellar mass in the $10^{10-10.7}\,M_\odot$ range. The prior distributions are in general well reconstructed by the cINN although a certain level of smoothing is visible especially in the less sampled higher-mass bin: see next Figure. The MAPs are in general heavily biased towards the peaks of the prior distributions; however, also the "point-predictions" from the cINN recover well the underlying true distributions (green vs. red contours) in the case of the stellar ex-situ fraction and the mean merger lookback time.

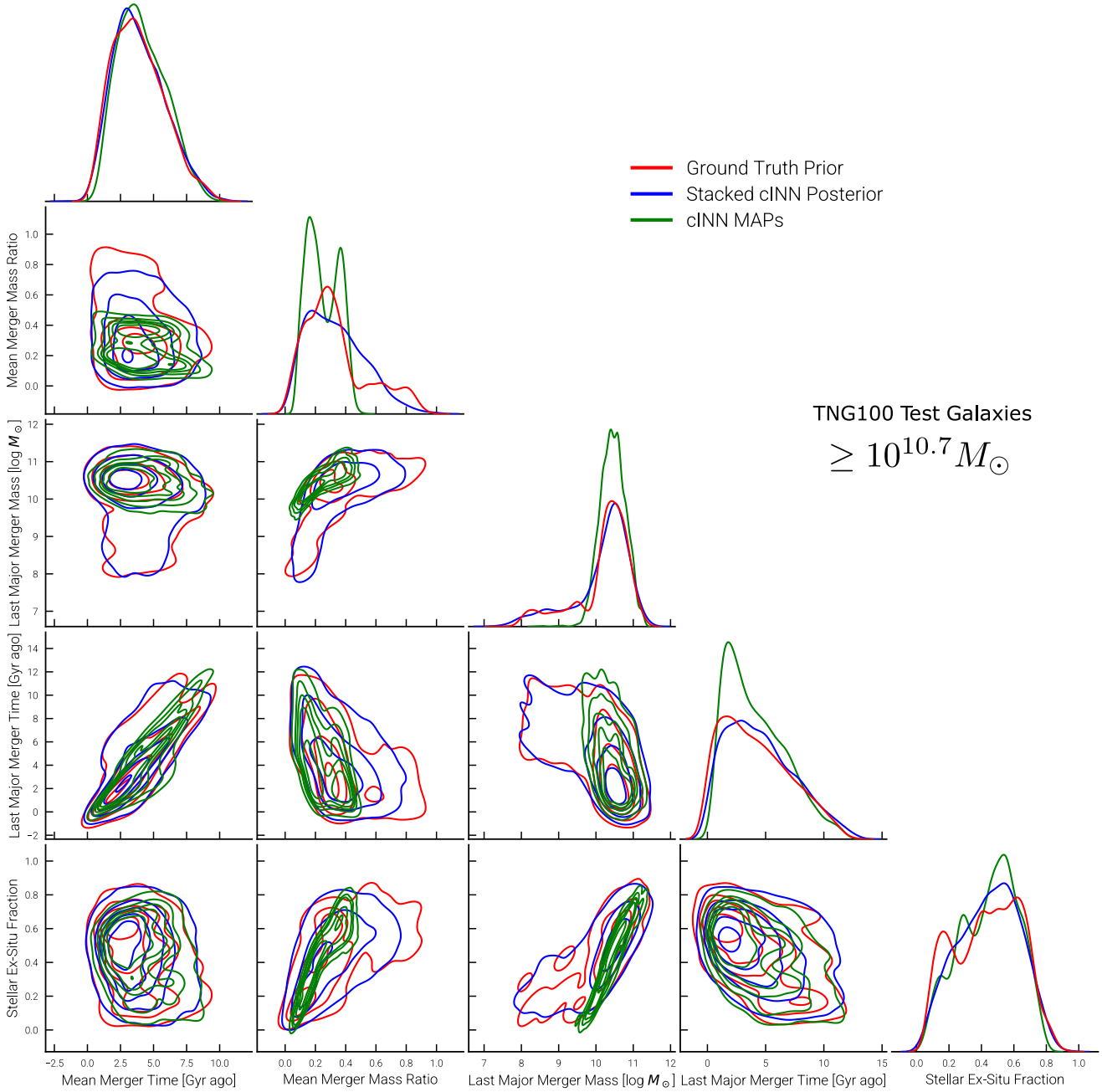


Figure 18: As in Fig. 17 but for high-mass galaxies, with stellar mass in the $10^{10.7-12}\,\mathrm{M}_\odot$ range.

tion steepen for higher-mass galaxies, with the exception of the mean merger lookback time that slightly decouple from the ex-situ fraction in high-mass galaxies.

- The mass and the time of the last major merger are connected, with more recent mergers being typically also more massive.

These correlations – extracted from the TNG100 simulation and successfully recovered by our cINN – are qualitatively in agreement with those predicted by the Illustris simulation and analyzed by [149] and [147], whose works have inspired our selection of stellar assembly and merger statistics.

However, whereas the integrated posteriors follow the general trends of the ground truth, this is not the case for their MAP values: these are somewhat biased and askew distributions, which naturally favour the peak of the posterior distributions (by construction). This is especially pronounced and critical for the mean merger mass ratio of low-mass galaxies, and less so for higher-mass galaxies, for which the MAPs seem to be able to roughly learn the correlation with e.g. the ex-situ fraction and to slightly grasp the trend towards higher-mass ratios. Also in the case of the cross-correlations and consistently with what seen in the cINN performance analysis so far, the MAPs work best for the stellar ex-situ fraction, the mean merger lookback time, and the time of the last major merger. For the other outputs, the MAPs have to be interpreted with caution, e.g. for the mass of the last major merger for massive galaxies with last major merger smaller than $10^{9.5}\,M_{\odot}$.

10.3 SENSITIVITY ANALYSIS: WHAT ARE THE MOST INFORMATIVE INPUT GALAXY PROPERTIES?

We conclude this analysis by looking into the connections that the networks actually learned, i.e. which input is most important to successfully infer which output. To do so, we use a Multilayer Perceptron (MLP), trained and tested on the same dataset like the cINN. We show in Appendix B that there are qualitatively no big differences between the MLP point predictions and the cINN MAPs: therefore, we here utilize the simpler MLP network to identify to what degree the predicted outputs are sensitive to input variations in the model. In practice, for each observable input parameter, we perform the following steps:

1. randomly shuffle the values of the observable across the whole set of test galaxies – in this way, we prevent this parameter from adding any information to the model without changing the overall prior distribution;

2. perform an MLP prediction on this modified sample, while all other observables are kept untouched;

3. measure the mean prediction error (in absolute values) of this modified prediction with respect to the ground truth;

4. subtract the mean error of the original prediction from the mean error of the modified prediction, hence producing a scalar that summarizes the importance of the observable under scrutiny.

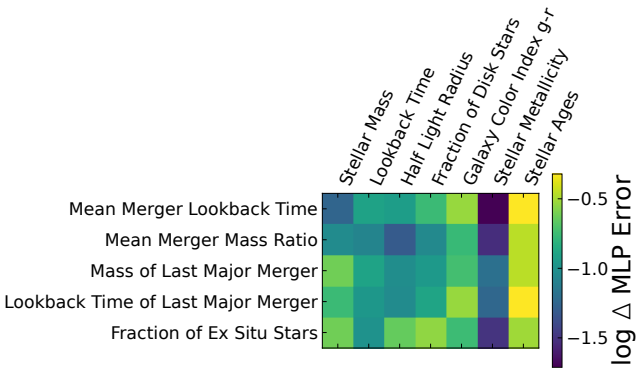


Figure 19: **What inputs are the most informative?** Sensitivity analysis from the MLP model. Each field and color denote the change in the mean prediction error (in absolute value) for the target properties (left) if an observable input (top) is randomly shuffled across the test sample. As the shuffling breaks any information given by the respective input, a shuffled input is considered to be important if the error between the original prediction and the shuffled prediction rises accordingly. As the change in error is evaluated target-wise, it is only comparable for each output property separately but not across targets (i.e. not across rows). We see for example a strong dependence of the ex-situ fraction on the stellar mass and on the morphological/structural parameters such as the disk-to-total ratio and the half light radius. Very prominent is also the influence of the (luminosity-weighted) stellar age on the prediction of most merger statistics.

The resulting quantity measures how the prediction absolute error changes, on average, if we randomly shuffle one of the inputs. As this shuffling breaks any learned connections, a large increase in error means that the input is important for the prediction capability of the model. In Fig. 19, we summarize our findings by evaluating the importance of each observable input, one at the time, and by color-coding the change in prediction errors for each unobservable output. As the changes in prediction errors are evaluated target-wise, the colors of Fig. 19 are comparable only for each target quantity separately but not across rows.

For example, we see that there is a substantial increase in the average prediction error for the stellar ex-situ fraction not only if the stellar mass is shuffled but also (and perhaps even slightly more) if the fraction of disc stars input is shuffled. The strong dependence on galaxy stellar mass is expected, given the correlation between ex-situ fraction and stellar mass already commented upon and shown in Fig. 12: mergers and satellites accretion play a progressively more important role in the stellar mass growth of progressively more massive galaxies [133, 134]. However, Fig. 19 shows that inputting a galaxy's morphology further improves the predictions. This is consistent with previous non-ML analyses of cosmological simulations, which have shown that, even at fixed stellar mass, galaxies with a higher fraction of disc stars (thus disc galaxies) have a lower ex-situ fraction i.e. their stellar inventory is dominated by in-situ star formation [147, 149]. Finally, the prediction of the ex-situ fraction is also sensitive to the galaxy stellar size, consistent with previous simulation-based findings whereby a large stellar half mass radius is also an indicator for a large ex-situ fraction, as ex-situ stars tend to populate larger galactocentric distances [48, 201].

The prediction of the time since the last major merger is strongly affected by the stellar age.

A similar dependence on stellar age is in place for the prediction of the mean merger lookback time. This in truth could also be caused by the fact that galaxies close to $z = 1$ have a more recent merger history and a higher stellar formation rate (thus younger stars). However, in general, as galaxies' average stellar age increase monotonically with galaxy stellar mass above $10^{10}\,M_{\odot}$ both in the Universe and in the TNG100 simulation [122], the merging with less massive objects is destined to reduce the stellar age of a galaxy. Furthermore, the merger of gas rich galaxies may also cause a burst in star formation, hence causing the (luminosity-weighted) stellar age to decrease.

Interestingly, the stellar metallicity of galaxies appears not to be important for the prediction of the chosen stellar assembly and merger statistics. This is at first glance in contradiction with results from simulations, where it has been seen that the merger history also affects the stellar metallicity [113]. We speculate that the same information provided by the metallicity is in fact also contained in the galaxy color and stellar age and we warn the reader that this may be different if focus is placed on the stellar haloes at low-surface brightnesses rather than on the central region of galaxies and on the the metallicity gradients rather than the average stellar metallicity (Zessner et al in prep.).

11 DISCUSSION AND OUTLOOK

In this paper, we have shown that it is in principle possible to infer important information about the past assembly and merger history of galaxies by solely relying on a handful of galaxy features as inputs, such as galaxy stellar mass and redshift, morphology, stellar size and average stellar age. However, we have also shown that not all chosen unobservable properties can be predicted by supervised deep neural networks to a good level of accuracy. Moreover, we have for now trained and tested the method on the data from one cosmological simulation only. In this Section, we hence discuss in more detail the more subtle and complex findings of our work and provide considerations for the future application of these ML models to real galaxy data.

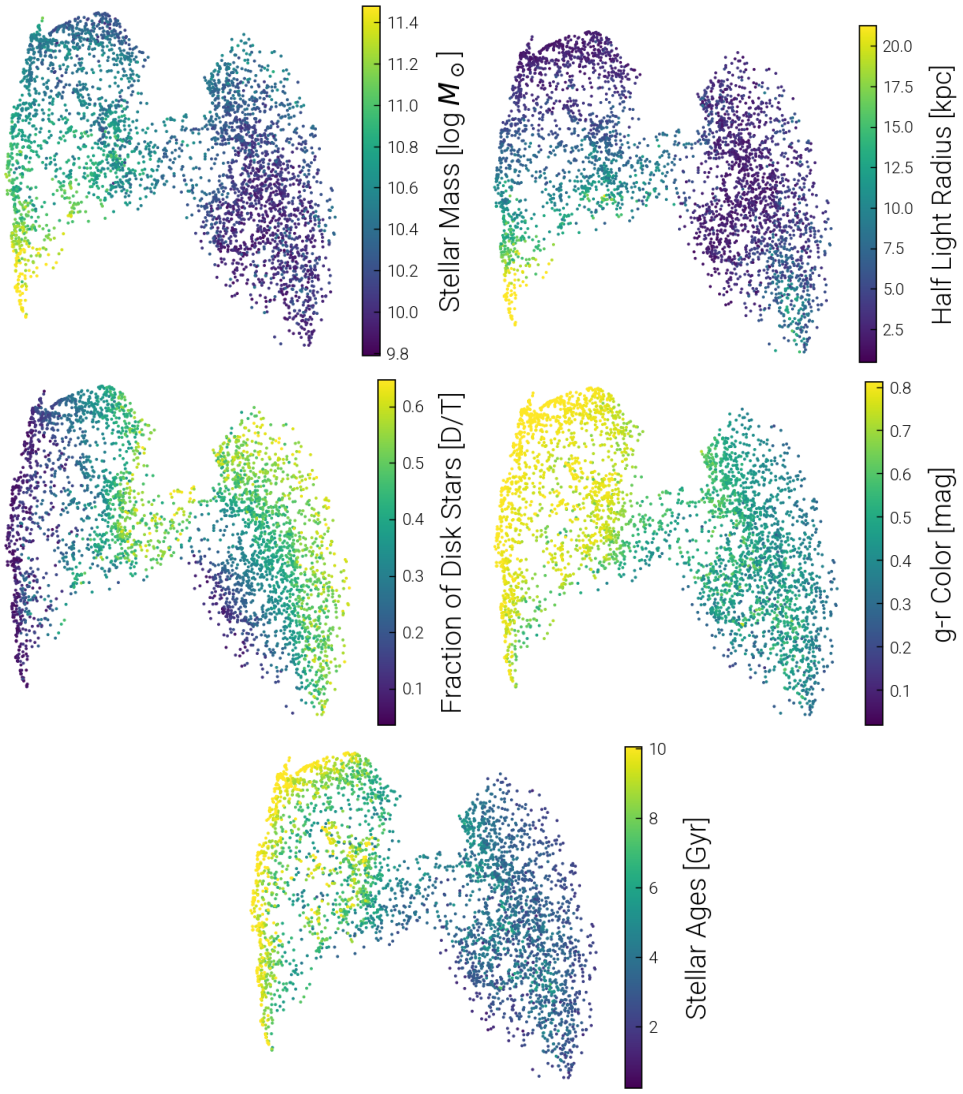
11.1 UNDERSTANDING THE CINN RESULTS

11.1.1 *The case of the mean merger mass ratio*

From the analysis in the previous sections, we have seen that the mean merger mass ratio of galaxies is not predicted well at all by the cINN. More specifically, we have seen that the cINN posterior of the mean merger mass ratio is often similar to the prior distribution. Are these results just caused by an insufficient training or modeling of the networks? Or is the information simply not contained in the list of observables we decided to use as inputs?

To get a rough idea about the nature of this problem, we use UMAP [Uniform Manifold Approximation and Projection for Dimension Reduction, 109] to visualize the 6-dimensional observational space of galaxies at redshift $z = 0$. With this, we get a non-linear embedding into an easily-visualizable 2D space, which is supposed to preserve the overall topology of the dataset as best as possible. We gauge the parameters of this method in such a way that the representation shows a clear bi-modality and also represents a smooth function in stellar mass. The first criterion is reasoned by the expected prominent difference in observables between star-forming and quenched galaxies; the second is justified by our fiducial demand that the representation should be a simple function for a quantity as basic and fundamental as the stellar mass of galaxies.

The UMAP results are shown in Fig. 20, where every data point corresponds to one $z = 0$ galaxy in our overall TNG100 sample. We deliberately omit axes to the panels as they have no immediate physical meaning. To show the connection with the quantities studied throughout, we color code each galaxy (= datapoint) according to one of the observable (top) and unobservable (bottom) properties, one property per panel. For example, looking at the galaxy g-r colors and the stellar ages (panels in the second row from the top), we can see that the 6D space reduces to two sub-regions: the red and old (quenched) galaxies on the left and the blue and young (star-forming) ones on the right. The galaxies with a dominant accretion


11.4
11.2
11.0
10.8
10.6
10.4
10.2
10.0
9.8
Stellar Mass [log $M_\odot$]
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Half Light Radius [kpc]
0.6
0.5
0.4
0.3
0.2
0.1
Fraction of Disk Stars [D/T]
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
g-r Color [mag]
10
8
6
4
2
Stellar Ages [Gyr]

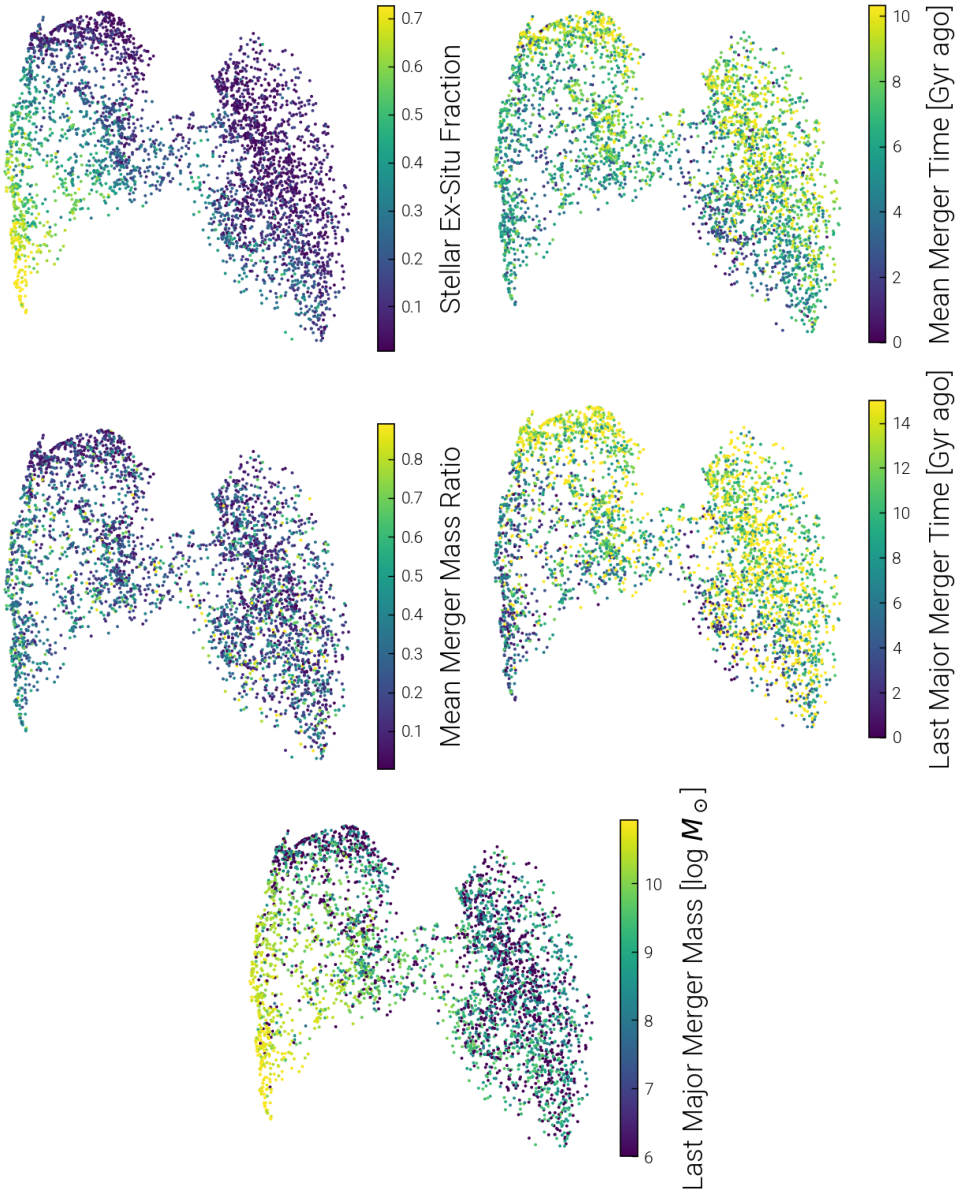


Figure 20: **UMAP representation of TNG100 galaxies at** $z = 0$, that embeds the six-dimensional space of the observable statistics (excluding the galaxies' redshift) into a 2D representation space. We deliberately omit axes to the panels as they have no immediate physical meaning. One dot corresponds to one galaxy. The colors denote, in each panel, the ground truth values from the simulation of the observable inputs (top) and the unobservable statistics of the stellar assembly and merger history (bottom). From top-left to bottom-right: stellar mass, stellar size, fraction of disc stars, g-r color, stellar age, ex-situ fraction, mean merger lookback time and mean mass ratios, and the lookback time and mass of the last major merger. For the last two quantities, galaxies with no major merger in their history are plotted with fixed (unphysical) masses and lookbacktimes according to Section 8.5. To obtain the UMAP representation, here we have used the whole TNG100 sample, including train, validation and test sets. In this representation space, the sample clearly decomposes into star-forming (right grouping) and quenched (left grouping) galaxies. It is remarkable how smoothly and coherently the ex-situ fraction is represented across this 2D space, even if only observable features have been used to create this representation.

and merger history (and hence high ex-situ fractions) are mainly located in the lower edge of the quenched population, with the highest stellar masses. Yet, non-negligible ex-situ fractions can be found also on the left edge of the blue and young galaxies "wing".

The very good cINN predictions of the ex-situ fraction of galaxies, for example, can be understood with the fact that the ex-situ fraction changes smoothly and coherently across this representation space – even if the ex-situ fraction was not used to create it! Similarly, the changes of stellar mass across the UMAP space align well with the ex-situ fraction, and indeed the stellar mass is an important predictor of the ex-situ fraction, as also shown in e.g. Fig. 19. However, in addition to this and as already demonstrated in the previous sections, the cINN is able to capture also more subtle and secondary dependencies. For example, the enhanced ex-situ galaxies in the blue and young regime coincide with lower fractions of disc stars, consistently with the findings of Fig. 19: the morphology of a galaxy, i.e. here the disc to total mass ratio, is especially important to distinguish low and high ex-situ galaxies in the blue regime.

Returning now to the original question of this section, we can see that galaxies with very different mean merger mass ratios can occupy drastically-different locations in the UMAP representation space, with little coherence: for example, low merger mass ratios can be found throughout the UMAP. In our view, this is a strong hint that the set of observables used in this work is not sufficient to fully determine the mean merger mass ratio across galaxies' histories. As also previously pointed out by [149] in the Illustris simulation, a weak correlation with the ex-situ fraction is in place; however there are still many galaxies with low merger mass ratios in the high ex-situ mass regime as well as many galaxies with high merger mass ratios in low ex-situ mass regime. Our cINN model describes all this by returning rather broad posteriors regarding this parameter, which is often similar to the overall prior.

In conclusion, the overall massiveness of all the merger events in a galaxy's history seems to leave a rather ambiguous imprint on the set of observables we have investigated. However, previous studies on stellar haloes in e.g. Illustris [140] have shown that the appearance of stellar halo features such as shells (as opposed to streams) heavily depends on the massiveness of the mergers. We are therefore confident that including light images of galaxies, and in particular images of their dim stellar haloes beyond the central luminous bodies, can further help to constrain the average mass ratio of a galaxy's mergers.

11.1.2 *Stellar mass and time of the last major merger*

In this work, we have aimed at inferring the stellar mass and time of the last *major* merger. These are of course very interesting statistics of the past of a galaxy, as a massive merger is expected to affect its evolution substantially. However, this choice adds great complexity to the analysis due to the fact that, as we have seen, not all galaxies have ever experienced a major merger (i.e. with stellar mass ratio $\geqslant 0.25$) throughout their lifetime. And so we have seen that the cINN posteriors of these merger summary statistics can be very complex (Fig. 14), often with wide plateaus or with multiple probability peaks.

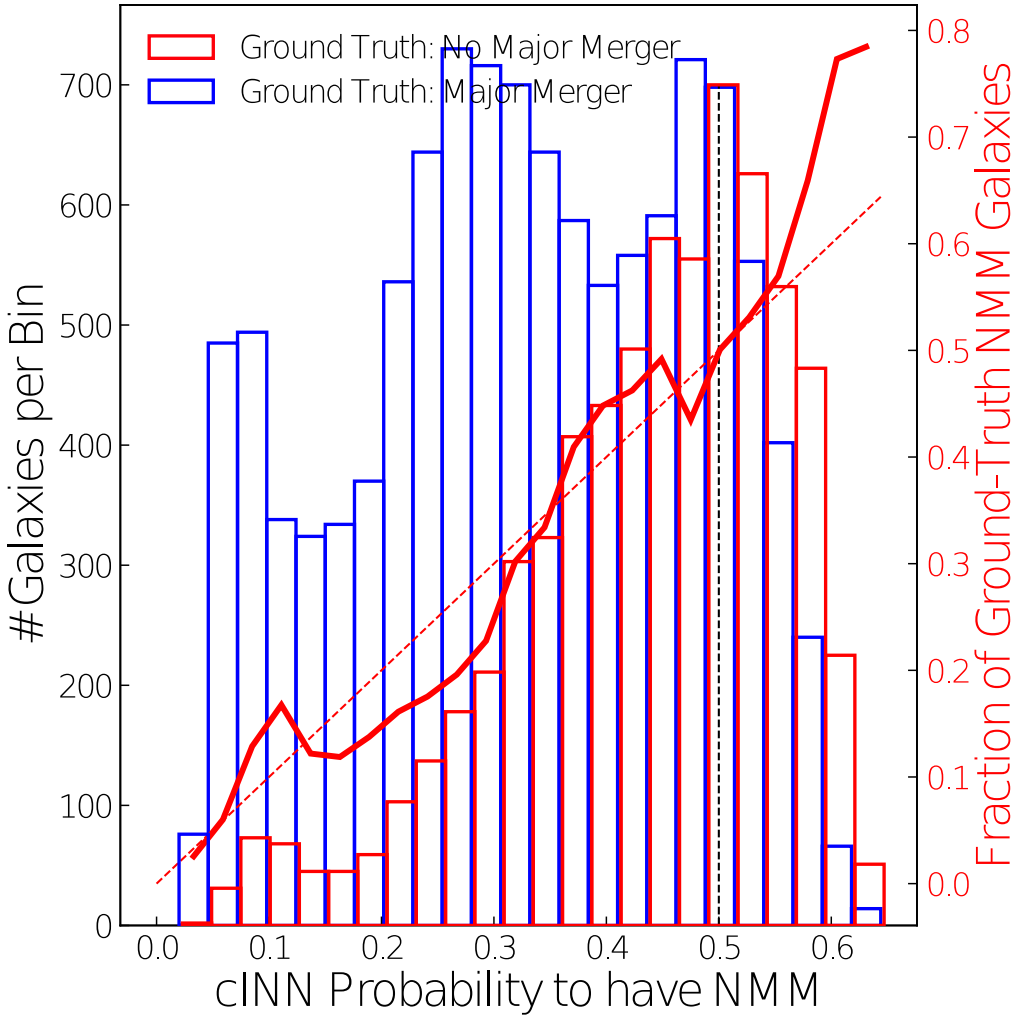


Figure 21: **Capability of the cINN to deal with galaxies with and without a major merger.** We investigate how the fraction of cINN samples that fall into the NMM (No Major Merger) bin, i.e. the probability, relates to the ground truth from the TNG100 simulation. We plot the histograms of these NMM probabilities for test galaxies with no major merger in their ground truth (red) and with at least one major merger in their ground truth (blue). A cINN probability of 0.5 (vertical dotted line) has been chosen as the boundary above which galaxies are *classified* by the cINN to have had no major merger, i.e. we set the MAP estimate to the NMM bin (unphysical time or mass of the major merger) only if more than 50 per cent of the posterior samples fall into this bin. Our expectation is that the fraction of cINN posterior samples in the NMM bin is a good proxy for the fraction of ground-truth galaxies that had no major merger. This is the case: the red thick curve denotes the number of ground-truth NMM galaxies divided by the total number of galaxies in bins of cINN probability. We see that this quantity (red curve) indeed follows the fraction of cINN NMM samples (red dotted line).

This complexity could have been easily avoided should we have chosen to determine, instead, the mass and time since the last merger of whatever type, i.e. with no constraints on the mass ratio or mass. However, we think this may have not been fully interesting from a scientific perspective and might have also revealed ambiguous without an imposition on the minimum stellar mass of the secondary, as a massive merger might be concealed by a more recent minor merger.

Throughout the analysis we have dealt with the fraction of galaxies that never had a major merger (NMM) by de facto including a classification scheme: firstly, we have set the ground-truth parameters of the mass and time of the "last major merger" to unphysical values in the case with no ground-truth major merger (NMM bin); secondly, we have set the cINN MAP estimate to the NMM bin only if more than 50 per cent of the posterior samples fall into this bin. But there are several open questions:

- Have the cINN posteriors successfully learnt the classification between galaxies with and without major mergers?
- For which galaxies is this ambiguity the largest?
- Are the MAP results more accurate if we use the second cINN peak for the galaxies wrongly classified to have never had a major merger?

For the first question, we look in Fig. 21 at the histogram of the probability inferred by the cINN that a galaxy is a NMM (No-Major-Merger) case: we show these probabilities for all TNG100 test galaxies divided in those that according to the ground truth had no major merger (red) and those with at least one major merger in their lifetime (blue). Firstly, it is noticeable that only a small fraction of the test galaxies have been predicted by the cINN to have a probability $\geqslant 50$ per cent to have had no major merger in their history; in fact there is no galaxy with such NMM probability > 70 per cent. Therefore there are no galaxies that had been identified by the cINN to unambiguously be NMM.

We can also see that roughly half of the galaxies with a predicted NMM probability of 50 per cent had indeed no major merger in their lifetime according to the ground truth: red vs. blue histograms at the location of the black dotted line. On the other hand, the galaxies with a low assigned NMM probability are overwhelmingly those with major mergers according to the ground truth, which hence the cINN seems to correctly identify. To further quantify this, we show as red thick curve the ground-truth fraction of the galaxies with no major merger with respect to the total number of galaxies (for each histogram bin). We clearly see that this ground-truth fraction is related to the cINN posterior probability, as the red dotted line denotes the 'ideal' 1:1 relation between cINN probability and the ground truth fraction.

Still, a non-negligible fraction of galaxies are wrongly classified by the cINN. In Fig. 22 we investigate which galaxies (in terms of observable features) are preferred to have no major mergers and which are instead ambiguous regarding this prediction. We split the test galaxies into three classes based on the NMM probabilities given by the cINN: namely, probability to have no major merger < 20 per cent (red), $20-50$ per cent (green), > 50 per cent (blue). The latter are those that in the MAP point prediction are classified as NMM, whereas the sum of the low- and

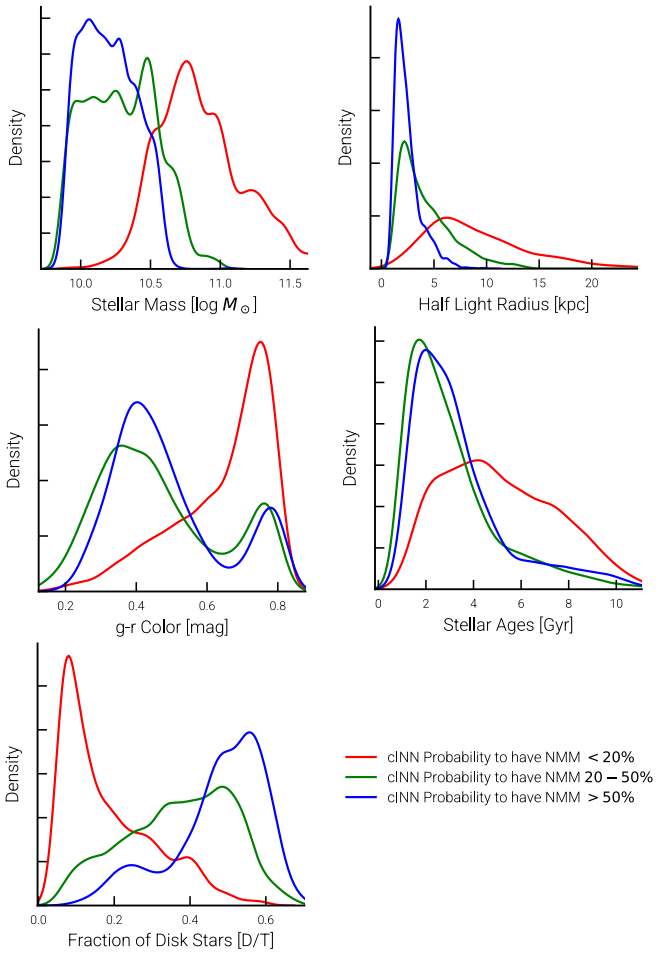


Figure 22: **Which galaxies (in terms of observable properties) have the least ambiguous prediction in relation to having had or not a major merger?** We plot how the TNG100 test galaxies populate the space of the observables when divided in three classes, based on the probability assigned by the cINN to be a No-Major-Merger (NMM) case: < 20 (red), $20-50$ (green), and > 50 (blue) per cent, respectively. The galaxies with higher probability to be classified as NMM (blue curves) are typically lower-mass, bluer and diskier. Galaxies with masses $\gtrsim 10^{11}$ $M_\odot$ are in general assigned by the cINN a very low probability that there was never a major merger in their history (red curves). However, the galaxies with ambiguous cINN probability (green curves) share similar physical properties with those with higher ones. The redshift and the stellar metallicity of galaxies do not exhibit significant biases across these three classes of galaxies, and so they are omitted.

intermediate-probability ones (red + green) gives the population that is predicted to have experienced a major merger. Following the comparison to the ground truth in Fig. 21 and given that there is no NMM probability $\geqslant 70$ per cent, we consider all galaxies with NMM probability in the $20-70$ per cent range to have a rather ambiguous prediction.

Fig. 22 shows that the observable properties of the galaxy populations with low vs. intermediate/high NMM probability are quite different: galaxies with a more ambiguous and higher NMM probability (green and blue curves) are in general lower-mass, disc-like, bluer and somewhat younger than those with very low NMM probability. We interpret that the cINN cannot distinguish the ambiguity between a) isolated disc-like and blue galaxies that never had any/recent major merger events vs. b) disc galaxies that had major mergers but were able to retain their disc shape thanks to the continuing star formation, the latter being sustained or even enhanced by the gas available during the merger (wet merger scenario). In contrast to that, we find that more massive, redder, and more elliptical galaxies are identified by the cINN to have overwhelmingly experienced major mergers, in agreement with the expectations from cosmological simulations.

In relation to the third aforementioned question, we notice that more accurate and less ambiguous results for the last major merger would have been obtained if we had chosen as cINN point prediction the best posterior peak, namely the one closest to the ground-truth value, instead of the highest peak of the posterior distribution (the MAP, used throughout). This alternative approach is justified by the fact that the relevant posteriors are often bimodal or with multiple peaks. In such a case, the picture of Fig. 15 turns into better fits. Namely, for the mass of the last major merger, we get a good median prediction down to progenitor masses $\gtrsim 10^{8.5}\,\mathrm{M}_\odot$ instead of the limit of $\gtrsim 10^{9.5}\,\mathrm{M}_\odot$ from the MAPs. For the time since the last major merger, the scatter in prediction accuracy is smaller, particularly for galaxies with more ancient (ground-truth) major merger. Clearly, when applying this model to observational data, there is no possibility to choose the "best peak". However, this analysis shows that also the complex major merger posterior distributions and the resulting major merger classification is feasible and physically interpretable.

11.1.3 *Dependence on the shape of the TNG100 prior*

Finally, a few points of discussion are needed in relation to the influence of the TNG100 prior distribution on the posteriors of the cINN. In a number of occasions, we have seen that the underlying prior strongly biases the cINN predictions. For example, firstly, the posteriors of the mean merger mass ratio are mostly equal to the prior itself (Figs. 14) and, as expected, the mean merger mass ratio is unconstrained. Moreover, the ex-situ fractions tend to be under-predicted and to be biased towards the prior peak, which peaks at lower ex-situ fractions. Finally, for most unobservable target quantities, the MAPs are in general shifted towards the prior peaks (Figs. 17 and 18).

Whether these behaviours are generally desirable is not clear: on the one hand, the cINN uses the knowledge of the unobservable parameters and their interrelationships from the overall TNG100 distributions (e.g. that low ex-situ galaxies

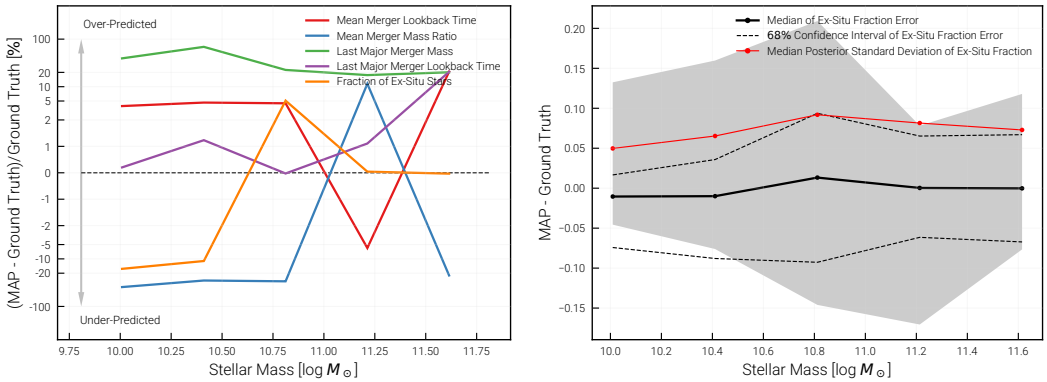


Figure 23: **Prediction performance as a function of galaxy mass.** We investigate how the MAP error of our cINN model (MAP - Ground Truth) depends on galaxy stellar mass, by using all galaxies in our TNG100 test set in 5 uniformly-spaced mass bins. **Left**: Median relative errors (i.e. normalized by the ground truth values on a galaxy-by-galaxy basis) for the five inferred unobservable statistics of the assembly and merger history of galaxies. For the time and mass of the last major merger, we neglect galaxies without a major merger in their history. **Right**: Errors (MAP - Ground Truth) for the ex-situ fraction as a function of galaxy stellar mass. The median error is drawn as solid black line while the area containing 68% of the galaxies (1σ) is included within dashed lines. For comparison, we also plot the median of the posterior standard deviations given by our cINN model as a solid red line. To quantify how our results improve upon the bare relationship between galaxy mass and ex-situ fraction that could be extracted directly from the TNG100 simulation, we plot as grey shaded area the 68 per cent confidence interval of the prior distribution in the ex-situ vs. mass plane (i.e. the 16th-84th percentiles in bins of stellar mass of Fig. 13, rightmost panel, having subtracted the median value).

are much more common than high ex-situ galaxies). On the other hand, however, this could be a problem when applying the resulting network to other cosmological simulations or even observations, i.e. when the shape of the underlying prior is not necessarily identical to the one used for the training. In this work, we have chosen to not re-normalize the resulting cINN posteriors to a flat prior; i.e. by dividing the cINN posteriors by the TNG100 prior. We leave the assessment of e.g. such an alternative approach as an open point for future work, where we will also compare the predictions from different cosmological simulations and therefore bracket the uncertainties associated to the galaxy-formation model and the numerical resolution.

11.2 FOR WHICH GALAXIES CAN WE GET THE BEST PREDICTIONS ON THEIR PAST HISTORY?

We have already quantified the overall performance and predictive capability of the cINN model in Section 10.2. However, we have not yet explicitly discussed whether these inferences are more promising for certain classes of galaxies than others. This is of relevance in view of the future application of this framework to

actual galaxy observations, which may be magnitude (i.e. mass) limited or which may reach different galaxies at different redshifts.

In terms of redshift dependence, Fig. 19 of Section 10.3 shows that the redshift of a galaxy (its lookback time) is not a crucially-informative input for the accuracy of the predictions. This statement applies to galaxies up to $z = 1$, as we have focused exclusively on the $z = 0 - 1$ range. In turn, this implies that we expect to be able to predict similarly well the past assembly and merger history of galaxies in the low-redshift Universe, where data from e.g. SDSS, DES, and HSC-SSP abounds, as well as up to 8 billion years ago.

In terms of mass dependence, we have seen by comparing Figs. 17 and 18 that the cINN is capable of recovering the overall cross-dependencies among predicted features to similar degrees for the low and high-mass end of the studied sample. We build upon this and quantify the errors on the predictions as a function of galaxy stellar mass in Fig. 23. As we could not find any significant variations with redshift (see above), here we show average trends for all galaxies at $z = 0 - 1$ combined. We show the errors (MAP - Ground Truth) normalized by the ground-truth values for all inferred statistics in the left panel, and the non-normalized (MAP - Ground Truth) for the ex-situ fraction only on the right. The medians are given as solid curves in bins of galaxy stellar mass; the area containing 68 per cent of the galaxies around the median is between dashed lines.

Firstly, Fig. 23 explicitly demonstrates that – except for the mean merger mass ratio and for the mass of the last major merger for low-mass galaxies – the cINN predicts the properties of the past assembly and merger histories of the average galaxies with a relative prediction error (compared to the ground truth) that is better than $10 - 20$ per cent throughout the $10^{10-12}\,M_\odot$ range. The accuracy of the prediction depends differently on galaxy stellar mass for different unobservable quantities, and e.g. weakly for the mass and time of the last major merger, which are typically overpredicted – see for comparison the absolute values of the prediction errors in Fig. 15.

The ex-situ fraction – the unobservable that is best predicted by the cINN based on the analysis thus far – in fact performs significantly worse for galaxies with stellar mass $\lesssim 10^{10.7}\,M_\odot$ and essentially perfectly (relative prediction errors smaller than a few percent) for galaxies with stellar mass $\gtrsim 10^{11}\,M_\odot$. This is not in contradiction with what we have seen so far: in terms of non-normalized prediction errors (right panel of Fig. 23 and Fig. 15), the cINN predicts on average the ex-situ fraction within $\pm$ 2.5 percentage points (solid black curve) independently of galaxy stellar mass. However, because of the strong correlation of ex-situ fraction with stellar mass, the median prediction error is much larger relatively to the ground truth for low-mass galaxies, where the ex-situ fraction is low and typically under predicted (see also Section 10.2.2, where we found the MAP to be smaller than the ground truth) .

Moreover, below $\lesssim 10^{10.7}\,M_\odot$, where typically in-situ star formation is the dominant stellar mass growth channel, the strong dependence between ex-situ fraction and galaxy mass is missing (see e.g. Fig. 13): although the cINN is still able to distinguish low- from high-ex-situ galaxies, in the case of low-mass galaxies the ambiguity hampers the ability to make an accurate point prediction and the MAP is influenced by the prior, which overall favours low values of ex-situ fractions. This

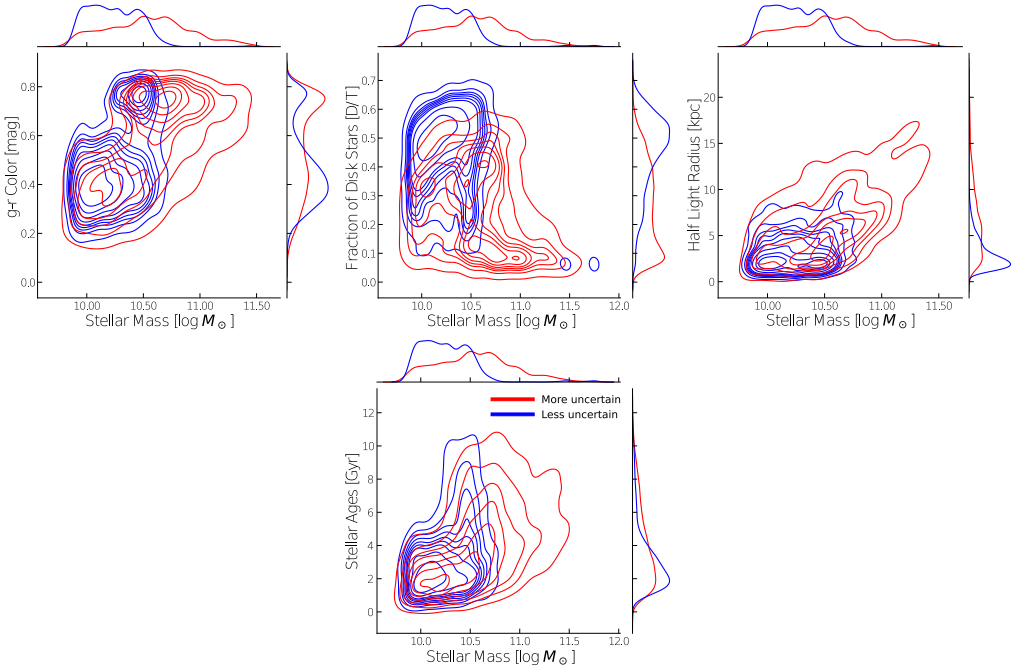


Figure 24: **Which galaxies (in terms of observable properties) have the least ambiguous assembly and merger history?** We answer this question by electing the ex-situ fraction as example. We separate the TNG100 test set in two subsets (stacking all redshifts): galaxies with a predicted ex-situ cINN posterior standard deviation that is below (blue) and above (red) the median of the whole test sample. The blue subset hence represents the galaxies whose predictions are less uncertain than those in the red subset. We show how these two sets populate the space of stellar mass vs. (from left to right) g-r colour, fraction of disc stars, half-light radius and average stellar age. In the plot edges, we show the projections of the two distributions on the x and y axis. We see that, for the ex-situ fraction, the cINN returns especially small error bars in the regime of low-mass, blue, and disc-like galaxies.

is also expressed in the confidence interval on the right panel of Fig. 23 (dashed curves), which is clearly biased towards an under-prediction of the ex-situ parameter at lower masses.

As there is a strong relationship between galaxy mass and ex-situ fraction visible in the rightmost panel of Fig. 13, we also investigate to what degree the cINN (using all 7 observables) performs better in comparison to what could be inferred from this bare connection. To to so, we also plot as grey shaded area the 68 per cent confidence interval of the prior distribution in the ex-situ vs. mass plane (i.e. the 16th-84th percentiles in bins of stellar mass). From this we see that the cINN posterior MAP error is significantly smaller than the scatter in the prior data.

To see for which galaxies, distinguished based on their observable properties, we get predictions with a low uncertainty and therefore low ambiguity, we move from quantifying the prediction error (i.e. the accuracy) to assessing the width or standard deviations of the cINN posterior distributions (i.e. the precision). We do this again focusing on the ex-situ fraction and plot in Fig. 24 how two subsets of the TNG100 test galaxies populate the observable space: i) the ones whose cINN posterior standard deviation for the ex-situ fraction is above the median of the test sample and whose predictions are hence affected by larger uncertainties (red) and ii) those with below-median standard deviation and whose prediction is hence less uncertain (blue).

It is manifest that the errorbars on the ex-situ fraction predictions tend to be indeed smaller for low-mass, blue, disc-like, and compact galaxies, which corresponds to the low ex-situ regime (see also Fig. 16 and our discussion in Section 10.2.3). Namely, it is easier for the cINN to identify the galaxies in the sample with a rather quiet merger history whereas the ambiguity rises with the level of past interactions.

These results would be qualitatively similar, albeit not identical, if we were to replace the posterior standard deviation (i.e. the prediction uncertainty) by the prediction error (MAP - Ground Truth). This aligns well with our findings in Fig.16 where we have shown that the standard deviation of the posterior is small for low ex-situ galaxies and that the standard deviation correlates well with the prediction error.

Finally, the distributions of Fig. 24, which focus on the ex-situ fraction, are qualitatively similar for the case of the mass of the last major merger: less ambiguous and uncertain predictions occur for lower-mass, bluer, diskier, more compact and younger galaxies. For the mean merger lookback time and the lookback time of the last major merger, on the other hand, we find a strong dependence on the redshift of the galaxy under scrutiny: for galaxies at earlier redshifts, the physical available timespan for merging is naturally shorter than for galaxies observed close to $z = 0$ and therefore the prediction errorbars are much smaller.

11.3 TOWARDS THE APPLICATION TO REAL OBSERVATIONS

Having demonstrated with the TNG100 simulation the feasibility of inferring unobservable properties of galaxies' past history from a handful of observable galaxy features, we plan in future work to apply these or similar types of ML models to real observational data. This will require quantifying and marginalising over the

systematic uncertainties due to adopting different cosmological galaxy simulations for the predictions, such Illustris, EAGLE, and others.

With the scalar-to-scalar models developed in this work, we are especially aiming for large-scale photometric galaxy surveys, such as the aforementioned SDSS, GAMA, or HSC-SSP, which can provide multi-band images and integral properties of large samples of galaxies. In fact, as discussed in Section 11.2, the framework can be used not only at low redshift, but similarly successfully also up to $z \sim 1$.

We have already designed our input observable parameters (stellar mass, g-r color, half-light radius, r-band weighted average age, etc) in such a way that they are (in principle) derivable from existing galaxy datasets, either via conventional modelling based on the images or via ML approaches as done by e.g. [128]. The only observable that we have used and whose operational definition is in fact far from what is possible with photometric surveys is the disk-to-total ratio, which proves to have important predictive power, especially for the stellar ex-situ fraction. Here we have obtained this "morphological estimator" from the kinematics of the stellar particles, whereas in reality either IFU data is needed or Sersic-index based morphological estimators shall be used.

We remind the reader that we train and evaluate the method with synthetic data without additional measurement uncertainties. However, real observations are affected by noise or systematic uncertainties. We will address this in detail in a follow-up study that includes realistic noise models. We note that observational uncertainties are connected to a loss in information and therefore lead to a more uncertain interference than the one presented here. In practice this means that the resulting posterior distribution functions become wider and that the intrinsic astrophysical degeneracies are more difficult to identify and resolve. The results in Section 10 can be therefore viewed as a best-case scenario: what information could one infer if the galaxy properties are exactly known? Realistic posteriors, including the observational noise, can be then constructed from the 'ideal' posterior by integrating over the uncertainty distributions of the observed parameters. This integration could e.g. be done by a Monte Carlo sampling: i.e. by drawing random samples for the conditional input of the cINN from a distribution describing the measurement uncertainty.

In upcoming work, we plan to completely bypass the issue of deriving and choosing summary statistics of galaxies (including their uncertainty distributions) by directly using the full images, possibly in multiple bands. Within our framework, this will require introducing an additional encoder head (e.g. a convolutional neural network) that translates the maps into an intermediate lower-dimensional representation space and which hence can then be used as conditional input to the cINN. In turn, this will allow us to exploit the full information content of the images, including those reaching low surface brightness levels and hence revealing stellar halo features (see e.g. deep surveys with the Dragonfly Array, with HSC, and LSST). Even more so, realistic mock images of the simulated galaxies will be of the essence, including the effects induced by the survey-specific resolution, point-spread function, redshift-effects, and artifacts. Albeit demanding, such stellar light mocks from simulated galaxies are now achievable [e.g. 24, 110, 151, 164, 180], and have already been used for diverse ML-based analyses [e.g. 22, 86, 200].

12 SUMMARY AND CONCLUSIONS

In this work, we have used the cosmological magnetohydrodynamical galaxy formation simulation TNG100 of the IllustrisTNG project to investigate the connection between integral properties of galaxies at one point in time, on the one side, and their past stellar assembly and merger history on the other. In particular, we have trained and tested a conditional Invertible Neural Network (cINN) to determine to what degree summary statistics of the unobservable past of galaxies can be predicted from a handful of galaxy features that are in principle available in large photometric surveys.

We have focused on TNG100 central galaxies in the $z = 0 - 1$ range with stellar masses of $10^{10-12}\,M_\odot$, for a total of 182'625 galaxies. We have made sure that the progenitors and descendants of each test and validation galaxy are not contained in the training set, to reduce the additional information that the tree-like structure of the simulation data may provide.

As observable inputs, we have chosen seven integral galaxy properties, namely stellar mass, redshift, stellar half-light radius, fraction of disc stars, integrated dust-attenuated galaxy color (SDSS g-r color index), luminosity-weighted stellar average metallicity and luminosity-weighted average stellar age (Fig. 12). To quantify the stellar assembly and merger history of galaxies, we have elected five summary statistics, namely the stellar ex-situ fraction (i.e. the fraction of accreted stars via mergers and satellite stripping), the mass-weighted average merger mass ratio and the mass-weighted average merger lookback time throughout a galaxy's history, and the stellar mass and time of the last major merger (if any; Fig. 13).

Thanks to the cINN, we have been able to evaluate not only the accuracy of the predictions (Figs. 14, 15, and 23) but also to estimate their uncertainty, via the width (standard deviation) of the posterior probability distributions (Fig. 16). We have dealt with galaxies that never had a major merger in their past by de facto including in the modeling a classification scheme. And we have interpreted our findings by exploring the relationships between inputs and outputs via a UMAP representation of the simulated galaxies (Fig. 20).

We have shown that it is in principle possible to infer important information about the past stellar assembly and merger history of galaxies by using solely scalar galaxy properties as inputs. In particular, we have shown that – except for the mean merger mass ratio – it is possible to reconstruct the studied unobservable properties well, i.e. with a median prediction error relative to the ground truth that is better than $10 - 20$ per cent throughout the considered mass and redshift range (Fig. 23; for the mass of the last major merger this statement applies only for galaxies with $\gtrsim 10^{10.7}\,M_\odot$). The fraction of ex-situ stars is recovered particularly successfully throughout the galaxy population: the cINN posterior maxima (Maximum A-Posteriori estimates or MAPs) align well around the ground truth values (Fig. 15); furthermore, the standard deviation of the ex-situ posterior behaves like a normal-distributed error (Fig. 16). In contrast, we have found that, given the

adopted set of observable inputs, it is not possible to predict the mean merger mass ratio (Fig. 15): the cINN posteriors are often close to the prior distributions of this parameter (Figs. 14, 17 and 18). Similarly as for the ex-situ fraction, the predicted average merger lookback time aligns well with the ground truth for the bulk of the galaxies. On the other hand, for the mass and lookback time of the last major merger, some caution is needed due to the shape of the posteriors, which are often complex or exhibit strong bimodalities – this being especially driven by the ambiguity with galaxies that never had any major mergers in their history. Yet, for those galaxies that the cINN correctly identifies as having experienced a major merger, the predicted time and stellar mass of the merger are within ± 1 billion years and ± 0.1dex from the true values for the average galaxy (but only if the merger is more massive than about $10^{9.5}\,M_{\odot}$ in stars; Fig. 15).

Importantly, the cINN is also able to learn the cross-correlations among the chosen target quantities (Figs. 17 and 18), with e.g. larger ex-situ fractions in galaxies with more massive and more recent last major mergers, also at fixed galaxy mass.

Using a Multilayer Perceptron (MLP), which is equivalent to the cINN in terms of point predictions (Figure 49), we can easily uncover which of the adopted galaxy properties are most informative for the prediction of the various target quantities (Fig. 19). We find that well-known correlations are learned by the networks, e.g. that the stellar ex situ fraction is larger for larger galaxy masses and hence stellar mass is a fundamental input. However, the networks also learn more subtle and residual dependencies, so that the stellar size and morphology of galaxies are also important predictors of the ex-situ fraction. The luminosity-weighted stellar age plays in general an important role for the inference of all merger statistics, whereas the luminosity-weighted average stellar metallicity is not essential.

Finally, our framework is able to predict similarly well the past assembly and merger history of galaxies both in the low-redshift Universe as well as up to $z \sim 1$. A current limitation of this work is that we have performed the evaluation of the model on data without any observational uncertainty. In future work, we plan to replace the set of pre-selected observables with survey-realistic images, which automatically also account for observational uncertainties. Hence, our tools have the potential to open up a powerful new avenue to quantify the stellar assembly and merger history of thousands of *observed* galaxies across cosmic epochs.

Part IV

COMPARING ILLUSTRISTNG AND HSC GALAXY IMAGES VIA CONTRASTIVE LEARNING

The following chapter is based on the work 'ERGO-ML: Comparing IllustrisTNG and HSC galaxy images via contrastive learning' which was published in MNRAS [61] as part of the PhD. Modern cosmological hydrodynamical galaxy simulations provide tens of thousands of reasonably realistic synthetic galaxies across cosmic time. However, quantitatively assessing the level of realism of simulated universes in comparison to the real one is difficult. In this paper of the ERGO-ML series (Extracting Reality from Galaxy Observables with Machine Learning), we utilize contrastive learning to directly compare a large sample of simulated and observed galaxies based on their stellar-light images. This eliminates the need to specify summary statistics and allows to exploit the whole information content of the observations. We produce survey-realistic galaxy mock datasets resembling real Hyper Suprime-Cam (HSC) observations using the cosmological simulations TNG50 and TNG100. Our focus is on galaxies with stellar masses between 10^9 and $10^{12} M_\odot$ at $z = 0.1 - 0.4$. This allows us to evaluate the realism of the simulated TNG galaxies in comparison to actual HSC observations. We apply the self-supervised contrastive learning method NNCLR to the images from both simulated and observed datasets (g, r, i - bands). This results in a 256-dimensional representation space, encoding all relevant observable galaxy properties. Firstly, this allows us to identify simulated galaxies that closely resemble real ones by seeking similar images in this multi-dimensional space. Even more powerful, we quantify the alignment between the representations of these two image sets, finding that the majority ($\gtrsim 70$ per cent) of the TNG galaxies align well with observed HSC images. However, a subset of simulated galaxies with larger sizes, steeper Sersic profiles, smaller Sersic ellipticities, and larger asymmetries appears unrealistic. We also demonstrate the utility of our derived image representations by inferring properties of real HSC galaxies using simulated TNG galaxies as the ground truth.

13 INTRODUCTION

The morphological structures of observed galaxies are rich in both complexity and diversity. In the nearby Universe, the stellar structures of some galaxies can be well-characterized by simple analytic models, e.g. smooth ellipticals and flattened disks. Meanwhile, the morphological menagerie also includes galaxies with tightly/loosely wound spiral structures, stellar streams and shells, lopsidedness, and other generally complicated structures. The categorization and characterization of these stellar morphologies, in conjunction with theoretical models, has yielded significant insight into their evolutionary pathways [31, 40, 108].

Modern cosmological hydrodynamical simulations of galaxies, like e.g. IllustrisTNG (see below), EAGLE [44, 155], HorizonAGN [91], and SIMBA [47], are nowadays able to produce diverse morphological galaxy populations that resemble those seen in observations. However, directly and quantitatively comparing the morphologies of simulated galaxies to observed ones is not a simple task.

Traditionally the comparison between simulated and observed galaxies has been done by deriving scalar properties or profiles that summarize the high dimensional observed data, e.g. images and spectra. These scalar properties can be measured for the simulated galaxies and then validated against the set of observed ones [e.g. 44, 73, 155, 188]. However, whereas the structures and summary statistics of observed galaxies are derived from images that are based on light and that necessarily include observational limitations (such as depth), additional steps are required to extract equivalent and self-consistently derived structural quantities from the simulation data. This has been achieved in recent years by constructing mocks of simulated galaxies as if they were observed in actual surveys [e.g. 18, 25, 150, 153, 163, 178]. In turn, these mocks have been used to extract scalar properties and summary statistics using the same methods as observers [e.g. 6, 8, 75, 79, 101, 110, 164, 203].

Notwithstanding the progress, the 'traditional way' of comparing simulated and observed galaxies still suffers from the limitation of restricting the comparisons to a limited set of scalar properties or summary statistics: important information content or cross correlations in the data may be missed. This caveat can be avoided at once by comparing simulations and observations directly at the map level: i.e. by contrasting full images of real galaxies to survey-realistic mocks of simulated galaxies. This ensures not only that all the available information contained in the images is used, but also bypasses the non-trivial choice and construction of summary statistics on which to frame the comparison: e.g., galaxy sizes [204], light profiles [e.g. Sersic profiles, 76, 158], morphological estimators [e.g. Gini-M20 parameters, 102], stellar colors [94], etc.

Recently, machine learning methods have been proposed to cope with the complex task of evaluating and comparing simulated and observed images [e.g. 85, 104, 131, 191, 199]. Especially of interest are the class of so called 'contrastive learning' methods [see 84, for an overview], which are self-supervised algorithms able to

identify features of interest from e.g. images without prior labeling of the galaxies [184]. In doing so, it is therefore possible to circumvent the difficult task of manually identifying and calculating labels overall.

In this paper, we build upon the prior successes of machine learning applied to galaxy images by implementing a novel contrastive learning method on a vastly larger set of observed and simulated galaxies. Our method accomplishes this by clustering images that have common features and repelling images which do not. As such, the model naturally identifies points of tension (i.e. contrasts) and commonalities within the imaging data. We simultaneously apply the method to images of galaxies from the Public Data Release 3 of the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP PDR3, [3]) and from HSC-SSP mock images of galaxies selected from the IllustrisTNG cosmological magneto-hydrodynamical simulations [TNG hereafter, 106, 116, 121, 122, 134, 136, 171]. By exposing the model to both real galaxy images and mock images of simulated galaxies simultaneously, instead of training the model on simulations and observations separately, we give the model every opportunity to focus on the potential differences between observations and simulations, as the model is designed to seek points of contrast between images. In practice, in our methodology, each galaxy image is placed into a high-dimension representation space that can then be explored for comparing the simulated and observed data sets. We do so by introducing an Out-of-Domain score for each image and by quantifying and comparing the score distributions across the galaxy populations. We also further reduce the representation spaces of the galaxy datasets into 2-dimensional UMAPs and visually inspect and compare them and connect them to known galaxy properties.

The method developed and discussed in this paper is especially important for the broader scope of the ERGO-ML (Extracting Reality from Galaxy Observables with Machine Learning) project, whereby we aim at inferring unobservable physical features of galaxies from observations. In two previous papers we have in fact shown that it is possible to infer posteriors for unobservable features, such as the mass fraction of accreted vs. in-situ stars [7, 160], by using a limited set of observable galaxy scalars [63]. We have shown that this is possible in the context of simulated galaxies, from e.g. the TNG100 simulation of the aforementioned TNG project. In order to extend this harvesting opportunity to real galaxies and to their image data, we first need to investigate and evaluate how similar images from simulations and observations are, i.e. if their domains are comparable.

The overall scope of this work is shown in Figure 25. First we introduce the data sets and the preparation steps in Section 14. We then introduce in Section 15 the Machine Learning Methods used throughout. We hence present the results of our study in the following Sections: an overview of the galaxy features that have been learnt by our self-supervised model in Section 16.1 and the quantitative comparison between TNG (simulated) and HSC (observed) galaxies in Sections 16.2 and 17. We summarize ideas and findings in Section 18.

In this work we use a cosmology consistent with the results from [137], $\Omega_{\Lambda,0} = 0.6911$, $\Omega_{m,0} = 0.3089$, $\Omega_{b,0} = 0.0486$, $\sigma_8 = 0.8159$, $n_s = 0.9667$ and $h = 0.6774$.

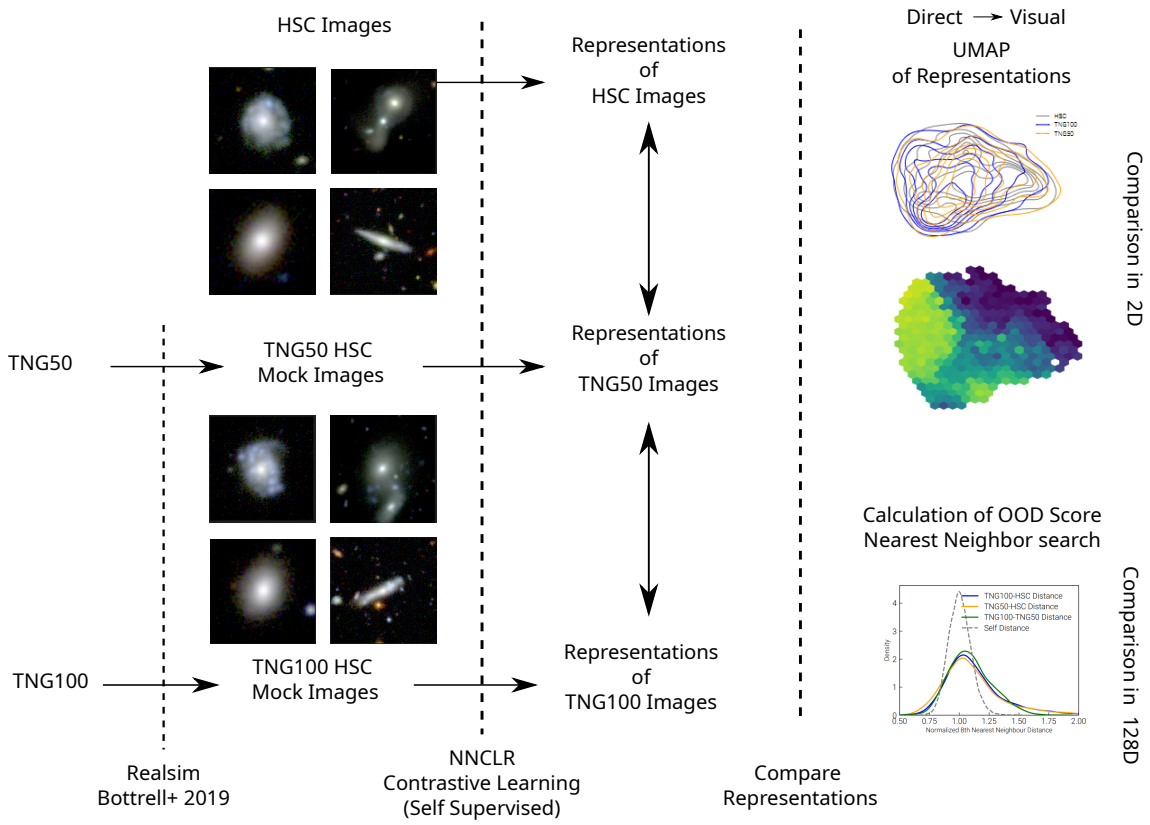


Figure 25: **Overview over the method proposed in this paper to compare observed and simulated galaxies at the image level.** We use a set of Hyper Suprime Cam (HSC) galaxy images and survey-realistic mocks of TNG50 and TNG100 simulated galaxies. We then use them together to train the contrastive learning method NNCLR, in order to map all images into the same representation space. This allows us in the following to compare the samples either in the representation space itself, via an Out-Of-Domain score and its distributions, or visually in a further dimensionally-reduced 2D UMAP representation.

14 GALAXY IMAGE DATA

Three distinct sets of galaxy images are used in this work:

1. Observed Hyper-Supreme-Cam images of real galaxies: Section 14.1.
2. Survey-realistic mock images of galaxies simulated within the TNG100 simulation: Section 14.2.
3. The same as above but using galaxies from the higher-resolution TNG50 simulation, which adds galaxies especially at the low-mass end of our sample: these are also explained in Section 14.2.

By using two flagship runs of the IllustrisTNG simulation model with different resolutions, we can a) build intuition about the meaning of the results via relative comparisons and b) investigate how simulation resolution affects the visual appearances of galaxies. In the following subsections, we detail the datasets and steps necessary to prepare the raw data for our machine learning model. It is worth noting that we apply exactly the same steps to all three datasets: this ensures that we do not introduce any artificial biases into the data.

14.1 SUBARU HYPER SUPRIME-CAM IMAGES

We obtain a large sample of galaxy images from the Public Data Release 3 of the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP; [3]).

The HSC-SSP is a 330 night observing program with the Subaru 8.2m telescope on Mauna Kea in Hawaii [68, 98, 111, 112]. The survey comprises several layers ranging in survey footprint size and depth. We focus on the Wide Layer, which has median i-band seeing of 0.61 arcseconds, 5σ point-source depth of 26.2 AB mag, and final survey area of 1200 square degrees – of these, 670 square degrees are provided in PDR3. Individual exposures are obtained in 5 optical broad-band filters *grizy* [92] and co-added to produce final depth images.

For this paper, we select a galaxy sample from HSC-SSP using a similar process to that described in [161] – which cross-matches sources in the HSC-SSP Wide Layer to their counterparts in the Sloan Digital Sky Survey (DR16, [2]). The selection first takes all SDSS DR16 objects that (1) have Petrosian magnitudes $r_{petro} < 20$ mag and (2) are within the HSC-SSP footprint. This list is then cross-matched to HSC-SSP sources with detections in each of the *grizy* broad bands using a 1 arcsec matching tolerance. Exactly as in [161], additional quality constraints are then placed on the cross-matched sample to e.g. remove objects whose photometry is contaminated by bright stars, bad pixels and cosmic rays [21, 43], and those that do not satisfy a conservative star/galaxy separation cut [14, 173]. This selection method is distinguished from the original one of [161] by their use of the SDSS galaxy spec-z sample [173] as a reference catalogue. We thereby lose the more reliable spec-z information from SDSS but gain a much larger dataset for our machine

learning approach. In this work we use redshifts mainly for matching the observed galaxy populations to the ones from the simulations at the same cosmic epoch – we focus throughout on the interval $0.1 \leqslant z \leqslant 0.4$. According to Figure 4 of [177], the estimated mean and dispersion of the photo-z bias, $(z_{\rm phot} - z_{\rm spec})/(1 + z_{\rm spec})$, over our chosen redshift interval are < 0.01 and approximately 0.06, respectively. We therefore adopt the less accurate, but to this aim *sufficiently* accurate, photo-zs from [177] for epoch-matching between simulated and observed galaxies.

We start from cutouts centered at each galaxy center with a fixed edge-size of 50 arcseconds including galaxies with photometric redshifts between $z = [0.1, 0.4]$. In total our sample contains r,g and i-band images of 768484 galaxies.

14.2 TNG SIMULATED GALAXIES AND THEIR HSC MOCK IMAGES

We use galaxies from the IllustrisTNG simulations [106, 116, 121, 122, 134, 136, 171] for our contrastive comparison to real galaxies from the HSC-SSP described above.

TNG is a suite of full cosmological magneto-hydro simulations of galaxies based on the moving-mesh code AREPO [167], which traces the formation and evolution of galaxies based on cold dark-matter particles and baryonic matter (gas cells, stars and black holes particles) by including forces between them, star formation, and feedback processes. The periodic-boundary cubic simulation boxes are evolved from redshift $z = 127$ to $z = 0$ with initial conditions given by the Zeldovich approximation [via the N-GenIC code 170]. The underlying cosmology is given by Planck [137].

In this work we use the outcome of the two full-physics runs called TNG100 and TNG50, which consist of simulation boxes with co-moving edge-lengths of roughly 100 Mpc and 50 Mpc [124]. While the smaller volume of TNG50 allows a higher resolution with respect to TNG100, the number of galaxies therein simulated is, for the same reason, smaller, especially at the high-mass end. Therefore we also incorporate TNG100 to boost the sample size especially for galaxies with stellar mass $\log(M_\star/M_\odot) > 10$.

TNG does not explicitly model light from stars nor the interaction of such light with gas and dust. It is nonetheless highly desirable to generate images based on light from the stellar/gas particle data that are compatible with images of observed galaxies taken with real instruments. Previous works showed the utmost importance of pushing images into the same domain as the observations [e.g. 22, 39, 85, 150, 175]. This is especially important for a contrastive learning application like the one pursued in this work, as the model is designed to find differences in images. For example, a non-realistic background could be identified as the most significant difference between images even if the imaged galaxies were fully consistent. We therefore use synthetic HSC-SSP images of TNG galaxies produced using dust radiative transfer and injected into real HSC-SSP cutouts: these have been developed by [26] and are succinctly described in the following.

14.2.1 *Forward modeling of simulated galaxies into HSC observations*

Galaxies from the TNG50 and TNG100 runs and found at the snapshots spanning $0 \leqslant z \leqslant 0.7$ were forward-modeled into synthetic HSC-SSP images by [26]. First,

dust radiative transfer post-processing with SKIRT[1] [12, 32, 33] was used to model the emission, absorption, and scattering of stellar light in the HSC grizy bands[2] [93] and several supplementary filters spanning $0.3 - 5$ microns. Light from particles representing stellar populations older than 10 Myr was modeled using the [28] stellar population synthesis (SPS) library assuming a [34] initial mass function. Continuum and nebular line emission from young stellar populations embedded in birth clouds (star particles ages < 10 Myr) was modeled using the MAPPINGS III library [78]. To model attenuation/scattering by dust, which isn't explicitly tracked in TNG, gas cells were ascribed dust densities using the post-processing method from [141]. In this dust model, the dust-to-gas mass ratio scales with gas metallicity [145] and further assumes a [196] Milky-Way dust grain composition and size distribution.

The images generated by SKIRT are *idealized*, i.e. they contain no noise/background and have been produced with high spatial resolution: 100 pc/pixel. The idealized SKIRT images were then inserted into real HSC-SSP fields using an adapted version of `RealSim`[3]. This step handles (1) cosmological surface-brightness dimming and physical-to-angular size scaling according to the snapshot redshift of each galaxy; and (2) injection into real HSC-SSP cutouts to match statistics of depth, seeing, and crowding by foreground/background objects and artefacts from the actual observations.[4].

14.2.2 *Simulated galaxy sample used in this work*

Throughout this paper, we use the HSC-realistic images of [26] generated for TNG50 and TNG100 galaxies in all 20 simulation snapshots between $z = 0.1$ and $z = 0.4$ (snapshots $72 - 91$). In particular, we focus on TNG50 galaxies with total stellar mass $\log(M_\star/M_\odot) > 9$ and TNG100 galaxies with $\log(M_\star/M_\odot) > 10$ over this redshift interval. These limits are chosen to ensure that all galaxies (1) comprise at least 10^4 stellar particles and hence have reasonably-resolved stellar structures and (2) are detectable with respect to the surface-brightness limits of the HSC-SSP. Each TNG50 and TNG100 galaxy was imaged along 4 lines of sight, oriented along the arms of a tetrahedron with respect to the simulation volume. In this way, the image sample is 4 times larger than the galaxy sample, giving us a synthetic image dataset of 234948 images for TNG50 and 498748 for TNG100.

14.3 DATA PREPARATION AND HOMOGENIZATION

While we ensure that the HSC mocks of TNG galaxies are as realistic as possible in comparison to real HSC images, a few further preparatory steps are needed to guarantee that the real and synthetic galaxy datasets are compared on an even ground. In particular, we have to tackle the following tasks:

1. Choosing a common approach to set the field of view for the images and resize them to the same dimensions.

1 https://skirt.ugent.be
2 https://www.tng-project.org/explore/gallery/bottrell23i
3 https://github.com/cbottrell/RealSim
4 https://www.tng-project.org/explore/gallery/bottrell23

Parameter	Description	HSC	TNG50 & TNG100
Redshift	Redshift of the galaxies relative to the observer	Photometric measurement [177]	Determined from the discrete snapshot number (i.e. simulation time) the galaxy is taken from
i-band Magnitude	Apparent I-Band magnitude of stars	HSC-SSP Photometric Pipeline [21]	Integrated stellar light incl. dust attenuation [123]
Petrosian Radius	Radius enclosing 90 per cent of the overall light that is originating from the respective galaxy. Determined from an isotropic Petrosian fit.	Derived via Petrofit in the FOV calculation	Derived via Petrofit in the FOV calculation

Table 4: Overview over the fundamental galaxy properties used to match the observed and simulated galaxy samples and for validation tests.

2. Matching the three galaxy populations in terms of basic observable properties so that we can implement an approximate common selection function for all datasets.

3. Deciding which bands to use and how to stretch them, i.e. to find a mapping between the raw image intensities and the normalized values the model works with.

4. Splitting the datasets into training, validation, and test sets.

It is of utmost importance to apply exactly the same preparation steps to all three sets of galaxy images, to ensure that we do not introduce any systematic differences that could then be picked up, spuriously, by our contrastive learning approach.

14.3.1 *FOV selection*

The FOV is an important factor when preparing images for any machine learning algorithm. In our case, if the fiducial FOV is too small, the model might miss important features around the galaxy, e.g. loosely-wound spiral arms, or extended stellar halo features. On the other hand, if the FOV is too large the model may pay more attention to background than to the target galaxy itself.

The field of view (FOV) for the HSC-realistic TNG mocks is large ($50 - 500$ kpc) and scales with the physical size of each simulated galaxy. Meanwhile the HSC cutouts have a fixed angular size. To ensure that we get good results across the full range of galaxy masses and sizes, we choose an adaptive approach to crop the large FOV of the simulated images. Namely, we perform a Petrosian fit with Petrofit [70] and measure the Petrosian radius r_{p90} containing 90 per cent of the light for the central object in each image. We then choose $4 \times r_{p90}$ as FOV and crop the images accordingly. The cropped images are then resized to a common grid of 256^2 pixels. Because the resolution of the original image data is fairly high (0.168 arcsec/pixel), the angular resolution of the cropped/resized images is almost always lower than

the resolution of the original data. While we did not perform any exhaustive tests regarding the Petrosian fitting, crosschecks with existing size measurements available for TNG and HSC suggests that our fitting procedure is consistent. Additionally, by applying the fitting to all datasets with the same parameters, we ensure that possible biases and errors are contained in similar fashions in all sets.

14.3.2 *Galaxy-population matching*

When comparing distinct sets of data, especially between observations and simulations, we have to ensure that the distribution of their intrinsic properties are the same. For example, observations are limited by exposure i.e. surface brightness cuts, in addition to survey area, whereas the galaxy populations from uniform-volume cosmological simulations are volume limited and limited e.g. by a mass cut. Therefore, the prior distributions of basic quantities of observed and simulated galaxy populations may be different even if the simulated galaxies were perfectly realistic.

To match the three galaxy data sets we first need to identify properties that are derivable from both the simulations and the observations. We choose photometric redshift, i-band Magnitude and galaxy stellar size i.e. Petrosian radius, as summarized in Table 4.

Then, to enforce a common selection function for both simulated and observed galaxies, we draw galaxies from the HSC-SSP set to match the redshifts, apparent magnitudes, and angular sizes of galaxies from the TNG50 and TNG100 samples. The specific algorithm is as follows. For each TNG(50/100) galaxy, we identify a match-tolerance window in the space of redshift ($\pm$0.035), apparent i-band magnitude ($\pm$0.05 mag) and Petrosian radius ($\pm$0.4 arcsec). These tolerances are chosen based on typical observational measurement uncertainties in each of these quantities. We then draw a random HSC galaxy from within the tolerance window and assign it uniquely to the TNG(50/100) galaxy being matched (without replacement). TNG(50/100) galaxies for which no HSC-SSP galaxy offers a match are discarded from our analysis. Because of the different volume and resolution of the parent simulations, the intrinsic galaxy distributions from TNG50 and TNG100 are different, and each is hence matched individually to HSC galaxies.

The three tolerances are simply conservative estimates and are practically free parameters. We chose them such that they are well within the stated errors for the HSC-SSP photometry pipeline [21] and MIZUKI photo-z measurements [177] while also ensuring that we don't restrict ourselves to too small of a window, which would lead to larger numbers of unmatched TNG galaxies and wasted data. This is especially problematic for the redshift, as the TNG galaxies are only available for a few discrete snapshots in time, rather than a smooth and continuous range.

We show the resulting distributions of matched galaxy samples in Figure 26. The unmatched i.e. intrinsic galaxy distributions are shown in the upper panels whereas the bottom panels show the galaxy distributions after the matching. As mentioned above, as the intrinsic distributions differ between TNG50 and TNG100, we show two sets of galaxy samples: one that has been matched to TNG50 and one that has been matched to TNG100.

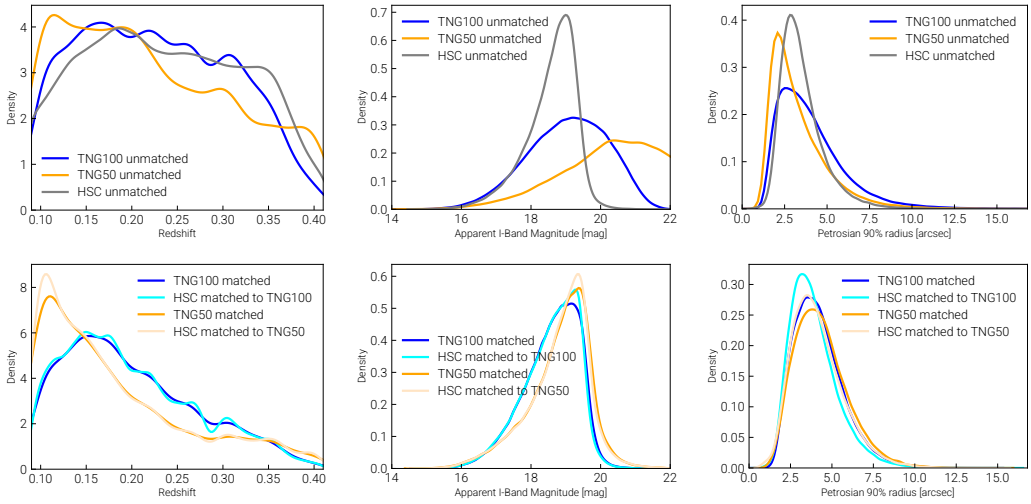


Figure 26: **Matching the observed HSC dataset to the two simulated TNG50 and TNG100 samples of galaxy images, respectively.** We show the distribution of all three sets in photometric redshift, apparent I-band magnitude and petrosian radius before (top) and after (bottom) the matching (refer to Table 4 for definitions). We split the set of matched HSC galaxies into two subsets: one that is matched to TNG100 and one that is matched to TNG50. This is necessary as the prior distributions between the two simulated image samples are different, because of the different volume and resolution of the simulations. Note that we apply a larger smoothing kernel for the redshift: this is necessary as the simulated TNG galaxies are given only for a set of discrete snapshots.

We see that, naturally, the main bias in the unmatched sets is the i-band magnitude, as there is a surface brightness cut for HSC but no such a limitation for the simulated galaxies. However the apparent brightness is heavily linked to the redshift: (1) At low redshift there are more bright TNG galaxies than HSC galaxies. This is a volume effect as the simulated volume is for low redshift larger than the observed. (2) At 'higher' redshifts (i.e. up to $z = 0.4$) there are fewer faint HSC than TNG galaxies owing to our HSC Petrosian magnitude and SDSS counterpart detection criteria for the HSC sample. Due to this offset, when matching the samples we remove many faint high-redshift TNG galaxies. The matched samples therefore lean towards the low-redshift regime. One could counteract this by decreasing the matching window for the redshift; however this would lead to an even more depleted TNG sample.

Moreover, there is a correlation between the luminosity and the Petrosian radius: TNG (50 and 100) galaxies tend to be increasingly larger than HSC galaxies for increasing luminosity. It is for this reason that we need to match simulated and observed samples also with respect to galaxy sizes. This offset was already found in previous studies that compared TNG100 and Pan-STARRS galaxies [150]. In fact, this is not visible in Figure 26 as the unmatched TNG population is biased towards smaller galaxies because of the large fraction of low-luminosity galaxies. As a consequence of the matching in stellar sizes, we improve the domain agreement also in terms of image resolution and FOV distribution. After the matching process, we have 628710 images (incl. HSC, TNG50 and TNG100) for training, validation and testing the machine learning model. From now on, we omit the specification

"matched" when referring to the samples of TNG100/TNG50 images matched to HSC.

14.3.3 *Band selection and stretching*

While in principle the model can work with any number of luminosity bands or channels, we decide to start with an RGB representation that is most intuitive for a visual inspection of the results. For the RGB images we use the respective HSC i, r, and g bands, respectively. For those bands we have, on the one hand, large amounts of HSC data with sufficient depth and data quality. Furthermore they roughly correspond to the RGB color scheme at the rest frame that covers most of the visual spectrum.

We then apply a stretching of the images based on [22] to ensure that galaxy features across the full dynamic range of surface brightness are well captured. Without this stretching a machine learning model might just focus on the much brighter galaxy core, neglecting features in the outskirts. To each channel we apply the following steps separately:

1. Perform a logarithm on the intensities.
2. Set the minimum brightness to the median of all pixels. We ignore all pixels with less than 10^{-7} for the calculation of the median.
3. Set the maximum brightness to the 99 per-cent quantile of the central 20×20 pixels.
4. Normalize the logarithmic images with the given minimal and maximal brightness.

The pixel values of the stretched images are therefore between 0 (background level) and 1 (center of the galaxies). We want to note that we experimented with multiple stretching mechanisms. For example, a arcsinh stretch better emphasizes the noise and thereby leads to a model putting more importance into the noise and, subsequently, the redshift of a given galaxy. The band selection and stretching therefore also depends sensitively on the scientific question at hand.

14.3.4 *Splitting of the sample(s)*

ML models may become so well-tuned to the dataset on which they are trained that they may generalize poorly to new data, even from the same domain (i.e. over-fitting). It is therefore common practice to split the full dataset in three fixed and distinct subsets. These are not to be mixed during the training and testing procedure. We split the combined HSC and TNG sets in (1) a training set (80 per cent of the overall sample), which is used for the training of the model only, (2) a validation set (10 per cent), which is used to assess the model performance, i.e. the generalization capability during the training, and (3) a test set (10 per cent) to which the model is never exposed in training and that is used to perform all analysis and tests (Section 16.1 and following).

As discovered in previous work [63], it is important to ensure that there are no progenitor/descendants of a (simulated) galaxy in multiple subsets. This is to avoid an artificial improvement of the test results later on. Although this was shown for scalar properties only, we keep this conservative approach and ensure that we split all TNG galaxies along with their root descendant; i.e. the galaxy they will finally merge into. In this way, all progenitors and descendants of a given galaxy are contained within only one of the three sets, exclusively.

15 THE MACHINE LEARNING METHODS

To capture the potentially complex features that distinguish galaxies from each other (including real and simulated galaxies), we use the self-supervised contrastive learning method. This model is designed to extract features from the images and encode them as high-dimensional, vectorized *representations*. Through training of the model, these representations should come to encode detailed information about each galaxy image. We use a Residual Network described in detail in Section 15.1 to map the images into this representation space. This network is then trained by the contrastive learning method NNCLR described in Section 15.2. The details of the training procedure is explained in Section 15.3.

15.1 E(2) EQUIVARIANT RESIDUAL NETWORKS

The input images have 256^2 pixels in 3 bands. This corresponds to 196608 degrees of freedom for each image. To efficiently learn features from this high dimensional image data, we use a (Wide-)ResNet Architecture [198] to extract information from the images and map this information into a lower, i.e. 256, dimensional representation space. Instead of standard convolutional layers, so called "E(2)-Equivariant Steerable Convolutional" layers [193] are used in our ResNet model. In doing so we explicitly pay attention to the fact that the orientation of the galaxies on the image plane should be irrelevant – all convolution layers, and therefore the mapping, are equivariant regarding transformations of the dihedral group D(8). Simply speaking, these are discrete rotations $360°/8 = 45°$ and reflections on the images that should not alter the properties of the galaxies shown in the images. Tests performed with a group of higher symmetry like $D(16)$ doesn't improve results but does increase the computation time. On the other hand, the model converges faster and towards a lower validation loss with D(8) than without any equivariance i.e. using conventional convolutional layers. For the fiducial model we use a ResNet depth of 16 and a width of 2. The ResNet layers are equivariant regarding D(8) and are regularized using a Dropout of 30 per cent of the nodes. The output from the ResNet is a flattened feature vector. To give the model additional flexibility regarding the representations, we add a fully-connected network, comprising 2 layers with 256 nodes each, at the end of the ResNet to map the feature vector into the final representation space.

15.2 CONTRASTIVE LEARNING: NNCLR

The ResNet architecture itself maps the images into a representation space – however, the task at hand is to ensure that these representations are meaningful for our purposes, i.e., they must encode the visible features of the galaxies in the images. To achieve this, the ResNet is trained using the self-supervised method 'Nearest neighbour Contrastive Learning' (NNCLR) [60] which is based on SimCLR [36].

SimCLR aims to 'teach' the ResNet to attract transformed images of one and the same galaxy and discriminate against transformed images of other galaxies. For this work we use the following (random) transformations:

- Reflections along the horizontal and vertical axis
- Rotations from $-15° - +15°$
- Central zooms of max 10 per cent i.e. the center of the image stays in the center of the crop while zooming
- Affine translations by up to 32 pixels
- Gaussian blur with a max std of 0.1 pixel
- Gaussian noise with a max std of 0.08 (note that the images are normalized to 1).

We perform these transformations on the original 256^2 pixel images as described in Section 14 and then rebin them to 128^2 pixel for the ResNet. As a result, the ResNet is automatically focusing at commonalities and differences between images and placing them in distinct spaces of the representation space: e.g. grouping elliptical or disk galaxies together.

While this describes the functioning of SimCLR on a first order, the NNCLR method we apply in this work takes another step. In addition to the transformations described above, we search for the nearest neighbour (hence the name) of that transformed image in the representation space and use that neighbour image as the transformed 'same' galaxy. In doing so, one can reach a higher generalization ability, as the network interpolates between the training galaxies through this additional augmentation step. We tested and compared the SimCLR method with NNLCR and choose NNCLR as the fiducial choice for this reason.

For the nearest neighbouring search we are *not* using all images of the training set as this would be too computational expensive. Instead we use a queue: at each training step we push the newest batch of input images and their representations into the queue and remove the oldest batch. In total we keep 512 batches of images and their representations in that queue from which to match nearest neighbours.

Finally, we do not run the contrastive learning directly on the representation space but we adding a so called 'Projection Network', which consists of 2 fully connected layers with 512 nodes each. This projects the representation space into a space with the same number of dimensions (i.e. 256) on which the contrastive loss is calculated and the network trained [refer for details to 60]. When we test/apply the network to data in the next sections, we remove this projection network and only use the representation space. This means that the projection network is just a part of the training.

15.3 TRAINING OF THE RESIDUAL NETWORK VIA NNCLR

For the training, we use the training sets of all galaxy samples at the same time: the simulated TNG100 and TNG50 galaxies and the observed HSC galaxies. In doing

so we allow the model to find differences and commonalities among all sets of images. We further elaborate on this choice in Section 17.

We use an Adam optimizer [95] to back-propagate the NNCLR loss of the common training set through the network. We do this in random batches of 32 images with 2 transformations for each image. After 350 random batches (which we call an epoch) we use the images in the validation set to validate the model. We set an initial learning rate of 0.001, which is reduced by a factor of 0.8 if the validation loss is not decreasing for 3 epochs. Training is stopped if the validation loss has not decreased for 15 epochs. All hyperparameters mentioned in this section (including the network architecture) have been optimized with respect of the validation loss using the Optuna framework [4]. We present the results of our model exploration and a discussion on parameter choices in Appendix C.

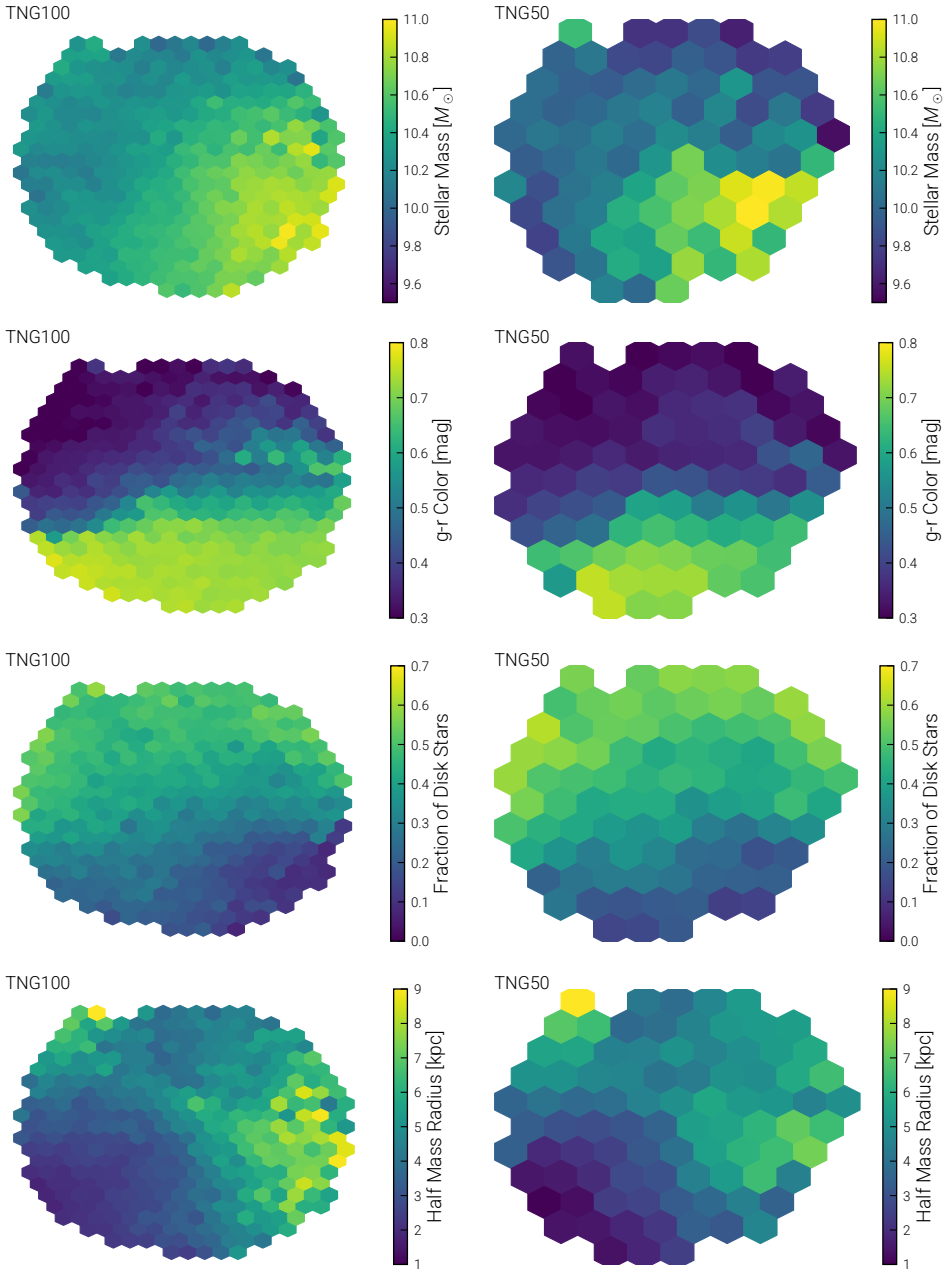


Figure 27: **Are the representations related to observable features?** Each panel shows a 2D hexbin histogram of TNG100 galaxy images (left panels) and TNG50 galaxy images (right panels) in the 2D UMAP parameter space. The UMAP mapping itself was trained using the 256-dimensional representations of TNG50/100 and HSC test galaxies. Note that because of the smaller sample size of TNG50, we choose a larger bin size in the right panels. The bins are coloured according to the median image/galaxy property in each bin. We show from top to bottom, and in the continuation figures: total stellar mass, fraction of disk stars, integrated galaxy colour index (g-r), stellar half-mass radius, Sérsic half-light radius, Sérsic index, Sérsic ellipticity, point asymmetry of the light distribution, concentration of the light, smoothness of the light distribution, Gini-M20 bulge parameter and Gini-M20 merger parameter. The UMAP, and therefore the representations upon which it is based, are clearly related to observable properties of the TNG50/100 galaxies.

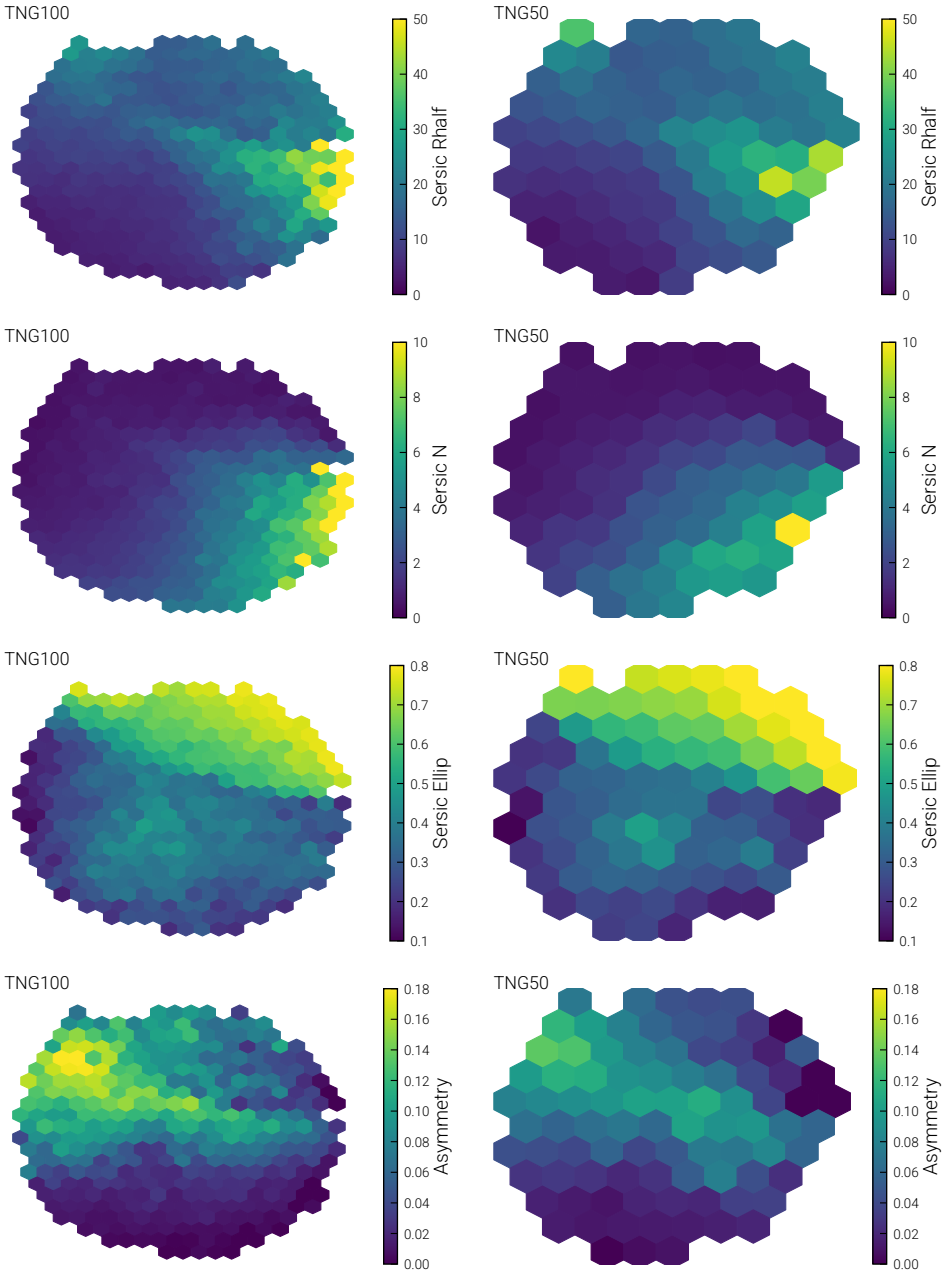


Figure 28: Continuation of Figure 27.

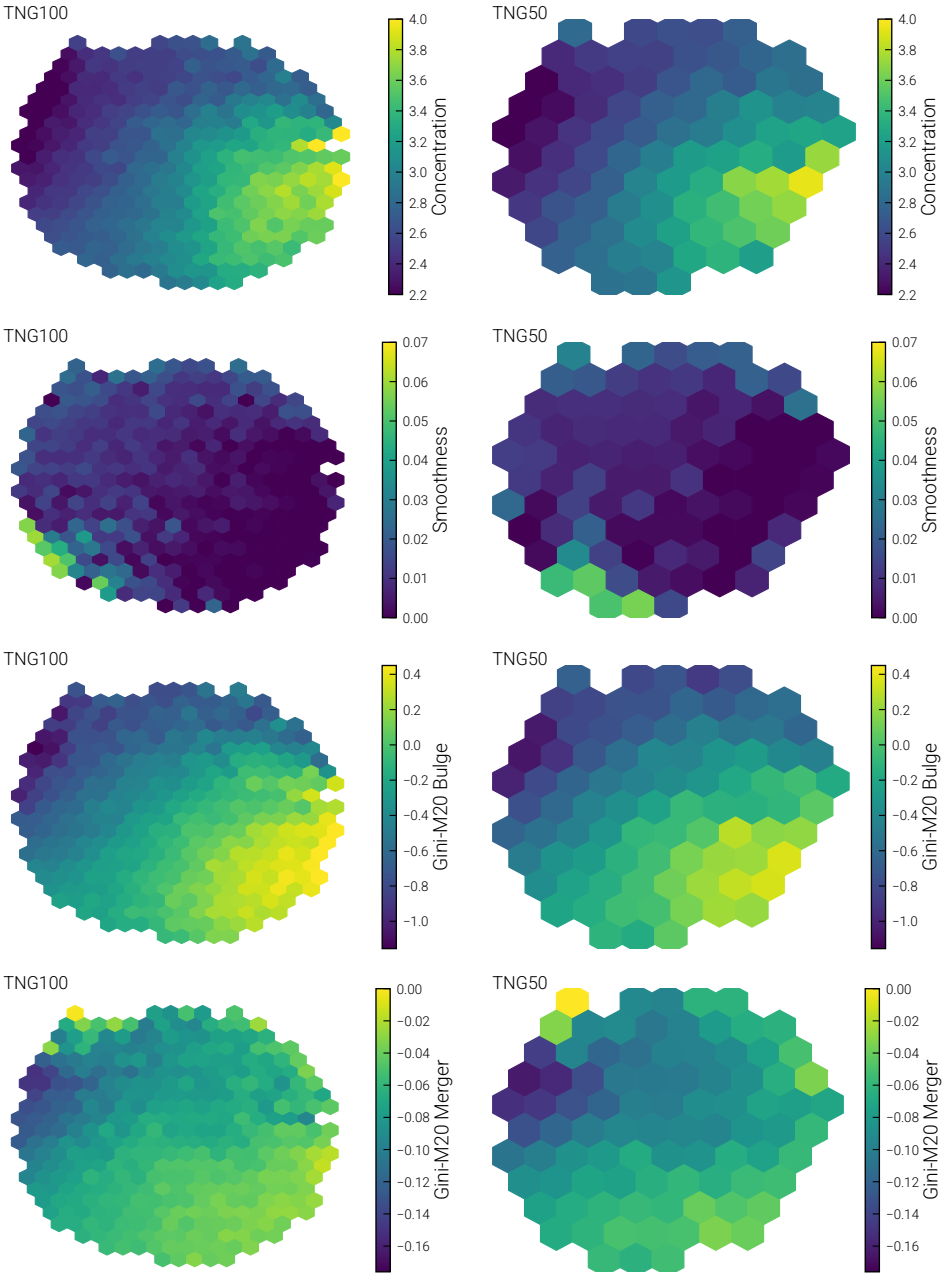


Figure 29: Continuation of Figure 27.

16 RESULTS

16.1 INTERPRETING THE REPRESENTATION SPACE WITH UMAPS

After training via the contrastive learning method NNCLR introduced above, the galaxy images from HSC and TNG50/100 should be arranged in a 'meaningful' way in representation space. Based on the choices described above, this means that each galaxy is assigned a 256 dimensional vector that is somehow related to its visual and observational properties in the image. However, a priori the definition of 'meaningful' might differ between a ML network and a human. For example, a model could pay more attention to the image background than to the galaxy in question. In this section, we check whether the representations indeed encode meaningful galaxy properties.

To visualize the 256 dimensional representation space, we utilize the dimensional reduction algorithm UMAP [109], which maps the 256 dimensions non-linearly into two dimensions while conserving the topology of the space as much as possible. To ensure that our network generalizes well to a previously unseen set of images, we train and apply the UMAP method to the representations of the images in the test set of both observations and simulations: HSC and TNG50/100 galaxies. During the training of UMAP, we gauge two parameters of the method: (1) the number of nearest neighbours and (2) the minimal distance between nodes. We do not find any large-scale difference when varying these parameters: namely, for no combination we see that the observed and simulated domains separate. However, the local structures in the UMAPs do differ strongly for different choices of those parameters. This has to be kept in mind when judging the representation space based on a single 2D UMAP representation. For this work we use a rather smooth UMAP representation, focusing more on the global than the local structure of the representation space, which leads to a fiducial choice of 100 neighbours and a minimal distance of 0.5.

In Figures 27, 28 and 29, we show the UMAPs of the TNG100 (left panels) and TNG50 (right panels) test galaxies, coloured according to a selected (but not exhaustive) set of observable galaxy properties. First, we show UMAPs coloured according to exactly known galaxy properties, i.e. measured from the stellar particle data of the TNG50 and TNG100 simulations considering all stars gravitationally bound to each galaxy according to SUBFIND [168]: total stellar mass, fraction of disk stars [71], integrated galaxy colour [122], half mass radius [refer for definition to 63]. We then examine morphological measurements that were derived from the survey-realistic mocks using STATMORPH [150] on the r-band of the uncropped TNG mock images: Sérsic half light radius, Sérsic index, Sérsic ellipticity [76, 158], point asymmetry of the light distribution, concentration of the light, smoothness of the light distribution, Gini-M20 bulge parameter and Gini-M20 merger parameter [refer for definition to 150]. In doing so we also incorporate statistics as they are measured and used by observers. Note that we remove 4352 (of overall 25896) im-

ages with unreliable STATMORPH-fits when plotting the UMAPs for the STATMORPH parameters.

We see that all these observable quantities are very well captured in the representation space: namely, galaxies with similar properties occupy the same regions within the representation space. Importantly, known and expected relations among properties are conserved, e.g. disk galaxies have on average a lower stellar mass while presenting themselves in a bluer colour due to ongoing star-formation (and therefore a lower g-r colour index). We also can see the expected relationship between the theoretical half-mass radius and the observational Sérsic half-light radius, one being derived just by summing up the known stellar particle masses and the other being based on a profile fit to galaxy light, respectively: despite the different definitions, both quantities align well in the representation. This is especially of interest as the FOV was chosen to be proportional to the Petrosian galaxy sizes, and hence the size is not directly observable from the images. It is also interesting to note that the Gini-M20 Bulge and Merger parameters are captured, which promises deeper insights in the structural composition of each galaxy and its history.

Comparing the results between TNG50 and TNG100, both sets show qualitatively similar trends: the model is therefore able to map two IllustrisTNG runs with different resolution into the same space. However, this statement is just based on the median of the quantities in a 2D projection. Although wedo not show any information about the scatter and point densities in each of the hexagonal bins, we have checked and indeed the standard deviation of some galaxy properties (e.g. galaxy colors) can be large and varying across the UMAPs, even though the medians exhibit clear trends. In the sections that follow, we therefore employ additional methods to compare the images based on the full representation space.

16.2 HOW SIMILAR ARE THE MOCKS FROM THE TNG SIMULATIONS TO THE OBSERVED HSC IMAGES?

The previous section showed that the ML-constructed representations of galaxy images naturally encode morphological properties and stellar structures. But how can we gauge whether simulated galaxies are realistic in comparison to observed ones? As argued in Section 13, this assessment is especially important for inferring unobservable properties from HSC data. This is only possible when the TNG galaxies that are used to learn the connection between light maps and galaxy properties are sufficiently realistic in comparison to the HSC observations in question. The task ahead is therefore to assess if the network has learned a sensible definition of similarity and to quantify this similarity.

In the following, we proceed with the underlying idea that, if two galaxy images are close in the 256-dimensional representation space, then they are also "similar" in appearance. If the representations of two galaxy samples have disjoint or matching subsets, we can use this to measure for which type of galaxies the mocks from TNG and the HSC images are different or similar.

16.2.1 *Comparing the representation space with UMAPs*

As a first step we reuse the UMAP method from Section 16.1 to investigate the overlap in 2D among the three image sets (TNG50, TNG100 and HSC).

The upper panel of Figure 30 shows the distribution of TNG50/100 and HSC representations in a 2D UMAP. A noteworthy success is that the sets are clearly not disjoint. In other words, there is significant overlap in the image-based representations of TNG50/TNG100/HSC galaxies. Meanwhile, we see that a residual fraction of TNG50/100 and HSC galaxies do not share a common domain in their representations. Also, the point densities do not align between the three sets. For example, the TNG sets both exhibit higher densities (i.e. more galaxy images) along a diagonal segment through the center of the UMAP, while being suppressed in the top right region.

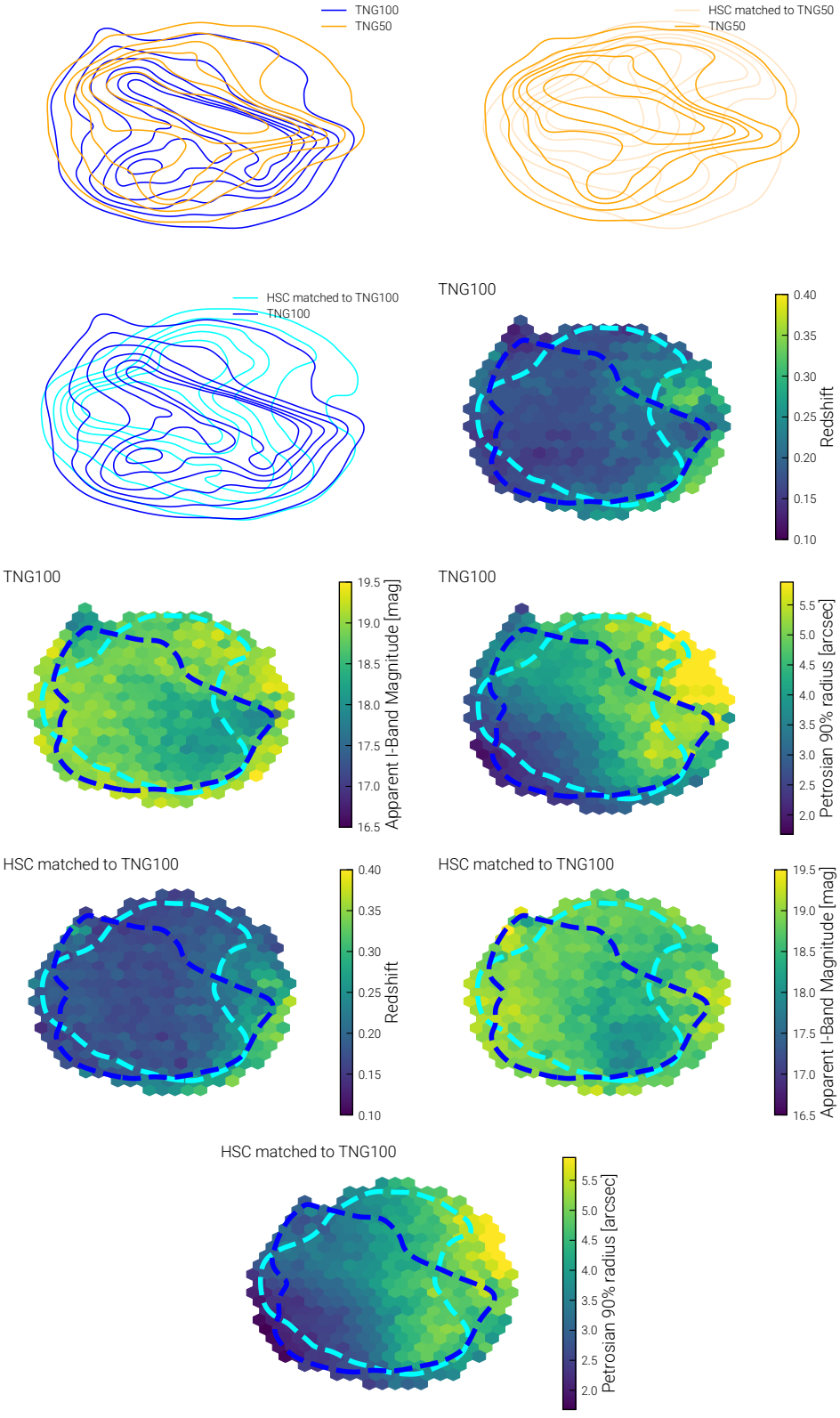
The next four rows of Figure 30 compare the UMAP embeddings of TNG50/100 images and their matched HSC ones, respectively, with histograms coloured according to the binned median of the properties used to match the galaxy samples (see Section 14.3.2). The first two rows show the UMAPs of the TNG100 galaxies and the UMAPs of the HSC galaxies matched to TNG100, respectively. Similarly, the latter two rows show the UMAPs of TNG50 galaxies and the UMAPs of the HSC galaxies matched to TNG50, respectively. From left to right, the color encodes (photometric) redshift, apparent i-band magnitude, and Petrosian radius.

Our main finding is that the trends in the UMAP embeddings of the observed quantities are visually well-reproduced across the three datasets – especially for the Petrosian radii. However for the redshift there is some discrepancy between TNG100/50 and their respective matched HSC sets. This is very pronounced for TNG50, where almost no clear trend between lower- and higher-redshift galaxies is visible. For TNG100 and the matched HSC sets there is however a strong correlation towards the bottom/right of the UMAPs. During training, we found that an alternative colour stretching, which puts more emphasis on the noise while removing the noise level of the augmentation, enhances the model's ability to distinguish between redshifts. However, this greater sensitivity to redshift compromises the model's sensitivity to other galaxy features. Furthermore, the distribution of redshifts in the sample is very askew. The model therefore might be blinded by the vast amount of galaxies with lower redshifts – especially for TNG50 for which the sample truncates at higher-redshifts due to a lack of appropriate matches in the magnitude-limited HSC sample (see Figure 26).

To summarize in relation to the areas where the distributions are not well aligned (see top most panel), we find the following: (1) for TNG100, the top right corner of the UMAPs suggests issues with low-redshift galaxies; (2) in contrast, in the bottom right corner of the UMAPs, high-redshift galaxies for the TNG50 set are missing.

In Figures 31 (TNG100), 32 (TNG50) and 33 (HSC, matched to TNG100) we show the images that are related to the representations of Figure 30: namely, on a grid in 2D UMAP space, we choose the closest data point of the representations and plot the affiliated galaxy image.

By comparing visually the trends for the three sets, we can see that the similarities are astonishing. For example, in all figures: (1) In the top-right, we see edge-on spiral galaxies; (2) in the top-left, we see spirals from the top, i.e. face on; (3) at the


TNG100
TNG50
HSC matched to TNG50
TNG50
HSC matched to TNG100
TNG100
TNG100
0.40
0.35
0.30
0.25
0.20
0.15
0.10
Redshift
TNG100
19.5
19.0
18.5
18.0
17.5
17.0
16.5
Apparent I-Band Magnitude [mag]
TNG100
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
Petrosian 90% radius [arcsec]
HSC matched to TNG100
0.40
0.35
0.30
0.25
0.20
0.15
0.10
Redshift
HSC matched to TNG100
19.5
19.0
18.5
18.0
17.5
17.0
16.5
Apparent I-Band Magnitude [mag]
HSC matched to TNG100
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
Petrosian 90% radius [arcsec]

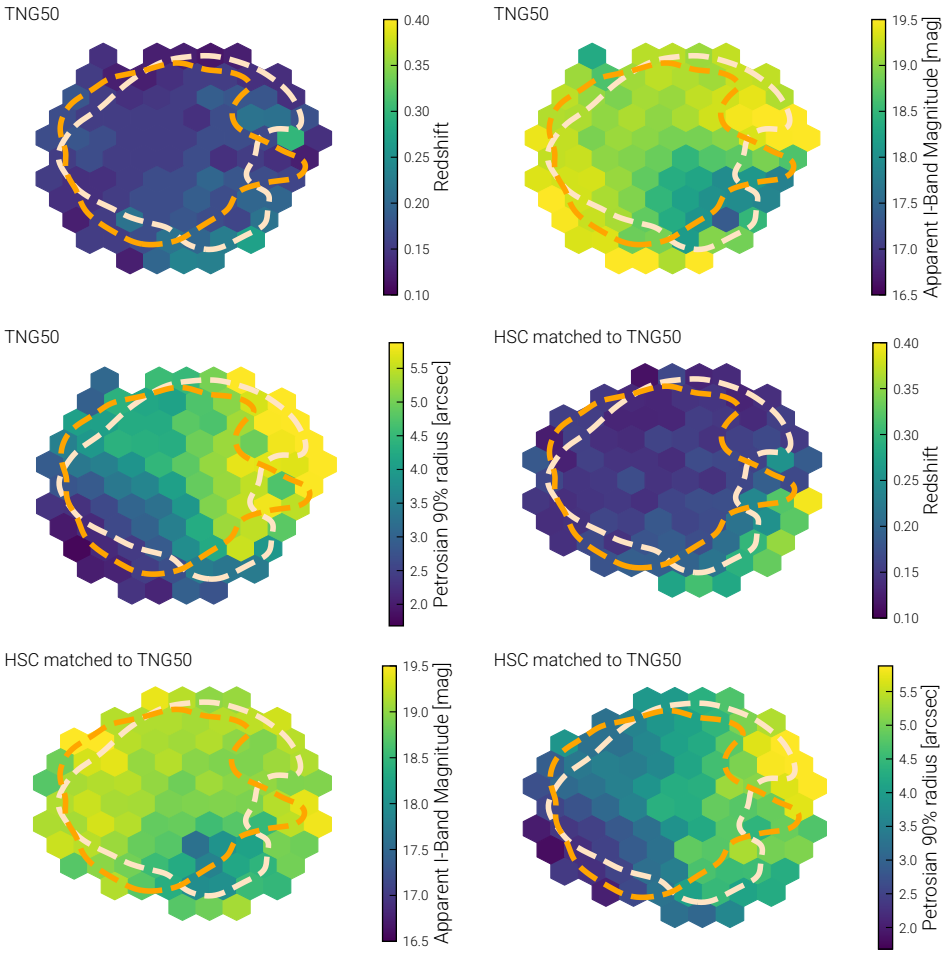


Figure 30: **How well do the representations of the observed and simulated galaxy images align to each other?** We compare the distributions of TNG50/100 and HSC images in the corresponding 2D-UMAP mapping of the 256-dimensional representations. In the uppermost panel we show KDE density plots of TNG100 (TNG50) images in blue (orange). Additionally the HSC sets that are matched against TNG100/TNG50, respectively, are shown in light blue/orange in two additional plots. We see that there is a large overlap among the three sets but also a slight offset and differences in point density. In the lower two panels we further investigate the offset in a visual way: we show the 2D hexbins of the UMAPs coloured bin-wise according to the median value of the three common properties used for the galaxy-sample matching: from left to right, redshift, i-band magnitude, Petrosian radius. In each row we show one of the datasets: from top to bottom, TNG100, HSC matched to TNG100, TNG50, HSC matched to TNG50. We also include the contours covering 80 per cent of the datasets from the uppermost panel: TNG100 in blue, TNG50 in orange and in light orange/light blue the subsets of HSC matched to TNG50/TNG100.

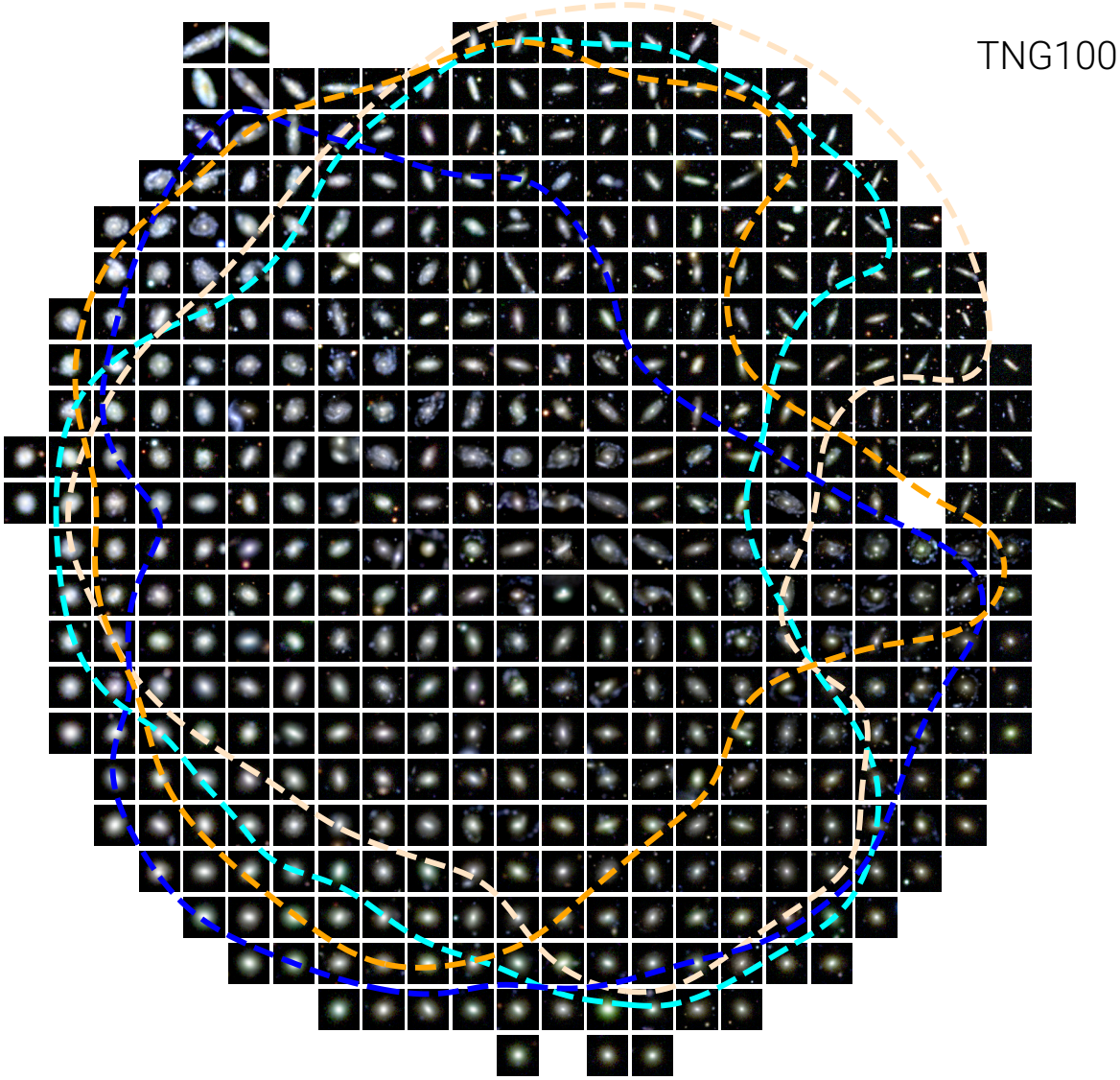


Figure 31: **What images underlie the UMAP representations?** We align TNG100 galaxy images to their positions in the UMAP representations from Figure 30. For each node on a rectangular grid in 2D UMAP space, we choose the closest data point (i.e. image representation) and plot its corresponding image. As a guideline, we report the contours as in Figure 30 covering 80 per cent of the datasets: TNG100 in blue, TNG50 in orange, and the corresponding HSC matched samples in light blue and light orange. With this visual inspection, we see that the galaxy structures are indeed well grouped together.

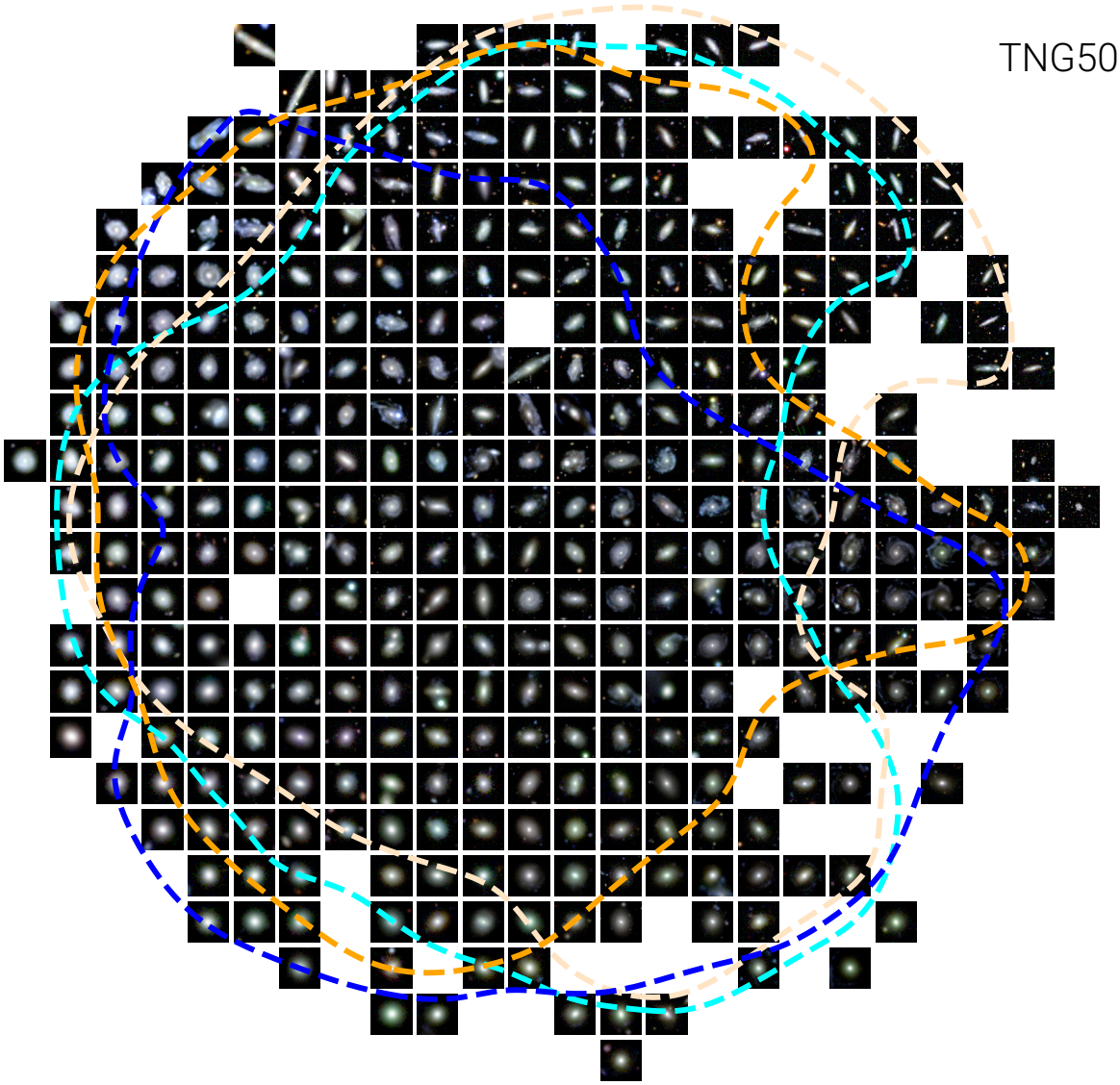


Figure 32: Same as Figure 31 but for TNG50 galaxies. Note that there are more empty regions due to the smaller sample size.

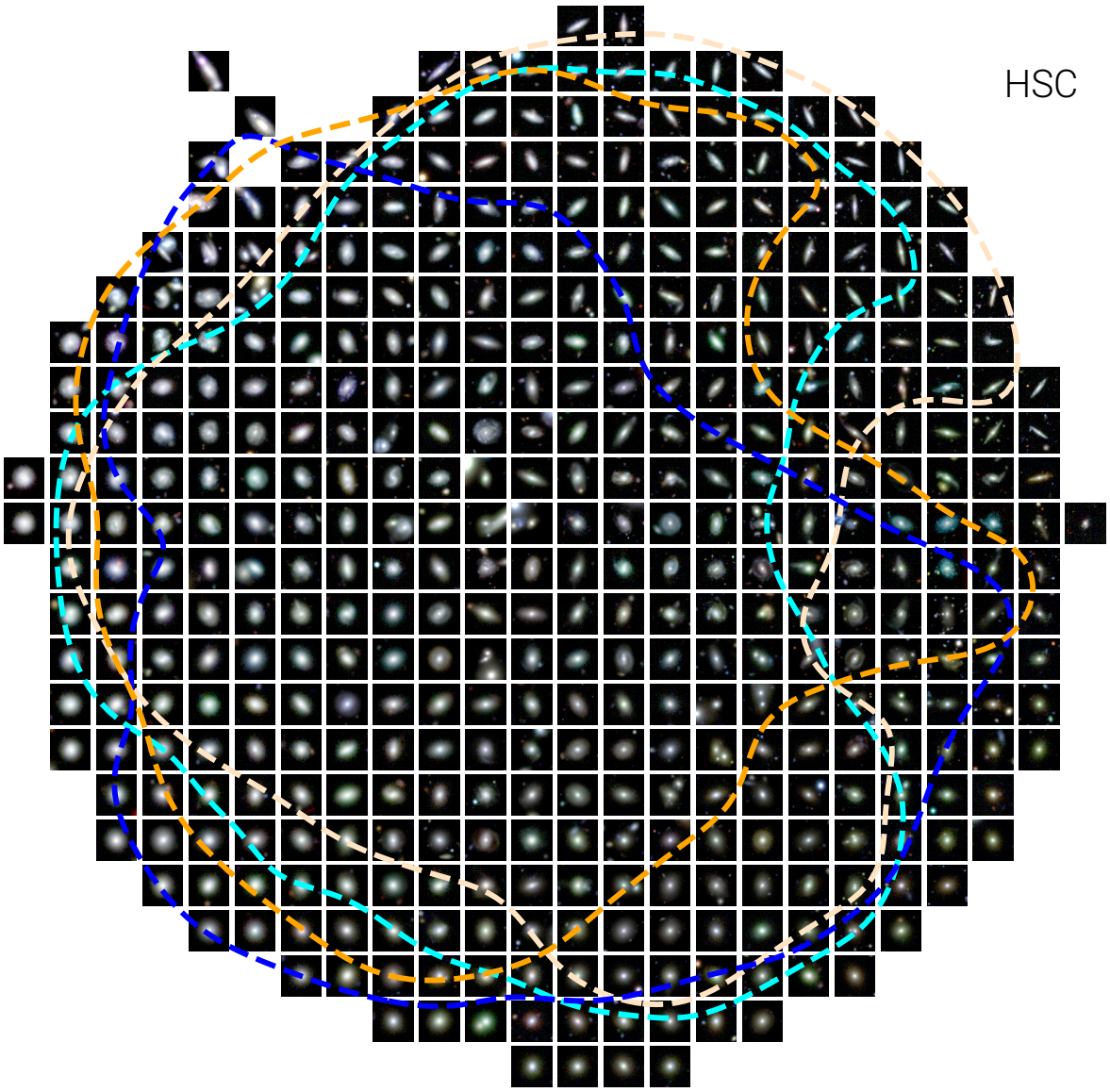


Figure 33: Same as Figure 31 but for HSC galaxies, specifically the HSC sample matched to TNG100. For brevity, we omit showing the analog figure for the HSC sample matched to TNG50, which would appear very similar.

bottom, we see elliptical galaxies. Comparing this to the results of the quantities in Figure 27 that where directly extracted from the simulated data, we can conclude that the overall concept of diskyness is well understood by the model.

However, there are also a clear differences between TNG50/100 and HSC: blue stellar clumps around and in disks of TNG galaxies are not prominently observed in HSC. Features like this have been previously noted and commented upon in stellar light mocks of TNG100 galaxies [150] and other simulations like Illustris and Eagle as well [178, 180]. It might be caused by young stellar populations that are insufficiently resolved by the particle scheme, i.e. a bright and fast evolving stellar population is treated by only a single particle. Furthermore the strength of these blue regions might be also affected by dust assumptions for the light-map creation with SKIRT. These simulated galaxies with blue stellar clumps are very prominent in the top-left to bottom-right diagonal, which corresponds to the 'central and diagonal segment' we have noticed in Figure 30. We find that it is somewhat model dependent if these blue clumps have a strong influence on the shape of the UMAPs. However, as the UMAP is just a 2D projection of the 256-dimensional representation space, the blue clumps might still be well outside the HSC sample. For the ML model adopted throughout, the blue clumps are expressed by a certain clustering of these galaxies in the aforementioned segment.

In the top-left, we see very extended edge-on spirals, some of which are partially even larger than the FOV. In Figure 30, we can see that the fitted Petrosian radii, used to determine the FOV of the images, are comparatively small in this corner. Contrary to this, the physical (simulation-based) mass radii in Figure 27 and the apparent Sersic sizes in Figure 28, suggest that the spirals are indeed large. This area is, therefore, populated by massive spirals whose Petrosian radii fit, and therefore FOV calculation, failed. Interestingly, this area is not heavily populated by HSC galaxies. We hypothesize that they are missing, as those large spirals have been removed from the HSC sample due to a crossmatch failure between HSC and SDSS sources [162].

16.2.2 *Quantifying the similarity via the normalized 8th-nearest-neighbour distance and the OOD Score*

While the visual inspection of the UMAPs is a rather intuitive way to compare two distributions, it only takes a compressed 2D map of the full 256 dimensions into account. This might let galaxies appear closer/farther to each other than they are in the higher-dimensional space. We see this problem, for example, by the fact that the simulated clumpy galaxies that clearly differ from the HSC observations are in the center of the UMAP. Furthermore, relative volumes are not conserved by the non-linear mapping. The areas enclosed by contours in Figure 30 are therefore not one to one transferable to the contours in the high dimensional representation space. In this section, we therefore proceed with a more quantitative way to compare simulated and observed galaxies by defining and calculating a scalar that tells us how far a TNG galaxy is from the HSC domain (or the other way round), i.e. an out-of-domain score.

Firstly, we introduce the normalized 8th-neighbour distance, which is obtained by the following steps:

1. Normalize the dimensions; i.e. scale each dimension of the representation to a mean of 0 and a variance of 1 to ensure that the variance in all dimensions have the same amplitude. We perform the normalization for all datasets (HSC and TNG50/100) together.

2. Calculate the distance to the 8th nearest neighbour for each galaxy image in its own set (d_{self}) and to the 8th nearest neighbour in the other set (d_{other}); here we use the Euclidean norm. Here we use these distances as a measure for the point density. To avoid too noisy statistics we smooth the distances out by using a k^{th} nearest neighbour instead of choosing the nearest.

3. Define the normalized 8th-neighbour distance as the ratio between the two distances d_{other}/d_{self}. The normalization by d_{self} is necessary to ensure that the distance is not mainly dominated by varying point densities. Note that each distance is calculated separately for each galaxy; it is therefore a quantity assigned to each galaxy separately and not for the distribution at whole.

For a given sample positioned somewhere in representation space, the normalized 8th-neighbour distance thereby describes the local ratio of point densities between another dataset and the sample's own dataset. For example, for a TNG galaxy that resides far outside of the HSC domain but still has many TNG neighbours, d_{self} will be small and d_{other} will be large, resulting in a normalized 8th-neighbour distance that is greater than unity. In contrast, for a TNG galaxy that resides in a space that is heavily populated by HSC galaxies and few other TNG galaxies, d_{other} will be smaller than d_{self} and the 8th-neighbour distance will be much smaller than unity. Very large distances therefore indicate discrepancies between the *individual images* and the populations to which they are compared to. Very small distances on the other hand indicate that the *individual images* are well within the domain of the population they are compared to. Note that this normalized 8th-neighbour distance is only meaningful if it is calculated between two sets of equal size. In the following we therefore always select a random subset from the larger set which matches the number of samples in the smaller set.

For two equal-sized sets with identical distributions this ratio is anticipated to be equal to 1. However we have to expect a certain intrinsic scatter which is caused by the fact that d_{self} itself is affected by a certain selection noise from the nearest neighbour search. To interpret the similarities of galaxy images, we hence need to quantify the amplitude of this scatter. Therefore, for each dataset, we also calculate the normalized 8th-neighbour distance between two equal-sized splits of that dataset to estimate the 'baseline' scatter. In other words, this is the scatter about unity that one would get when calculating the scores between two equal-sized samples drawn from the same population. In total, we measure the distances for the following scenarios:

1. Between the TNG100 set and the HSC set

2. Between the HSC set and the TNG100 set

3. Between the TNG50 set and the HSC set

4. Between the HSC set and the TNG50 set

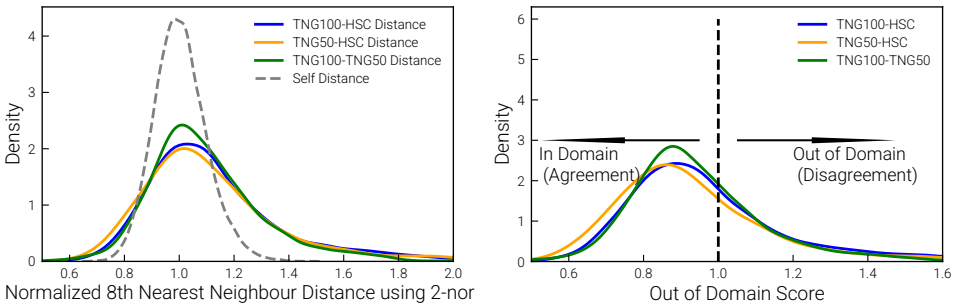


Figure 34: **Our measure to quantify the discrepancy between the individual galaxies and a target dataset: the Out of Domain Score (OOD).** Here we show the distributions for the measures that we have introduced and that aim at quantifying the overlap between the domains of the observed (HSC) and simulated (TNG50 and TNG100) datasets. Left: Distributions of the 'raw' normalized 8th nearest neighbour distance. We measure this quantity for various scenarios: between pairs of the three data sets and also between two random splits of each dataset to investigate the intrinsic scatter that persists even if two distributions are drawn from the same population. For convenience we only show the HSC self-distance as the self-distances are qualitatively similar for all three sets. Note that there is a difference from which 'side' this quantity is measured: E.g. TNG100-HSC is the distribution of the distances for the TNG100 galaxies relative to the HSC galaxies. HSC-TNG100 is the distribution of the distances for the HSC galaxies relative to the TNG100 galaxies. However, for convenience we only show the distances in one direction as the distances are also qualitatively similar for both directions. Right: The measure that we use in the rest of the paper: the Out of Domain Score (OOD). It is overall equal to the left panel but normalized by the 95th percentile of the self distances. In doing so we define galaxies with scores larger than 1 as out-of-domain while scores smaller than 1 as in-domain. With this definition we measure that 33 per cent of the TNG100 and 30 per cent of the TNG50 galaxies are out-of-domain relative to the HSC images. Overall we measure that the comparison in 256 dimensions yield similar results as the ones in the 2D UMAP representations in Figure 30. There is a large overlap between all three datasets. However, it is also noteworthy that there are fewer extreme outliers between TNG50 and TNG100 than between the simulations and HSC. This is a hint for a general offset between TNG and reality.

5. Between the TNG100 set and the TNG50 set
6. Between the TNG50 set and the TNG100 set
7. Between two random same-size splits of the TNG50/100 sets
8. Between two random same-size splits of the HSC set

In all cases, by HSC set we mean the one matched to the corresponding simulation sample, as per Section 14.3.2. The last three scenarios are aimed at estimating the baseline scatter for each dataset. Note, that there is a priory a difference in which direction this quantity is calculated; e.g. for 'Between the TNG100 set and the HSC set' we calculate d_{self} using the TNG100 galaxies and d_{other} using the HSC galaxies. While we derived the quantities in both directions, the differences between the two directions are qualitatively negligible and we therefore focus in the following only on the scores that are calculated for the simulations relative to HSC (cases i and ii above).

We plot the distributions of the obtained measures in the left panel of Figure 34. The modal value is indeed close to 1 for all scenarios. Looking at the variance, we see that the distributions of the distances between HSC and TNG galaxy images (blue and yellow curves) extend to values larger than 1, as the underlying samples are not identical, as we already saw qualitatively with the UMAPs. We have checked and this picture remains unchanged if we had used a distance to the nth nearest neighbour with $n = 4, 16, 32$ instead of 8 (Figure 51).

Now, comparing the distributions between HSC and TNG galaxy images with the distribution between the three self-splits of TNG/HSC galaxies (all roughly equivalent and indicated with the one dashed gray curve), we can now quantify the intrinsic scatter. We incorporate this in the definition of the so called out-of-domain score (OOD), defined by dividing each distance by the 95th percentile of the self-distances. The resulting distributions are shown in the right panel of Figure 34.

We define galaxies with an OOD Score smaller than 1 as "within the domain" of the target distribution. Galaxies with an OOD Score larger than 1 are "outside of the domain". Naturally, this definition is arbitrary but should serve for now as a start point to classify the galaxies into the two regimes. However, we should ensure that the OOD is not too dependent on the model and the choice of the number of neighbours taken into account. Otherwise the main takeaways of the following sections would not reflect properties of the galaxy samples but mainly features of our fiducial model. We investigate and discuss this topic in Appendix D. All in all, there is a large overlap between all three datasets. With our definitions, the large majority of the simulated galaxies are well within the domain of HSC observed galaxies: we measure that 33 per cent of the TNG100 and 30 per cent of the TNG50 images are out-of-domain relative to the HSC images, or conversely that about 70 per cent of the simulated galaxies are realistic, with little difference between the two simulations. However, it is also noteworthy that there are fewer extreme outliers between TNG50 and TNG100 than between the simulations and HSC. This is a hint for an offset between TNG and reality.

To connect the OOD Score also to the UMAPs of the previous Sections, we plot in Figure 35 the UMAPs coloured according to the twice-normalized 8th-neighbour

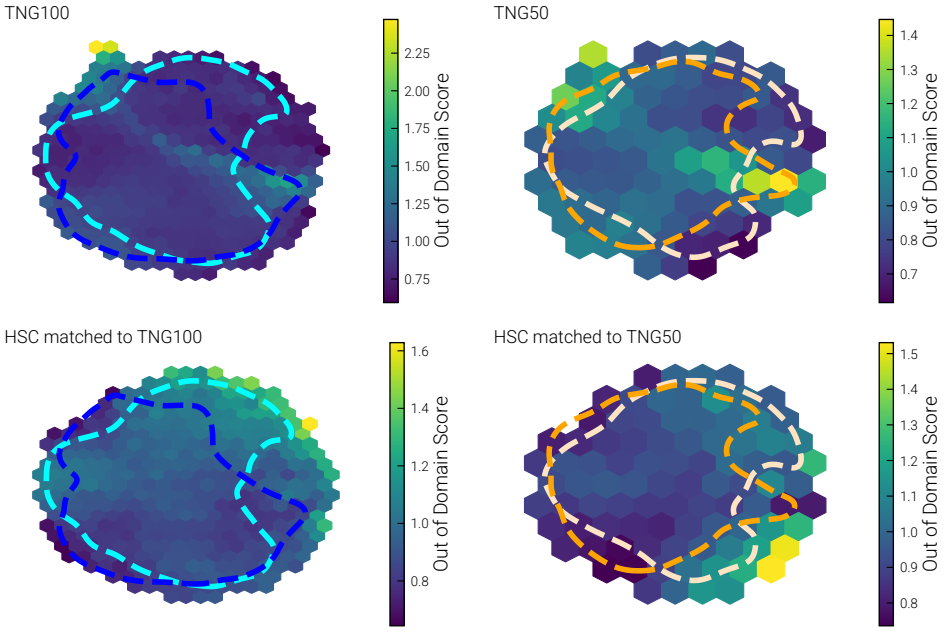


Figure 35: **The Out of Domain Score as measurement for the offset between the TNG and HSC datasets.** Top-Left: UMAP of the representations of TNG100 galaxies. Bottom-Left: UMAP of the representations of HSC galaxies matched to TNG100. Top-Right: UMAP of the representations of TNG50 galaxies. Bottom-Right: UMAP of the representations of HSC galaxies matched to TNG50. In contours we show the 80 per-cent coverings of TNG50 (orange), TNG100 (blue), HSC matched to TNG50 (light orange) and HSC matched to TNG100 (light blue) galaxies. The UMAPs are coloured according to the bin-wise median of the Out of Domain Score assigned to the galaxies relative to the matched HSC galaxies (for the TNG sets) and relative to the TNG50/100 galaxies (for the HSC sets). As a reminder, a OOD Score larger than 1 denotes discrepancy.

distance, i.e. the OOD Score, for the TNG and the HSC sets. The measure can be easily used to identify the subsets that are disjoint/matching. Here it is noteworthy that for TNG there is a diagonal strip of galaxies that are not comparable to HSC; looking at the images of Figures 31 and 32, we see that they might overlap with the clumpy TNG galaxies identified earlier.

16.2.3 *Connecting the OOD Score to galaxy properties*

With the OOD measure defined and calculated in Section 16.2.2, we have now a tool to identify which kind of galaxies are similar and which ones differ between simulations and observations based on meaningful image-based representations. But so, which TNG galaxies are more or less realistic compared to HSC galaxies?

In Figure 36 the OOD score is plotted against each of the galaxy properties introduced in Section 16.1 and Figures 27,28, and 29. As those quantities are derived from the TNG simulations, these plots include TNG50 (orange) and TNG100 (blue) representations only. The following trends are visible: (1) The OOD scores correlate with: larger radii (half mass and sersic half light), steeper sersic profiles, smaller sersic ellipticities and larger asymmetries. (2) There is also a correlation with the

fraction of disk stars, however especially prominent for TNG50. We see a general trend for both simulations that brighter/more massive galaxies fall outside of the HSC domain. There may be multiple reasons for this result, including ones that are not related to the simulation output itself: (1) Brighter and more massive galaxies might have also a more unique halo including stellar halo features, mergers and/or large satellite galaxies. (2) Overall the region of higher mass/lower surface brightness is less well-sampled and could therefore cause that the model is less well suited for this regime. The same reasoning is also valid to some extent to the Sersic radius (which is a relative measure for the galaxy size) and the stellar half-mass radius (which is an absolute measure for the galaxy size).

The visually-observed clumpiness discussed above is captured by the trend in asymmetry but not by the smoothness. We assume that this is caused by the fact that the blue clumps have less of an effect on the r-band derived STATMORPH features with the smoothness focusing only on the central part of the galaxy in question. Furthermore as we have removed unreliable STATMORPH fits, we also might have removed images showing a very unique morphology.

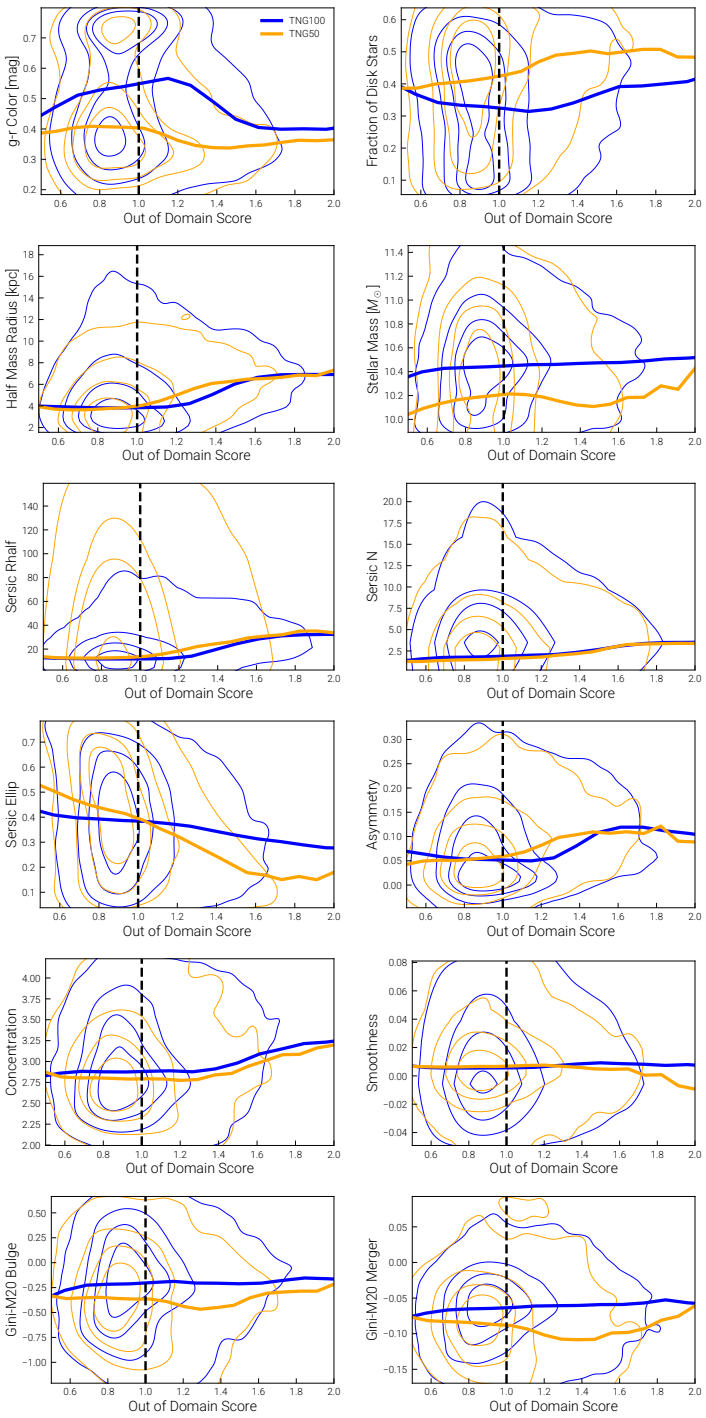


Figure 36: **What galaxy observables drive the OOD Scores of the TNG images relative to the HSC images?** We plot 'in-principle' observable galaxy properties against the OOD Score; these are: total stellar mass, fraction of disk stars, integrated galaxy colour index (g-r), half mass radius, Sérsic half light radius, Sérsic index, Sérsic ellipticity, asymmetry of the light distribution, concentration of the light, smoothness of the light distribution, Gini-M20 bulge parameter and Gini-M20 merger parameter (these are the same as in Section 16.1). We use these distributions to investigate which types of TNG galaxies fall into the domain of HSC galaxies and which types do not. We perform this study on TNG100 galaxies (blue) and TNG50 galaxies (orange), relative to their respective matched HSC samples. The dashed lines show the distributions while the solid lines show the median in 10 fixed bins of OOD Score. We define all galaxies with OOD Scores smaller than 1 as within the domain, while all galaxies larger than 1 outside of the domain. Note that the TNG100 and TNG50 samples shown here represent distinct galaxy image populations: e.g. the stellar mass of the TNG50 sample is significantly lower than the one of TNG100.

17

FIRST APPLICATIONS, INSIGHTS, DISCUSSION AND OUTLOOK

We conclude this work with two initial applications of our trained model and a discussion about our findings. Again, all analyses discussed so far have been performed on the test set only. First, we use the model to identify counterparts between TNG and HSC galaxies. This will serve as a visual validation of the work we presented in the previous sections. Second, we infer the redshift, Petrosian radius and i-band-magnitude for the HSC galaxies, by using only the information from TNG, to get a first idea of the performance of this method for simulation-based inference.

17.1 IDENTIFYING COUNTERPARTS BETWEEN TNG AND HSC GALAXIES

It is possible to use the representation spaces constructed in this paper to identify counterparts between sets of galaxy images. In particular, to identify the most similar TNG/HSC galaxy, we choose the nearest neighbour (using again the sum-norm) in representation space for the HSC/TNG galaxy in question. To visualize the variance in potential counterparts we not only show the closest neighbours but the three closest neighbours. To also investigate the importance of the OOD Score we show multiple examples of counterparts for various ranges of OOD Scores. The following is a compilation of visuals:

- In Figure 37 we show HSC galaxies with OOD Scores < 1 and their TNG50/100 counterparts; i.e. each HSC galaxy is well within the domain of the TNG galaxies (i.e. within the domain, by our definition).
- In Figure 38 we show TNG50/100 galaxies with OOD Scores < 1 and their HSC counterparts; i.e. each TNG50/100 galaxy is well within the domain of the HSC galaxies (i.e. within the domain, by our definition).
- In Figure 39 we show HSC galaxies with OOD Scores > 1; i.e. each HSC galaxy is in a region that is significantly underpopulated by TNG galaxies (i.e. out of domain, by our definition).
- Finally, in Figure 40 we show examples as in Figure 39 but with OOD Scores > 1.5, to focus on counterparts of HSC galaxies that come from a region really outside the domain of the simulated ones.

As the three bands relating to the images' RGB have been stretched independently in the preprocessing, the colours of the images might appear different from the original data: differences in the apparent colour should be therefore not a direct concern.

To first order, we see a good agreement among the galaxy images identified as counterparts, in terms of their type and shape. In fact, based on the visual inspection of Figure 39, choosing an OOD Score value of 1 to discriminate between in-

and out-of-domain may be rather conservative: as shown in Figure 39, we can find a number of galaxy counterparts with OOD Scores > 1 (and usually smaller than about 1.5) that actually appear rather similar. By placing the out-of-domain threshold at e.g. 1.3 instead of 1.0, we find that only 8 (instead of 33) per cent of the TNG100 and 6 (instead of 30) per cent of the TNG50 images are out-of-domain relative to the HSC images, or conversely that as many as $\approx$ 90 per cent of the simulated galaxies could be considered realistic.

This said, for TNG50 and TNG100, we are clearly able to identify problematic cases by using the OOD Score, particularly for cases with OOD Score > 1.5 as those of Figure 40: the method and metric developed in this paper can therefore be applied as a first cleaning step for simulation-based inference.

17.2 A PREVIEW: ARE WE ABLE TO INFER THE REDSHIFT, PETROSIAN RADII AND I-BAND-MAGNITUDE FROM THE HSC IMAGES?

In the previous sections we have mainly focused on how to *compare* the survey-realistic mocks of TNG simulated galaxies with HSC galaxies. However, the ultimate goal of our project ERGO-ML is to infer physical parameters from observed galaxies by using the information provided by simulated mock galaxies. In future work we want to expand on this idea by training an ML model on the representation space to transfer the properties from the simulated galaxies to the observed ones. Here, as a preview, we want to present a simpler application: we perform a nearest neighbour interpolation among the TNG galaxies.

For each HSC galaxy (in the test set) we choose the nearest TNG galaxy (in the test set) and use the property of that TNG galaxy as a 'prediction'. We consider the same galaxy properties we used for the sample matching in Section 14.3.2: redshift, Petrosian radius and i-band-magnitude. We already ensured that their definition is as similar as possible between the two domains of simulated and observed images. Furthermore we ensured with the galaxy-sample matching, that their prior distributions are the same for both datasets. And although these galaxy parameters are still very basic, their prediction from the images is not trivial as, for example, we normalize the images in brightness and FOV.

We show the results of this exercise in Figure 41. In the top we show the interpolated quantities from TNG100 against the ground truth quantities from the HSC survey. We see that overall the prediction follows the ground truth for all three properties of interest, albeit with a rather large scatter. However, the reader is to be reminded that we did not train on any labels and this prediction is fully self-supervised, i.e. this prediction is just based on the positioning of galaxies in representation space according to their appearance. By incorporating these representations into a supervised machine learning method, which is able to perform higher order interpolations, we can expect much better results. We also emphasize that the photometric redshifts used for the HSC galaxy sample have a comparatively large error of $\sigma[\Delta z_{\mathrm{phot}}/(1+z_{\mathrm{phot}})] \approx 0.05$ [177].

In the bottom we show the absolute prediction error against the similarity score of each HSC galaxy, i.e. the OOD Score. In doing so we check if the scores we introduced and derived in Section 16.2.2 are meaningful in terms of prediction capability. HSC galaxies that are far out of the domain populated by the TNG

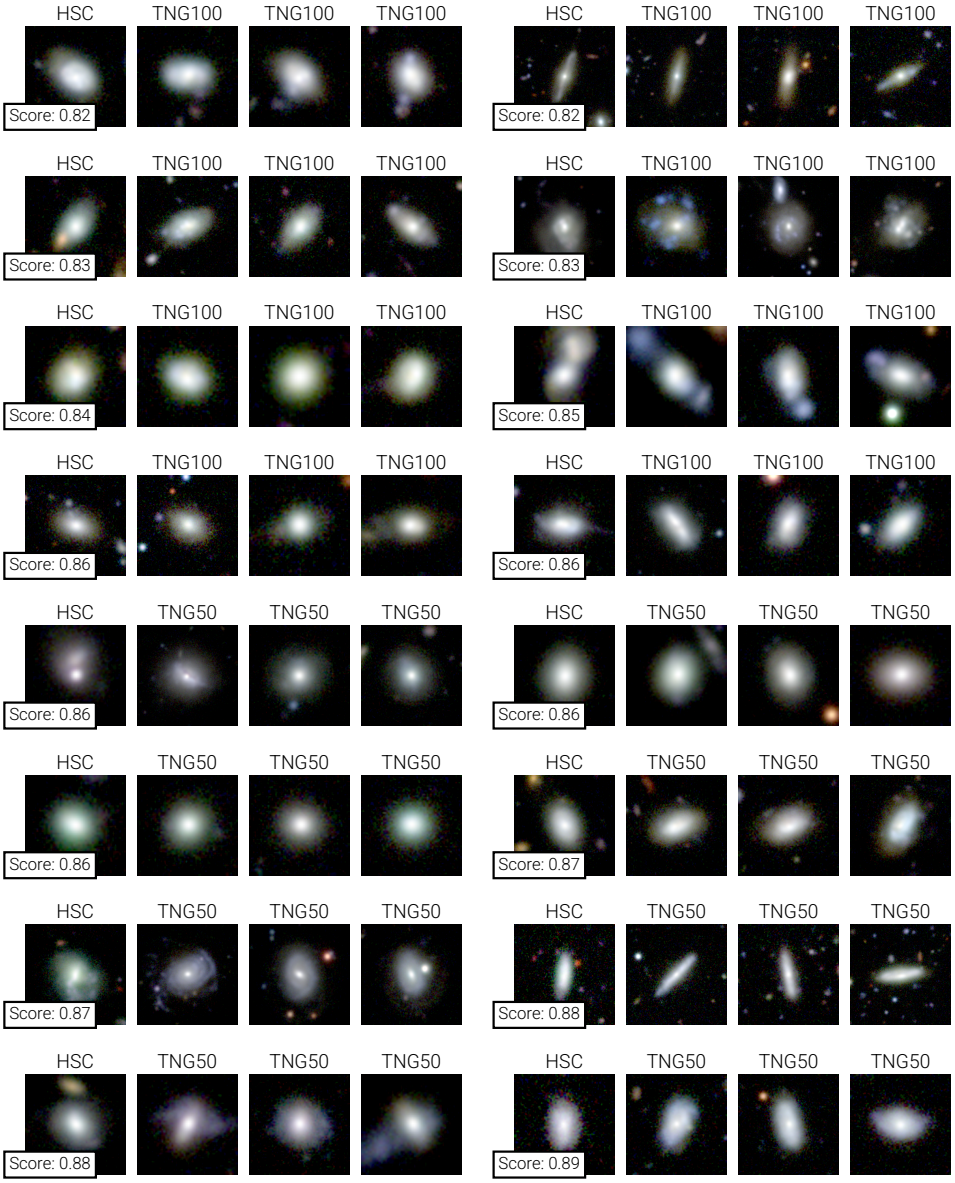


Figure 37: **Identified TNG50/100 counterparts of HSC galaxy images.** We show TNG50/100 counterparts of randomly selected HSC galaxies which have been identified as the closest neighbour in the representation space (using the sum-norm). Additionally we show the OOD Score for each HSC galaxy. We see overall a good visual match among the galaxies.

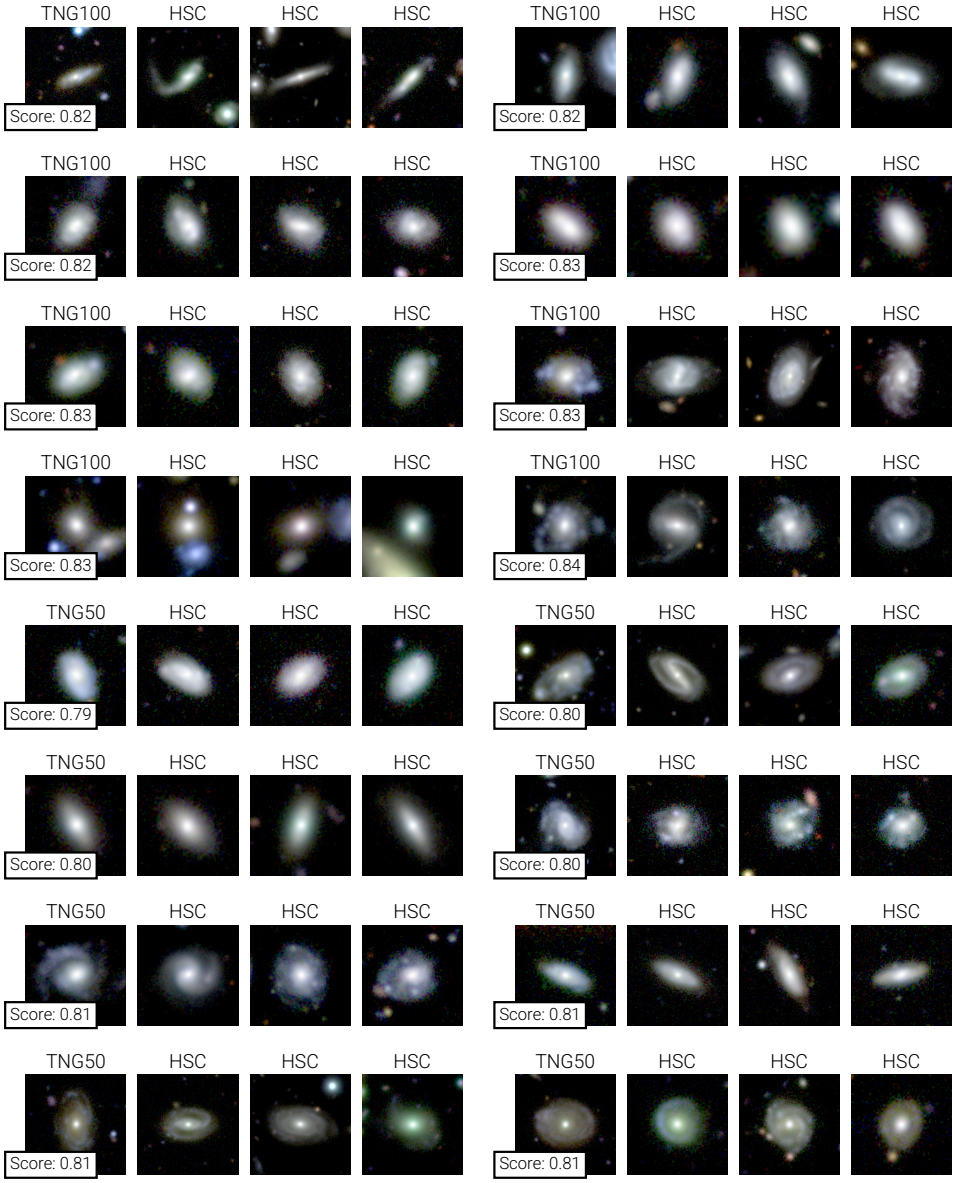


Figure 38: **Identified HSC counterparts of TNG50/100 galaxy images.** Same as Figure 37 but with HSC counterparts for randomly selected TNG50/TNG100 galaxies.

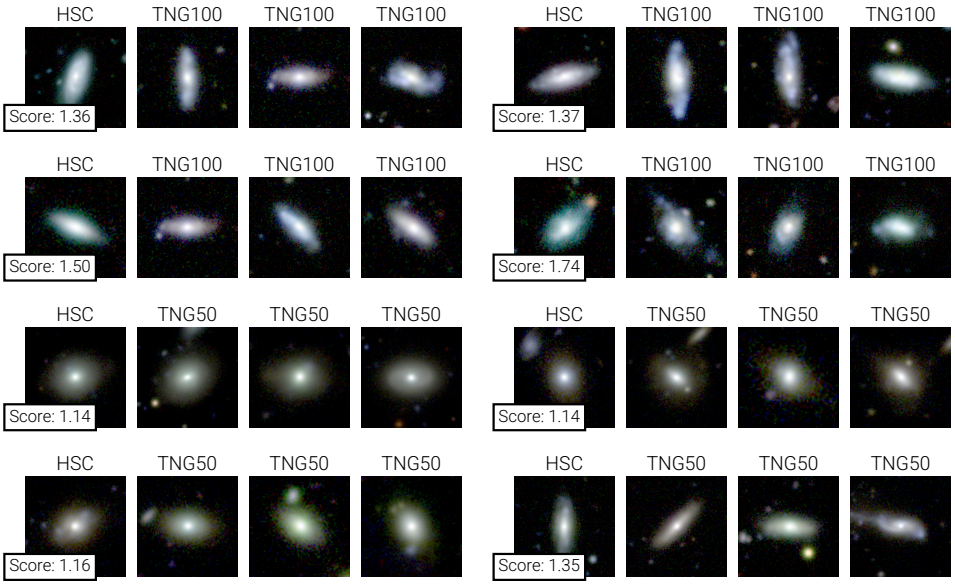


Figure 39: Same as Figure 37 but with randomly selected HSC galaxies with high OOD Scores. We do not see any significant differences, which instead would have been expected given the high OOD Scores. So, based on this visual comparison, it appears that, in fact, the TNG sets seem to extrapolate well beyond their own domains.

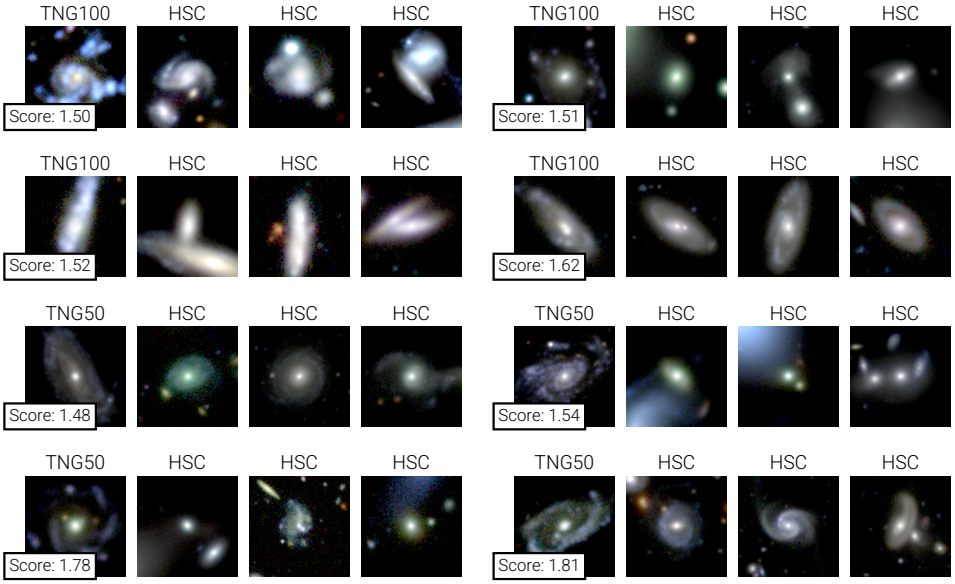


Figure 40: Same as Figure 38 but with very high OOD scores. Here we clearly see that the identification of out-of-domain galaxies works. In particular, we see galaxies that are in a very crowded environment and/or for which the Petrosian fitting did not work i.e. captured not only the light profile of the central galaxy but a whole cluster leading therefore to a very large FOV. We also find cases of very large neighbouring galaxies and galaxies with a very clumpy appearance. These galaxies could be excluded successfully for scientific applications by using the OOD Score.

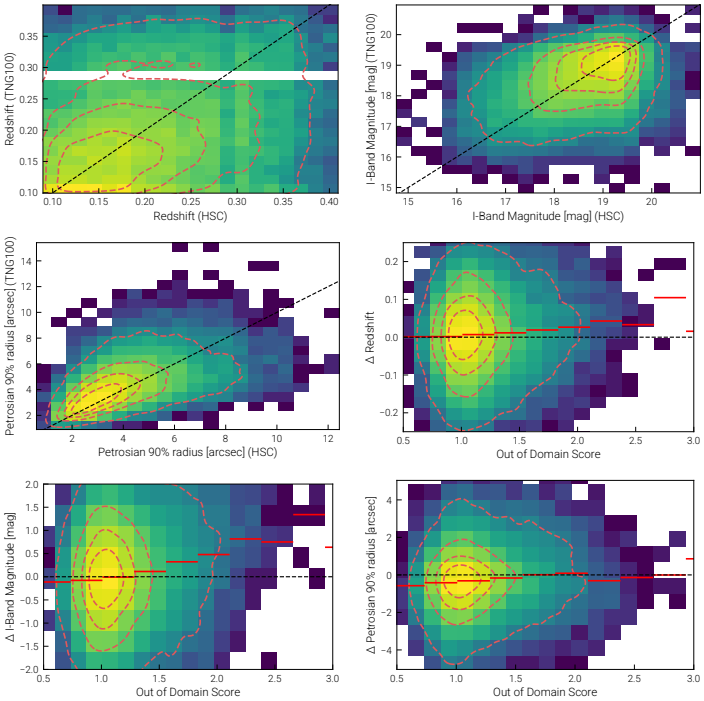


Figure 41: **How well can we reproduce galaxy properties of HSC by interpolating from TNG100 galaxies in representation space?** For this exercise we assign to each HSC galaxy the properties of the closest TNG100 neighbour in the 256-dimensional representation space and we focus on redshift, i-band magnitude and Petrosian radius. In the top panel we show the 2D histogram of the results for each property: on the x axis is the HSC ground truth while the y axis shows the quantities derived from the interpolation from TNG100. In the lower panel, we show the dependence of the prediction error (HSC Ground Truth - TNG100-based prediction) on the Out of Domain Score. Additional to the histogram we show in each plot the 'ideal' result as black dotted line and outline the distribution of the histogram also with red dotted lines. We see that with our model we can nicely reproduce the i-band magnitude and the Petrosian radius while the quality of the prediction is also dependent on the OOD Score. Note that this result is fully unsupervised – at no point was any active training performed to reproduce these quantities. The large scatter in the prediction error is especially caused by the coarse interpolation to the nearest neighbor in a high dimensional – and sparsely populated - space: the results of this figure represent only a lower estimation for the overall capabilities of an inference model which is based on these representations.

galaxies, i.e. have no similar TNG counterparts, can not be expected to get a meaningful prediction for their properties (although an extrapolation might be possible of course). We see that this is indeed the case, and especially for the i-band magnitude, which is systematically over-predicted for larger OOD Scores. However, for the redshift and the radii the method extrapolates to the HSC domain without significant systematic biases. We also checked, although we do not show, that the inference works also between the two simulations, TNG50 and TNG100: the prediction error is much smaller than between TNG100 and HSC, which is expected as the galaxies have on average a smaller OOD Score and as we actually compare the exactly-known properties from the simulations. In conclusion, the extraction of galaxy properties based on the learned representations is promising: we will further improve when the unsupervised learning is augmented with a supervised training towards quantities of interest.

17.3 DISCUSSION

17.3.1 *Are IllustrisTNG galaxies sufficiently realistic for simulation-based inference?*

In this paper we have compared three distinct galaxy image datasets: one observed and two simulated. As this work is meant as a step towards a simulation-based inference framework, the relevant question now is: how realistic are the galaxy-image mocks from TNG?

On one hand, this comparison is important insofar as it will establish the degree to which the TNG simulations can produce realistic galaxy morphologies and populations. In other words, are the model assumptions, numerical recipes, and numerical resolution sufficient for this task? But it is also important for us to ensure that the identified TNG counterparts of HSC galaxies are indeed similar enough so that we can safely assume that also their (not directly observable) physical properties are the 'same'. If this is the case, then we can map the knowledge from the simulation on to the observations.

The problem of this task, determining whether two galaxies are similar or not, is threefold: (1) Two images which are (nearly) exactly the same still show two different galaxies with not necessarily the same properties. I.e. there can be multiple formation pathways that can lead to similar galaxy morphologies. We showed this already for an inference based on scalars [63]. However, also with 2D images this intrinsic uncertainty is present. (2) On the other hand, two viewing-angles of the same galaxy can look completely different. As our model is not aware of this, spirals e.g. split themselves into a face-on and edge-on region (see Figure 33). (3) The number of TNG galaxies is limited: We might not find for a given HSC image a very similar TNG galaxy. The model must therefore learn to interpolate among the various galaxy images and their morphological features. The quality of the model is therefore another source of uncertainty.

Previous works have shown that TNG galaxies reproduce, to first approximation, many fundamental properties and statistics of observed galaxies [e.g. 57, 74, 85, 150, 179]. In the previous Section 16.2.2, we have also shown that the largest majority of the simulated galaxies from TNG are well within the domain of observed HSC galaxies, without any remarkable difference between TNG100 and TNG50

[differently than what found by 199, with an ML-based comparison of galaxy images from SDSS]. However, details of the stellar structures of TNG galaxies have also been shown to deviate in certain aspects and regimes from observations. In particular:

- Tension in the morphology-colour-size relation was found between observed Pan-STARRS and simulated TNG100 galaxies [150].
- It was shown with previous machine learning approaches that the fine-grained details in the structures of certain classes of galaxies, specifically small, more spheroidal, and quenched galaxies, are different between observed SDSS and simulated TNG galaxies [199].
- Higher concentrations and asymmetries were found in low-mass TNG50 galaxies compared to Kilo-Degree Survey (KiDS) observations [79].
- Finally, in this work we have seen that certain TNG galaxies exhibit blue clumps in the images that are not visible for HSC (Figures 31 and 32).

For the purposes of a simulation-based inference framework, what needs to be done is therefore to develop a method that only maps from similar and realistic-enough galaxies and that allows us to identify and exclude non-similar galaxies. Or even better, what is needed is a metric that allows to place a confidence score on the predictions based on some identifiable galaxy properties.

In this work, by matching the datasets according to redshifts, radii and luminosities, we have already ensured that the three datasets have the same priors in terms of very basic parameters. We have then trained the contrastive learning method NNCLR on all datasets (simulated and observed) at the same time: in this way, the model is agnostic about the fact that it is looking at images from different domains and can find similarities and differences among the images. We have then compared the images based on the location of their respective representations derived by this model. To this aim, we have introduced an Out of Domain (OOD) Score that describes the local difference in point density in the high dimensional-representation space. In doing so we are able to identify regions in the representation space where galaxies of a dataset are over- or under-represented. As a success we have demonstrated that simulated galaxies with TNG and observed galaxies with HSC overall land roughly in the same space. The simulation methods and the mocking procedure are therefore able to produce large samples of sufficiently realistic galaxies. We can for example also see that our model is e.g. able to understand the concept of a disk galaxy across the domains (see the counterpart examples of Figure 37 and following). We can further use the OOD Score to gauge the level of realism, as we expand upon next.

17.3.2 *The relation between OOD Score and the degree of realism*

We now want to further elaborate about the meaning of the OOD Score introduced in this work in terms of realism of the TNG set: is a low OOD Score equivalent to a high degree of realism? And vice versa?

A low OOD Score means that a TNG galaxy is mapped together with HSC galaxies to the same location in the representation space. Based on our checks, we know

that there is an overall correlation between representation space and observable properties (e.g. Figures 27, 28, 29). Furthermore our search for counterparts (Section 17.1) and our basic interference exercise (Section 17.2) showed consistent results.

However, one of the uncertainties in the proposed approach are the features that the model looks at: are those the image properties and galaxy features that we are actually interested in? For example, there is a small minority of cases whereby a low OOD Score is actually assigned to galaxies that do not resemble their neighbours in representation space. We have found galaxies with a very low OOD Score that are not in fact similar, upon visual inspection, to their identified counterparts. This occurs for example when the Petrosian fit fails and the FOV is too large or too small (Figure 52). The ensuing difference in galaxy size with respect to the FOV is then identified as a major feature by the model and HSC and TNG galaxies are grouped according to it, mostly ignoring other observable features. This grouping then leads to a small OOD Score as the failing FOV determination is the common and defining issue.

Keeping these uncertainties in mind, we can provide an answer to the opening question ("is a low OOD Score for simulated galaxies equivalent to a high degree of realism?") with a conservative yes. But it is always advisable to validate the OOD Score against basic galaxy properties and parameters that are available from the simulations, to double check the learned features and to eliminate outliers.

On the other hand, is a high OOD Score for simulated galaxies equivalent to a low degree of realism? A high OOD Score means that a TNG galaxy is in an area with few or no HSC galaxies. It is therefore identified as peculiar by the model. For TNG, we have seen that this is the case for e.g. galaxies with larger radii (half mass and sersic half light), steeper sersic profiles, smaller sersic ellipticities and larger asymmetries in comparison to the rest of the simulated galaxy population. Even though we should remember that, because of unreliable fits with Statmorph, we might lack some peculiar galaxies, we have confirmed that the high OOD Scores typically correspond to peculiar and non-realistic cases (Figure 40). Interestingly, as noted above in the discussion of Figure 39, we have also found that HSC galaxies that have been assigned an OOD Score larger than 1 can still be seemingly similar to their TNG counterparts. This could be caused by the fact that some disky TNG galaxies with clumpy features are repelled from the smoother HSC disks, leaving in their region more HSC than TNG galaxies and causing therefore an OOD Score of 1 or more: there may still be realistic TNG galaxies in that region but ultimately we do not grasp the full diversity of galaxy formation as not all TNG disks land in the same region.

Keeping these uncertainties in mind, we can answer to the second question ("is a high OOD Score for simulated galaxies equivalent to a low degree of realism?") also with a conservative yes. But it is advisable to consider various threshold values in OOD Score, i.e. various operational definitions of OOD Score, to discriminate between within and out of domain.

17.3.3 *Implications for simulation-based inference*

Finally, we want to comment about the implications of our results on the broader goal of implementing simulation-based inference from galaxy images. The starting point of this paper was the idea to replace the scalar set of inputs of the inference model presented in [63] with images. We can now replace the scalar inputs with the results of the ResNet pre-trained and validated in this paper, as we know that the resulting representation space faithfully encodes observational features of the imaged galaxies. Most notably, this is the case not only for TNG galaxies but also HSC galaxies. Therefore, with this paper, we have not only skipped the definition and calculation of summary statistics, but we also have a tool to gap the bridge between observations and simulations.

In future applications, to deal, on an first order, with the discrepancies between TNG and HSC galaxies, we can e.g. exclude all TNG/HSC galaxies from training that have an OOD Score above 1. Furthermore, we suggest to validate this method also by including mock images of other cosmological simulations like e.g. EAGLE and SIMBA. This would provide additional sets of a ground truth we can test against, similar to our inference test between TNG50 and TNG100 in Section 17.2, and would greatly lower the chance to overfit towards unique details of the TNG mocks. Moreover, this would allow us to automatically marginalize over multiple galaxy formation and evolution models, as the posterior distributions inferred for any given parameter or property would also include the uncertainty caused by different numerical realizations.

Another possible application and avenue allowed by the tools developed in this work could be a check of the inferred values against scalars that are derived 'traditionally' i.e. by analytical modelling. This can for example be done for properties like the photometric redshift (like done in Section 17.2 or the galaxy stellar mass. The ability to derive those statistics from stellar light images with a sufficiently high precision would give the model a higher credibility and a broad applicability.

18

SUMMARY AND CONCLUSIONS

Modern cosmological hydrodynamical simulations for galaxy formation and evolution are now able to return tens of thousands of synthetic, reasonably realistic galaxies across cosmic time. However, quantitatively assessing the level of realism of simulated universes in comparison to the real one is difficult and, to a certain extent, open ended. The main goal of this work was to develop a framework to compare simulated (IllustrisTNG) galaxies with observed (HSC: Hyper Suprime-Cam) galaxies at the image/map level without the need to define and calculate summary statistics. In particular, we have used a large sample of hundreds of thousands synthetic HSC-SSP images of galaxies from the TNG50 and TNG100 simulations [26] and compared them with real HSC-SSP galaxies over $0.1 \leqslant z \leqslant 0.4$ in the galaxy stellar mass range of $10^{9-12} M_{\odot}$.

To conduct the comparison, we have used imaging data from both HSC and TNG50/100 sets to train a Wide-ResNet model, by using the NNCLR contrastive learning method. Our model maps the images onto a common 256 dimensional representation space, whereby representations of galaxy images that are identified to be similar/different are close/far to each other. Using these representations, we have examined the degree of similarity among the datasets. In Section 16.1, we showed that the representations correlate strongly with physical galaxy properties derived from the simulations. We showed that they also correlate with image-derived morphological properties like Sersic parameters. The network is therefore able to encode the morphological structures of galaxies in the representations and is by that also able to distinguish galaxies based on their morphological properties. In Section 16.2 we have investigated how the representations of mocks from TNG and HSC galaxies are related to each other. We have shown that there is overall a large intersection between the TNG and HSC datasets in the representation space, which also implies that the mocking procedure is overall working and that the simulation itself produces sufficiently realistic galaxies.

To quantify the level of agreement between simulated and observed galaxy images, we have introduced in Section 16.2.3 a score, the so-called Out Of Domain Score, that aims at quantifying the similarity of a given galaxy relative to the galaxies in another set. We have connected the predicted values of the OOD Score of galaxies to a selection of –in principle observable– galaxy properties.

In Section 17, we have shown first applications of our framework and methodology: for example, it can be used to find similar TNG galaxies of HSC galaxies; or it could be used in the opposite direction, to find HSC galaxies with certain properties, for example disk galaxies. As a first-order approach, we have shown that it is possible to infer redshift, i-band-mag and Petrosian radii from HSC galaxy images by taking the prediction from their closest TNG neighbour in representation space.

Our main findings are:

1. Contrastive learning is an efficient way to learn morphological features of galaxy images (Figure 27 and following) without the need for prior labeling.

2. HSC galaxies cluster close to the TNG galaxies in representation space: our model is not able to distinguish between simulated and observed galaxies for a large subset ($\gtrsim$ 70 per cent) of the overall adopted galaxy population (Figures 30 and 34). This indicates that the galaxies produced in the IllustrisTNG simulations are overall realistic also at the stellar-light image level.

3. At the same time, we can identify types and subsets of TNG galaxies that have not been observed by HSC (Figures 36 and 40), and that hence are not realistic. For example, especially problematic are blue clumps in and around TNG50 and TNG100 galaxies that are less common in HSC observations. This outlier, or 'out-of-distribution', detection allows us to assess the realism of the simulated galaxies and their stellar light images.

4. We are able to identify galaxy counterparts between HSC observations and the TNG simulations (Figure 37 and following).

5. It is possible and practically feasible to transfer galaxy properties from the simulated galaxies on the observed HSC galaxy images (Figure 41).

In the next work of our ERGO-ML series, we will use the representation space obtained in this work as input to a conditional Invertible Neural Network (cINN) to infer properties of observed galaxies by using the underlying connections predicted by the cosmological galaxy simulations. We will aim to predict posteriors for observable quantities (e.g. magnitudes, redshift, galaxy sizes/profiles), semi-observable quantities (e.g. stellar mass, fraction of disk stars) and unobservable quantities (e.g. ex-situ fraction, merger mass and times) directly from the images for observed HSC galaxies. The Out-of-Domain Score that we have introduced and studied in this paper will allow us to e.g. excluding from the training subsets of simulated galaxies that are deemed to be not sufficiently realistic.

Part V

THE ASSEMBLY HISTORIES OF HSC GALAXY IMAGES VIA INVERTIBLE NEURAL NETWORKS, CONTRASTIVE LEARNING, AND COSMOLOGICAL SIMULATIONS

The following chapter is based on the work 'ERGO-ML: The assembly histories of HSC galaxy images via invertible neural networks, contrastive learning, and cosmological simulations' which is currently in preparation and yet not published. In the framework of ΛCDM cosmology, massive galaxies grow over time by both forming their own stars as well as by gradually accumulating stars (so-called ex-situ or accreted) from other merging or orbiting galaxies. This process follows a hierarchical pattern, whereby smaller galaxies merge to form bigger ones. However, since we can only observe galaxies at particular points in cosmic history, we cannot directly witness the dynamics of those (past) merger events. In this paper on ERGO-ML (Extracting Reality from Galaxy Observables with Machine Learning), we develop an algorithm that is able to infer the posterior distributions of summary statistics of the assembly history of observed galaxies directly from Hyper Suprime-Cam (HSC) images. To do so, we apply the self-supervised contrastive learning framework NNCLR on a set of realistic HSC mock images (g,r,i - bands) produced from galaxies simulated within TNG50 and with stellar masses between $10^{9-12}\mathrm{M}_\odot$ and from TNG100 galaxies with stellar masses between $10^{10-12}\mathrm{M}_\odot$. The resulting representation is then used as conditional input for a cINN (conditional Invertible Neural Network) to gain posteriors for the merger/assembly statistics, particularly the lookback time and stellar mass of the last major merger and the fraction of ex-situ stars. Through validation against the ground truth available for simulated TNG galaxies, we assess the performance of our model, achieving good accuracy in inferring the stellar ex-situ fraction ($\leqslant \pm 20$ per cent) and the mass of the last major merger for masses above $9.5\log\mathrm{M}_\odot$ ($\leqslant \pm 0.5\log\mathrm{M}_\odot$). The information content about the lookback time is, instead, limited. For the first time, we use our simulation-based model to infer the assembly-history properties of galaxies observed with HSC, in particular from HSC images that are close to the domain of TNG ones. Despite challenges such as information loss (e.g. due to projection effects and surface brightness limits of observations) and domain shifts (between simulation and observation), our results demonstrate the feasibility of extracting merger quantities from galaxy images.

19 INTRODUCTION

Understanding the assembly history of galaxies is of paramount importance in unraveling the mysteries of cosmic evolution. How galaxies accumulate their mass and undergo mergers shapes their present-day properties and holds key insights into their past. Traditionally, relations such as the Tully-Fisher relation have provided valuable empirical connections between observable galaxy properties and their underlying assembly histories. Recent advancements in machine learning (ML) have opened new avenues for probing galaxy evolution, particularly in inferring unobservable properties from observable data. In our previous work [63], we demonstrated the feasibility of simulation-based inference using scalar features, laying the foundation for exploring galaxy assembly histories. Similarly, [7] utilized a similar approach to infer the ex-situ fraction from Manga IFU data, although limited by the availability of high-quality IFU data. However, the question remains: to what extent can we investigate galaxy merger and assembly histories using visual images, especially considering the sparse availability of high-quality IFU data? In our recent paper [61], we addressed the challenge of ensuring comparability between simulated and observed galaxy images through representation learning, setting the stage for a deeper exploration of galaxy evolution using visual data. In this paper, we take the final step towards bridging the gap between visual images of galaxies and merger statistics. Using IllustrisTNG mocks as ground truth, we employ ML techniques to learn the relationship between visual features and assembly histories. Subsequently, we extend our methodology to infer the same quantities for observed galaxies from the HSC-SSP Survey, providing unprecedented insights into the merger and assembly histories of real galaxies.

20 GALAXY DATA

In this paper, we investigate the connection between galaxy morphology, represented by images in visual bands, and their merger and assembly history. To accomplish this, we utilize a comprehensive dataset consisting of highly resolved simulated galaxies from the TNG50 and TNG100 simulations within the IllustrisTNG project. These simulated galaxies are meticulously tailored to mimic the characteristics of galaxies captured by the Hyper Suprime-Cam (HSC). Additionally, our analysis incorporates observational data from the HSC galaxies. It is important to note that the data used in this study is the same as that employed in our previous works [61, 63]. Here, we provide a succinct overview of the data; for more detailed information, readers are encouraged to refer to the referenced papers.

20.1 MERGER/ASSEMBLY STATISTICS

To summarize a galaxy's past, we select three key statistics listed in Table 5. The definition and selection follow [63], however, we choose to remove the average merger ratio and lookback time due to the relatively poor prediction results in the previous study and the rather theoretical nature of average quantities derived from the simulated merger tree. Nevertheless, galaxy merger histories are derived from the SUBLINK tree of progenitor subhaloes/galaxies from which then meaningful scalars are derived.

Cosmological simulations reveal two primary channels for star assembly: in-situ star formation and ex-situ star accretion, the latter contributing to the "ex-situ" stellar mass fraction distributed across galaxy bodies [e.g., 126, 133]. Given its significance, we employ the ex-situ stellar mass fraction as a primary indicator of a galaxy's past merger history. This fraction has been previously examined in simulations [e.g., 49], observational discussions [e.g., 87] and was also inferred by our previous (scalar-based) model [63].

Recent major mergers can significantly influence galaxy properties. Therefore, we also consider the lookback time and stellar mass of the last major merger undergone by a galaxy [147]. To distinguish galaxies without major mergers, we set a lower mass threshold for major mergers at $8.5 \log \mathrm{M}_{\odot}$. Mergers falling below this threshold are not considered major mergers, ensuring our sample is not contaminated by low-mass 'major mergers', particularly prevalent in TNG50 simulations.

20.2 GALAXY IMAGE DATA

20.2.1 *Observations: Subaru Hyper Suprime-Cam images*

We utilize galaxy images from the Public Data Release 3 (PDR3) of the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP [3]) as described in detail in [61]. The HSC-SSP, conducted over 330 nights with the Subaru telescope on Mauna

Name	Description	Note
Last Major Merger Mass	Total stellar mass of the last merger with a stellar mass ratio $\geqslant 1/4$ and a minimum total stellar mass of $8.5 \log \mathrm{M}_\odot$.	The mass is defined as the maximum total stellar mass the secondary had during its lifetime. If there is no major merger, the parameter is set to the value of $10^4\,\mathrm{M}_\odot$, i.e. below the resolution limit of TNG100 and TNG50.
Last Major Merger Time	Lookback time of the last merger with a stellar mass ratio $\geqslant 1/4$ and a minimum total stellar mass of $8.5 \log \mathrm{M}_\odot$.	The lookback time is relative to the redshift of the galaxy under consideration. If there is no major merger, the parameter is set to the unphysical value of 15 Gyr.
Stellar Ex-Situ Fraction	Ratio between ex-situ to total stellar mass.	We track individual stellar particles through cosmic time using their unique particle ID: more specifically, we call in-situ the stars of a galaxy that have formed within its main progenitor branch and call ex-situ those that formed in galaxies on secondary branches of the galaxy merger tree, following the definitions and results of [134, 147]. This mass ratio includes all stellar particles identified by SUBFIND to belong to the respective subhalo i.e. all gravitationally-bound stellar particles without using any aperture.

Table 5: **Unobservable galaxy properties used in this work and predicted for observed HSC galaxies.** We list and define the unobservable properties of galaxies' past assembly and merger history that are used in this work and that are obtained by applying a ResNet and cINN to HSC stellar-light galaxy images. Quantities are adapted from [63], note however the changed definition of a major merger, which now includes a mass cut at $8.5 \log \mathrm{M}_\odot$ of the secondary.

Kea, Hawaii, provides a comprehensive survey with various layers of data [68, 98, 111, 112]. The focus here is on the Wide Layer, offering extensive coverage and depth.

To construct a galaxy sample, a process similar to the one outlined in [161] is used. This involves cross-referencing sources in the HSC-SSP Wide Layer with those in the Sloan Digital Sky Survey (SDSS DR16 [2]). Selection criteria are applied, including limiting Petrosian magnitudes and ensuring objects fall within the HSC-SSP footprint. Additional constraints are then used to refine the sample, removing objects affected by various factors like bright stars, bad pixels, and cosmic rays.

In contrast to the original method of [161] which used SDSS galaxy spec-z data for reference, this work opts for a larger dataset, sacrificing the reliability of spec-z information in favor of a broader dataset for machine learning applications. The redshifts are primarily utilized for matching observed galaxy populations with simulations within a specific cosmic epoch range in Section 20.3.

The galaxy sample is comprised of cutouts centered on each galaxy, with a consistent size and covering a range of photometric redshifts. In total, the sample contains images of 768,484 galaxies in the r, g, and i bands, facilitating further analysis and investigation into galaxy properties and evolution.

20.2.2 *Simulations: TNG50 and TNG100*

We translate galaxies from the TNG50 and TNG100 simulations into synthetic observations that mimic the HSC (Hyper Suprime-Cam) survey, covering a redshift range of $0 \leqslant z \leqslant 0.7$. These HSC-realistic images are closer described in [26, 61]. Throughout this process, we undertake complex modeling steps: utilizing SKIRT for dust radiative transfer, which allows us to account for emission, absorption, and scattering of stellar light across various wavelengths; distinguishing between light emissions from stellar populations older than 10 million years and younger populations embedded in birth clouds; and incorporating dust attenuation by assigning dust densities to gas cells using a post-processing method.

Furthermore, we generate idealized images with high spatial resolution using SKIRT. Subsequently, we merge these images into real HSC-SSP fields using adapted techniques to consider cosmological effects and physical-to-angular size scaling based on the snapshot redshift of each galaxy. It's worth noting that our study focuses on galaxies meeting specific stellar mass thresholds to ensure they contain a minimum number of stellar particles and are detectable within the sensitivity limits of the HSC-SSP.

The simulated galaxy sample is drawn from snapshots between $z = 0.1$ and $z = 0.4$, with stellar masses above $10^{10} M_\odot$ (TNG100) and $10^{9} M_\odot$ (TNG50). We image each galaxy along four lines of sight, resulting in an extensive dataset for analysis.

20.3 DATA PREPARATION

To ensure a fair comparison between real and synthetic galaxy datasets, several preparatory steps are taken:

1. Field of View (FOV) Selection and Resizing: The FOV is based on the apparent Petrosian radius which contains 90 per-cent of the light r_{p90} calculated with the same algorithm from observed and mocks as well. With this, we maintain consistency across different galaxy sizes and across the domains of simulated and observed images. Images are cropped to $4 \times r_{p90}$ and resized to a common grid of 256^2 pixels.

2. Galaxy Population Matching: Observational and simulated galaxy samples are matched based on redshift, apparent magnitude, and Petrosian radius to enforce a common selection function, addressing differences in intrinsic distributions [150]. The matching algorithm and results are presented in detail in [61]. This matching limits our sample to the depth of the HSC sample which is $< 20\mathrm{mag}$ in the i-band.

3. Band Selection and Stretching: RGB representation using HSC i, r, and g bands is adopted, with image stretching to capture galaxy features across varying surface brightness levels [22].

4. Dataset Splitting: The dataset is divided into training, validation, and test sets to prevent overfitting. Progenitor/descendant galaxies are segregated to avoid biased test results [63].

These steps ensure uniformity in preparation, preventing systematic differences that could affect subsequent analysis. The total number of images in our sample after the preparation process, TNG50/100 and HSC including 4 projections for each TNG image, is 628,710.

21

FROM IMAGES TO POSTERIORS: MACHINE LEARNING METHODS

In the past, simple relations like Tully-Fisher could be described with analytical models. However, inferring posteriors (i.e., 1D functions) for unobservable quantities from images (2D data grids) is a complex task. To do so, we use the power of machine learning, to be precise, neural networks to cope with this problem. We use the pre-trained ResNet from [61], which is not only trained to convey morphological features into a high-dimensional representation space but is also able to test if mock images from TNG and observed HSC galaxies are similar enough to transfer the learned knowledge from TNG to HSC. These representations are now input for a conditional Invertible Neural Network to model posteriors of interest. We proved in previous work that this task is feasible using a limited set of integral properties [63], which are now replaced by the high-dimensional image representation. In the following, we will shortly reintroduce the basic concepts and describe in detail the overall training procedure.

21.1 RESNET AND CONTRASTIVE LEARNING

To translate the images into a low-dimensional representation that we can use as a condition for the cINNs, we employ a Wide-ResNet Architecture [198]. Within the ResNet, instead of standard convolutional layers, we utilize so-called "E(2)-Equivariant Steerable Convolutional" layers [193]. With this, we explicitly account for the fact that the orientation of the objects on the image plane should be irrelevant; i.e., all convolutional layers are equivariant regarding transformations of the Dihedral Group D(16). In simpler terms, this includes discrete rotations of $360°/16 = 22.5°$ and reflections.

With the aim of speeding up training and avoiding a domain shift between observed and simulated galaxies, we train the ResNet using the self-supervised method 'Nearest Neighbor Contrastive Learning' (NNCLR) [60]. Here, we create two transformed images for each image and teach the ResNet to recognize transformed images of the same galaxy and discriminate between transformed images of different galaxies. The random transformations we use include:

- Reflections along the horizontal and vertical axes
- Small rotations ranging from $-10°$ to $+10°$
- Small affine shifts with a maximal displacement of 10 per cent of the overall image side length
- Central zoom-ins, i.e., the center of the image stays at the center of the crop
- Gaussian noise

We also continue to use these transformations later during cINN training. However, during validation and testing, we use the original images. The NNCLR method was used to pre-train the ResNet used in this work in [61] but is also utilized in this work to ensure that the training of the cINN inference head is not leading to a domain shift when unfixing the ResNet weights.

21.2 CINN

In this paper, we adopt the same architecture as in [63], utilizing Conditional Invertible Neural Networks (cINNs) to infer the unobservable merger history. However, unlike [63], we employ image representations from a ResNet as conditional input. Conditional Invertible Neural Networks (cINNs) as introduced by [10], a generalization of Invertible Neural Networks [9]. These networks map, in our use case, merger-history statistics of interest x to a tuple of observable image representations y and latent variables z, offering a mechanism for inferring the posterior distribution $p(x|y)$. This is performed by enforcing a multivariate Gaussian distribution for the latent variables z through negative log-likelihood loss during the training. The well-known Gaussian distribution can then be used to sample the posterior of x, conditioned by y. Implemented via the FrEIA framework [11] which are based on PyTorch [129], cINNs have already demonstrated success in astronomical applications [e.g., 63, 99].

21.3 TRAINING PROCEDURE

The overall model is represented by the cINN with the last layer of the ResNet serving as conditional input. We initialize the ResNet with the weights of the pre-trained model obtained via contrastive learning from [61]. This approach allows us to begin with a network that has already demonstrated an understanding of the basic morphological features of galaxies. Furthermore, we keep the weights of the ResNet fixed in this initial step. This ensures that backpropagation focuses solely on modeling the posterior.

We utilize an Adam optimizer [95] to back-propagate the negative log-likelihood (NLL) loss of the TNG50 and TNG100 training sets through the network. This is done in random batches of 256 images. After processing 300 randomly chosen batches (which we refer to as an epoch), we use the images in the validation set to assess the model's performance. We set an initial learning rate of 0.005, which is reduced by a factor of 0.5 if the validation loss does not decrease for 5 epochs. Once the learning rate falls below a specified level of 0.0005, we proceed to the second step of training: the ResNet weights are unfixed and are now also updated during training. This allows the ResNet to specialize for the inference of the quantities in question. Our tests indicate that significantly lower validation losses can be achieved with unfixed ResNet weights.

However, there is a risk associated with unfixed ResNet weights, as the model may become overly focused on the peculiarities of the TNG mocks. To mitigate this risk, we augment the NLL loss (based on the output of the cINN) with a Contrastive Loss (based on the output of the ResNet). This contrastive loss is calculated on an additional batch of training images from the TNG50, TNG100, and HSC sam-

ples. This approach actively prevents the domain of simulated and observed galaxy images from diverging.

Therefore, the overall loss of the second step is expressed as:

$$L = L_{NLL}(I_{TNG}) + L_{CL}(I_{TNG}, I_{HSC})$$

where I_{TNG} represents the simulated TNG50 and TNG100 images, and I_{HSC} represents the observed HSC images. We continue the training until the validation loss L_{NLL} has not improved for 20 epochs.

22 RESULTS

In the following, we utilize the test set defined in Section 20.3 for multiple tasks: first, we validate the prediction capability of our model. We then retest whether the predictions obtained by training on TNG galaxies are indeed transferable to observed HSC galaxies. Finally, we apply the model to HSC galaxies identified as sufficiently similar to TNG galaxies to uncover, in part, their merger and assembly history.

22.1 GENERATING POSTERIORS AND MAPS

The cINN is trained to provide sample points of the learned posterior distribution, collectively representing the uncertainty associated with the non-invertible nature of the problem at hand. For each galaxy in our test sample, we draw 512 points that represent this distribution. To obtain a continuous representation of the posterior distribution from the samples, we utilize kernel density estimation (KDE). This yields a smooth, continuous function that approximates the underlying probability density function of the data. This continuous representation enables a more detailed analysis of the uncertainty associated with the predicted quantities, thereby enhancing the interpretability and utility of the model's outputs. However, to compare the model's predictions to the ground truth of the test TNG galaxies, we require a single quantity. While one could also use the mean or median of the posterior, we choose the MAP (Maximum A-Posteriori) estimation, which corresponds to the maximum of the posterior distribution. To find the MAP for each unobservable quantity, we employ a KDE of the posterior based on the 512 samples. We evaluate the estimated density on a 512-point grid along each dimension. Although this approach has a discrete nature, limiting prediction accuracy, it is significantly lower than the posterior widths and therefore negligible. It's important to note that for the lookback time and mass of the last major merger, we introduce an additional bin for galaxies or predictions with no major merger history (NMMs). If at least 50 per cent of posterior samples fall into this bin, the cINN MAP is set accordingly. Taking additional information about the posterior into account, we calculate the standard deviation of the posterior point cloud along each dimension. This provides insight into the uncertainty or precision of the MAP predictions. We display some examples of inferred posteriors in Figure 42 for a subset of test TNG100 galaxies. Initially, we observe that the ex-situ inference appears to be quite accurate, with higher precision towards the lower ex-situ end and greater ambiguity towards the higher mass end. The lookback time of the last major mergers sometimes closely resembles the prior, indicating that the model cannot extract any information from the image for those specific galaxies. Additionally, similar to the inference from scalar properties [63], we notice that the posteriors are often non-Gaussian, justifying the use of a method capable of modeling arbitrarily shaped posteriors.

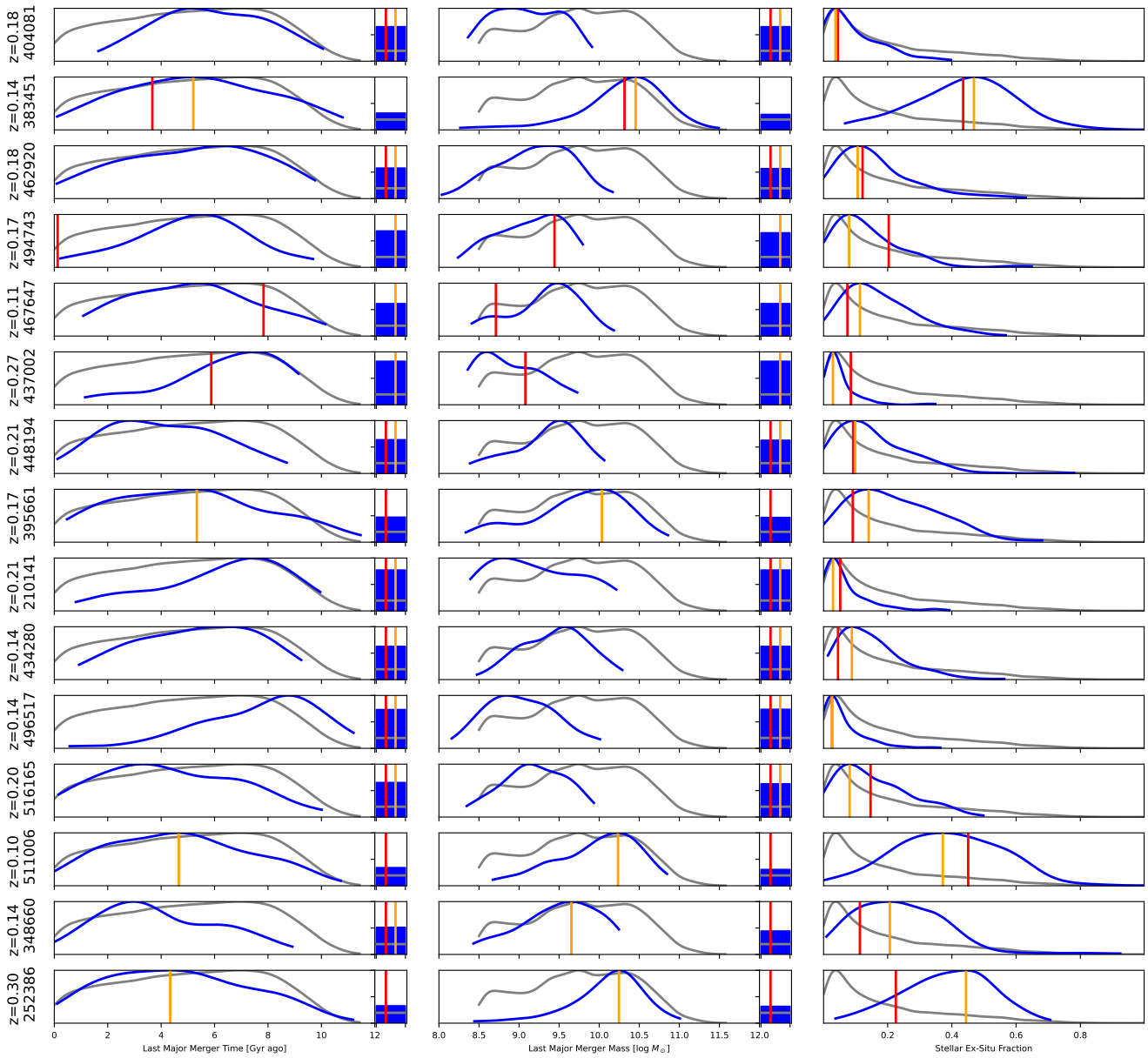


Figure 42: **Posterior distributions of the three assembly history statistics for 15 randomly selected galaxies from the TNG100 simulation test set.** Each row represents a single galaxy, with its redshift and unique SUBFIND ID listed on the left, while each column displays the inferred unobservable statistics of its assembly and merger history. The distributions are depicted for the entire test galaxy sample, i.e. the prior (in grey), our cINN model predictions (in blue), Maximum A-Posteriori (MAP) estimates (in yellow), and the ground truth from the TNG simulations (in red). These distributions are normalized so that the maximum value equals 1. Additionally, we include an extra bin for the last major merger quantities, indicating the fraction of posterior samples falling outside the valid regime, representing the probability that the galaxy has not experienced a major merger in its history. The MAP estimate for this bin is utilized only if the fraction of posterior samples exceeds 50 per cent, indicating a greater than 50 per cent likelihood that no major merger occurred during the lifetime of the respective galaxy.

22.2 VALIDATION OF INFERENCE RESULTS

To validate the model's predictive power on an actual population basis, we plot the MAPs defined and calculated in the previous section for the test set consisting of TNG50 and TNG100 galaxies. We then compare these predictions against the ground truth. The results are depicted in Figure 43: firstly, we display MAPs vs. Ground truth and see a very good match for the ex-situ fraction, as previously mentioned in Figure 42. However, we notice that the predictions tend to favor lower ex-situ masses due to the prior, which heavily favors low ex-situ fractions. For the mass of the last major merger, we find a good match towards the high-mass end, while the prediction diminishes towards the lower-mass mergers. The predicted MAPs of lookback times of the last major merger exhibit two maxima at around 2 Gyr and 8 Gyr. It's important to mention that we display the results of galaxies from a range between $z = 0.1$ and $z = 0.4$, with the distribution heavily leaning towards the low redshift end. The fraction of correctly identified galaxies that have had no major merger in their lifetime (or with a major merger mass below $8.5 \log M_\odot$) stands at 46 per cent. The fraction of falsely identified no major merger galaxies is at 20 per cent. However, we observe how the number of false classifications grows with less massive ground truth major merger mass. Our model produces a posterior function also for galaxies identified to have no major merger in their lifetime. Hence, we can not only plot the MAPs but also alternative peaks that might be closer to the ground truth. By doing this exercise, we observe a significant improvement in the prediction capability towards lower mass major mergers and the corresponding lookback time. To incorporate the information contained in the posteriors, we display in Figure 44 the standard deviation of the posteriors against the ground truth values and the respective absolute errors of the MAPs (MAP - Ground Truth). Overall, we observe that the qualitative trends are similar to those observed when using a set of scalar observables [63]: the posteriors' widths for the ex-situ fraction are very low for low ex-situ galaxies and increase towards higher ex-situ fractions. Additionally, there is a noticeable connection between the standard deviation and the absolute error for the ex-situ fraction, although the errors appear to be less Gaussian than the posteriors based on scalars.

22.3 ARE THE RESULTS TRANSFERABLE TO THE DOMAIN OF OBSERVED HSC IMAGES?

After assessing the model's prediction capability with the ground truth, our next step is to identify the observed HSC galaxies that fall into the same region in representation space as the training TNG galaxies. This process is crucial to circumvent a domain shift that would render the application of the model to the domain of observed galaxies pointless. We largely follow the approach introduced in [61] to identify the distances between galaxies in the representation space. From this, we calculate the Out of Domain Score (OOD) on a per-galaxy basis to determine how dissimilar a galaxy is from another dataset. We use the quotient d_{other}/d_{self}, where d_{self} represents the distance to the 8th nearest neighbor for each galaxy image in its own set, and d_{other} represents the distance to the 8th nearest neighbor in the other set. It's important to note that the representation space changes during training as

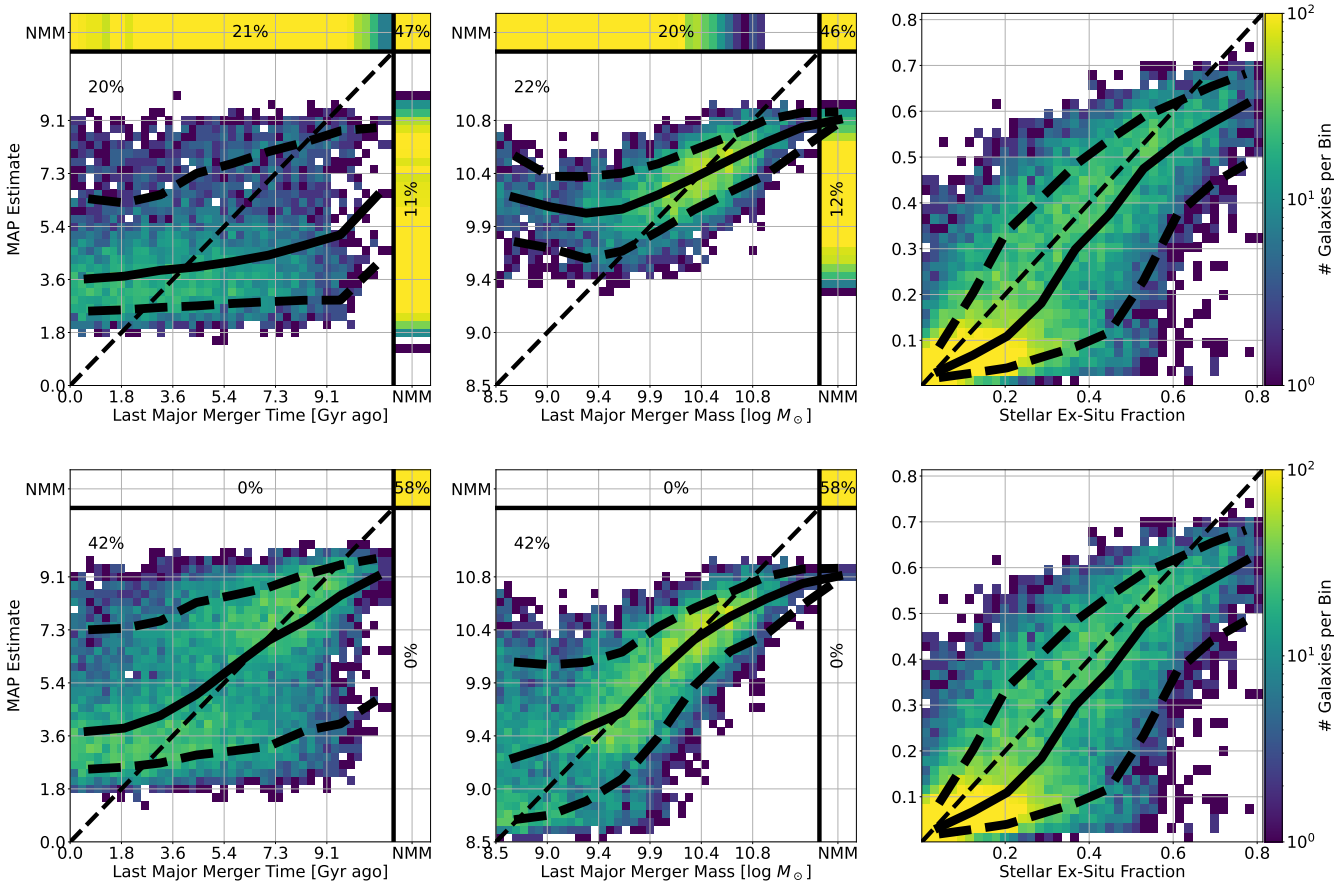


Figure 43: **Validation of our cINN's predictive accuracy based on testing against ground truth data from the TNG50 and TNG100 simulations.** The columns display results for three inferred but unobservable statistics concerning the assembly and merger history of galaxies, assessed using TNG50 and TNG100 test galaxies. **Top:** Two-dimensional histograms between cINN Maximum A-Posteriori (MAP) predictions (y-axis) and the simulation ground truth (x-axis) for each of the three output quantities. The solid black line represents the median, while the area encompassed by the black dashed lines encompasses 80 per cent of the data points. The ideal scenario, where MAP prediction and ground truth align perfectly, is depicted by a dotted black diagonal line. **Bottom:** Similar to the top panel, but instead of MAPs, we use the predicted value of the posterior peak closest to the ground truth. This adjustment allows for the visualization of inferred values for misclassified non-major merger galaxies of the mass and lookback time of the last major merger. Additionally, for the time and mass of the last major merger (2 panels on the left), we show galaxies identified by the cINN as having no major merger (NMM) (top), or according to the ground truth (right), and both (upper right corner). The percentage denotes the overall fraction of galaxies in each of these three categories.

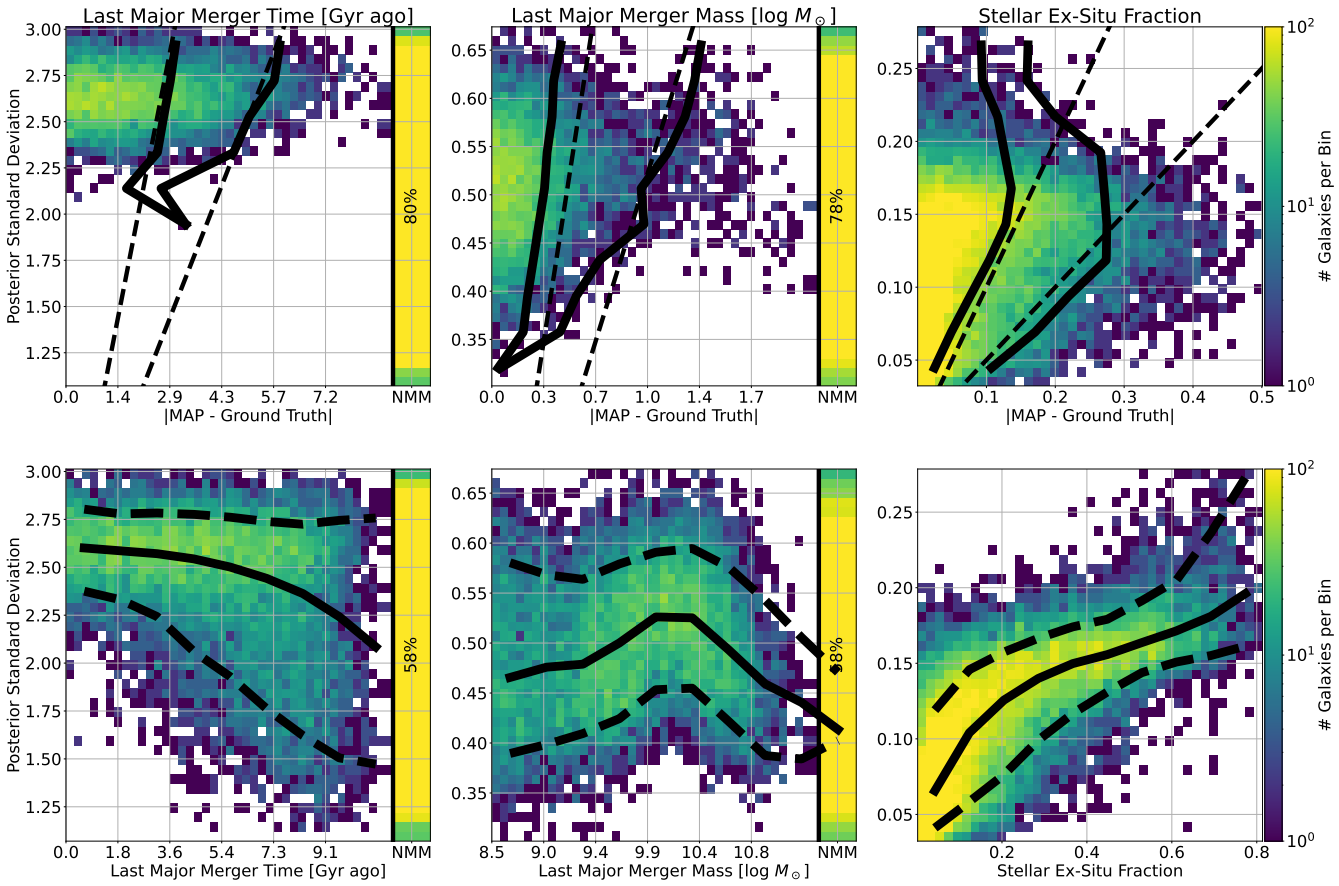


Figure 44: **Uncertainties of the cINN Predictions. Top:** Two-dimensional histograms of the standard deviation of cINN posterior distributions (y-axis) versus the absolute value of MAP prediction errors (|MAP predictions - ground truth|, x-axis) across all TNG50 and TNG100 test galaxies. Thick black curves represent the 68th and 95th percentiles of test galaxies within bins of posterior standard deviation. Dashed lines illustrate the expected outcome under the assumption of normal distributed posteriors. **Bottom:** Two-dimensional histogram of the standard deviation of cINN posterior distributions (y-axis) versus the simulation ground truth (x-axis). The black curves indicate the median, with 80 per cent of test galaxies falling within the area enclosed by dashed lines. For the time and mass of the last major merger, galaxies identified by the ground truth as never having experienced a major merger (NMM) are represented in an additional bin. The fraction of galaxies falling into this bin is denoted. In such cases, standard deviations are calculated solely from posterior samples within the physical range (i.e., lookback time $\leqslant$ 13.7 Gyr and stellar mass $\geqslant 10^{8.5}, \mathrm{M}_\odot$).

we unfreeze the ResNet weights and proceed with end-to-end training, including a contrastive learning step. Comparing Figure 45 to the results of [61], we observe that the overlap of the two IllustrisTNG simulations and the observed data has qualitatively not changed significantly toward the representations achieved by a ResNet trained only using contrastive learning. This indicates that adding the recursive cINN head does not lead to a significant domain shift that renders the model inapplicable to the population of observed HSC galaxy images. Examining the OOD scores calculated for HSC galaxies relative to their matched counterparts (HSC-TNG100 and HSC-TNG50, dashed lines), we notice that the OODs tend to be larger than their counterparts calculated for TNG towards HSC. This suggests that the fraction of HSC galaxy images not as well covered by the TNG sample is larger than the fraction of TNG galaxies not well covered by the HSC sample. Since we can only reliably infer (merger) properties for HSC galaxies whose images are close to images in the TNG sample, we introduce a cutoff. To assess the influence of this cutoff, we display in Figure 46 the distributions of OOD scores against the MAPs of the three inferred merger statistics. We present this separately for HSC galaxies matched to TNG50 (orange) and HSC galaxies matched to TNG100 (blue). Cutting off larger OOD scores would particularly remove a significant fraction of the low ex-situ fraction, driven by the high frequency of low ex-situ fraction galaxies in the sample. Interestingly, the high ex-situ fraction of the HSC galaxies matched towards TNG100 has OODs around 1, while the OODs for HSC galaxies matched towards TNG50 increase when approaching the high ex-situ end. This could be caused either by the dissimilarity of the TNG50 mocks towards HSC or simply by the fact that TNG50 has fewer high mass and therefore high ex-situ galaxies in the sample. Similar trends can also be observed for the mass and lookback time of the last major merger towards recent and more massive mergers, which naturally correlate with a larger ex-situ fraction. We do not explicitly display the distribution of OODs for the fraction of HSC galaxies identified to have no major merger; they largely resemble the distribution of OODs at the low mass/lookback time last major merger end.

22.4 RECOVERING THE ASSEMBLY HISTORY OF HSC GALAXIES VIA IMAGES

In this section, we take the next step and apply the model to observed HSC galaxies. As mentioned in the previous section, we must ensure that the images for which we want to infer information are close enough to the set of simulated images we trained on to ensure that (1) the observed image in question has a corresponding TNG image and (2) the area in the representation space around that image is sufficiently sampled to allow training of the cINN. For now, we choose to only use HSC galaxies with an OOD smaller than 1, although we observed by visual inspection in the previous study [61] that galaxies with OOD > 1 might still be sufficiently similar. In Figure 47, we present the resulting distributions for all three merger statistics: the total HSC sample, the HSC sample with OOD smaller or equal to 1, and the TNG50 and TNG100 samples for comparison. Overall, the relations seem to be well imprinted from TNG to the HSC galaxies. The most intriguing difference is the much larger number of low ex-situ fraction galaxies in HSC. However, this number decreases slightly when removing the predictions of HSC galaxies

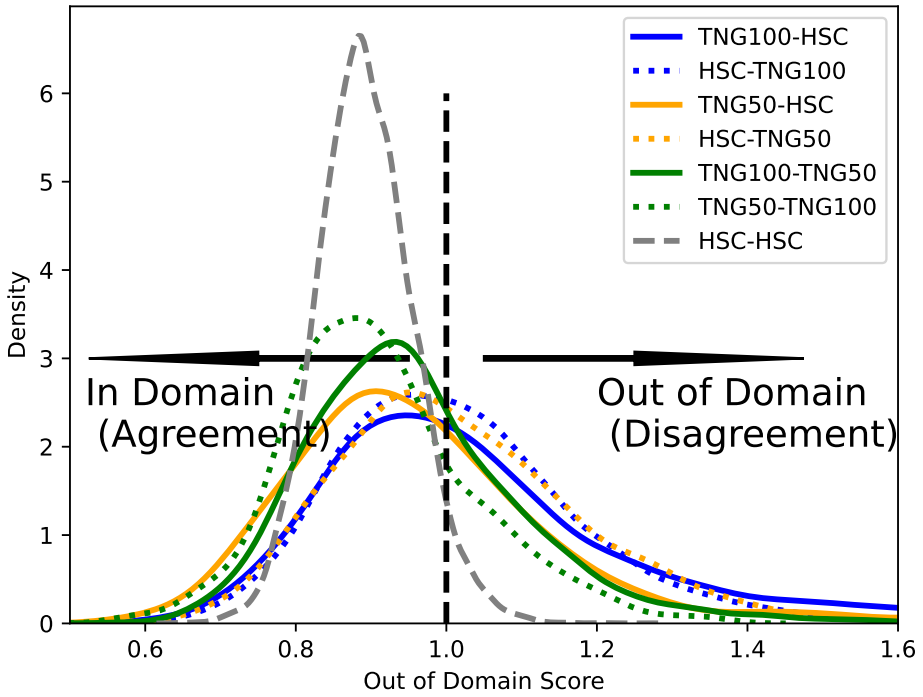


Figure 45: **The Distribution in Out of Domain Score (OOD) after retraining the pre-trained ResNet with a cINN head.** We show the overlap between the domains of the observed (HSC) and simulated (TNG50 and TNG100) datasets after re-training the ResNet of [61] with the cINN head to infer merger properties. This exercise is important to ensure that the inference has not lead to a strong domain shift between simulations and observations. The plot shall therefore be compared to Figure 10 in [61]. We evaluate the Out of Domain Score (OOD) for various scenarios, including comparisons between pairs of the three datasets and between two random splits of each dataset. This investigation aims to understand the inherent scatter that persists even when two distributions are drawn from the same population. For simplicity, we present only the self-distance for the HSC dataset, as the self-distances exhibit qualitatively similar trends for all three sets. It's important to note that there is a distinction in which 'side' this quantity is measured: For instance, TNG100-HSC represents the distribution of distances for TNG100 galaxies relative to HSC galaxies, while HSC-TNG100 represents the distribution of distances for HSC galaxies relative to TNG100 galaxies.

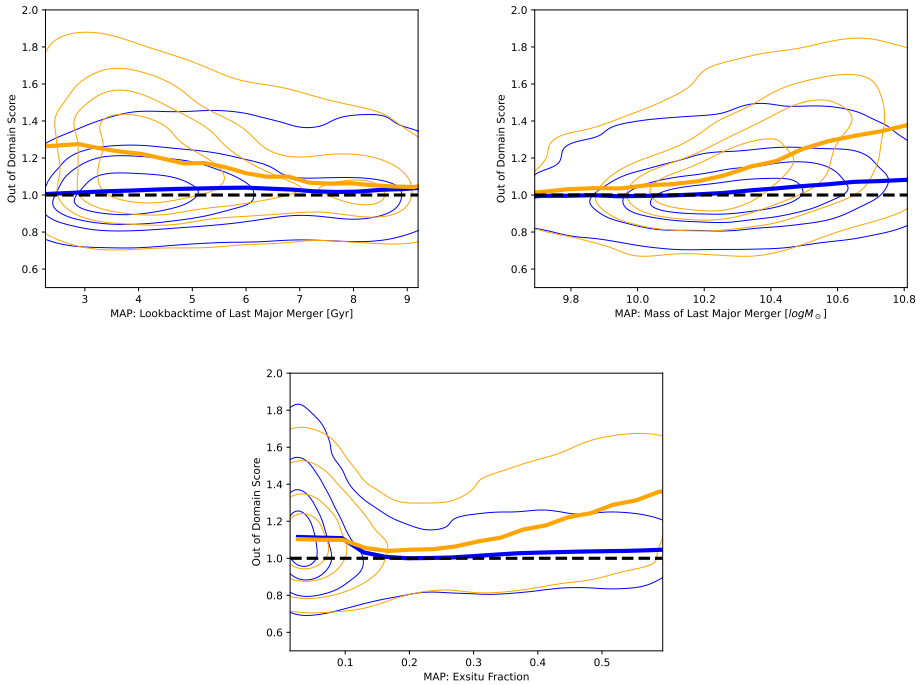


Figure 46: **For which HSC galaxies can we trustfully infer merger properties from the TNG simulations?** We plot the inferred merger parameters of HSC galaxies introduced in Table 5 against the OOD score of HSC galaxies. With this, we can identify for which subset of HSC galaxies we can perform a trusted inference based on the TNG simulations. We perform this study on TNG100 galaxies (blue) and TNG50 galaxies (orange), relative to their respective matched HSC samples. The dashed lines show the distributions while the solid lines show the median in 10 fixed bins of properties.

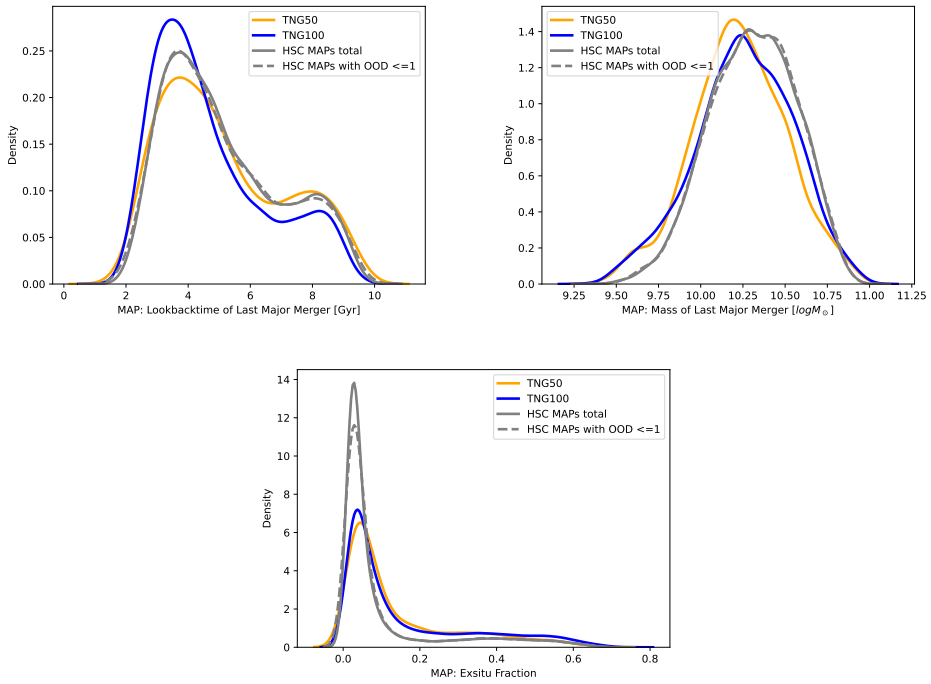


Figure 47: **Merger History of HSC galaxies based on galaxy images:** We show the distribution of MAP predictions for the tree merger quantities (Lookback time and Mass of the Last Major Merger, Ex-situ Fraction) for the galaxies in our HSC test image set (grey line). We also show the distribution for HSC galaxies that are well within the domain of TNG images by choosing an OOD cutoff $\leqslant 1$ (dashed grey line). Here the OOD is calculated for all HSC galaxies relative to the combined set of TNG50 and TNG100 galaxies. For comparison, we also show the distribution for TNG50 and TNG100 galaxies. The fraction of HSC galaxies that are predicted to have –most likely– no major merger is $\approx$ 80 per cent.

that are far out of the domain of the TNG galaxies. Another interesting feature is that the number of galaxies with ex-situ fractions around 0.2 is slightly suppressed. However, the cut-off of the OOD score has no significant influence on the mass and lookback time of the last major merger. Our model infers that around 80 per cent of the HSC galaxies in our sample had no major merger in their lifetime (stellar mass ratio $> 1/4$ and stellar mass above $8.5 \log M_\odot$). The inference for the HSC galaxies tends slightly more towards larger lookback times and higher merger masses. This is an interesting result, considering that mergers with higher lookback times usually have lower merger masses.

23 DISCUSSION

In this work, we performed simulation-based inference of merging parameters from HSC galaxy images using TNG50 and TNG100 as the training set. Based on the validation via the ground truth available for the simulated TNG galaxies, we are able to judge the performance of our model. The inference seems to deliver a high amount of information content, especially for the stellar ex-situ fraction. Additionally, the mass of the last major merger can be well recovered for higher masses above $9.5 \log M_{\odot}$. Comparing this to our initial approach with scalar properties [63], we observe that although the results are similar, the overall errors seem to be larger. How can this be, even though the images should contain more information than the scalars inferred from them? Two main explanations can be given here: (1) the scalars we used in the earlier work were not all observable. Especially the fraction of disc stars is a very theoretical quantity not inferable from images but heavily benefits the determination of the ex-situ fraction. (2) We add observable realism to the images; therefore, we introduce a significant amount of information loss that is unavoidable in real-world applications. Overall, we still conclude that the determination of merger quantities based on images seems feasible. Regarding the domain shift between the dataset we train on (TNG50 and TNG100) and the dataset, we apply the model to, we can report that simultaneous training of the cINN with a log-likelihood loss and the ResNet via contrastive learning/NNCLR keeps the domains significantly overlapping. In the following sections, we will discuss more closely the options to improve the capability of the model and how trustworthy our inference results are.

23.1 WHAT CAN BE DONE TO FURTHER IMPROVE THE INFERENCE RESULTS?

In the preparation of the dataset, we already use observable quantities like the redshift, the i-band luminosity, and the Petrosian radius. Furthermore, the actual luminosity is scaled away by normalizing the pixel values to a range of $[0, 1]$. Additionally, the radius of the galaxies is lost during the Field of View (FOV) selection. Funneling this information, alongside the image representations, into the cINN model might lead to a significant improvement in the inference results. However, this is not a trivial task, as the redshift and i-band luminosities are exactly known for the TNG galaxies while they have to be determined from images for the HSC galaxies. This adds a large uncertainty to those quantities, which would have been accounted for when training the network with the exact simulation ground truth. One way to do so would be the calculation of the photometric redshift and luminosity from the mock images in the same way as for the real HSC images. The training scalars should (!) then be connected to the same observer bias.

Alternatively (or additionally), one could constrain the HSC set to galaxies with a known spectroscopic redshift whose uncertainty is, in general, much smaller and well-known. Another improvement could be introduced by more informa-

tive merger and galaxy properties which are inferred alongside: Significant cross-correlations, like with stellar mass, might add additional information to the training of the network without adding the necessity to add more observable data into the model during the inference phase. These additions will be addressed in the final version of this paper.

Finally, it is important to consider how the uncertainty in Figure 42 is formed: by the intrinsic information limitation of the input, insufficient realism of the simulation, or insufficient modeling? We will delve into this in the next section.

23.2 CAN WE TRUST THE INFERENCE?

By using contrastive learning including simulated and observed galaxy images, we ensured that the model is not focusing too much on simulation-specific features. This approach is overall successful, as we are able to imprint information from TNG on the set of HSC data. Nevertheless, we know that our mocks are not perfect and there are small and large deviations between the domains. As in previous work, the definition of similarity is arbitrary, and it is also not ensured that two identical images of galaxies represent identical objects. In the future release of this paper, we will therefore also add the results of the EAGLE simulations: HSC EAGLE mocks created with the same algorithm as the TNG mocks we use up to now. With this, we can test our model with a second dataset with a ground truth: Will the inference results significantly deviate from the TNG results? And how big is the difference between the two simulations? That said, the simulation-based inference from cosmological simulations is still at an early stage. One can expect that with each generation of simulations, the mock surveys become more and more similar to their observed counterparts. Otherwise, if a significant gap persists, one could of course also question our current understanding of the cosmos.

24

SUMMARY AND CONCLUSIONS

In this paper, we undertook simulation-based inference of the assembly and merger-history properties of galaxies observed with HSC, by using visual HSC galaxy images (r,g,i -band) as input and survey realistic mock images of TNG50 and TNG100 simulated galaxies as the training set.

In particular, we infer posterior distributions for the mass and the lookback time of the last major merger and the ex-situ fraction of each galaxy by using the combination of a ResNet trained via self-supervised contrastive learning on TNG and HSC galaxies and of a cINN trained on the ground truth of the TNG50 and TNG100 simulations. We use galaxies in the stellar mass range between $10^{9-12}\mathrm{M}_{\odot}$ and redshifts between $z = [0.1, 0.4]$. We match the galaxies between TNG and HSC by the redshift, the integrated i-band luminosity, and the petrosian radius which contains 90 per cent of the galaxys light. With this, we train our models with overall 628,710 galaxy images (simulated and observed). Through validation against the ground truth available for the simulated TNG galaxies, we assessed the performance of our model. The inference demonstrated a high information content of the images regarding the stellar ex-situ fraction. This means, that we can constrain the ex-situ fraction well within ± 20 per cent for 80 per cent of the sample, allowing a clear distinction between low and high ex-situ galaxies. Additionally, we achieved accurate recovery of the mass of the last major merger for higher masses above $9.5 \log \mathrm{M}_{\odot}$ ($\pm 0.5 \log \mathrm{M}_{\odot}$ for 80 per cent of the sample). The lookback time of the last major merger on the other hand is much harder to constrain with the available information.

Comparing our approach to our earlier work focusing on scalar properties [63], we find similar results but with a larger overall width of the posteriors due to the increased observational realism, and therefore information loss, in this work. Despite these challenges, we conclude that the determination of merger and assembly-history quantities based on galaxy images is feasible and yields valuable information that has been, up to now, impossible to derive directly from observations. Our model maintains a significant overlap between the training and application domains, supported by the simultaneous training of the cINN and ResNet with a log-likelihood loss and NNCLR, respectively. This means, that we are able to infer merger quantities for almost 50 per cent of the observed HSC galaxies.

In the future, there are several directions for improvement. Incorporating additional observational information such as redshift, luminosity, and galaxy radius into the model may enhance the inference accuracy by reducing the posterior widths. Furthermore, constraining the dataset to galaxies with known spectroscopic redshifts could reduce uncertainties associated with the photometric redshift measurements. In terms of trustworthiness, our contrastive learning approach effectively mitigates the model's focus on simulation-specific features. However, the possible imperfections of our mock data and the inherent uncertainty in defining the similarity between simulated and observed galaxies call for further ways to

validate the results. Future work will therefore explore the use of additional simulation datasets for training, such as HSC EAGLE mocks, to further validate the inference with ground truths drawn from a different domain than the one used in the training. Furthermore, the influence of different simulation models on the inference can be assessed.

In summary, while cosmological simulation-based inference applied to galaxy evolution is still in a very early stage, our work demonstrates promising capabilities for extracting quantities describing the unobservable merger and assembly history from observed galaxy images. In fact, every quantity that can be defined and measured from simulations can in principle be transferred to observations. This yields a powerful tool as one can now compare observations and simulations with the same means. This allows for a better understanding of unobservable processes, like merging, in our universe. On the other hand, it allows also the identification of conflict points between simulations and observations by contrasting the inference results of 'semi-observables' like stellar mass against traditional models. Because of this, simulation-based inference from full cosmological simulations promises to heavily benefit observers in the interpretation of their data, as well as theorists in the development of future galaxy evolution models.

Part VI

DISCUSSION AND SUMMARY

25 DISCUSSION

We will now discuss the overall results of this thesis, but first, we want to remind ourselves of what has been done from a statistical viewpoint. The topic of this thesis is simulation-based inference of unobservable galaxy properties from galaxy observables, by learning the connection between observables and unobservables from cosmological simulations. Our machine learning model of choice (conditional Invertible Neural Networks, cINNs) learns the conditional posterior function $p_{TNG}(x|c)$, where x represents merger and assembly statistics and c represents the conditional inputs which are in this thesis either scalar observables (Part iii) or images (Part v). When training the model with TNG data, it is important to note that x and c are sampled from the domain of TNG. Furthermore, the posterior probability will be modeled relative to the prior $p_{TNG}(x)$ of the TNG dataset. We showed in Part iii that this concept is feasible, for example, to derive the ex-situ fraction from a limited set of observable integral galaxy statistics within TNG100. Since the actual first-order results of galaxy surveys are not summary statistics (which are furthermore hard to define across simulations and observations), we focus in the next two chapters (Part iv and v) on using visual images from HSC as observational input c.

The challenge with this is, that applying the learned posterior $p_{TNG}(x|c)$ to observables c drawn from a different domain than the one trained on, namely the domain of simulated survey realistic TNG mock images versus the domain of observed HSC images, will lead to non-meaningful results as we are applying the model out of the training domain. We therefore have to ensure that the samples drawn from the marginal distribution $p_{HSC}(c)$ (i.e. the distribution of images) used for the application, are not too far from the marginal distribution $p_{TNG}(c)$ used for the training.

In theory, one could directly compare the images in the space spanned by the pixel values, in this case, a $3 \times 256 \times 256 = 196608$ dimensional space. However, in such a space, even rotating a galaxy in the image space would lead to a completely different position in this space, which is not desirable. Because of this, in Part iv, we apply a contrastive learning method (Nearest Neighbor Contrastive Learning, NNCLR) that maps the images from both domains into a common 256-dimensional representation space. Contrastive learning is able to identify similarities and differences between the images in a self-supervised way. Furthermore, it may neglect minor or unimportant differences, such as background sources in the images, and helps to interpolate between the images (e.g. rotations of the galaxy in the image plane).

Assuming that the simulated mocks would be a perfect mirror of the real universe, the image representations of both sets would perfectly align with each other; however, we still see that there is a certain difference. The exact reason for this is unknown, but it is likely a combined effect of insufficient realism of the simulated galaxy, insufficient mocking of the simulated galaxies, insufficient modeling of the

representations, and also simply the fact that the marginal distributions of $p_{HSC}(c)$ and $p_{TNG}(c)$ are not identical a-priori (e.g. the HSC sample is subject to the selection bias of observations, which is not the case for the simulated TNG galaxies). Nevertheless, we were able to identify a fraction of samples from $p_{HSC}(c)$ which are close to the domain of $p_{TNG}(c)$, such that we can trust in applying them to $p_{TNG}(x|c)$ in a further step. To quantify this, we introduced the Out-of-Domain score (OOD), which is based on the distances between an HSC galaxy and its 8th nearest neighbor in the TNG representations, normalized by the distance to the 8th nearest neighbor in the HSC dataset.

With this, we are able to bridge the gap in Part v from simulations to observations: we train $p_{TNG}(x|c)$ with the TNG images $p_{TNG}(c)$ and then select from the domain of HSC images $p_{HSC}(c)$ the samples which are close to $p_{TNG}(c)$. The result of the inference for a single HSC galaxy can be interpreted as the posterior $p_{TNG}(x|c)$ of the most similar TNG galaxy to the HSC galaxy in question.

In the following, we are commenting on open questions and challenges of the approach taken in this thesis and also state future ideas for improvement.

25.1 LEARNING THE POSTERIOR AND THE ROLE OF THE PRIOR

With traditional inference models, like a Multi-Layer Perceptron trained with a Mean Squared Error, one can predict the mean of the posterior. By using an appropriate loss function, one can also predict the standard deviation of a Gaussian approximation of the posterior function. However, we find in Figures 14 and 42 posteriors with multiple peaks and askew posteriors. The power of our approach is the modeling of arbitrary posterior shapes: we can learn the distributions as provided in the simulation. On the other hand, a challenge we encountered throughout the thesis is the selection of the prior which will naturally influence the shape of the posterior. The cINN automatically learns the prior from the training data: In the scalar case (Part iii), this training set contains all central TNG100 galaxies within a mass cut of $10^{10} M_{\odot}$. When applying this model to observations, the posterior is then based on this simulation-based prior, neglecting any observer bias (luminosity cut, volume effects) which is not favorable.

But how do we choose the prior? As the prior is in the space of unobservables, we cannot match observations and simulations in that space. This is also not wanted, as a priori the merging and assembly history of simulations might differ from the real Universe. Instead, we match the marginal distribution of the simulated images $p_{TNG}(c)$ to the one observed $p_{HSC}(c)$. Simply speaking, we ensure that the images we use for the training are similar to the images we later apply the model to.

For the inference based on HSC images (Part v), we do this by matching the TNG galaxies to the HSC sample beforehand, using basic observable scalars: redshift, i-band luminosity, and the galaxy size (Figure 26). By this, we ensure that the galaxy population we train on are mimicking the population of HSC galaxies. However, the limitation of this matching to three quantities naturally also limits the quality of the matching. A future addition to this approach could use the learned image representation which encodes a much wider set of observable galaxy properties (Figure 27, 28 and 29). So before training the cINN on the TNG images, one could apply a second matching based on the representation space: for each HSC

galaxy, we choose the nearest TNG galaxy to construct a matched sample. This training sample and therefore the posteriors, as returned by the cINN, would then encode the prior $p_{TNG}(x)$ connected to the images $p_{TNG}(c)$ which are best possibly matched to the HSC images we want to apply the model to.

A question one could ask now: do we want an informative prior in the first place? One can normalize the posteriors shown in this work by the prior of the training data to obtain posteriors relative to a flat prior. We, however, decided that it is advantageous to carry the population-wise information of the simulation (e.g., low ex-situ galaxies are much more frequent than high ex-situ galaxies) into the inference.

25.2 THE REPRESENTATION SPACE: CHALLENGES AND OPPORTUNITIES

The idea behind representation learning is to map the input images into a meaningful representation that retains almost all of the important information but is easier to handle than the original high-dimensional data.

Beyond NNCLR, which was used in this work, there are other methods like SimCLR, BYOL [77], or Variational Autoencoders which can be used to construct such a representation space. In this thesis, we used and compared NNCLR and SimCLR, favoring NNCLR over SimCLR because of a better interpolation performance between galaxy images. This said, differences between NNCLR and SimCLR are not as substantial, such that the selection of the representation learning method appears crucial. Nevertheless, we want to emphasize that the resulting representation space might improve when using other methods and so also the capability of the model to distinguish and sort galaxies according to their appearance.

The representations we calculated in this work, representing an informative description of galaxy appearances, can have important applications also outside of ML: we were, for example, able to easily define outliers and group galaxies according to visual appearance (Figure 33). Until now, relations between galaxy properties can only be studied in bins of known statistics like redshift or stellar mass. With the representations, one could also identify correlations in bins of 'appearance'. This would be a natural extension to relations like e.g. Tully-Fisher, which is only valid in the 'appearance class' of disk galaxies. Furthermore, we were also able to identify simulated analogs of observed galaxies and vice versa. Observers/Theorists can therefore use the representations to construct a matched sample of Simulated/Observed galaxies.

A difficulty is that the learned representations also depend on the data which is used in the training. If there is high diversity in the data, the model might focus more on large-scale differences but less on small-scale differences or differences between simulations and observations. A possible adjustment for this could be an iterative approach: We could remove galaxies from the training set, that are far outside of the respective domain of the other set. By doing this, the actual intersection of the marginal distributions $p_{TNG}(c) \cup p_{HSC}(c)$ can be found, and regions of the representation space will be removed which are too sparsely populated to infer something. On the downside, this could also lead to an over-sensitivity of contrastive learning on negligible details like noise properties.

25.3 THE OUT OF DOMAIN SCORE AND THE CURSE OF DIMENSIONALITY

We showed in Figures 30 and 34 that the OOD is able to identify density differences in the representation space between the observed and simulated images. Nevertheless, we want to briefly emphasize the problem of using nearest-neighbor distances in high-dimensional spaces the OOD is based on.

The point density ρ in the representation space is given as $\rho = 8/V_8$, with V_8 as the volume of the N-dimensional ball containing the 8 nearest neighbors. The volume of an N-dimensional ball scales with $V_8 \propto d_8^N$, where d_8 is the distance to the 8th nearest neighbor and N is the dimensionality of the representation space. This relationship holds for any L-Norm, including the Euclidean norm utilized in this thesis. This implies that the 8th nearest neighbor distance is related to the actual point density via $d_8 \propto \rho^{-1/N}$. In high dimensions, the 8th nearest neighbor distances, and consequently the OOD, are only sensitive to very large changes in the point density. This should be taken into account when interpreting the OODs and this might also explain why we still see very similar galaxies even with an OOD larger than 1 (Figure 39).

A workaround for this problem could be to calculate the OOD in a lower-dimensional representation space, which can be for example achieved with UMAP. However, such a dimensional reduction with UMAP could potentially lead to a loss in information which might let the image domains appear closer than they are in the representation space.

25.4 INFORMATION CONTENT OF THE DATA

One open question is the dependency between the inference capability and the specifics of the data we provide. What is the influence of the projection of the galaxy on the image plane? Can we add more information with better-resolved images or other filters to better constrain the assembly and merger history? What is the improvement we can expect by adding more types of data, especially spectroscopic information?

These are interesting questions, especially as they can give us an idea of what we can expect from surveys in terms of constraining complex astrophysical processes like galaxy mergers. For the image-based model in Part v, we e.g. have removed information regarding the total luminosity (by normalizing the images) and the galaxy sizes (by choosing a FOV relative to the apparent galaxy size). In the full paper release, we plan to funnel back this information into the model, as we already measured those quantities directly from the images. Furthermore, future surveys like LSST promise large samples of galaxy images with depth down to almost 26.9mag (10 years combined in r-band) [90]. With this depth, stellar halo features like stellar streams and shells will be visible for a relevant-sized population of galaxies. As those features are theorized to originate from mergers, one can expect a large contribution from them in constraining quantities describing recent mergers. Furthermore, we want to point to the successes using IFU data like MaNGA [7] which allows us to also take stellar and gas kinematics and metallicity into account. Nevertheless, while adding more and better data might improve the inference results, one should also take care that this also limits the number of

galaxies to which such inference is applicable. For example, LSST 'only' images in 6 broadband filters. Galaxies observed by this survey and not cross-matched with spectroscopic observations (e.g., in the COSMOS field) can only be judged based on the imaging data. Another limitation is the resolution of the cosmological simulations used. Images (and other types of observable data) cannot be sufficiently mocked to arbitrary fine detail by current cosmological simulations to be compared to highly resolved images, such as those from the JWST.

25.5 CAN WE TRUST THE RESULTING POSTERIORS?

In this thesis, we developed an ML model that returns a posterior of unobservable quantities. But how can we ensure that those posteriors are correct? When testing such a model with simulated galaxies, one can use the ground truth and compare the inferred quantities against them.

As we do not predict point estimates but continuous posteriors, judging the accuracy of the model becomes even more complicated. This problem can be exemplified by looking at the MAP predictions of the mean merger ratios in Figure 15 on the one hand, suggesting that the model training failed. On the other hand, the training was successful in learning that the mean merger ratios are not well constrained by the observables available to the model. We want to note at this point, that we tackled this problem for the scalar case in Appendix 48 by calculating the calibration error.

It becomes more complicated if we specify the question: given an HSC galaxy, can we trust the posterior the model returns? The posterior we infer for simulated TNG galaxies in Part v contains the uncertainty introduced by the nature of the inverse problem. Due to the information loss connected to the imaging of a galaxy, the physical properties cannot be exactly inferred. This includes unobservables like the merger and assembly history but also 'semi-observables' like the stellar mass as well. Furthermore, the posterior contains the prior information learned from the simulation, such as the fact that galaxies with low ex-situ fractions are way more frequent than ones with high ex-situ fractions. The accuracy of this can be, again, tested using the ground truth of the simulation (Figure 43 and 44).

Hence the question, could there be an unrecognized domain shift between TNG and HSC while training the model on TNG? We tackle this question by simultaneously training the cINN and the ResNet via contrastive learning with both simulated and observed galaxies. By doing this, we ensure that the ResNet does not start to focus solely on simulation-specific details in the image. However, the mapping of the ResNet from images to representations is in fact not bijective. The reason for this is that in theory, differences between two galaxies could be ignored by contrastive learning and mapped to the same point in representation space.

How can we test if this is the case? We would need another set of galaxy images with ground truth, which are completely independent from the TNG image set. This could be, for example, image mocks from the EAGLE simulations treated like the HSC images. If the inference results from the model trained on TNG work also for images from EAGLE, this would be a strong argument that transfer of knowledge works. If the inference results differ, this could be caused either by a domain shift or by the fact that EAGLE produces different galaxy properties.

Another option could be the inference of semi-observables for which estimates from other, more traditional methods exist for observed galaxies. If, for example, the stellar mass is correctly inferred alongside the merger history of a galaxy, this would suggest that the transfer of information from the simulation to the observation is correctly adjusted.

Despite this, one should always keep in mind that we are just reproducing the posteriors as learned from the simulation. When inferring, for example, the ex-situ fraction with the model trained on TNG mocks, it is the distribution of ex-situ fractions as it would be expected for similar-looking galaxies in TNG. If the simulation changes, the inference results also change. In the long term, it will be of interest to marginalize over multiple simulations. When training the cINN with galaxies from multiple simulations, the posterior would automatically also encode the uncertainty due to differences in the simulation code. This way, we can provide observers with a reliable tool to infer galaxy properties from observations, for which this thesis lays the groundwork.

26 SUMMARY AND OUTLOOK

Galaxy mergers are an important, and for massive galaxies dominant, mechanism of growth. They can quench and restart star formation, transform the morphology, and influence the stellar kinematics of galaxies. Because of this, the merger history of a galaxy is of profound interest in astrophysics. But a challenge is, that the merging of galaxies usually happens on time scales of 100s of Myrs. The large-scale structure of galaxies therefore appears practically static and the merging process is defacto unobservable. What can observations, like visual images, tell us about this unobservable merger and assembly history of a galaxy? This is the central question of this thesis.

To answer this, we apply machine learning models to learn the connection between observables and the unobservable merger and assembly history from full cosmological galaxy simulations, namely TNG50 and TNG100 of the IllustrisTNG project. The idea is to then apply those models to actually observed galaxies to transfer the information from the simulations toward the observations. First (Part iii) we train and test a machine learning model (cINN) on scalar observable galaxy properties, derived from central galaxies ($z = 0 - 1$ and stellar masses of $10^{10-12} M_{\odot}$) within TNG100, to infer unobservable quantities describing the galaxies past merger and assembly history. By using just seven observable integral properties of galaxies (including the total stellar mass, redshift, color, stellar size, morphology, metallicity, and ages) the analysis demonstrates that these observables can effectively constrain parameters like the stellar ex-situ fraction and the time of the last major merger (Figure 15). Furthermore, the cINN returns the actual posterior distributions, i.e. the probability distribution of a galaxy for having a certain merger and assembly history, given the observables. This is especially important as we have to deal with a non-invertible problem: the observations contain limited information and can not arbitrarily accurately recover the quantities (Figure 14).

Using scalar properties, however, has many drawbacks: they have to be consistently defined and extracted from simulations and observations first. Furthermore, they might limit the information content from the original observations by selecting only a subset of relevant properties. On the other hand, they might be over-idealized if the properties are not sufficiently realistic to be measured from observations. This is more or less the case for all properties used in Part iii which are all exactly known properties from the simulation and the results can be therefore taken as a 'best case'. To overcome this, we target in the following work visual images of the Hyper-Suprime-Cam (HSC) in r,g,i - bands. From the simulated galaxies in TNG50 and TNG100, we generate and use survey realistic mock images incorporating the same limitations as the observations: limiting depth, image resolution, realistic background sources, and the same PSF as the HSC instrument.

As an intermediate step, we investigate in Part iv if the TNG mocks are sufficiently realistic toward the sample of real HSC observations. To do so, we employ self-supervised contrastive learning (NNCLR) on both observed and simulated

galaxy images to detect similarities and contrasts between the two sets. Our study targets galaxies with redshift $z = 0.1 - 0.4$ and an effective i-band luminosity cut of $20\mathrm{mag}$. We find that a majority of simulated galaxies closely resemble observed ones, but also identify a subset with unrealistic characteristics (Figures 37-52). Furthermore, the derived image representations, in which similar galaxies are close and dissimilar are distant, successfully encode galaxy observables (Figures 27-29) which promises the application of merger history inference from them. Based on those image representations we can define a quantity, the Out of Domain score (OOD), to measure the dissimilarity of an HSC image towards the domain of TNG images. With this, we can identify those galaxies to which we can transfer knowledge from the simulations (Figure 34).

In Part v we then use those representations obtained by self-supervised contrastive learning on realistic HSC mock images generated from TNG50 and TNG100 galaxies to train a conditional Invertible Neural Network (cINN) to estimate the posterior of merger/assembly statistics. Namely, we are deriving the lookback time and stellar mass of the last major merger and the fraction of ex-situ stars. Validation against simulated TNG galaxies demonstrates high accuracy in inferring properties like the stellar ex-situ fraction and the mass of the last major merger for certain mass ranges, showcasing the potential for extracting merger quantities from observed galaxy images despite challenges like observational limitations (Figure 15 and Figure 43). Finally, we also apply the model to previously unseen HSC galaxies to infer for the first time merger statistics for real galaxies (Figure 47).

We summarize the highlights of this thesis as follows:

- The inference of unobservable merger parameters from observations is possible, either from summary statistics or directly from images. The stellar ex-situ fractions can be, for example, constrained to $\leqslant \pm 10$ per cent in the scalar case and $\leqslant \pm 20$ per cent in the visual-HSC-image case (Figure 15 and Figure 43).
- The possibility of modeling arbitrary posteriors for the merger history is advantageous over predicting single moments of a fixed distribution as the posteriors are in general non-gaussian (Figure 14 and Figure 42).
- The survey-realistic mock images of simulated TNG galaxies are to a large amount ($\gtrsim 70$ per cent) realistic enough to use them in the simulation-based inference of unobservable quantities (Figure 34).
- As a valuable byproduct of this thesis, we construct a representation space that encodes morphological information contained in simulated and observed images (Figures 27-29).
- We are, for the first time, able to infer the merger history of HSC galaxy images based on the knowledge learned from TNG50/100 with machine learning (Figure 47).

These results are very exciting because they open a window into previously unobservable quantities that were only accessible to theorists. This not only includes the merger and assembly history of a galaxy but every quantity that can be defined and calculated from cosmological simulations. Simulation-based definitions can be

therefore directly applied to observations. Relations, found in cosmological simulations, can be transferred to observations and tested against traditional non-ML methods. Also, the representations are a useful tool as they encode all meaningful observational information contained in an image. They can be used to directly compare and match simulated and observed galaxies without the previously exhausting task of defining and calculating common summary statistics based on them. Lastly, large upcoming (galaxy) surveys like LSST and Euclid, which will produce many PB on data, can be analyzed efficiently. While e.g. for SDSS citizen science project was used to classify the found galaxies, galaxy classifications can now be theoretically performed based on the perfectly known galaxies in cosmological simulations.

Overall, this thesis highlights the potential of cosmological simulation-based inference for extracting the unobservable merger and assembly history quantities from observed galaxy images, enabling a deeper understanding of galaxy evolution and also enhancing the comparison between observations and cosmological simulations.

Part VII

APPENDIX

A

CALIBRATION ERRORS OF THE CINN MODEL

While our analysis in the main text of this paper has mainly focused on the performance of the cINN in terms of MAPs and posterior standard deviations, we show here more quantitatively to what degree the posterior distribution itself is related to the distribution of ground truth galaxies. To do so, we calculate the calibration error for each output: this is the deviation between the ground truth *distribution* of the TNG100 galaxy test sample and the respectively-predicted cINN posterior distribution. To obtain these calibration errors, we measure for the whole sample how often the ground truth falls into a certain confidence interval of the cINN posteriors, for each target quantity.

For example, we expect that we find within the 50 per cent interval of the sampled posterior $p(x|c)$ also 50 per cent of the ground truth test galaxies. A positive value therefore denotes an under-confident posterior (the interval contains more test galaxies than displayed by the posterior) and a negative value therefore denotes an overconfident posterior (the interval contains less test galaxies than displayed by the posterior).

We show the calibration errors as function of the posterior confidence level for each of the 5 outputs in Fig. 48. Note that we perform these calculations on the raw posteriors i.e. without using the previously-utilized "No-Major-Merger" bin. Especially for the ex-situ fraction, our networks are characterized by a low calibration error: $\lesssim 2$ per cent. The over-confidence for all 5 statistics is smaller than ≈ 2.5 per cent. We can conclude that the cINN posteriors therefore follow the ground truth distribution quite well, as we have also already verified in Figs. 17 and 18.

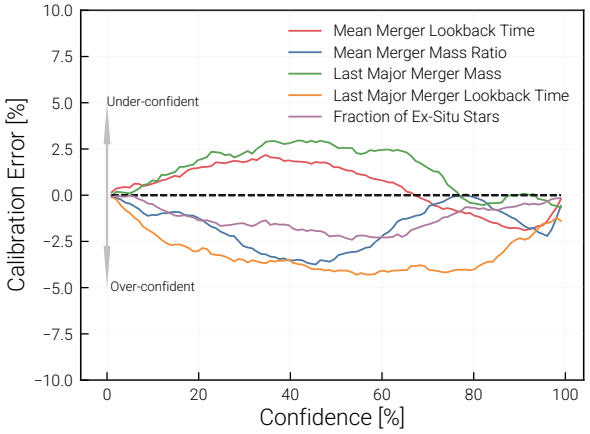


Figure 48: **Calibration error of the cINN model as function of the confidence level for each of the 5 target quantities.** The calibration error is the deviation between the fraction of test galaxy ground truths contained within a certain posterior confidence interval and the expected fraction according to the respective confidence level. A positive calibration error therefore denotes an under-confident posterior while a negative value denotes an over-confident posterior. We see that especially the ex-situ fraction posterior are well within $\pm 2\%$, whereas the mean merger mass ratio and the lookback time of the last major merger tends to be slightly over-confident. We exclude galaxies with no major merger in their history for the calculation of the last major merger calibration error.

B

THE BASELINE MODEL: A MULTILAYER PERCEPTRON (MLP)

Additional to the cINNs, which should be the main focus of this work, we also utilize a simpler Multilayer Perceptron (MLP) neural network to tackle the problem outlined in Fig. 11. We will use it as a baseline for the cINN and as method to perform the sensitivity analysis in Section 10.3. A MLP consists of multiple layers of artificial neurons (i.e. linear transformations) and a nonlinear rectifier in between. The 7 galaxy observables are given as input for the first linear layer, whereas the 5 merger-statistics unobservables are given by the output of the last linear layer. With this, we get a regression model that maps the observables to the unobservable parameters. The model is trained on the simulation-based dataset by minimizing the mean squared error loss

$$L_{MSE} = \frac{1}{N}\sum_{i}^{N}(\hat{x}_i - x_i)^2 \tag{2}$$

with x denoting the ground truth merger-history set of parameters from the TNG simulations, $\hat{x}$ being the MLP model predictions, and N the number of samples in the training batch. For the implementation and optimization of this network, we utilize the framework KERAS [37].

B.1 TRAINING OF THE MLP

The basic architecture of a MLP is a series of dense neuron layers (i.e. each neuron is connected to all neurons of the previous and next layer) and of activation and regularization layers. To do a recursion on 5 unobservable quantities, we need an output layer with 5 neurons (one for each unobservable quantity) and also an input layer with 7 values. There is no straightforward way to determine the configuration of the layers between the input and the output layers. To reduce the number of possible architectures to explore, we use some basic constrains: a) We keep the number of neurons constant for all dense layers. Thus we only have to find the number of neurons per layer (the width of the MLP) and the numbers of layers (the depth of the MLP). b) Between each dense layer, we use one activation layer and one regularization layer. And c) we use a rectified linear unit as activation function.

The optimizer ADAM [96] proves to be a stable and (sufficiently) fast converging method and performs well with the default values given by the KERAS framework implementation. We therefore use this optimizer in all of the following training procedures. If not stated otherwise, the learning rate is set to the default value $\eta = 10^{-3}$.

We use the mean squared error (MSE) as loss function, which leads to a faster and more stable convergence than the mean absolute error when starting from random weights.

Layer	Output Shape	Parameters
Input	7	0
Dense	256	2048
Batch Normalization	256	1024
ReLU Activation	256	0
Dense	256	65792
Batch Normalization	256	1024
ReLU Activation	256	0
Dense	256	65792
Batch Normalization	256	1024
ReLU Activation	256	0
Dense	256	65792
Batch Normalization	256	1024
ReLU Activation	256	0
Dense	5	1285

Table 6: **Layer configuration used for the Multilayer Perceptron neural network adopted in this work** (MLP, Sections B and B.1). We give the layer type (left column), the output shape of the respective layer (middle column), and the number of trainable parameters per layer (right column). Within our physical problem at hand, the 7 input parameters of the MLP are the 7 observable galaxy properties, whereas the target/output quantities are 5 summary statistics of the galaxies' assembly and merger history; both chosen and defined in Section 8.5. In total, this model has 202'757 trainable parameters.

As regularization method, we include a Batch Normalization [88] between each dense neuron layer and activation function. Because we deal with a very unevenly distributed sample (especially as a function of galaxy stellar mass, with much fewer galaxies at the high-mass end), we have to take special care of an appropriate batch size.

Regarding the node configuration, we choose to use 3 hidden layers (in addition to the in- and output layer) with 256 nodes each.

The resulting configuration of the MLP adopted in this work is summarized in Table 6. After the validation loss has converged with this training procedure, an additional training step is performed: we change the loss function to the mean absolute error (MAE) and halve the learning rate to $\eta = 0.5 \times 10^{-3}$. We do so because the MSE loss function leads to a better convergence when starting from a set of randomly initialized weights. However, we expect that our data set is contaminated by ambiguous outliers, i.e. of case galaxies whose merger statistics are not properly described by the set of 7 input quantities. We therefore focus more on the optimization regarding the more "well-behaved" galaxies in the sample.

As for the cINN (refer to Section 9.2.2), we also perform an ensemble averaging to deal with random model to model variations: We train 7 equivalent networks with random initial weights. The ensemble prediction is then given by the galaxy-

wise median of the 7 single network predictions. We choose 7 models because the additional training and validation effort from this is still computationally feasible.

B.2 COMPARISON BETWEEN MLP AND CINN: DO THEY YIELD THE SAME RESULTS?

How well does the much simpler MLP architecture perform in comparison to the cINN (Sections 8.6 and 9.2.1)? We quantify this in Fig. 49, by noting that exactly the same set of training/validation and test galaxies have been used for both MLP and cINN. We can hence directly compare the galaxy-wise point predictions by the MLP with the MAPs of the cINN (Section 10.1).

In Fig. 49, we show the difference in absolute prediction error between the two models, namely $|\text{MLP} - \text{Ground Truth}| - |\text{MAP} - \text{Ground Truth}|$, where here MLP represents the MLP point-prediction value and MAP is the cINN Maximum A-Posteriori estimation, which we have already quantified and analyzed e.g. in Figs. 15 and 16. Therefore, the cINN works better on galaxies where this quantity is positive, the MLP works better where it is negative. If the error of both methods is comparable, this value is close to 0.

In the top panels of Fig. 49, we compare the difference between the prediction errors against the true values of the target quantities. We see that this metric follows overall a rather symmetric distribution around 0 for all the assembly and merger statistics, implying that the MLP and the cINN (evaluated via the MAPs) perform similarly well.

We find larger deviations for the mean merger mass ratio, where the MLP performs slightly better than the cINN. The point-recursion with a mean squared error loss is able to track the dependence better than the MAP of the cINN posteriors, which instead tend to broaden towards higher mass ratios but keep their peak (and therefore the MAP) at a similar position. However, this over-performance of the MLP is on average not larger than ± 0.1: therefore, not even the MLP is capable of inferring this parameter.

We also see that the lookback time of the last major merger has a significant fraction of galaxies that do not work well with the MLP model. This might be caused by our choice to set this parameter to 15 Gyr for those galaxies with no major mergers. This choice does not work well with point predictions, as the MLP is not able to express ambiguities and therefore fall in between the peaks of the potentially multi-modal posterior distributions. The same problem might be the case for the mass of the last major merger, where the cINN MAPs perform substantially better than the MLP predictions.

We hence conclude that both the MLP and the cINN methods are rather equivalent regarding their point prediction capability, particularly so for inferring the ex-situ-fractions and the mean merger lookback time and the time of the last major merger. This is a confirmation of our implementation of the two methods, but it is not too surprising, as the minimization of the mean squared error (which is used to train the MLP) corresponds to the squared term in the log-likelihood loss (used to train the cINN).

Further deviations (in terms of scatter) between the two models might be explained by galaxies that have an intrinsically non-well defined point prediction i.e.

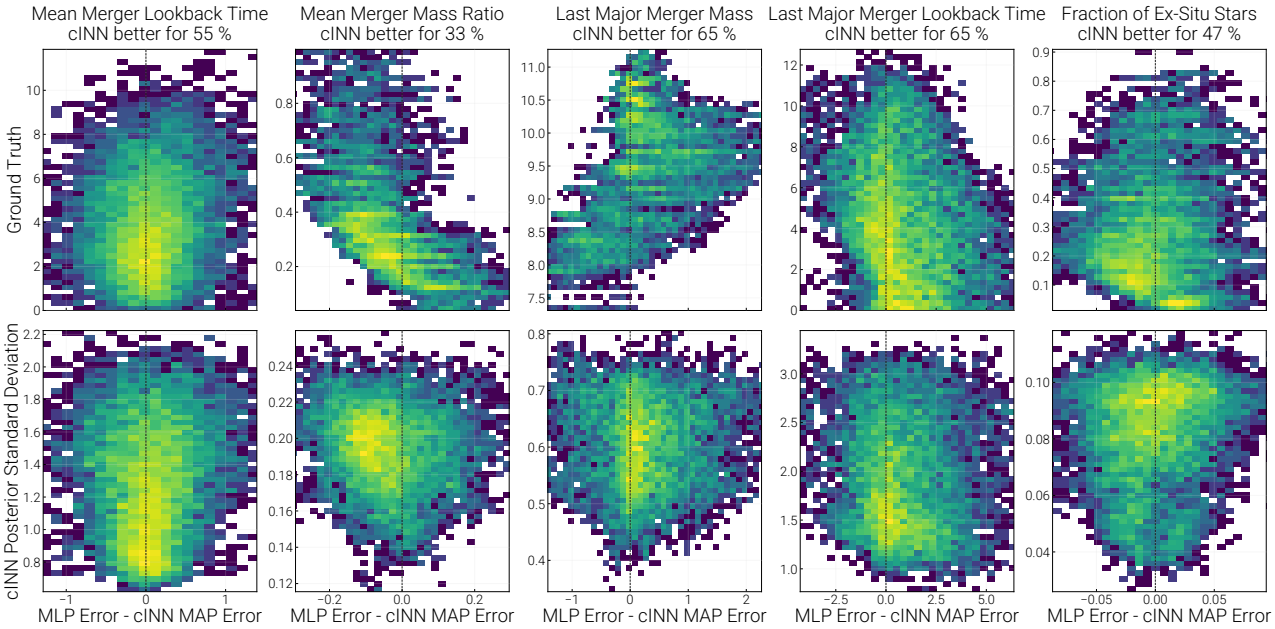


Figure 49: **Comparison of the MLP with the cINN MAP predictions**. On the x-axis, we plot the difference between the prediction errors from the two methods: $|\text{MLP} - \text{Ground Truth}| - |\text{MAP} - \text{Ground Truth}|$, where here MLP represents the MLP point-prediction value and MAP is the cINN Maximum A-Posteriori estimation. Therefore, a positive value of this quantity means that the cINN performs better in the predition than the MLP. **Top:** ground truth vs. difference between the errors of the MLP and cINN methods. We observe that the distributions are in general symmetrical around 0, but for the case of the mean merger mass ratio (second column from the left). The performance of the two models is therefore similar in terms of their point predictions, with a slightly better performance for the cINN. **Bottom:** standard deviation of the cINN posterior vs. difference between the errors of the MLP and cINN methods. The galaxies where the MLP and cINN MAPs disagree more substantially are also predicted to have a broader posterior distribution. This is especially pronounced for the ex-situ fraction. We speculate that this is an effect caused by the intrinsic uncertainty in the prediction of the selected unobservables. In these plots, for the time and mass of the last major mergers, we exclude galaxies with no major merger in their history.

a broad posterior distribution. To test this hypothesis, we also plot the difference between the prediction errors from the MLP and the cINN against the cINN posterior standard deviation for each merger statistic: lower panels of Fig. 49. We see that the difference in performance between the MLP and the cINN indeed increases with the width of the inferred posterior distributions, especially for the stellar ex-situ fraction and the mean merger lookback time, i.e. the two statistics we identified to have gaussian like posteriors in the previous Section. The intrinsic uncertainty of the prediction therefore plays an important role in the diverging results between the two methods.

C

HYPERPARAMETERSEARCH WITH OPTUNA

To find the best working set of hyperparameters for the NNCLR machine learning method, we perform a hyperparameter search using Optuna [4]. Optuna is a bayesian hyperparametersearch framework that allows to identify the best set of parameters by iteratively probing and narrowing the posterior of the parameterspace to find the set of parameters that minimizes the final validation NNCLR i.e. contrastive-loss for the model.

We optimize the model architecture by training each model with random parameters (Batch Size, Dropout-rate of the ResNet, ResNet Depth and Width, Linear Layer Depth and Width, dimension of the representation space) given by the Optuna optimizer. We also optimize the model for the learning rate, learning rate patience and decay, patience of the early stopping, the augmentation parameters and the L_2 Decay. The parameters are summarized in Table 7 together with the search range and the fiducial choices. Parameters that are not mentioned here are fixed. Because of the large number of images, we perform the validation after every 10000 seen images. Furthermore we stop the training if the validation loss has not improved for twice the Learning Rate patience, i.e. when the learning rate has decreased twice without an effect.

This said, as the training of a single model can take more than 24 hours on a single GPU, it is computationally very expensive to probe the whole space at once. We therefore iterate with our optimization run:

1. As the first step we use a fixed training time of 2h (By previous testing, we ensured that the model start to converge within this time-range). In doing so we get an overall rough idea about how the performance of the model is in general related to the choices without spending too much time for single runs.

2. At a second step we narrow the search space based on the previous step and train until convergence is reached i.e. the validation loss is not improving anymore for a certain interval (this interval is part of the optimized parameters) or is pruned because it is under-performing. Under performance is triggered by an above median validation loss after 2h.

3. Additionally we run a selection of runs with larger architectures as they might be deprived by the time-based training/pruning.

We perform each model training on a single GPU with 40GB Memory, which limits the possible model size, batch size and NNCLR queue size. We therefore can't ensure that larger models might not perform even better.

In the following we will briefly comment on our findings. First, the Batchsize has to be equal or larger than approx 32 to give good results; however with larger batchsizes we impose restrictions on the NNCLR Queue Size and model architecture because of limited GPU Memory. The length of the NNCLR Queue should

Parameter	Searchspace	Best Value/-Fiducial Choice
Batch Size	[4, 128]	32
Depth of the ResNet	[10,16]	16
Width of the ResNet	[1, 3]	2
Dropout of the ResNet	[0.0, 0.5]	0.3
Representation Dimension	[64, 512]	256
Representation MLP Depth	[1, 3]	2
Projection MLP Dimension	[64, 512]	256
Projection MLP Depth	[1, 3]	2
Initial Learning Rate	[0.0001, 0.005]	0.001
Learning Rate Decay	[0.2, 0.9]	0.8
Learning Rate Patience	[2, 10]	3
L2 Decay	[0.000001, 0.001]	0.000001
NCE Temperature	[0.01, 0.1]	0.04
NNCLR Queue Size (in number of Batches)	[64, 4096]	512
NNCLR Translation Transformation	[0.05, 0.5]	0.25
NNCLR Scaling Transformation	[1.0, 2.5]	1.1
NNCLR Gaussian Noise Transformation	[0.01, 0.1]	0.08
NNCLR Gaussian Blur Transformation	[0.001, 10]	0.1

Table 7: **Overview over the model and training parameters that have been optimized by using the hyperoptimization tool Optuna.**

be 256 (although the queue could be larger we see no significant improvement anymore from this point on. I.e. with $256 \times 32 = 8192$ images in the queue, our dataset is sufficiently sampled). We find that we should favour larger and wider ResNet model although again a bigger model is again limited by the GPU Memory. The dropout rate of the ResNet layers should be around 0.3 although the loss is not very sensitive to this parameter. For the linear layers we see for the depth not a very big sensitivity, we choose to use 2 layers for each. The dimension of the projection network clearly favours 256 dimensions. For the dimension of the representation space we see best validation losses for 128 or 256. We suspect that a smaller representation space prunes the information content too much while a too large representation space probably benefits over-fitting because of the additional degrees of freedom. However, because of the results of previous study [84], which favour larger representation spaces, we settle for a dimension of 256. Regarding the transformations we see that we better use small zoom ins of 1.1, i.e. 10 per cent. This might be due to the already quite narrow FOV. The translation parameter should be between $0.2 - 0.3$ (as fraction of the image edge length). The sigma for the Gaussian blur is rather small with 0.1 pixels while the added noise is considerably larger with an upper standard deviation of 0.8 (considering pixel values normalized to 1). With an initial learning rate of 0.001 and a decay by a factor of 0.8 every 3 epochs without improvement we train the model with a slow but steady decrease in learning rate.

D

DEPENDENCE OF THE OUT OF DOMAIN SCORE ON THE MODEL AND PARAMETERS

Each model training is a stochastic process and the resulting machine learning model is dependent on the architectural and parametric choices made. Here we investigate how much the results of this paper, especially the Out of Domain Score, depend on these choices and stochastic influences.

To do so, we use the 15 best runs (in terms of validation loss) made for the hyper parameter optimization in Appendix D and calculate representations with their resulting models like the fiducial run presented in this work. We show the variation in OOD Score among the runs in Figure 50. Although there is scatter among the 15 runs, the overall trends are clearly visible; e.g. there are more high OOD Score galaxies in TNG100 relative to HSC than in TNG50 relative to HSC, while the agreement between the two TNG runs is in general identified to be higher. Looking at the OOD Score variations for each galaxy individually, we see that the OOD calculation for most galaxies does not have a variance above ± 0.2 across the 15 runs. I.e. galaxies with a large OOD Score are almost always consistently identified as OOD by the 15 models. //

For the calculation of the OOD Score we also perform a nearest neighbour search. Our fiducial choice is to use the 8th nearest neighbour; as validation, we also re-calculate the distances using the 4th, 16th and 32th nearest neighbour in Figure 51. While such a parameter has an influence, i.e. especially causes a smoothing of the curves, we do not see a significant change among the choices.

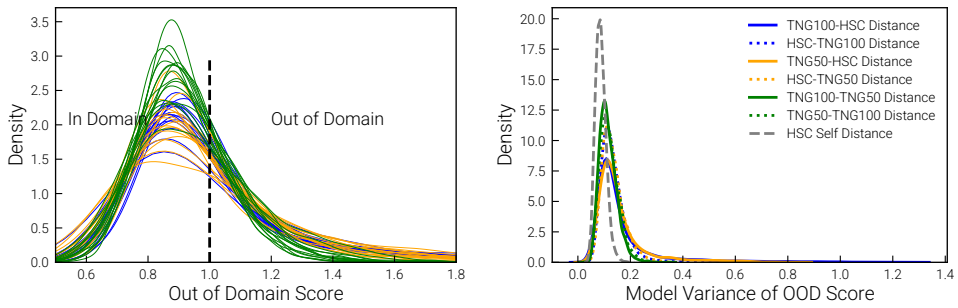


Figure 50: We show the influence of the model to model variation on the OOD Score defined and calculated in Section 16.2.2. To do so we recalculate the OOD Scores for the 15 best Optuna runs and compare them at the galaxy-population level (left plot) and for individual galaxies (right plot). On the left we reproduce Figure 34 but showing the curves for all 15 runs. We see that the run to run scatter is too small to actually influence the plot qualitatively. On the right we show the model to model variance in OOD Score for each individual galaxy. The most galaxies have a variance below 0.2, which is small relative to the overall range in OOD values.

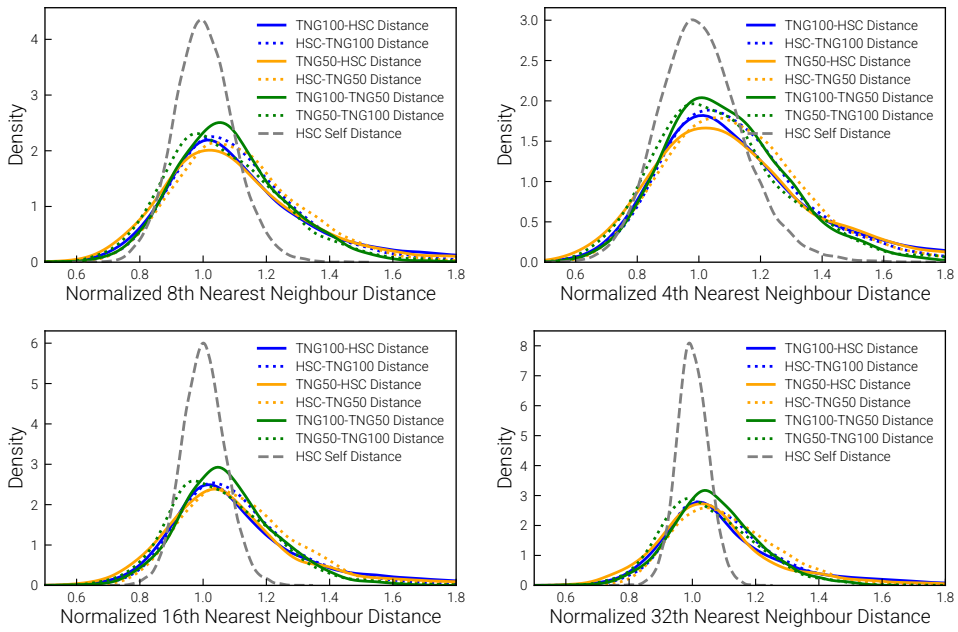


Figure 51: We show the influence of the choice in nearest neighbour on the calculation of the Normalized Nearest Neighbour Distance introduced in Section 16.2.2. We show the same plot for (from top-left to bottom-right) 4 choices in nearest neighbour: 8 (our fiducial choice), 4, 16 and 32. While the variation has an influence on the curve, the differences are too marginal to influence the claims made in this work.

E

EXAMPLE FOR FALSE FRIENDS BECAUSE OF A FAILING FOV DETERMINATION

We find galaxies with very small values of the OOD Score, which we have introduced in the main text of this paper as an indicator for high similarity between any given galaxy and a set of galaxies for comparison. However, after closer inspection, we found that some of those imaged galaxies are not similar at all: rather, the FOV of these galaxy images are much larger than the average of the images. An example for such a case can be seen in Figure 52. The model captured this large FOV as a significant similarity/point of tension with the other galaxies. This is a good example of how the definition of similarity by a machine learning model can differ from a human definition. Furthermore it underlines the importance to use validation checks and to not completely rely on the contrastive learning in finding a meaningful representation.

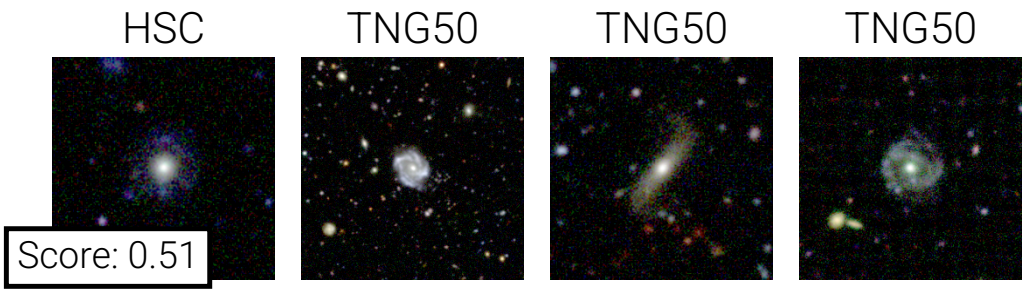


Figure 52: Examples for "False Friends" because of a failing FOV determination. The depicted HSC galaxy on the left is given a very low score as it got closely grouped together with other galaxies (from HSC and TNG) whose FOV is significantly larger than the average. It is possible to identify such failing cases either by visual inspection or by comparing the known galaxy properties for the TNG mock images.

BIBLIOGRAPHY

[1] Kevork N. Abazajian et al. "The Seventh Data Release of the Sloan Digital Sky Survey." In: *ApJS* 182.2 (June 2009), pp. 543–558. DOI: 10.1088/0067-0049/182/2/543. arXiv: 0812.0649 [astro-ph].

[2] Romina Ahumada et al. "The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra." In: *ApJS* 249.1, 3 (July 2020), p. 3. DOI: 10.3847/1538-4365/ab929e. arXiv: 1912.02905 [astro-ph.GA].

[3] Hiroaki Aihara et al. "Third data release of the Hyper Suprime-Cam Subaru Strategic Program." In: *PASJ* 74.2 (Apr. 2022), pp. 247–272. DOI: 10.1093/pasj/psab122. arXiv: 2108.13045 [astro-ph.IM].

[4] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A Next-generation Hyperparameter Optimization Framework." In: *arXiv e-prints*, arXiv:1907.10902 (July 2019), arXiv:1907.10902. DOI: 10.48550/arXiv.1907.10902. arXiv: 1907.10902 [cs.LG].

[5] Andrés Almeida et al. "The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V." In: *ApJS* 267.2, 44 (Aug. 2023), p. 44. DOI: 10.3847/1538-4365/acda98. arXiv: 2301.07688 [astro-ph.GA].

[6] Pedro Alonso, Wenting Wang, Jun Zhang, Hekun Li, Shi Shao, Qi Guo, Yanqin He, Cai-Na Hao, and Rui Shi. "Dark against Luminous Matter around Isolated Central Galaxies: A Comparative Study between Modern Surveys and IllustrisTNG." In: *ApJ* 947.1, 19 (Apr. 2023), p. 19. DOI: 10.3847/1538-4357/acbf4a. arXiv: 2212.04702 [astro-ph.GA].

[7] Eirini Angeloudi, Jesús Falcón-Barroso, Marc Huertas-Company, Regina Sarmiento, Annalisa Pillepich, Daniel Walo-Martín, and Lukas Eisert. "ERGO-ML: towards a robust machine learning model for inferring the fraction of accreted stars in galaxies from integral-field spectroscopic maps." In: *MNRAS* 523.4 (Aug. 2023), pp. 5408–5429. DOI: 10.1093/mnras/stad1669. arXiv: 2306.01056 [astro-ph.GA].

[8] Felipe Ardila et al. "Stellar and weak lensing profiles of massive galaxies in the Hyper-Suprime Cam survey and in hydrodynamic simulations." In: *MNRAS* 500.1 (Jan. 2021), pp. 432–447. DOI: 10.1093/mnras/staa3215. arXiv: 2010.00123 [astro-ph.GA].

[9] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W. Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. "Analyzing Inverse Problems with Invertible Neural Networks." In: *arXiv e-prints*, arXiv:1808.04730 (Aug. 2018), arXiv:1808.04730. arXiv: 1808.04730 [cs.LG].

[10] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. "Guided Image Generation with Conditional Invertible Neural Networks." In: *arXiv e-prints*, arXiv:1907.02392 (July 2019), arXiv:1907.02392. arXiv: 1907.02392 [cs.CV].

[11] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. "Guided Image Generation with Conditional Invertible Neural Networks." In: *arXiv e-prints*, arXiv:1907.02392 (July 2019), arXiv:1907.02392. arXiv: 1907.02392 [cs.CV].

[12] Maarten Baes, Joris Verstappen, Ilse De Looze, Jacopo Fritz, Waad Saftly, Edgardo Vidal Pérez, Marko Stalevski, and Sander Valcke. "Efficient Three-dimensional NLTE Dust Radiative Transfer with SKIRT." In: *ApJS* 196.2, 22 (Oct. 2011), p. 22. DOI: 10.1088/0067-0049/196/2/22. arXiv: 1108.5056 [astro-ph.CO].

[13] I. K. Baldry, M. L. Balogh, R. Bower, K. Glazebrook, and R. C. Nichol. "Color bimodality: Implications for galaxy evolution." In: *The New Cosmology: Conference on Strings and Cosmology*. Ed. by Roland E. Allen, Dimitri V. Nanopoulos, and Christopher N. Pope. Vol. 743. American Institute of Physics Conference Series. AIP, Dec. 2004, pp. 106–119. DOI: 10.1063/1.1848322. arXiv: astro-ph/0410603 [astro-ph].

[14] I. K. Baldry et al. "Galaxy And Mass Assembly (GAMA): the input catalogue and star-galaxy separation." In: *MNRAS* 404.1 (May 2010), pp. 86–100. DOI: 10.1111/j.1365-2966.2010.16282.x. arXiv: 0910.5120 [astro-ph.CO].

[15] Rachael L. Beaton et al. "The Carnegie-Chicago Hubble Program. I. An Independent Approach to the Extragalactic Distance Scale Using Only Population II Distance Indicators." In: *ApJ* 832.2, 210 (Dec. 2016), p. 210. DOI: 10.3847/0004-637X/832/2/210. arXiv: 1604.01788 [astro-ph.CO].

[16] K. G. Begeman. "HI rotation curves of spiral galaxies. I. NGC 3198." In: *A&A* 223 (Oct. 1989), pp. 47–60.

[17] V. Belokurov, D. Erkal, N. W. Evans, S. E. Koposov, and A. J. Deason. "Co-formation of the disc and the stellar halo." In: *MNRAS* 478.1 (July 2018), pp. 611–619. DOI: 10.1093/mnras/sty982. arXiv: 1802.03414 [astro-ph.GA].

[18] Lucas A. Bignone, Susana E. Pedrosa, James W. Trayford, Patricia B. Tissera, and Leonardo J. Pellizza. "Non-parametric morphologies of galaxies in the EAGLE simulation." In: *MNRAS* 491.3 (Jan. 2020), pp. 3624–3642. DOI: 10.1093/mnras/stz3014. arXiv: 1908.10936 [astro-ph.GA].

[19] Michael R. Blanton and John Moustakas. "Physical Properties and Environments of Nearby Galaxies." In: *ARA&A* 47.1 (Sept. 2009), pp. 159–210. DOI: 10.1146/annurev-astro-082708-101734. arXiv: 0908.3017 [astro-ph.GA].

[20] Ana Bonaca, Charlie Conroy, Phillip A. Cargile, Rohan P. Naidu, Benjamin D. Johnson, Dennis Zaritsky, Yuan-Sen Ting, Nelson Caldwell, Jiwon Jesse Han, and Pieter van Dokkum. "Timing the Early Assembly of the Milky Way with the H3 Survey." In: *ApJ* 897.1, L18 (July 2020), p. L18. DOI: 10.3847/2041-8213/ab9caa. arXiv: 2004.11384 [astro-ph.GA].

[21] James Bosch et al. "The Hyper Suprime-Cam software pipeline." In: *PASJ* 70, S5 (Jan. 2018), S5. DOI: 10.1093/pasj/psx080. arXiv: 1705.06766 [astro-ph.IM].

[22] Connor Bottrell, Maan H Hani, Hossen Teimoorinia, Sara L Ellison, Jorge Moreno, Paul Torrey, Christopher C Hayward, Mallory Thorp, Luc Simard, and Lars Hernquist. "Deep learning predictions of galaxy merger stage and the importance of observational realism." In: *Monthly Notices of the Royal Astronomical Society* 490.4 (Oct. 2019), pp. 5390–5413. DOI: 10.1093/mnras/stz2934. URL: https://doi.org/10.1093%2Fmnras%2Fstz2934.

[23] Connor Bottrell, Maan H. Hani, Hossen Teimoorinia, David R. Patton, and Sara L. Ellison. "The combined and respective roles of imaging and stellar kinematics in identifying galaxy merger remnants." In: *MNRAS* (Dec. 2021). DOI: 10.1093/mnras/stab3717. arXiv: 2201.03579 [astro-ph.GA].

[24] Connor Bottrell, Paul Torrey, Luc Simard, and Sara L. Ellison. "Galaxies in the Illustris simulation as seen by the Sloan Digital Sky Survey - II. Size-luminosity relations and the deficit of bulge-dominated galaxies in Illustris at low mass." In: *MNRAS* 467.3 (May 2017), pp. 2879–2895. DOI: 10.1093/mnras/stx276. arXiv: 1701.08206 [astro-ph.GA].

[25] Connor Bottrell, Paul Torrey, Luc Simard, and Sara L. Ellison. "Galaxies in the Illustris simulation as seen by the Sloan Digital Sky Survey - II. Size-luminosity relations and the deficit of bulge-dominated galaxies in Illustris at low mass." In: *MNRAS* 467.3 (May 2017), pp. 2879–2895. DOI: 10.1093/mnras/stx276. arXiv: 1701.08206 [astro-ph.GA].

[26] Connor Bottrell et al. "IllustrisTNG in the HSC-SSP: image data release and the major role of mini mergers as drivers of asymmetry and star formation." In: *arXiv e-prints*, arXiv:2308.14793 (Aug. 2023), arXiv:2308.14793. DOI: 10.48550/arXiv.2308.14793. arXiv: 2308.14793 [astro-ph.GA].

[27] Michael Boylan-Kolchin, Volker Springel, Simon D. M. White, Adrian Jenkins, and Gerard Lemson. "Resolving cosmic structure formation with the Millennium-II Simulation." In: *MNRAS* 398.3 (Sept. 2009), pp. 1150–1164. DOI: 10.1111/j.1365-2966.2009.15191.x. arXiv: 0903.3041 [astro-ph.CO].

[28] G. Bruzual and S. Charlot. "Stellar population synthesis at the resolution of 2003." In: *MNRAS* 344.4 (Oct. 2003), pp. 1000–1028. DOI: 10.1046/j.1365-8711.2003.06897.x. arXiv: astro-ph/0309134 [astro-ph].

[29] Kevin Bundy et al. "OVERVIEW OF THE SDSS-IV MaNGA SURVEY: MAPPING NEARBY GALAXIES AT APACHE POINT OBSERVATORY." In: *The Astrophysical Journal* 798.1 (Dec. 2014), p. 7. ISSN: 1538-4357. DOI: 10.1088/0004-637x/798/1/7. URL: http://dx.doi.org/10.1088/0004-637X/798/1/7.

[30] Kevin Bundy et al. "Overview of the SDSS-IV MaNGA Survey: Mapping nearby Galaxies at Apache Point Observatory." In: *ApJ* 798.1, 7 (Jan. 2015), p. 7. DOI: 10.1088/0004-637X/798/1/7. arXiv: 1412.1482 [astro-ph.GA].

[31] Ronald J. Buta. "Galaxy Morphology." In: *Planets, Stars and Stellar Systems: Volume 6: Extragalactic Astronomy and Cosmology*. Dordrecht: Springer Netherlands, 2013, pp. 1–89. ISBN: 978-94-007-5609-0. DOI: 10.1007/978-94-007-5609-0_1. URL: https://doi.org/10.1007/978-94-007-5609-0_1.

[32] P. Camps and M. Baes. "SKIRT: An advanced dust radiative transfer code with a user-friendly architecture." In: *Astronomy and Computing* 9 (Mar. 2015), pp. 20–33. DOI: 10.1016/j.ascom.2014.10.004. arXiv: 1410.1629 [astro-ph.IM].

[33] P. Camps and M. Baes. "SKIRT 9: Redesigning an advanced dust radiative transfer code to allow kinematics, line transfer and polarization by aligned dust grains." In: *Astronomy and Computing* 31, 100381 (Apr. 2020), p. 100381. DOI: 10.1016/j.ascom.2020.100381. arXiv: 2003.00721 [astro-ph.GA].

[34] Gilles Chabrier. "Galactic Stellar and Substellar Initial Mass Function." In: *PASP* 115.809 (July 2003), pp. 763–795. DOI: 10.1086/376392. arXiv: astro-ph/0304382 [astro-ph].

[35] K. C. Chambers et al. "The Pan-STARRS1 Surveys." In: *arXiv e-prints*, arXiv:1612.05560 (Dec. 2016), arXiv:1612.05560. DOI: 10.48550/arXiv.1612.05560. arXiv: 1612.05560 [astro-ph.IM].

[36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. DOI: 10.48550/ARXIV.2002.05709. URL: https://arxiv.org/abs/2002.05709.

[37] François Chollet et al. *Keras*. https://keras.io. 2015.

[38] A. Ćiprijanović, D. Kafkes, K. Downey, S. Jenkins, G. N. Perdue, S. Madireddy, T. Johnston, G. F. Snyder, and B. Nord. "DeepMerge - II. Building robust deep learning algorithms for merging galaxy identification across domains." In: *MNRAS* 506.1 (Sept. 2021), pp. 677–691. DOI: 10.1093/mnras/stab1677. arXiv: 2103.01373 [astro-ph.IM].

[39] A. Ćiprijanović, D. Kafkes, K. Downey, S. Jenkins, G. N. Perdue, S. Madireddy, T. Johnston, G. F. Snyder, and B. Nord. "DeepMerge - II. Building robust deep learning algorithms for merging galaxy identification across domains." In: *MNRAS* 506.1 (Sept. 2021), pp. 677–691. DOI: 10.1093/mnras/stab1677. arXiv: 2103.01373 [astro-ph.IM].

[40] Christopher J. Conselice. "The Evolution of Galaxy Structure Over Cosmic Time." In: *ARA&A* 52 (Aug. 2014), pp. 291–337. DOI: 10.1146/annurev-astro-081913-040037. arXiv: 1403.2783 [astro-ph.GA].

[41] B. A. Cook, C. Conroy, A. Pillepich, V. Rodriguez-Gomez, and L. Hernquist. "THE INFORMATION CONTENT OF STELLAR HALOS: STELLAR POPULATION GRADIENTS AND ACCRETION HISTORIES IN EARLY-TYPE ILLUSTRIS GALAXIES." In: *The Astrophysical Journal* 833.2 (Dec. 2016), p. 158. ISSN: 1538-4357. DOI: 10.3847/1538-4357/833/2/158. URL: http://dx.doi.org/10.3847/1538-4357/833/2/158.

[42] Luca Costantin et al. "Expectations of the Size Evolution of Massive Galaxies at $3 \leqslant z \leqslant 6$ from the TNG50 Simulation: The CEERS/JWST View." In: *ApJ* 946.2, 71 (Apr. 2023), p. 71. DOI: 10.3847/1538-4357/acb926. arXiv: 2208.00007 [astro-ph.GA].

[43] Jean Coupon, Nicole Czakon, James Bosch, Yutaka Komiyama, Elinor Medezinski, Satoshi Miyazaki, and Masamune Oguri. "The bright-star masks for the HSC-SSP survey." In: *PASJ* 70, S7 (Jan. 2018), S7. DOI: 10.1093/pasj/psx047. arXiv: 1705.00622 [astro-ph.IM].

[44] Robert A. Crain et al. "The EAGLE simulations of galaxy formation: calibration of subgrid physics and model variations." In: *MNRAS* 450.2 (June 2015), pp. 1937–1961. DOI: 10.1093/mnras/stv725. arXiv: 1501.01311 [astro-ph.GA].

[45] Darren J. Croton, Volker Springel, Simon D. M. White, G. De Lucia, C. S. Frenk, L. Gao, A. Jenkins, G. Kauffmann, J. F. Navarro, and N. Yoshida. "The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies." In: *MNRAS* 365.1 (Jan. 2006), pp. 11–28. DOI: 10.1111/j.1365-2966.2005.09675.x. arXiv: astro-ph/0508046 [astro-ph].

[46] Shany Danieli et al. "The Dragonfly Wide Field Survey. I. Telescope, Survey Design, and Data Characterization." In: *ApJ* 894.2, 119 (May 2020), p. 119. DOI: 10.3847/1538-4357/ab88a8. arXiv: 1910.14045 [astro-ph.GA].

[47] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H. Rafieferantsoa, and Sarah Appleby. "SIMBA: Cosmological simulations with black hole growth and feedback." In: *MNRAS* 486.2 (June 2019), pp. 2827–2849. DOI: 10.1093/mnras/stz937. arXiv: 1901.10203 [astro-ph.GA].

[48] Thomas A. Davison, Mark A. Norris, Joel L. Pfeffer, Jonathan J. Davies, and Robert A. Crain. "An EAGLE's view of ex situ galaxy growth." In: *MNRAS* 497.1 (Sept. 2020), pp. 81–93. DOI: 10.1093/mnras/staa1816. arXiv: 2006.08590 [astro-ph.GA].

[49] Thomas A Davison, Mark A Norris, Joel L Pfeffer, Jonathan J Davies, and Robert A Crain. "An EAGLE's view of ex situ galaxy growth." In: *Monthly Notices of the Royal Astronomical Society* 497.1 (June 2020), pp. 81–93. ISSN: 0035-8711. DOI: 10.1093/mnras/staa1816. eprint: https://academic.oup.com/mnras/article-pdf/497/1/81/33519876/staa1816.pdf. URL: https://doi.org/10.1093/mnras/staa1816.

[50] Martín Emilio De Los Rios, Mihael Petač, Bryan Zaldivar, Nina R. Bonaventura, Francesca Calore, and Fabio Iocco. "Determining the Dark Matter distribution in galaxies with Deep Learning." In: *arXiv e-prints*, arXiv:2111.08725 (Nov. 2021), arXiv:2111.08725. arXiv: 2111.08725 [astro-ph.GA].

[51] G. M. De Silva et al. "The GALAH survey: scientific motivation." In: *MNRAS* 449.3 (May 2015), pp. 2604–2617. DOI: 10.1093/mnras/stv327. arXiv: 1502.04767 [astro-ph.GA].

[52] Li-Cai Deng et al. "LAMOST Experiment for Galactic Understanding and Exploration (LEGUE) — The survey's science plan." In: *Research in Astronomy and Astrophysics* 12.7 (July 2012), pp. 735–754. DOI: 10.1088/1674-4527/12/7/003. arXiv: 1206.3578 [astro-ph.GA].

[53] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP." In: *arXiv e-prints*, arXiv:1605.08803 (May 2016), arXiv:1605.08803. arXiv: 1605.08803 [cs.LG].

[54] S. Djorgovski and Marc Davis. "Fundamental Properties of Elliptical Galaxies." In: *ApJ* 313 (Feb. 1987), p. 59. DOI: 10.1086/164948.

[55] K. Dolag, S. Borgani, G. Murante, and V. Springel. "Substructures in hydrodynamical cluster simulations." In: *MNRAS* 399.2 (2009), pp. 497–514. DOI: 10.1111/j.1365-2966.2009.15034.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2966.2009.15034.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2966.2009.15034.x.

[56] Martina Donnari, Annalisa Pillepich, Dylan Nelson, Federico Marinacci, Mark Vogelsberger, and Lars Hernquist. "Quenched fractions in the IllustrisTNG simulations: comparison with observations and other theoretical models." In: *MNRAS* 506.4 (Oct. 2021), pp. 4760–4780. DOI: 10.1093/mnras/stab1950. arXiv: 2008.00004 [astro-ph.GA].

[57] Martina Donnari, Annalisa Pillepich, Dylan Nelson, Mark Vogelsberger, Shy Genel, Rainer Weinberger, Federico Marinacci, Volker Springel, and Lars Hernquist. "The star formation activity of IllustrisTNG galaxies: main sequence, UVJ diagram, quenched fractions, and systematics." In: *MNRAS* 485.4 (June 2019), pp. 4817–4840. DOI: 10.1093/mnras/stz712. arXiv: 1812.07584 [astro-ph.GA].

[58] Simon P. Driver et al. "GAMA: towards a physical understanding of galaxy formation." In: *Astronomy and Geophysics* 50.5 (Oct. 2009), pp. 5.12–5.19. DOI: 10.1111/j.1468-4004.2009.50512.x. arXiv: 0910.5123 [astro-ph.CO].

[59] Simon P. Driver et al. "Galaxy And Mass Assembly (GAMA): Data Release 4 and the z < 0.1 total and z < 0.08 morphological galaxy stellar mass functions." In: *MNRAS* 513.1 (June 2022), pp. 439–467. DOI: 10.1093/mnras/stac472. arXiv: 2203.08539 [astro-ph.GA].

[60] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. *With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations*. 2021. DOI: 10.48550/ARXIV.2104.14548. URL: https://arxiv.org/abs/2104.14548.

[61] Lukas Eisert, Connor Bottrell, Annalisa Pillepich, Rhythm Shimakawa, Vicente Rodriguez-Gomez, Dylan Nelson, Eirini Angeloudi, and Marc Huertas-Company. "ERGO-ML: comparing IllustrisTNG and HSC galaxy images via contrastive learning." In: *MNRAS* 528.4 (Mar. 2024), pp. 7411–7439. DOI: 10.1093/mnras/stae481. arXiv: 2310.19904 [astro-ph.GA].

[62] Lukas Eisert, Zoller Julian, Beckmann Jan, Varding Miriam, and Häberle Maximilian. *Distant Luminosity - Astrophotographer Collective — distant-luminosity.com*. https://www.distant-luminosity.com/gallery.html. [Accessed 04-05-2024].

[63] Lukas Eisert, Annalisa Pillepich, Dylan Nelson, Ralf S. Klessen, Marc Huertas-Company, and Vicente Rodriguez-Gomez. "ERGO-ML I: inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks." In: *MNRAS* 519.2 (Feb. 2023), pp. 2199–2223. DOI: 10.1093/mnras/stac3295. arXiv: 2202.06967 [astro-ph.GA].

[64] Sara L. Ellison, J. Trevor Mendel, David R. Patton, and Jillian M. Scudder. "Galaxy pairs in the Sloan Digital Sky Survey - VIII. The observational properties of post-merger galaxies." In: *MNRAS* 435.4 (Nov. 2013), pp. 3627–3638. DOI: 10.1093/mnras/stt1562. arXiv: 1308.3707 [astro-ph.CO].

[65] S. M. Faber and R. E. Jackson. "Velocity dispersions and mass-to-light ratios for elliptical galaxies." In: *ApJ* 204 (Mar. 1976), pp. 668–683. DOI: 10.1086/154215.

[66] Onsi Fakhouri and Chung-Pei Ma. "The nearly universal merger rate of dark matter haloes in ΛCDM cosmology." In: *MNRAS* 386.2 (May 2008), pp. 577–592. DOI: 10.1111/j.1365-2966.2008.13075.x. arXiv: 0710.4567 [astro-ph].

[67] Leonardo Ferreira, Christopher J. Conselice, Kenneth Duncan, Ting-Yun Cheng, Alex Griffiths, and Amy Whitney. "Galaxy Merger Rates up to z ≈ 3 Using a Bayesian Deep Learning Model: A Major-merger Classifier Using IllustrisTNG Simulation Data." In: *The Astrophysical Journal* 895.2 (June 2020), p. 115. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ab8f9b. URL: http://dx.doi.org/10.3847/1538-4357/ab8f9b.

[68] Hisanori Furusawa et al. "The on-site quality-assurance system for Hyper Suprime-Cam: OSQAH." In: *PASJ* 70, S3 (Jan. 2018), S3. DOI: 10.1093/pasj/psx079.

[69] Gaia Collaboration et al. "Gaia Data Release 2. Summary of the contents and survey properties." In: *A&A* 616, A1 (Aug. 2018), A1. DOI: 10.1051/0004-6361/201833051. arXiv: 1804.09365 [astro-ph.GA].

[70] Robel Geda, Steven M. Crawford, Lucas Hunt, Matthew Bershady, Erik Tollerud, and Solohery Randriamampandry. "PetroFit: A Python Package for Computing Petrosian Radii and Fitting Galaxy Light Profiles." In: *The Astronomical Journal* 163.5 (Apr. 2022), p. 202. DOI: 10.3847/1538-3881/ac5908. URL: https://doi.org/10.3847/1538-3881/ac5908.

[71] Shy Genel, S. Michael Fall, Lars Hernquist, Mark Vogelsberger, Gregory F. Snyder, Vicente Rodriguez-Gomez, Debora Sijacki, and Volker Springel. "Galactic Angular Momentum in the Illustris Simulation: Feedback and the Hubble Sequence." In: *The Astrophysical Journal Letters* 804.2, L40 (May 2015), p. L40. DOI: 10.1088/2041-8205/804/2/L40. arXiv: 1503.01117 [astro-ph.GA].

[72] Shy Genel, Reinhard Genzel, Nicolas Bouché, Thorsten Naab, and Amiel Sternberg. "The Halo Merger Rate in the Millennium Simulation and Implications for Observed Galaxy Merger Fractions." In: *ApJ* 701.2 (Aug. 2009), pp. 2002–2018. DOI: 10.1088/0004-637X/701/2/2002. arXiv: 0812.3154 [astro-ph].

[73] Shy Genel, Mark Vogelsberger, Volker Springel, Debora Sijacki, Dylan Nelson, Greg Snyder, Vicente Rodriguez-Gomez, Paul Torrey, and Lars Hernquist. "Introducing the Illustris project: the evolution of galaxy populations across cosmic time." In: *MNRAS* 445.1 (Nov. 2014), pp. 175–200. DOI: 10.1093/mnras/stu1654. arXiv: 1405.3749 [astro-ph.CO].

[74] Shy Genel et al. "The size evolution of star-forming and quenched galaxies in the IllustrisTNG simulation." In: *MNRAS* 474.3 (Mar. 2018), pp. 3976–3996. DOI: 10.1093/mnras/stx3078. arXiv: 1707.05327 [astro-ph.GA].

[75] Aritra Ghosh et al. "Morphological Parameters and Associated Uncertainties for 8 Million Galaxies in the Hyper Suprime-Cam Wide Survey." In: *ApJ* 953.2, 134 (Aug. 2023), p. 134. DOI: 10.3847/1538-4357/acd546. arXiv: 2212.00051 [astro-ph.GA].

[76] Alister W. Graham and Simon P. Driver. "A Concise Reference to (Projected) Sérsic $R^{1/n}$ Quantities, Including Concentration, Profile Slopes, Petrosian Indices, and Kron Magnitudes." In: *Publ. Astron. Soc. Australia* 22.2 (Jan. 2005), pp. 118–127. DOI: 10.1071/AS05001. arXiv: astro-ph/0503176 [astro-ph].

[77] Jean-Bastien Grill et al. "Bootstrap your own latent: A new approach to self-supervised Learning." In: *arXiv e-prints*, arXiv:2006.07733 (June 2020), arXiv:2006.07733. DOI: 10.48550/arXiv.2006.07733. arXiv: 2006.07733 [cs.LG].

[78] Brent Groves, Michael A. Dopita, Ralph S. Sutherland, Lisa J. Kewley, Jörg Fischera, Claus Leitherer, Bernhard Brandl, and Wil van Breugel. "Modeling the Pan-Spectral Energy Distribution of Starburst Galaxies. IV. The Controlling Parameters of the Starburst SED." In: *ApJS* 176.2 (June 2008), pp. 438–456. DOI: 10.1086/528711. arXiv: 0712.1824 [astro-ph].

[79] Alejandro Guzmán-Ortega, Vicente Rodriguez-Gomez, Gregory F. Snyder, Katie Chamberlain, and Lars Hernquist. "Morphological signatures of mergers in the TNG50 simulation and the Kilo-Degree Survey: the merger fraction from dwarfs to Milky Way-like galaxies." In: *MNRAS* 519.4 (Mar. 2023), pp. 4920–4937. DOI: 10.1093/mnras/stac3334. arXiv: 2211.05785 [astro-ph.GA].

[80] HST. *A Hubble Atlas of Interacting Galaxies*. 2008. URL: https://sci.esa.int/web/hubble/-/42690-a-hubble-atlas-of-interacting-galaxies. (accessed: 31.01.2022).

[81] F. Hammer, Y. B. Yang, J. L. Wang, M. Puech, H. Flores, and S. Fouquet. "Does M31 Result from an Ancient Major Merger?" In: *ApJ* 725.1 (Dec. 2010), pp. 542–555. DOI: 10.1088/0004-637X/725/1/542. arXiv: 1010.0679 [astro-ph.GA].

[82] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In: *arXiv e-prints*, arXiv:1512.03385 (Dec. 2015), arXiv:1512.03385. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385 [cs.CV].

[83] Amina Helmi, Carine Babusiaux, Helmer H. Koppelman, Davide Massari, Jovan Veljanoski, and Anthony G. A. Brown. "The merger that led to the formation of the Milky Way's inner stellar halo and thick disk." In: *Nature* 563.7729 (Oct. 2018), pp. 85–88. DOI: 10.1038/s41586-018-0625-x. arXiv: 1806.06038 [astro-ph.GA].

[84] Marc Huertas-Company, Regina Sarmiento, and Johan Knapen. "A brief review of contrastive learning applied to astrophysics." In: *arXiv e-prints*, arXiv:2306.05528 (June 2023), arXiv:2306.05528. DOI: 10.48550/arXiv.2306.05528. arXiv: 2306.05528 [astro-ph.IM].

[85] Marc Huertas-Company et al. "The Hubble Sequence at z ~ 0 in the IllustrisTNG simulation with deep learning." In: *MNRAS* 489.2 (Oct. 2019), pp. 1859–1879. DOI: 10.1093/mnras/stz2191. arXiv: 1903.07625 [astro-ph.GA].

[86] Marc Huertas-Company et al. "The Hubble Sequence at z ≈ 0 in the IllustrisTNG simulation with deep learning." In: *MNRAS* 489.2 (Aug. 2019), 1859–1879. ISSN: 1365-2966. DOI: 10.1093/mnras/stz2191. URL: http://dx.doi.org/10.1093/mnras/stz2191.

[87] Héctor J. Ibarra-Medel et al. "SDSS IV MaNGA: the global and local stellar mass assemby histories of galaxies." In: *Monthly Notices of the Royal Astronomical Society* 463.3 (Aug. 2016), pp. 2799–2818. ISSN: 0035-8711. DOI: 10.1093/mnras/stw2126. eprint: https://academic.oup.com/mnras/article-pdf/463/3/2799/18238704/stw2126.pdf. URL: https://doi.org/10.1093/mnras/stw2126.

[88] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *Computing Research Repository* abs/1502.03167 (2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

[89] Tomoaki Ishiyama et al. "The Uchuu simulations: Data Release 1 and dark matter halo concentrations." In: *MNRAS* 506.3 (Sept. 2021), pp. 4210–4231. DOI: 10.1093/mnras/stab1755. arXiv: 2007.14720 [astro-ph.CO].

[90] Željko Ivezić et al. "LSST: From Science Drivers to Reference Design and Anticipated Data Products." In: *ApJ* 873.2, 111 (Mar. 2019), p. 111. DOI: 10.3847/1538-4357/ab042c. arXiv: 0805.2366 [astro-ph].

[91] S. Kaviraj, C. Laigle, T. Kimm, J. E. G. Devriendt, Y. Dubois, C. Pichon, A. Slyz, E. Chisari, and S. Peirani. "The Horizon-AGN simulation: evolution of galaxy properties over cosmic time." In: *MNRAS* 467.4 (June 2017), pp. 4739–4752. DOI: 10.1093/mnras/stx126. arXiv: 1605.09379 [astro-ph.GA].

[92] Satoshi Kawanomoto, Fumihiro Uraguchi, Yutaka Komiyama, Satoshi Miyazaki, Hisanori Furusawa, François Finet, Takashi Hattori, Shiang-Yu Wang, Naoki Yasuda, and Naotaka Suzuki. "Hyper Suprime-Cam: Filters." In: *PASJ* 70.4, 66 (Aug. 2018), p. 66. DOI: 10.1093/pasj/psy056.

[93] Satoshi Kawanomoto, Fumihiro Uraguchi, Yutaka Komiyama, Satoshi Miyazaki, Hisanori Furusawa, François Finet, Takashi Hattori, Shiang-Yu Wang, Naoki Yasuda, and Naotaka Suzuki. "Hyper Suprime-Cam: Filters." In: *PASJ* 70.4, 66 (Aug. 2018), p. 66. DOI: 10.1093/pasj/psy056.

[94] Lalitwadee Kawinwanichakij et al. "Hyper Suprime-Cam Subaru Strategic Program: A Mass-dependent Slope of the Galaxy Size-Mass Relation at z < 1." In: *ApJ* 921.1, 38 (Nov. 2021), p. 38. DOI: 10.3847/1538-4357/ac1f21. arXiv: 2109.09766 [astro-ph.GA].

[95] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *arXiv e-prints*, arXiv:1412.6980 (Dec. 2014), arXiv:1412.6980. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs.LG].

[96] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[97] Diederik P. Kingma and Prafulla Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions." In: *arXiv e-prints*, arXiv:1807.03039 (July 2018), arXiv:1807.03039. arXiv: 1807.03039 [stat.ML].

[98] Yutaka Komiyama et al. "Hyper Suprime-Cam: Camera dewar design." In: *PASJ* 70, S2 (Jan. 2018), S2. DOI: 10.1093/pasj/psx069.

[99] Victor F Ksoll, Lynton Ardizzone, Ralf Klessen, Ullrich Koethe, Elena Sabbi, Massimo Robberto, Dimitrios Gouliermis, Carsten Rother, Peter Zeidler, and Mario Gennaro. "Stellar parameter determination from photometry using invertible neural networks." In: *MNRAS* 499.4 (Nov. 2020), pp. 5447–5485. ISSN: 0035-8711. DOI: 10.1093/mnras/staa2931. eprint: https://academic.oup.com/mnras/article-pdf/499/4/5447/34141391/staa2931.pdf. URL: https://doi.org/10.1093/mnras/staa2931.

[100] Cedric Lacey and Shaun Cole. "Merger rates in hierarchical models of galaxy formation." In: *MNRAS* 262.3 (June 1993), pp. 627–649. DOI: 10.1093/mnras/262.3.627.

[101] Johannes U. Lange, Frank C. van den Bosch, Andrew R. Zentner, Kuan Wang, and Antonia Sierra Villarreal. "Maturing satellite kinematics into a competitive probe of the galaxy-halo connection." In: *MNRAS* 482.4 (Feb. 2019), pp. 4824–4845. DOI: 10.1093/mnras/sty2950. arXiv: 1810.10511 [astro-ph.CO].

[102] Jennifer M. Lotz et al. "The Evolution of Galaxy Mergers and Morphology at z < 1.2 in the Extended Groth Strip." In: *ApJ* 672.1 (Jan. 2008), pp. 177–197. DOI: 10.1086/523659. arXiv: astro-ph/0602088 [astro-ph].

[103] Steven R. Majewski et al. "The Apache Point Observatory Galactic Evolution Experiment (APOGEE)." In: *AJ* 154.3, 94 (Sept. 2017), p. 94. DOI: 10.3847/1538-3881/aa784d. arXiv: 1509.05420 [astro-ph.IM].

[104] Berta Margalef-Bentabol, Marc Huertas-Company, Tom Charnock, Carla Margalef-Bentabol, Mariangela Bernardi, Yohan Dubois, Kate Storey-Fisher, and Lorenzo Zanisi. "Detecting outliers in astronomical images with deep generative networks." In: *MNRAS* 496.2 (Aug. 2020), pp. 2346–2361. DOI: 10.1093/mnras/staa1647. arXiv: 2003.08263 [astro-ph.GA].

[105] Federico Marinacci, Rüdiger Pakmor, and Volker Springel. "The formation of disc galaxies in high-resolution moving-mesh cosmological simulations." In: *MNRAS* 437.2 (Nov. 2013), 1750–1775. ISSN: 0035-8711. DOI: 10.1093/mnras/stt2003. URL: http://dx.doi.org/10.1093/mnras/stt2003.

[106] Federico Marinacci et al. "First results from the IllustrisTNG simulations: radio haloes and magnetic fields." In: *MNRAS* (Aug. 2018). ISSN: 1365-2966. DOI: 10.1093/mnras/sty2206. URL: http://dx.doi.org/10.1093/mnras/sty2206.

[107] David Martínez-Delgado et al. "Stellar Tidal Streams in Spiral Galaxies of the Local Volume: A Pilot Survey with Modest Aperture Telescopes." In: *The Astronomical Journal* 140.4 (Oct. 2010), pp. 962–967. DOI: 10.1088/0004-6256/140/4/962. arXiv: 1003.4860 [astro-ph.CO].

[108] Karen L. Masters and Galaxy Zoo Team. "Twelve years of Galaxy Zoo." In: *Galactic Dynamics in the Era of Large Surveys*. Ed. by Monica Valluri and J. A. Sellwood. Vol. 353. Jan. 2020, pp. 205–212. DOI: 10.1017/S1743921319008615. arXiv: 1910.08177 [astro-ph.GA].

[109] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML].

[110] Allison Merritt, Annalisa Pillepich, Pieter van Dokkum, Dylan Nelson, Lars Hernquist, Federico Marinacci, and Mark Vogelsberger. "A missing outskirts problem? Comparisons between stellar haloes in the Dragonfly Nearby Galaxies Survey and the TNG100 simulation." In: *MNRAS* 495.4 (July 2020), pp. 4570–4604. DOI: 10.1093/mnras/staa1164. arXiv: 2004.11402 [astro-ph.GA].

[111] Satoshi Miyazaki et al. "Hyper Suprime-Cam." In: *Ground-based and Airborne Instrumentation for Astronomy IV*. Ed. by Ian S. McLean, Suzanne K. Ramsay, and Hideki Takami. Vol. 8446. International Society for Optics and Photonics. SPIE, 2012, 84460Z. DOI: 10.1117/12.926844. URL: https://doi.org/10.1117/12.926844.

[112] Satoshi Miyazaki et al. "Hyper Suprime-Cam: System design and verification of image quality." In: *PASJ* 70, S1 (Jan. 2018), S1. DOI: 10.1093/pasj/psx063.

[113] Antonela Monachesi et al. "The Auriga stellar haloes: connecting stellar population properties with accretion and merging history." In: *Monthly Notices of the Royal Astronomical Society* 485.2 (Feb. 2019), pp. 2589–2616. ISSN: 0035-8711. DOI: 10.1093/mnras/stz538. eprint: https://academic.oup.com/mnras/article-pdf/485/2/2589/28052181/stz538.pdf. URL: https://doi.org/10.1093/mnras/stz538.

[114] Rohan P. Naidu et al. "Reconstructing the Last Major Merger of the Milky Way with the H3 Survey." In: *arXiv e-prints*, arXiv:2103.03251 (Mar. 2021), arXiv:2103.03251. arXiv: 2103.03251 [astro-ph.GA].

[115] Rohan P. Naidu et al. "Reconstructing the Last Major Merger of the Milky Way with the H3 Survey." In: *ApJ* 923.1, 92 (Dec. 2021), p. 92. DOI: 10.3847/1538-4357/ac2d2d. arXiv: 2103.03251 [astro-ph.GA].

[116] Jill P. Naiman et al. "First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium." In: *MNRAS* 477.1 (June 2018), pp. 1206–1224. DOI: 10.1093/mnras/sty618. arXiv: 1707.03401 [astro-ph.GA].

[117] Lorenza Nanni et al. "iMaNGA: mock MaNGA galaxies based on IllustrisTNG and MaStar SSPs - I. Construction and analysis of the mock data cubes." In: *MNRAS* 515.1 (Sept. 2022), pp. 320–338. DOI: 10.1093/mnras/stac1531. arXiv: 2203.11575 [astro-ph.GA].

[118] Lorenza Nanni et al. "iMaNGA: mock MaNGA galaxies based on IllustrisTNG and MaStar SSPs - II. The catalogue." In: *MNRAS* 522.4 (July 2023), pp. 5479–5499. DOI: 10.1093/mnras/stad1337. arXiv: 2211.13146 [astro-ph.GA].

[119] D. Nelson et al. "The illustris simulation: Public data release." In: *Astronomy and Computing* 13 (Nov. 2015), pp. 12–37. DOI: 10.1016/j.ascom.2015.09.003. arXiv: 1504.00362 [astro-ph.CO].

[120] Dylan Nelson, Annalisa Pillepich, Mohammadreza Ayromlou, Wonki Lee, Katrin Lehle, Eric Rohr, and Nhut Truong. "Introducing the TNG-Cluster Simulation: overview and physical properties of the gaseous intracluster medium." In: *arXiv e-prints*, arXiv:2311.06338 (Nov. 2023), arXiv:2311.06338. DOI: 10.48550/arXiv.2311.06338. arXiv: 2311.06338 [astro-ph.GA].

[121] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rüdiger Pakmor, Rainer Weinberger, Shy Genel, Paul Torrey, Mark Vogelsberger, Federico Marinacci, and Lars Hernquist. "First results from the TNG50 simulation: galactic outflows driven by supernovae and black hole feedback." In: *MNRAS* 490.3 (Aug. 2019), 3234–3261. ISSN: 1365-2966. DOI: 10.1093/mnras/stz2306. URL: http://dx.doi.org/10.1093/mnras/stz2306.

[122] Dylan Nelson et al. "First results from the IllustrisTNG simulations: the galaxy colour bimodality." In: *MNRAS* 475.1 (Mar. 2018), pp. 624–647. DOI: 10.1093/mnras/stx3040. arXiv: 1707.03395 [astro-ph.GA].

[123] Dylan Nelson et al. "First results from the IllustrisTNG simulations: the galaxy colour bimodality." In: *MNRAS* 475.1 (Mar. 2018), pp. 624–647. DOI: 10.1093/mnras/stx3040. arXiv: 1707.03395 [astro-ph.GA].

[124] Dylan Nelson et al. "The IllustrisTNG simulations: public data release." In: *Computational Astrophysics and Cosmology* 6.1, 2 (May 2019), p. 2. DOI: 10.1186/s40668-019-0028-x. arXiv: 1812.05609 [astro-ph.GA].

[125] Danail Obreschkow, Pascal J. Elahi, Claudia del P. Lagos, Rhys J. J. Poulton, and Aaron D. Ludlow. "Characterizing the structure of halo merger trees using a single parameter: the tree entropy." In: *MNRAS* 493.3 (Apr. 2020), pp. 4551–4569. DOI: 10.1093/mnras/staa445. arXiv: 1911.11959 [astro-ph.GA].

[126] Ludwig Oser, Jeremiah P. Ostriker, Thorsten Naab, Peter H. Johansson, and Andreas Burkert. "The Two Phases of Galaxy Formation." In: *ApJ* 725.2 (Dec. 2010), pp. 2312–2323. DOI: 10.1088/0004-637X/725/2/2312. arXiv: 1010.1381 [astro-ph.CO].

[127] Rüdiger Pakmor et al. "The MillenniumTNG Project: the hydrodynamical full physics simulation and a first look at its galaxy clusters." In: *MNRAS* 524.2 (Sept. 2023), pp. 2539–2555. DOI: 10.1093/mnras/stac3620. arXiv: 2210.10060 [astro-ph.CO].

[128] Johanna Pasquet, E. Bertin, M. Treyer, S. Arnouts, and D. Fouchez. "Photometric redshifts from SDSS images using a convolutional neural network." In: *A&A* 621 (Dec. 2018), A26. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201833617. URL: http://dx.doi.org/10.1051/0004-6361/201833617.

[129] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035.

[130] Karl Pearson. *LIII. On lines and planes of closest fit to systems of points in space*. Sept. 2018. DOI: 10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

[131] W. J. Pearson, L. Wang, J. W. Trayford, C. E. Petrillo, and F. F. S. van der Tak. "Identifying galaxy mergers in observations and simulations with deep learning." In: *A&A* 626, A49 (June 2019), A49. DOI: 10.1051/0004-6361/201935355. arXiv: 1902.10626 [astro-ph.GA].

[132] Annalisa Pillepich, Piero Madau, and Lucio Mayer. "BUILDING LATE-TYPE SPIRAL GALAXIES BY IN-SITU AND EX-SITU STAR FORMATION." In: *The Astrophysical Journal* 799.2 (Jan. 2015), p. 184. ISSN: 1538-4357. DOI: 10.1088/0004-637x/799/2/184. URL: http://dx.doi.org/10.1088/0004-637X/799/2/184.

[133] Annalisa Pillepich et al. "Halo mass and assembly history exposed in the faint outskirts: the stellar and dark matter haloes of Illustris galaxies." In: *MNRAS* 444.1 (Aug. 2014), 237–249. ISSN: 1365-2966. DOI: 10.1093/mnras/stu1408. URL: http://dx.doi.org/10.1093/mnras/stu1408.

[134] Annalisa Pillepich et al. "First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies." In: *MNRAS* 475.1 (Jan. 2018), 648–675. ISSN: 1365-2966. DOI: 10.1093/mnras/stx3112. URL: http://dx.doi.org/10.1093/mnras/stx3112.

[135] Annalisa Pillepich et al. "Simulating galaxy formation with the IllustrisTNG model." In: *MNRAS* 473.3 (Jan. 2018), 4077–4106. ISSN: 1365-2966. DOI: 10.1093/mnras/stx2656. URL: http://dx.doi.org/10.1093/mnras/stx2656.

[136] Annalisa Pillepich et al. "First results from the TNG50 simulation: the evolution of stellar and gaseous discs across cosmic time." In: *MNRAS* 490.3 (Sept. 2019), 3196–3233. ISSN: 1365-2966. DOI: 10.1093/mnras/stz2338. URL: http://dx.doi.org/10.1093/mnras/stz2338.

[137] Planck Collaboration et al. "Planck 2015 results. XIII. Cosmological parameters." In: *A&A* 594, A13 (Sept. 2016), A13. DOI: 10.1051/0004-6361/201525830. arXiv: 1502.01589 [astro-ph.CO].

[138] Planck Collaboration et al. "Planck 2018 results - VI. Cosmological parameters." In: *A&A* 641 (2020), A6. DOI: 10.1051/0004-6361/201833910. URL: https://doi.org/10.1051/0004-6361/201833910.

[139] Erwin Platen, Rien van de Weygaert, Bernard J. T. Jones, Gert Vegter, and Miguel A. Aragón Calvo. "Structural analysis of the SDSS Cosmic Web - I. Non-linear density field reconstructions." In: *MNRAS* 416.4 (Oct. 2011), pp. 2494–2526. DOI: 10.1111/j.1365-2966.2011.18905.x. arXiv: 1107.1488 [astro-ph.CO].

[140] Ana-Roxana Pop, Annalisa Pillepich, Nicola C Amorisco, and Lars Hernquist. "Formation and incidence of shell galaxies in the Illustris simulation." In: *MNRAS* 480.2 (July 2018), 1715–1739. ISSN: 1365-2966. DOI: 10.1093/mnras/sty1932. URL: http://dx.doi.org/10.1093/mnras/sty1932.

[141] Gergö Popping, Annalisa Pillepich, Gabriela Calistro Rivera, Sebastian Schulz, Lars Hernquist, Melanie Kaasinen, Federico Marinacci, Dylan Nelson, and Mark Vogelsberger. "The dust-continuum size of TNG50 galaxies at z = 1-5: a comparison with the distribution of stellar light, stars, dust, and H_2." In: *MNRAS* 510.3 (Mar. 2022), pp. 3321–3334. DOI: 10.1093/mnras/stab3312. arXiv: 2101.12218 [astro-ph.GA].

[142] Douglas Potter, Joachim Stadel, and Romain Teyssier. "PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys." In: *Computational Astrophysics and Cosmology* 4.1, 2 (May 2017), p. 2. DOI: 10.1186/s40668-017-0021-1. arXiv: 1609.08621 [astro-ph.IM].

[143] C. Pulsoni, O. Gerhard, M. Arnaboldi, A. Pillepich, V. Rodriguez-Gomez, D. Nelson, L. Hernquist, and V. Springel. "The stellar halos of ETGs in the IllustrisTNG simulations. II. Accretion, merger history, and dark halo connection." In: *A&A* 647, A95 (Mar. 2021), A95. DOI: 10.1051/0004-6361/202039166. arXiv: 2009.01823 [astro-ph.GA].

[144] P. J. Quinn, Lars Hernquist, and D. P. Fullagar. "Heating of Galactic Disks by Mergers." In: *ApJ* 403 (Jan. 1993), p. 74. DOI: 10.1086/172184.

[145] A. Rémy-Ruyer et al. "Gas-to-dust mass ratios in local galaxies over a 2 dex metallicity range." In: *A&A* 563, A31 (Mar. 2014), A31. DOI: 10.1051/0004-6361/201322803. arXiv: 1312.3442 [astro-ph.GA].

[146] Vicente Rodriguez-Gomez et al. "The merger rate of galaxies in the Illustris simulation: a comparison with observations and semi-empirical models." In: *MNRAS* 449.1 (Mar. 2015), 49–64. ISSN: 0035-8711. DOI: 10.1093/mnras/stv264. URL: http://dx.doi.org/10.1093/mnras/stv264.

[147] Vicente Rodriguez-Gomez et al. "The stellar mass assembly of galaxies in the Illustris simulation: growth by mergers and the spatial distribution of accreted stars." In: *MNRAS* 458.3 (May 2016), pp. 2371–2390. DOI: 10.1093/mnras/stw456. arXiv: 1511.08804 [astro-ph.GA].

[148] Vicente Rodriguez-Gomez et al. "The stellar mass assembly of galaxies in the Illustris simulation: growth by mergers and the spatial distribution of accreted stars." In: *MNRAS* 458.3 (May 2016), pp. 2371–2390. DOI: 10.1093/mnras/stw456. arXiv: 1511.08804 [astro-ph.GA].

[149] Vicente Rodriguez-Gomez et al. "The role of mergers and halo spin in shaping galaxy morphology." In: *MNRAS* 467.3 (May 2017), pp. 3083–3098. DOI: 10.1093/mnras/stx305. arXiv: 1609.09498 [astro-ph.GA].

[150] Vicente Rodriguez-Gomez et al. "The optical morphologies of galaxies in the IllustrisTNG simulation: a comparison to Pan-STARRS observations." In: *MNRAS* 483.3 (Mar. 2019), pp. 4140–4159. DOI: 10.1093/mnras/sty3345. arXiv: 1809.08239 [astro-ph.GA].

[151] Vicente Rodriguez-Gomez et al. "The optical morphologies of galaxies in the IllustrisTNG simulation: a comparison to Pan-STARRS observations." In: *MNRAS* 483.3 (Mar. 2019), pp. 4140–4159. DOI: 10.1093/mnras/sty3345. arXiv: 1809.08239 [astro-ph.GA].

[152] Suman Sarkar and Biswajit Pandey. "Unravelling the cosmic web: an analysis of the Sloan Digital Sky Survey Data Release 14 with the local dimension." In: *MNRAS* 485.4 (June 2019), pp. 4743–4753. DOI: 10.1093/mnras/stz499. arXiv: 1812.03661 [astro-ph.CO].

[153] Regina Sarmiento, Marc Huertas-Company, Johan H. Knapen, Héctor Ibarra-Medel, Annalisa Pillepich, Sebastián F. Sánchez, and Alina Boecker. "MaNGIA: 10 000 mock galaxies for stellar population analysis." In: *A&A* 673, A23 (May 2023), A23. DOI: 10.1051/0004-6361/202245509. arXiv: 2211.11790 [astro-ph.GA].

[154] M. Sarzi et al. "Fornax3D project: Overall goals, galaxy sample, MUSE data analysis, and initial results." In: *A&A* 616, A121 (Aug. 2018), A121. DOI: 10.1051/0004-6361/201833137. arXiv: 1804.06795 [astro-ph.GA].

[155] Joop Schaye et al. "The EAGLE project: simulating the evolution and assembly of galaxies and their environments." In: *MNRAS* 446.1 (Jan. 2015), pp. 521–554. DOI: 10.1093/mnras/stu2058. arXiv: 1407.7040 [astro-ph.GA].

[156] Joop Schaye et al. "The EAGLE project: simulating the evolution and assembly of galaxies and their environments." In: *MNRAS* 446.1 (Jan. 2015), pp. 521–554. DOI: 10.1093/mnras/stu2058. arXiv: 1407.7040 [astro-ph.GA].

[157] P. Schechter. "An analytic expression for the luminosity function for galaxies." In: *ApJ* 203 (Jan. 1976), pp. 297–306. DOI: 10.1086/154079.

[158] J. L. Sérsic. "Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy." In: *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina* 6 (Feb. 1963), pp. 41–43.

[159] Rui Shi, Wenting Wang, Zhaozhou Li, Jiaxin Han, Jingjing Shi, Vicente Rodriguez-Gomez, and Yingjie Peng. "A machine learning approach to infer the accreted stellar mass fractions of galaxies." In: *arXiv e-prints*, arXiv:2112.07203 (Dec. 2021), arXiv:2112.07203. arXiv: 2112.07203 [astro-ph.GA].

[160] Rui Shi, Wenting Wang, Zhaozhou Li, Jiaxin Han, Jingjing Shi, Vicente Rodriguez-Gomez, Yingjie Peng, and Qingyang Li. "A machine learning approach to infer the accreted stellar mass fractions of central galaxies in the TNG100 simulation." In: *MNRAS* 515.3 (Sept. 2022), pp. 3938–3955. DOI: 10.1093/mnras/stac1541. arXiv: 2112.07203 [astro-ph.GA].

[161] Rhythm Shimakawa, Masayuki Tanaka, Connor Bottrell, Po-Feng Wu, Yu-Yen Chang, Yoshiki Toba, and Sadman Ali. "Passive spiral galaxies deeply captured by Subaru Hyper Suprime-Cam." In: *PASJ* 74.3 (June 2022), pp. 612–624. DOI: 10.1093/pasj/psac023. arXiv: 2203.00969 [astro-ph.GA].

[162] Rhythm Shimakawa, Masayuki Tanaka, Kei Ito, and Makoto Ando. "GALAXY CRUISE: Spiral and ring classifications for bright galaxies at z=0.01-0.3." In: *arXiv e-prints*, arXiv:2309.14830 (Sept. 2023), arXiv:2309.14830. DOI: 10.48550/arXiv.2309.14830. arXiv: 2309.14830 [astro-ph.GA].

[163] Gregory F. Snyder, Theodore Peña, L. Y. Aaron Yung, Caitlin Rose, Jeyhan Kartaltepe, and Harry Ferguson. "Mock galaxy surveys for HST and JWST from the IllustrisTNG simulations." In: *MNRAS* 518.4 (Feb. 2023), pp. 6318–6324. DOI: 10.1093/mnras/stac3397. arXiv: 2211.09677 [astro-ph.GA].

[164] Gregory F. Snyder et al. "Galaxy morphology and star formation in the Illustris Simulation at z = 0." In: *MNRAS* 454.2 (Dec. 2015), pp. 1886–1908. DOI: 10.1093/mnras/stv2078. arXiv: 1502.07747 [astro-ph.GA].

[165] Sho Sonoda and Noboru Murata. "Neural Network with Unbounded Activation Functions is Universal Approximator." In: *arXiv e-prints*, arXiv:1505.03654 (May 2015), arXiv:1505.03654. DOI: 10.48550/arXiv.1505.03654. arXiv: 1505.03654 [cs.NE].

[166] Volker Springel. "The cosmological simulation code gadget-2." In: *MNRAS* 364.4 (Dec. 2005), 1105–1134. ISSN: 1365-2966. DOI: 10.1111/j.1365-2966.2005.09655.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2005.09655.x.

[167] Volker Springel. "E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh." In: *MNRAS* 401.2 (Jan. 2010), 791–851. ISSN: 1365-2966. DOI: 10.1111/j.1365-2966.2009.15715.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2009.15715.x.

[168] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. "Populating a cluster of galaxies - I. Results at z=0." In: *MNRAS* 328.3 (Dec. 2001), pp. 726–750. DOI: 10.1046/j.1365-8711.2001.04912.x. arXiv: astro-ph/0012055 [astro-ph].

[169] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. "Populating a cluster of galaxies - I. Results at z=0." In: *MNRAS* 328.3 (Dec. 2001), pp. 726–750. DOI: 10.1046/j.1365-8711.2001.04912.x. arXiv: astro-ph/0012055 [astro-ph].

[170] Volker Springel et al. "Simulations of the formation, evolution and clustering of galaxies and quasars." In: *Nature* 435.7042 (June 2005), pp. 629–636. DOI: 10.1038/nature03597. arXiv: astro-ph/0504097 [astro-ph].

[171] Volker Springel et al. "First results from the IllustrisTNG simulations: matter and galaxy clustering." In: *MNRAS* 475.1 (Mar. 2018), 676–698. ISSN: 1365-2966. DOI: 10.1093/mnras/stx3304. URL: http://dx.doi.org/10.1093/mnras/stx3304.

[172] Volker Springel et al. "First results from the IllustrisTNG simulations: matter and galaxy clustering." In: *MNRAS* 475.1 (Mar. 2018), pp. 676–698. DOI: 10.1093/mnras/stx3304. arXiv: 1707.03397 [astro-ph.GA].

[173] Michael A. Strauss et al. "Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample." In: *AJ* 124.3 (Sept. 2002), pp. 1810–1824. DOI: 10.1086/342343. arXiv: astro-ph/0206225 [astro-ph].

[174] M. J. Stringer, A. J. Benson, K. Bundy, R. S. Ellis, and E. L. Quetin. "Mock observations with the Millennium Simulation: cosmological downsizing and intermediate-redshift observations." In: *MNRAS* 393.4 (Mar. 2009), pp. 1127–1140. DOI: 10.1111/j.1365-2966.2008.14186.x. arXiv: 0806.2232 [astro-ph].

[175] Y. Su et al. "A deep learning view of the census of galaxy clusters in IllustrisTNG." In: *MNRAS* 498.4 (Nov. 2020), pp. 5620–5628. DOI: 10.1093/mnras/staa2690. arXiv: 2007.05144 [astro-ph.CO].

[176] Dylan Nelson for the TNG Collaboration. *IllustrisTNG - Infinite Gallery — tng-project.org*. https://www.tng-project.org/explore/gallery/bottrell23/. [Accessed 27-04-2024].

[177] Masayuki Tanaka, Jean Coupon, Bau-Ching Hsieh, Sogo Mineo, Atsushi J. Nishizawa, Joshua Speagle, Hisanori Furusawa, Satoshi Miyazaki, and Hitoshi Murayama. "Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1." In: *PASJ* 70, S9 (Jan. 2018), S9. DOI: 10.1093/pasj/psx077. arXiv: 1704.05988 [astro-ph.GA].

[178] Paul Torrey et al. "Synthetic galaxy images and spectra from the Illustris simulation." In: *MNRAS* 447.3 (Mar. 2015), pp. 2753–2771. DOI: 10.1093/mnras/stu2592. arXiv: 1411.3717 [astro-ph.GA].

[179] Paul Torrey et al. "The evolution of the mass-metallicity relation and its scatter in IllustrisTNG." In: *MNRAS* 484.4 (Apr. 2019), pp. 5587–5607. DOI: 10.1093/mnras/stz243. arXiv: 1711.05261 [astro-ph.GA].

[180] James W. Trayford, Peter Camps, Tom Theuns, Maarten Baes, Richard G. Bower, Robert A. Crain, Madusha L. P. Gunawardhana, Matthieu Schaller, Joop Schaye, and Carlos S. Frenk. "Optical colours and spectral indices of z = 0.1 eagle galaxies with the 3D dust radiative transfer code skirt." In: *MNRAS* 470.1 (Sept. 2017), pp. 771–799. DOI: 10.1093/mnras/stx1051. arXiv: 1705.02331 [astro-ph.GA].

[181] R. B. Tully and J. R. Fisher. "A new method of determining distances to galaxies." In: *A&A* 54 (Feb. 1977), pp. 661–673.

[182] A. J. Turnbull, T. J. Bridges, and D. Carter. "Imaging of the shell galaxies NGC 474 and 7600, and implications for their formation." In: *MNRAS* 307.4 (Aug. 1999), pp. 967–976. DOI: 10.1046/j.1365-8711.1999.02724.x. arXiv: astro-ph/9905041 [astro-ph].

[183] "Variational Autoencoder." In: *Variational Methods for Machine Learning with Applications to Deep Networks* (2021). URL: https://api.semanticscholar.org/CorpusID:240802776.

[184] J. Vega-Ferrero et al. "On the nature of disks at high redshift seen by JWST/CEERS with contrastive learning and cosmological simulations." In: *arXiv e-prints*, arXiv:2302.07277 (Feb. 2023), arXiv:2302.07277. DOI: 10.48550/arXiv.2302.07277. arXiv: 2302.07277 [astro-ph.CO].

[185] Francisco Villaescusa-Navarro et al. "Multifield Cosmology with Artificial Intelligence." In: *arXiv e-prints*, arXiv:2109.09747 (Sept. 2021), arXiv:2109.09747. DOI: 10.48550/arXiv.2109.09747. arXiv: 2109.09747 [astro-ph.CO].

[186] Francisco Villaescusa-Navarro et al. "The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations." In: *ApJ* 915.1, 71 (July 2021), p. 71. DOI: 10.3847/1538-4357/abf7ba. arXiv: 2010.00619 [astro-ph.CO].

[187] M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, S. Bird, D. Nelson, and L. Hernquist. "Properties of galaxies reproduced by a hydrodynamic simulation." In: *Nature* 509.7499 (May 2014), pp. 177–182. DOI: 10.1038/nature13316. arXiv: 1405.1418 [astro-ph.CO].

[188] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Greg Snyder, Dylan Nelson, and Lars Hernquist. "Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe." In: *MNRAS* 444.2 (Oct. 2014), pp. 1518–1547. DOI: 10.1093/mnras/stu1536. arXiv: 1405.2921 [astro-ph.CO].

[189] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. "Cosmological simulations of galaxy formation." In: *Nature Reviews Physics* 2.1 (Jan. 2020), pp. 42–66. DOI: 10.1038/s42254-019-0127-2. arXiv: 1909.07976 [astro-ph.GA].

[190] Rodrigo Von Marttens, Luciano Casarini, Nicola R. Napolitano, Sirui Wu, Valeria Amaro, Rui Li, Crescenzo Tortora, Askery Canabarro, and Yang Wang. "Inferring galaxy dark halo properties from visible matter with Machine Learning." In: *arXiv e-prints*, arXiv:2111.01185 (Nov. 2021), arXiv:2111.01185. arXiv: 2111.01185 [astro-ph.GA].

[191] L. Wang, W. J. Pearson, and V. Rodriguez-Gomez. "Towards a consistent framework of comparing galaxy mergers in observations and simulations." In: *A&A* 644, A87 (Dec. 2020), A87. DOI: 10.1051/0004-6361/202038084. arXiv: 2009.02974 [astro-ph.GA].

[192] Risa H. Wechsler and Jeremy L. Tinker. "The Connection Between Galaxies and Their Dark Matter Halos." In: *ARA&A* 56 (Sept. 2018), pp. 435–487. DOI: 10.1146/annurev-astro-081817-051756. arXiv: 1804.03097 [astro-ph.GA].

[193] Maurice Weiler and Gabriele Cesa. *General* E(2)*-Equivariant Steerable CNNs*. 2021. arXiv: 1911.08251 [cs.CV].

[194] Rainer Weinberger et al. "Simulating galaxy formation with black hole driven thermal and kinetic feedback." In: *MNRAS* 465.3 (Mar. 2017), 3291–3308. ISSN: 1365-2966. DOI: 10.1093/mnras/stw2944. URL: http://dx.doi.org/10.1093/mnras/stw2944.

[195] Rainer Weinberger et al. "Supermassive black holes and their feedback effects in the IllustrisTNG simulation." In: *MNRAS* 479.3 (Sept. 2018), pp. 4056–4072. DOI: 10.1093/mnras/sty1733. arXiv: 1710.04659 [astro-ph.GA].

[196] Joseph C. Weingartner and B. T. Draine. "Dust Grain-Size Distributions and Extinction in the Milky Way, Large Magellanic Cloud, and Small Magellanic Cloud." In: *ApJ* 548.1 (Feb. 2001), pp. 296–309. DOI: 10.1086/318651. arXiv: astro-ph/0008146 [astro-ph].

[197] Po-Feng Wu et al. "Toward Precise Galaxy Evolution: A Comparison between Spectral Indices of z 1 Galaxies in the IllustrisTNG Simulation and the LEGA-C Survey." In: *AJ* 162.5, 201 (Nov. 2021), p. 201. DOI: 10.3847/1538-3881/ac20d6. arXiv: 2108.10455 [astro-ph.GA].

[198] Sergey Zagoruyko and Nikos Komodakis. *Wide Residual Networks*. 2017. arXiv: 1605.07146 [cs.CV].

[199] Lorenzo Zanisi et al. "A deep learning approach to test the small-scale galaxy morphology and its relationship with star formation activity in hydrodynamical simulations." In: *MNRAS* 501.3 (Mar. 2021), pp. 4359–4382. DOI: 10.1093/mnras/staa3864. arXiv: 2007.00039 [astro-ph.GA].

[200] Lorenzo Zanisi et al. "A deep learning approach to test the small-scale galaxy morphology and its relationship with star formation activity in hydrodynamical simulations." In: *MNRAS* 501.3 (Mar. 2021), pp. 4359–4382. DOI: 10.1093/mnras/staa3864. arXiv: 2007.00039 [astro-ph.GA].

[201] Ling Zhu, Annalisa Pillepich, Glenn van de Ven, Ryan Leaman, Lars Hernquist, Dylan Nelson, Ruediger Pakmor, Mark Vogelsberger, and Le Zhang. "The mass of the dynamically-hot inner stellar halo predicts the ancient accreted satellite(s) mass." In: *arXiv e-prints*, arXiv:2110.13172 (Oct. 2021), arXiv:2110.13172. arXiv: 2110.13172 [astro-ph.GA].

[202] Ling Zhu et al. "Disentangling the formation history of galaxies via population-orbit superposition: method validation." In: *MNRAS* 496.2 (June 2020), pp. 1579–1597. DOI: 10.1093/mnras/staa1584. arXiv: 2003.05561 [astro-ph.GA].

[203] Anna de Graaff, James Trayford, Marijn Franx, Matthieu Schaller, Joop Schaye, and Arjen van der Wel. "Observed structural parameters of EAGLE galaxies: reconciling the mass-size relation in simulations with local observations." In: *MNRAS* 511.2 (Apr. 2022), pp. 2544–2564. DOI: 10.1093/mnras/stab3510. arXiv: 2110.02235 [astro-ph.GA].

[204] A. van der Wel et al. "3D-HST+CANDELS: The Evolution of the Galaxy Size-Mass Distribution since z = 3." In: *ApJ* 788.1, 28 (June 2014), p. 28. DOI: 10.1088/0004-637X/788/1/28. arXiv: 1404.2844 [astro-ph.GA].