Rare Galaxy Classes Identified In Foundation Model Representations

Mike Walmsley[*]
Dunlap Institute for Astronomy and Astrophysics
University of Toronto
Toronto, ON M5S 3H4, Canada
`m.walmsley@utoronto.ca`

Anna M.M Scaife
Jodrell Bank Centre for Astrophysics
Department of Physics & Astronomy
University of Manchester
Manchester M13 9PL, UK
`anna.scaife@manchester.ac.uk`

Abstract

We identify rare and visually distinctive galaxy populations by searching for structure within the learned representations of pretrained models. We show that these representations arrange galaxies by appearance in patterns beyond those needed to predict the pretraining labels. We design a clustering approach to isolate specific local patterns, revealing groups of galaxies with rare and scientifically-interesting morphologies.

1 Introduction

Astronomers often focus on specific classes of galaxy that have been particularly affected (or unaffected) by a process of interest [1, 2, 3, 4]. These classes may be extremely rare; 'jellyfish' galaxies, for example, are only reliably identified in dense cluster environments [5, 6].

It has recently become possible to automatically search for rare galaxy classes by manually identifying a small set of examples and then finetuning a pretrained model to search for more examples at scale. The most frequently-used galaxy model library is Zoobot [7], a set of models trained on 92M responses by Galaxy Zoo volunteers [8] answering a diverse set of classification tasks. By jointly predicting responses to all of these tasks, Zoobot's encoder learns to assign visually similar galaxies to similar points in representation space [9]. One can then straightforwardly (e.g. with a linear head) identify areas of representation space containing known examples of the target rare galaxy class.

But what if there are galaxy classes which are so rare we don't know to look for them? Here, we reverse the finetuning workflow and instead identify rare classes of galaxies directly from Zoobot's representation, without any known examples.

Walmsley et al. [9] suggested that the representation learned by Zoobot may arrange galaxies in patterns beyond those required to predict the supervised labels. Our approach here will both prove this thesis and demonstrate that, beyond being grouped by similarity, variations in the density of galaxies within that representation can be exploited to reveal new subclasses.

Our aim of finding rare galaxy classes is related to anomaly-finding. Unlike standard anomaly-finding [10, 11, 12], we explicitly aim to identify anomalies with multiple examples; a one-off strange galaxy is a puzzle, but a set of them is a new population.

Our aim is also related to the topic of data-driven galaxy classification schema. Astronomers often divide all galaxies into broad (i.e. not rare) classes and then compare their features. The most 'meaningful' criteria with which to define these broad classes is a matter of decades-long debate

---

[*]Second affiliation: Jodrell Bank Centre for Astrophysics, Department of Physics & Astronomy, University of Manchester, Manchester M13 9PL, UK

[13]. Much work has been done to uncover alternative data-driven criteria both before [14] and with [15, 16] machine learning, but these alternative criteria are rarely used in practice. Our method is not intended to describe galaxies in general - less than 2% of all galaxies studied here fall in one of our classes - but rather aims to isolate rare and visually distinctive classes within the bulk sample.

2 Data

We begin with the public pretrained models available from the Zoobot model library. We run all experiments with both the EfficientNetB0 and MaxViT Tiny architectures (224px, three-channel) to test if adding multi-axis attention blocks alters the representation learned [17, 18, 19]. These models have been referred to [7] as galaxy 'foundation' models, in that they are pretrained on a diverse task and extensive training data in order to learn an adaptable representation, but we highlight that have relatively few parameters; EfficientNetB0 and MaxViT Tiny are both low-parameter variants of their respective model families.

We construct our initial representations by applying the pretrained MaxViT Tiny and EfficientNetB0 Zoobot models to images from the DESI Legacy Surveys [20]. Our base catalog is the 8.7M galaxies of suitable brightness ($r < 19.0$) and angular extent (Petrosian radius > 3") to have well-resolved features [21].

We use the predicted morphology labels (from the EfficientNetB0 model, for consistency) to automatically remove featureless ('Smooth' response probability $\geq 50\%$) and edge-on ('edge-on disk' $\geq 30\%$) galaxies before applying our search method. Naively applying our search method to this sample returns many clusters of artifacts (e.g. images contaminated by telescope issues). We are primarily interested in rare galaxies, not rare image issues, so (outside of a final experiment, Sec. 4) we remove these ('artifact $\geq 30\%$) but note that flagging contaminated images may be a further useful application of our search method. Our final sample is 632k DESI Legacy Survey images of featured galaxies.

3 Methods

The representation dimensionality of the published MaxViT Tiny and EfficientNetB0 architectures is 1024 and 1280, respectively. Motivated by a PCA analysis of the representation variance in the original networks, showing that the representations are highly redundant, we find we can retrain these networks using a lower representation dimension of 128 without any measurable performance loss. We refer to our adjusted networks as 'bottlenecked' networks.

We then apply dimensionality reduction to further compress our representation before searching for clusters in the compressed space. This approach is more obviously useful for an unsupervised or semi-supervised representation, because, if you have plentiful labels, one can instead simply train an end-to-end supervised model to predict those labels. However, here, our labels describe the features that each galaxy has (spiral arms, bars, etc.) but do not assign them classes. Accurately predicting these labels is therefore not sufficient to identify rare galaxy populations; they need to be clustered into classes.

Clustering within the representation is non-trivial. Supervised representations of galaxies have previously been noted as broadly smooth and naive clustering as unsuccessful [9]. We hypothesise that this global smoothness masks small-scale detailed structure and design our clustering approach to identify this structure.

We first reduce the bottlenecked representation dimensionality with UMAP [22]. We then construct a tree of candidate clusters with HDBSCAN [23] and post-process that tree for our final clusters.

UMAP's `n_neighbours` sets the local connectivity of the force-directed graph. Each point (node) has edges to the closest `n_neighbours`, and those edges apply forces weighted by distance. To find rare subclasses, `n_neighbours` must be large enough that small clusters can form but small enough that those clusters are not 'torn apart' by numerous non-cluster neighbours. We find `n_neighbours`$= 50$ provides a good balance for our goal of finding rare galaxy subtypes.

UMAP's `n_components` sets the final dimensionality and hence the typical separation of galaxies. Reducing to $D = 2$ preserves only the most noteable clusters (though is useful as a visualisation aid).

Reducing to $D = 15$ leads to qualitatively less persuasive clusters, likely because of sparsity[2]. We find $D = 5$ provides a good balance for our dataset size.

We next apply HDBSCAN to identify clusters in the reduced representation. HDBSCAN is typically used to identify well-separated clusters of significant size. To identify only confident clusters, we adjust the denoising distance transformation ('core distance') by setting `min_samples`$= 300$. We also set `min_cluster_size`$= 10$ to allow for rare classes. To allow for variable size clusters, we define clusters as the leaf nodes in HDBSCAN's condensed tree rather than the typical 'stable' (excess-of-mass) nodes.

This configuration is able to identify small clusters within the smooth bulk sample, but at the cost of further splitting larger clusters. A hyperparameter search identifies no single λ or ϵ threshold that selects small clusters without splitting larger clusters, and so instead we simply note that manually grouping visually similar clusters is trivial if required (all clusters presented here are shown as-found without manual intervention).

4 Results

Fig 1 shows classes identified directly from Zoobot's representation using our search method. Our classes reveal visually cohesive groups of galaxies with rare morphologies. Our data-driven search identifies these without expert astronomer direction.

Several of our classes have not previously been automatically detected or studied at scale. Most notably, class D shows a population of galaxies with dominant bars and irregular clumpy spiral structure (likely low-mass and starforming).

The classes are split in ways not obviously motivated by the supervised labels. For example, with class D (described above), the Galaxy Zoo labels on which Zoobot was trained include bar and spiral arm measures but do not specify if those spirals are irregular or clumpy. Similarly, class A shows 'shredded' spiral galaxies which have been mistakenly recorded by the DESI pipeline as several sources - a distinction not present in the Galaxy Zoo labels.

To prove that the representation does indeed group galaxies independently of the training labels, we run our search method on our sample prior to removing galaxy images (Sec. 2). The response 'Artifact' is a leaf in the Galaxy Zoo decision tree and so the labels include no further information on which type of artifact an image might be. Nonetheless, our search method reveals (Fig 2) that Zoobot's representation is arranging artifacts by visual appearance. This directly demonstrates the internal emergence of classes beyond those needed for predicting training labels.

5 Discussion

Quantitative comparisons with existing methods are not possible because, to our knowledge, there are no previous works attempting to extract rare galaxy subclasses from a deep representations. Further, there is no commonly-accepted definition for what might constitute anomalies or anomaly subclasses. Standard benchmarks for this task (or for similarity metrics in general) would help advance the field.

Our representation is broadly resistant to clustering. Most galaxies (95%+) cannot be confidently assigned to clusters, even with generous HDBSCAN mutual reachability distance settings. Adding a loss term to encourage preserving density ('densmap') further smooths the compressed representation. Whether this limited clustering is a consequence of the nature of supervised representations (and perhaps self-supervised representations e.g. [24]) or induced by our compression process remains to be seen. Altering the training process (e.g. with a hybrid self/supervised approach, [25]) or the compression process might lead to more easily clustered representations. Alternatively, a continuous user-guided search algorithm may work better.

6 Conclusion

We have proven that the representation learned by the Zoobot pretrained galaxy models groups visually similar galaxies together - even where those patterns are definitively independent of the

---

[2]A 15D latin hypercube of 630k galaxies would have approx. 2.4 galaxies per dimension.
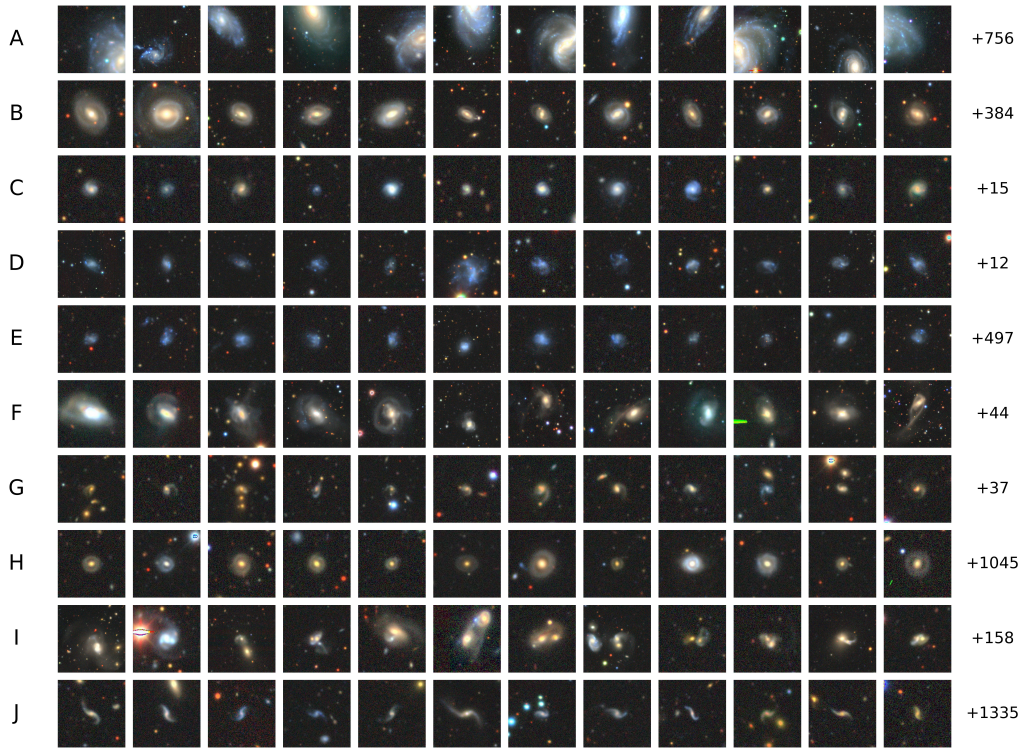
Figure 1: Selected galaxy classes identified within Zoobot's representation (EffNetB0 architecture). Notable examples include low-mass barred (D) and unbarred (E) galaxies, galaxies with clear low surface brightness structure (F), one-armed spirals (G) and merging pairs (I). Class A is reveals a DESI pipeline issue where extended clumpy spirals are 'shredded' into many sources.
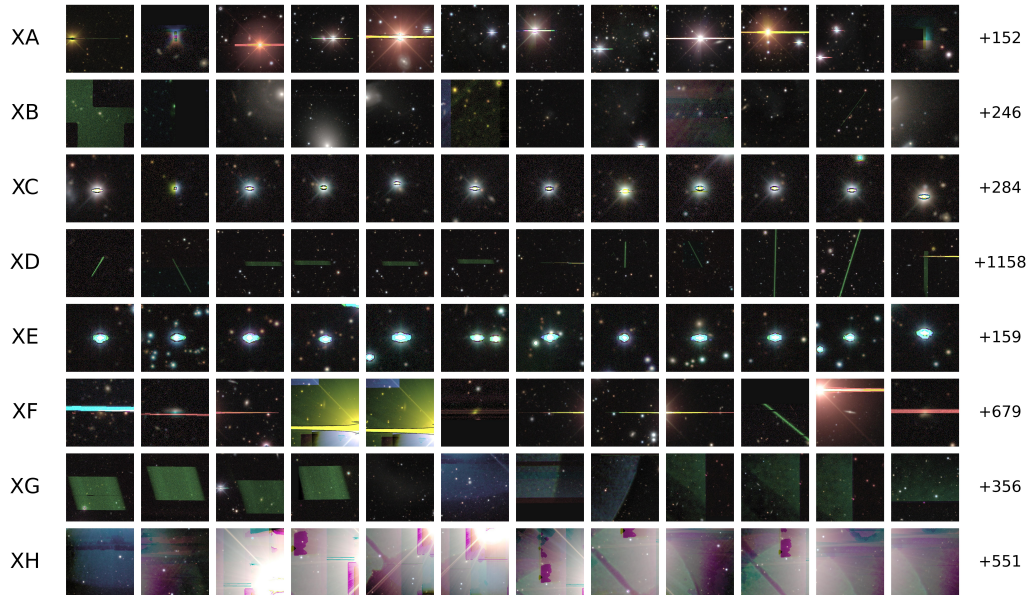

Figure 2: Selected artifact classes identified within Zoobot's representation (MaxViT architecture). Zoobot's representation places artifacts into visually distinct classes, even though Zoobot's pretraining labels do not distinguish between different types of artifact.

4

pretraining labels. Further, we have shown that this grouping of similar galaxies leads to patterns (overdensities) which can be exploited to identify new subclasses. We applied UMAP and tuned HDBSCAN clustering to extract these patterns. Our clustering reveals groups of galaxies with rare and distinctive morphologies. Some of these groups have not previously been automatically identified or studied at scale; for example, low-mass barred galaxies with clumpy irregular spiral structure. We hope this simple search technique will help astronomers discover new unexpected populations of galaxies that shed light on the processes driving galaxy formation.
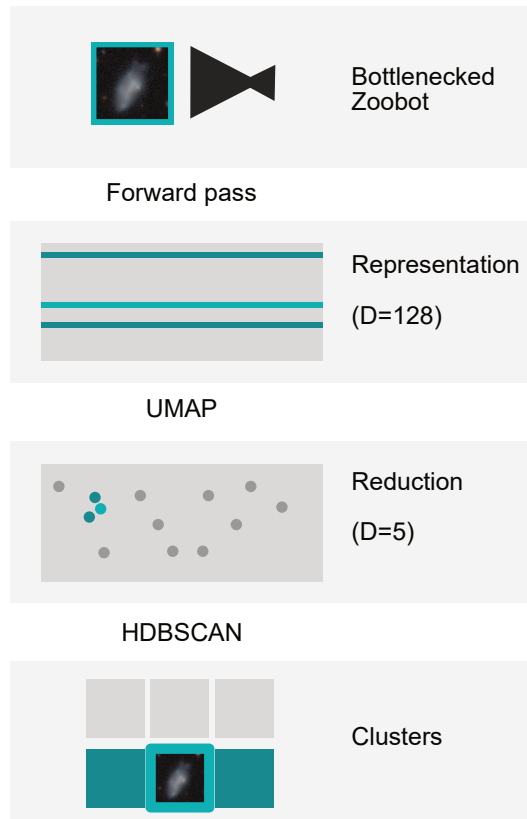

Figure 3: Schematic of search approach. Galaxy images are passed through pretrained Zoobot model, modified to use a D=128 feature bottleneck. Resulting representation is compressed via UMAP then clustered with HDBSCAN.

References

[1] Brooke Simmons, Chris Lintott, Kevin Schawinski, Edward C. Moran, Anna Han, Sugata Kaviraj, Karen L. Masters, Megan M. Urry, Kyle W. Willett, Steven P. Bamford, and Robert C. Nichol. Galaxy zoo: Bulgeless galaxies with growing black holes. *Monthly Notices of the Royal Astronomical Society*, 429(3):2199–2211, 3 2013. ISSN 13652966. doi: 10.1093/mnras/sts491. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/sts491.

[2] R. Grouchy and R. Buta. A photometric and spectroscopic study of non-barred ringed galaxies. *Proceedings of the International Astronomical Union*, 2(S235):100, 8 2006. ISSN 17439213. doi: 10.1017/S1743921306005357.

[3] Alex Fitts, Michael Boylan-Kolchin, James S. Bullock, Daniel R. Weisz, Kareem El-Badry, Coral Wheeler, Claude André Faucher-Giguère, Eliot Quataert, Philip F. Hopkins, Dušan Kereš, Andrew Wetzel, and Christopher C. Hayward. No assembly required: Mergers are mostly irrelevant for the growth of low-mass dwarf galaxies. *Monthly Notices of the Royal Astronomical Society*, 479(1):319–331, 9 2018. ISSN 13652966. doi: 10.1093/mnras/sty1488. URL https://academic.oup.com/mnras/article/479/1/319/5033708.

[4] G. Martin, S. Kaviraj, C. Laigle, J. E.G. Devriendt, R. A. Jackson, S. Peirani, Y. Dubois, C. Pichon, and A. Slyz. The formation and evolution of low-surface-brightness galaxies. *Monthly Notices of the Royal Astronomical Society*, 485(1):796–818, 2 2019. ISSN 13652966. doi: 10.1093/mnras/stz356. URL `http://arxiv.org/abs/1902.04580`.

[5] Conor McPartland, Harald Ebeling, Elke Roediger, and Kelly Blumenthal. Jellyfish: The origin and distribution of extreme ram-pressure stripping events in massive galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 455(3):2994–3008, 1 2016. ISSN 13652966. doi: 10.1093/mnras/stv2508. URL `https://academic.oup.com/mnras/article/455/3/2994/2892617`.

[6] Callum Bellhouse, Bianca Poggianti, Alessia Moretti, Benedetta Vulcani, Ariel Werle, Marco Gullieuszik, Mario Radovich, Yara Jaffé, Jacopo Fritz, Alessandro Ignesti, Cecilia Bacchini, Neven Tomičić, Johan Richard, and Geneviève Soucail. Locations and Morphologies of Jellyfish Galaxies in A2744 and A370. *The Astrophysical Journal*, 937(1):18, 9 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/AC8B6E. URL `https://iopscience.iop.org/article/10.3847/1538-4357/ac8b6ehttps://iopscience.iop.org/article/10.3847/1538-4357/ac8b6e/meta`.

[7] Mike Walmsley, Campbell Allen, Ben Aussel, Micah Bowles, Kasia Gregorowicz, Inigo Val Slijepcevic, Chris J Lintott, Anna M m. Scaife, Maja Jabłońska, Kosio Karchev, Denise Lanzieri, Devina Mohan, David O'Ryan, Bharath Saiguhan, Crisel Suárez, Nicolás Guerra-Varas, and Renuka Velu. Zoobot: Adaptable Deep Learning Models for Galaxy Morphology. *Journal of Open Source Software*, 8(85):5312, 2023. doi: 10.21105/joss.05312. URL `https://doi.org/10.21105/joss.05312`.

[8] Karen L. Masters. Twelve years of Galaxy Zoo. In *Proceedings of the International Astronomical Union*, volume 14, pages 205–212, 10 2019. doi: 10.1017/S1743921319008615. URL `http://arxiv.org/abs/1910.08177`.

[9] Mike Walmsley, Anna M.M. Scaife, Chris Lintott, Michelle Lochner, Verlon Etsebeth, Tobias Géron, Hugh Dickinson, Lucy Fortson, Sandor Kruk, Karen L. Masters, Kameswara Bharadwaj Mantha, and Brooke D. Simmons. Practical galaxy morphology tools from deep supervised representation learning. *Monthly Notices of the Royal Astronomical Society*, 513(2):1581–1599, 5 2022. ISSN 13652966. doi: 10.1093/mnras/stac525. URL `https://academic.oup.com/mnras/article/513/2/1581/6539338`.

[10] Marc Henrion, Daniel J. Mortlock, David J. Hand, and Axel Gandy. Classification and Anomaly Detection for Astronomical Survey Data. In *Springer Series in Astrostatistics*, volume 1, pages 149–184. Springer, New York, NY, 2013. ISBN 9781461435082. doi: 10.1007/978-1-4614-3508-2{_}8. URL `https://link.springer.com/chapter/10.1007/978-1-4614-3508-2_8`.

[11] Daniel Giles and Lucianne Walkowicz. Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection. *Monthly Notices of the Royal Astronomical Society*, 484(1):834–849, 3 2019. ISSN 0035-8711. doi: 10.1093/mnras/sty3461. URL `https://academic.oup.com/mnras/article/484/1/834/5256651`.

[12] Kate Storey-Fisher, Marc Huertas-Company, Nesar Ramachandra, Francois Lanusse, Alexie Leauthaud, Yifei Luo, Song Huang, and J. Xavier Prochaska. Anomaly detection in Hyper Suprime-Cam galaxy images with generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 508(2):2946–2963, 5 2021. ISSN 13652966. doi: 10.1093/mnras/stab2589. URL `http://arxiv.org/abs/2105.02434`.

[13] Allan Sandage. Classification and Stellar Content of Galaxies Obtained from Direct Photography. *Galaxies and the Universe. Edited by Allan Sandage, Mary Sandage, and Jerome Kristian, with an index prepared by Gustav A. Tammann. Published by the University of Chicago Press (Stars and Stellar Systems. Volume 9), Chicago, IL USA., 1975, p.1*, page 1, 1 1975. URL `https://ui.adsabs.harvard.edu/abs/1975gaun.book....1S/abstract`.

[14] B. C. Whitmore, Whitmore, and B. C. An objective classification system for spiral galaxies. I. The two dominant dimensions. *ApJ*, 278:61–80, 3 1984. ISSN 0004-637X. doi: 10.1086/161768. URL `https://ui.adsabs.harvard.edu/abs/1984ApJ...278...61W/abstract`.

[15] Alex Hocking, James E. Geach, Yi Sun, and Neil Davey. An automatic taxonomy of galaxy morphology using unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society*, 473(1):1108–1129, 2018. ISSN 13652966. doi: 10.1093/mnras/stx2351.

[16] Ting Yun Cheng, Marc Huertas-Company, Christopher J. Conselice, Alfonso Aragón-Salamanca, Brant E. Robertson, and Nesar Ramachandra. Beyond the hubble sequence - Exploring galaxy morphology with unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society*, 503(3):4446, 2021. ISSN 13652966. doi: 10.1093/mnras/stab734. URL http://arxiv.org/abs/2009.11932.

[17] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th International Conference on Machine Learning, ICML 2019*, pages 10691–10700, Long Beach, CA, 5 2019. Proceedings of Machine Learning Research. ISBN 9781510886988. URL http://arxiv.org/abs/1905.11946.

[18] Ross Wightman. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models, 2019.

[19] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-Axis Vision Transformer. In *European Conference of Computer Vision (ECCV)*, 4 2022. doi: 10.48550/arxiv.2204.01697. URL https://arxiv.org/abs/2204.01697v4.

[20] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, Martin Landriau, Michael Levi, Ian McGreer, Aaron Meisner, Adam D. Myers, John Moustakas, Peter Nugent, Anna Patej, Edward F. Schlafly, Alistair R. Walker, Francisco Valdes, Benjamin A. Weaver, Christophe Yèche, Hu Zou, Xu Zhou, Behzad Abareshi, T. M. C. Abbott, Bela Abolfathi, C. Aguilera, Shadab Alam, Lori Allen, A. Alvarez, James Annis, Behzad Ansarinejad, Marie Aubert, Jacqueline Beechert, Eric F. Bell, Segev Y. BenZvi, Florian Beutler, Richard M. Bielby, Adam S. Bolton, César Briceño, Elizabeth J. Buckley-Geer, Karen Butler, Annalisa Calamida, Raymond G. Carlberg, Paul Carter, Ricard Casas, Francisco J. Castander, Yumi Choi, Johan Comparat, Elena Cukanovaite, Timothée Delubac, Kaitlin DeVries, Sharmila Dey, Govinda Dhungana, Mark Dickinson, Zhejie Ding, John B. Donaldson, Yutong Duan, Christopher J. Duckworth, Sarah Eftekharzadeh, Daniel J. Eisenstein, Thomas Etourneau, Parker A. Fagrelius, Jay Farihi, Mike Fitzpatrick, Andreu Font-Ribera, Leah Fulmer, Boris T. Gänsicke, Enrique Gaztanaga, Koshy George, David W. Gerdes, Satya Gontcho A Gontcho, Claudio Gorgoni, Gregory Green, Julien Guy, Diane Harmer, M. Hernandez, Klaus Honscheid, Lijuan (Wendy) Huang, David J. James, Buell T. Jannuzi, Linhua Jiang, Richard Joyce, Armin Karcher, Sonia Karkar, Robert Kehoe, Jean-Paul Kneib, Andrea Kueter-Young, Ting-Wen Lan, Tod R. Lauer, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Michael Lesser, Laurence Perreault Levasseur, Ting S. Li, Justin L. Mann, Robert Marshall, C. E. Martínez-Vázquez, Paul Martini, Hélion du Mas des Bourboux, Sean McManus, Tobias Gabriel Meier, Brice Ménard, Nigel Metcalfe, Andrea Muñoz-Gutiérrez, Joan Najita, Kevin Napier, Gautham Narayan, Jeffrey A. Newman, Jundan Nie, Brian Nord, Dara J. Norman, Knut A. G. Olsen, Anthony Paat, Nathalie Palanque-Delabrouille, Xiyan Peng, Claire L. Poppett, Megan R. Poremba, Abhishek Prakash, David Rabinowitz, Anand Raichoor, Mehdi Rezaie, A. N. Robertson, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Gregory Rudnick, Sasha Safonova, Abhijit Saha, F. Javier Sánchez, Elodie Savary, Heidi Schweiker, Adam Scott, Hee-Jong Seo, Huanyuan Shan, David R. Silva, Zachary Slepian, Christian Soto, David Sprayberry, Ryan Staten, Coley M. Stillman, Robert J. Stupak, David L. Summers, Suk Sien Tie, H. Tirado, Mariana Vargas-Magaña, A. Katherina Vivas, Risa H. Wechsler, Doug Williams, Jinyi Yang, Qian Yang, Tolga Yapici, Dennis Zaritsky, A. Zenteno, Kai Zhang, Tianmeng Zhang, Rongpu Zhou, and Zhimin Zhou. Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal*, 157(5):168, 4 2019. ISSN 00046256. doi: 10.3847/1538-3881/ab089d. URL http://arxiv.org/abs/1804.08657.

[21] Mike Walmsley, Tobias Géron, Sandor Kruk, Anna M M Scaife, Chris Lintott, Karen L Masters, James M Dawson, Hugh Dickinson, Lucy Fortson, Izzy L Garland, Kameswara Mantha, David O', Ryan 10, Jürgen Popp, Brooke Simmons, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DESI: Detailed Morphology Measurements for 8.7M Galaxies in the DESI Legacy Imaging Surveys. *MNRAS*, 000:1–18, 9 2023. URL https://arxiv.org/abs/2309.11425v1.

7

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, 2 2018. ISSN 2475-9066. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[23] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017. ISSN 2475-9066. doi: 10.21105/joss.00205.

[24] George Stein, Jacqueline Blaum, Peter Harrington, Tomislav Medan, and Zarija Lukić. Mining for Strong Gravitational Lenses with Self-supervised Learning. *The Astrophysical Journal*, 932 (2):107, 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/ac6d63. URL http://arxiv.org/abs/2110.00023.

[25] Mike Walmsley, Inigo Val Slijepcevic, Micah Bowles, and Anna M. M. Scaife. Towards Galaxy Foundation Models with Hybrid Contrastive Learning. *Machine Learning for Astrophysics Workshop at the Thirty Ninth International Conference on Machine Learning (ICML 2022)*, 6 2022. doi: 10.48550/arxiv.2206.11927.