Publications of the Astronomical Society of Australia (2023), **40**, e001, 53 pages
doi:10.1017/pasa.2022.55

Dawes Review

The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys

M. Huertas-Company[1,2,3] and F. Lanusse[4]

[1]Instituto de Astrofísica de Canarias, c/ Vía Láctea sn, 38025 La Laguna, Spain, [2]Universidad de La Laguna. Avda. Astrofísico Fco. Sanchez, La Laguna, Tenerife, Spain, [3]LERMA, Observatoire de Paris, CNRS, PSL, Université Paris-Cité, Paris, France and [4]AIM, CEA, CNRS, Université Paris-Saclay, Université Paris-Cité, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France

Abstract

The amount and complexity of data delivered by modern galaxy surveys has been steadily increasing over the past years. New facilities will soon provide imaging and spectra of hundreds of millions of galaxies. Extracting coherent scientific information from these large and multi-modal data sets remains an open issue for the community and data-driven approaches such as deep learning have rapidly emerged as a potentially powerful solution to some long lasting challenges. This enthusiasm is reflected in an unprecedented exponential growth of publications using neural networks, which have gone from a handful of works in 2015 to an average of one paper per week in 2021 in the area of galaxy surveys. Half a decade after the first published work in astronomy mentioning deep learning, and shortly before new big data sets such as Euclid and LSST start becoming available, we believe it is timely to review what has been the real impact of this new technology in the field and its potential to solve key challenges raised by the size and complexity of the new datasets. The purpose of this review is thus two-fold. We first aim at summarising, in a common document, the main applications of deep learning for galaxy surveys that have emerged so far. We then extract the major achievements and lessons learned and highlight key open questions and limitations, which in our opinion, will require particular attention in the coming years. Overall, state-of-the-art deep learning methods are rapidly adopted by the astronomical community, reflecting a democratisation of these methods. This review shows that the majority of works using deep learning up to date are oriented to computer vision tasks (e.g. classification, segmentation). This is also the domain of application where deep learning has brought the most important breakthroughs so far. However, we also report that the applications are becoming more diverse and deep learning is used for estimating galaxy properties, identifying outliers or constraining the cosmological model. Most of these works remain at the exploratory level though which could partially explain the limited impact in terms of citations. Some common challenges will most likely need to be addressed before moving to the next phase of massive deployment of deep learning in the processing of future surveys; for example, uncertainty quantification, interpretability, data labelling and domain shift issues from training with simulations, which constitutes a common practice in astronomy.

Keywords: methods: data analysis – cosmology: observations – cosmology: theory – galaxies: evolution – galaxies: formation

(Received 18 March 2022; revised 13 October 2022; accepted 17 October 2022)

1. Introduction

Most fields in astronomy are rapidly changing. Unprecedentedly large observational data exists or will soon become available. Modern spectro-photometric surveys such as the Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) or Euclid (Laureijs et al. 2011) will provide high quality spectra and images for hundreds of millions of galaxies. Integral field spectroscopic surveys at low and high redshift are reaching statistically relevant sizes (e.g. MaNGA—Bundy et al. 2015) enabling to resolve the internal structure of galaxies beyond integrated properties. In addition, new facilities like the James Webb Space Telescope (JWST) are opening the window to a completely new redshift and stellar mass regime both in imaging and spectroscopy and we will be able to witness the emergence of the first galaxies in the universe. X-ray and radio facilities (e.g. SKA, Athena) will probe cold and hot gas in galaxies with improved resolution. On the theory side, computing power has evolved to the extent that we can now generate realistic simulations of galaxies in a cosmological context spanning most of the Universe's history (e.g. TNG—Pillepich et al. 2018) which properly reproduce a large number of observable properties. In this context of growing complexity and rapid increase of data volumes, it has become a new challenge for the community to combine and accurately extract scientifically relevant information from these datasets.

Although Machine Learning applications to astronomy exist since at least thirty years ago (see Section 2), the past years have witnessed an unprecedented increase of deep learning methods translated on an exponential increase of publications (Figure 1). This *revival* is fuelled by significant breakthroughs in the field of Machine Learning since the popularisation of Convolutional Neural Networks (CNNs) a decade ago (Krizhevsky, Sutskever, & Hinton 2012). The first published work mentioning deep learning in astronomy is from 2015 in which CNNs were applied for the classification of galaxy morphology. Since then, the number of works using deep learning in astrophysics has been growing

Corresponding author: M. Huertas-Company, Email: mhuertas@iac.es

Cite this article: Huertas-Company M and Lanusse F. (2023) The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys. *Publications of the Astronomical Society of Australia* **40**, e001, 1–53. https://doi.org/10.1017/pasa.2022.55

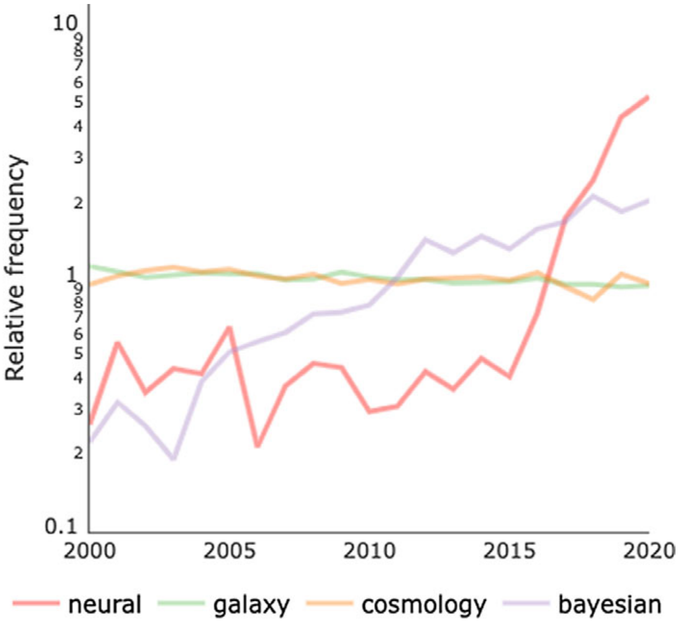

Figure 1. Relative change of the number of papers on arXiv:astro-ph with different keywords in the abstract as a function of time. The number of works mentioning neural networks in the abstract has experienced an unprecedented growth in the last ~6 yr, significantly steeper than other topic in astrophysics. Source: ArXivSorter.

exponentially, being the fastest growth of other topics in the field (Figure 1). The generalisation of deep learning represents to some extent a change of paradigm in the way we approach data analysis. By using gradient-based optimisation techniques to extract meaningful features directly from the data, we move from an approach based on specific algorithms and features to a fully data-driven one. It has potentially profound implications for astronomy and science in general.

After half a decade of this new wave of applications of deep learning to astrophysics, we thus think it is timely to look back at the impact that this new technology has had in our field. Given the large amount of existing publications, we will only review works focusing on the analysis of galaxy surveys and we will restrict to recent works after the so-called deep learning boom. We believe however that most of the lessons learned can be extrapolated to other areas of astrophysics.

It is not the goal of this review to provide technical details about how deep learning techniques work, but to describe its applications to cosmology and galaxy formation in a unique reference document. Over the past years, deep learning has been used for a large variety of tasks such as classification, object detection, but also to derive physical properties of galaxies such as photometric redshifts, to identify anomalous objects or to accelerate simulations and constrain cosmology among many others. The review is thus structured following major areas of application. We provide a quick description and some keywords about the technical solutions adopted for each science case but we emphasise it is not the goal of this work to focus on technical aspects. The reader is encouraged to read the publications referring to each method—which are provided in a best effort basis—to obtain a complete and formal description of the deep learning methods. For completeness and easy access, we provide in the Appendix A a list of different methods mentioned in the review with the corresponding references. A list of current and future galaxy surveys which are referred in this work is also provided in Appendix A.

Following this idea, we have divided the applications of deep learning into four major categories:

- Deep learning for general computer vision tasks. These are applications that we consider closest to standard computer vision applications for natural images for which deep learning has been shown to generally outperform other traditional approaches. It typically includes classification and segmentation tasks.
- Deep learning to derive physical properties of galaxies (both posteriors and point estimates). These are applications in which deep learning is used to estimate galaxy properties such as photometric redshifts or stellar populations properties. Neural Networks are typically used to replace existing algorithms with a faster and more efficient solution, hence more suited for large data volumes. In addition, we also review applications in which deep learning is employed to derive properties of galaxies which are not directly accessible with known observables, that is, to find new relations between observable quantities and physical properties of galaxies from simulations.
- Deep learning for assisted discovery. Neural networks are used here for data exploration and visualisation of complex datasets in lower dimension. We include also in this category, efforts to automatically identify potentially interesting new objects, that is, anomalies or outliers.
- Deep learning for cosmology. Cosmological simulations including baryonic physics are computationally expensive. Deep learning can be used as a fast emulator of the galaxy-halo connection by populating dark matter halos. In addition, a second major application is cosmological inference. Cosmological models are traditionally constrained using summary statistics (e.g. 2 point statistics). Deep learning has been used to bypass these summary statistics and constrain models using all available data.

This is of course a subjective division of the applications of deep learning to cosmology and galaxy formation. There necessarily exist overlaps between the different categories. The review is organised such that, for each family of applications, we review the state-of-the-art and key publications, highlight where the limitations are and what could be the most promising research lines for the future (Sections 3–6). In the final sections (Section 7), we assess the impact of deep learning for galaxy surveys and extract some global lessons learned. We have tried to provide a fair and complete description of the different works. However, as previously stated, the field has exploded in the last years and it has become more and more difficult to keep track of all new publications. This is partly one of the motivations for this review. It is also implies that we might easily miss some relevant works. We apologise in advance.

2. A very brief historical overview—or what we are not covering in this review

Before we start discussing the most relevant results and applications, we would like to clarify that this review focuses on recent applications of deep learning, essentially after the first applications of CNNs to astronomy. As described in the introduction section, we consider as *deep learning* all recent developments around

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

neural networks which have arisen in the last decade approximately, since the first practical application of convolutional neural networks for image classification. Deep learning generally designates gradient-based optimisation techniques of modular architectures of varying complexity; it is therefore a sub field of the more general machine learning discipline. There is a long history of machine learning applications in astronomy which started since well before the more recent deep learning boom. Different types of machine learning algorithms including early Artificial Neural Networks (ANNs), Decision Trees (DTs), Random Forests (RFs) or kernel algorithms such as Support Vector Machines (SVMs) have been applied to different areas of astrophysics since the second half of the past century. For example ANNs, decision trees and Self-Organising Maps (SOMs) have been extensively applied to the classification of stars and galaxies (e.g. Odewahn et al. 1992; Weir, Fayyad, & Djorgovski 1995; Miller & Coe 1996; Bazell & Peng 1998; Andreon et al. 2000; Qin et al. 2003; Ball et al. 2006); an ANN being the primary way of identifying point sources in the commonly used SExtractor software (Bertin & Arnouts 1996) for segmentation of astronomical images. The problem of galaxy morphology classification has also been subject to a significant amount of machine learning related works led in particular by the group of O. Lahav and collaborators using ANNs and DTs (e.g. Storrie-Lombardi et al. 1992; Lahav et al. 1995, 1996; Odewahn et al. 1996; Naim et al. 1997; Madgwick 2003; Cohen et al. 2003). Ball et al. (2004) is likely the first work to use ANNs to classify galaxies in the SDSS. In the first decade of the present century SVMs became more popular and were also used to provide catalogs of galaxy morphology (e.g. Huertas-Company et al. 2008, 2011). Decision Trees have also been applied to other classification tasks such as AGN/galaxy separation (e.g. White et al. 2000; Gao, Zhang, & Zhao 2008). Beyond classification, machine learning, and especially ANNs have been extensively applied to the problem of estimating photometric redshifts (e.g. D'Abrusco et al. 2007; Li et al. 2007; Banerji et al. 2008). This review will not describe these works though. We refer the reader to Ball & Brunner (2010) and Baron (2019) for a complete and extensive review of *pre-deep learning* machine learning techniques applied to astronomy. This obviously does not mean that other machine learning approaches are less interesting for astrophysics. There have been recently very relevant applications of RFs for example for anomaly detection (see the works by Baron & Poznanski 2017) and to assess the main causes of star formation quenching in galaxies (e.g. Bluck et al. 2022) among many others. However, we have made the choice not to include a detailed description of these works in this review. We will focus essentially on how deep learning has changed the landscape in the past half decade.

What is different with deep learning and why a dedicated review? In many aspects, deep learning represents a change in the way we approach data analysis. Because, we now have access to large datasets and the computing resources are powerful enough—especially thanks to Graphic Processing Units—we can move from an algorithmic-centred approach relying on manually engineered features to a fully data-driven unsupervised feature learning approach. This implies that instead of developing advanced domain specific algorithms for each task, we rely on a generic optimisation algorithm to extract the most meaningful features in an end-to-end training loop. This is a new approach to data in astrophysics and in science in general. This change of paradigm has enabled in fact tremendous progress in the computer vision community, especially for image classification, but also for many other tasks such as translation, speech recognition or image segmentation over the past ten years. The purpose of this review is therefore to assess what has been the impact so far of this new approach for data processing in the fields of galaxy formation and cosmology.

3. Deep learning for computer vision tasks in astronomy

We begin by reviewing applications close to standard computer vision problems, for which deep learning approaches have been demonstrated to be very efficient. We focus on classification and source detection.

3.1. Classification

Source classification is a basic first order processing step in most deep surveys, for which deep learning has had a noticeable impact in the past years. The rapid penetration of deep learning can be naturally explained because it is arguably the most straightforward out-of-the box application. Deep learning started in fact to attract the attention of the computer vision community, when convolutional neural networks first won the ImageNet contest of image classification (Krizhevsky et al. 2012).

One of the first tasks scientists do when confronted with a complex problem is to identify objects that look morphologically similar. In extra galactic astronomy, object classification can be of different flavours. For imaging, the most common applications which we review in the remaining of this section, are galaxy morphology classification, star-galaxy separation, and strong lenses detection. We understand this is not an exhaustive list of all image classification applications, however the techniques and approaches used are representative. From the spectroscopic point of view, there have been some works attempting to classify galaxy spectra. However, this remains less common than images. Finally, the classification of transients is something which has been extensively explored over the last years, especially in view of the LSST survey from the Rubin Observatory.

3.1.1. Optical/NIR galaxy morphology

Galaxy morphological classification is a paradigmatic example of a science case where deep learning has rapidly become the state-of-the-art. This task was first done by E. Hubble who classified galaxies in the well-known Hubble sequence. The classification scheme, which is now more than a 100 yr old, establishes that galaxies in the Universe today come essentially in two flavours. On one side, there are disk galaxies, like our own Milky Way; on the other side elliptical like galaxies. We now know that, besides morphology, the broad classes present different physical properties. Understanding the origin of the morphological diversity remains an open issue in the field of galaxy formation.

Therefore, galaxy morphological classification is still performed in almost all extra galactic imaging surveys. The traditional way to estimate galaxy morphology has been through visual inspection. However, this approach became prohibitively time consuming in the last decade with the advent of large extra galactic imaging surveys such as the Sloan Digital Sky Survey (SDSS). Approaches to overcome this limitation came in two fronts essentially. Citizen science approaches, of which the Galaxy Zoo project is the more popular example (Lintott et al. 2008), were developed to classify large samples of galaxies. Automation through Machine Learning has been always on the table as early as 1990s

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

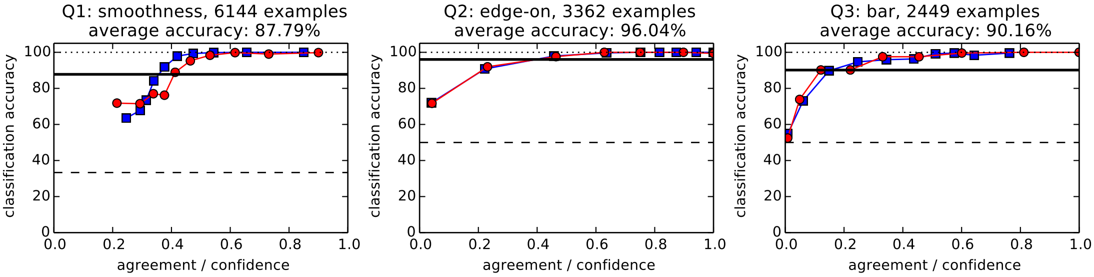

Figure 2. Example of level of agreement (red circles) and model confidence (blue squares) versus classification accuracy. Each panel shows a different question in the Galaxy Zoo classification tree (smoothness, edge-on, bar). The authors quote an unprecedented accuracy of >90%. This is the first work that uses CNNs in astrophysics. The figure is adapted from Dieleman et al. (2015)

(e.g. Spiekermann 1992) and continued in the 2000s (e.g. Huertas-Company et al. 2008). Huertas-Company et al. (2011) was the first to provide the community with a Machine Learning based classification of SDSS galaxies. However the accuracy reached by these early approaches based on manually engineered features remained moderate—especially when dealing with detailed morphological features such as bars or spiral arms—hampering their penetration in the community. The main limitation is that the features typically used by these methods present only weak correlations with detailed morphological structures and are also very dependent on noise and spatial resolution.

In this context, it is not a surprise that the first works using Convolutional Neural Networks in astronomy focus on galaxy morphology (Dieleman, Willett, & Dambre 2015; Huertas-Company et al. 2015). The first one used labelled images from the Galaxy Zoo2 sample and trained a supervised CNN to estimate the morphological properties of SDSS galaxies going from global morphology to more detailed properties such as the number of spiral arms. The work by Dieleman et al. (2015) was the winner of a public challenge on the Kaggle platform.[a] It achieved unprecedented classification accuracy of >90% in most of the tasks (Figure 2).

This is a major improvement compared to previous approaches and marks the beginning of the penetration of deep learning techniques in astrophysics.

Soon after, Huertas-Company et al. (2015) applied a similar architecture to high redshift galaxies observed with the Hubble Space Telescope (HST), demonstrating again a similar improvement as compared to other approaches, including feature-based Machine Learning. Using CNNs to classify galaxy images represents in some sense a change of paradigm in the way we approach the classification problem, analogous to what happened with natural images. Instead of manually trying to identify specific features that correlate with the classes to distinguish, the features are learnt simultaneously with the classification process. This of course entails loosing some interpretability, since the features learned by the network are no longer directly associated with physical properties. Interpretability is a major issue in deep learning applications to natural sciences that we will address in Subsection 7.3.

CNNs as state-of-the-art approach In the past years, the number of works using deep learning to classify galaxies based on their morphology has exploded, and CNNs have been used to classify the morphologies of galaxies in a variety of optical/Near-Infrared surveys (we address the radio domain in Section 3.1.2). It demonstrates that deep learning has fast become the state-of-the-art approach to estimate galaxy morphology in big datasets. Figure 3 illustrates the most common approach used for morphological classification. Images are fed into a series of convolutional layers which extract some summary statistics. The extracted statistics are then fed into a Multi-Layer Perceptron which maps them into a class. The employed loss function is usually a cross-entropy loss. We notice that in the original approach by Dieleman et al. (2015), a series of siamese networks were introduced to add rotational invariance. This approach has not been used in other works. Doménguez Sánchez et al. (2018) revisited the proof-of-concept work by Dieleman et al. (2015) by carefully cleaning the training sets and released the first deep learning catalog of galaxy morphologies in SDSS. Ghosh et al. (2020) explored the classification of distant galaxies from the CANDELS survey based on their bulge-to-total ratios. Goddard & Shamir (2020) applied a similar strategy for Pan-STARRS. Vega-Ferrero et al. (2021) and Cheng et al. (2021b) used CNNs to classify galaxies in the Dark Energy Survey. Bom et al. (2021) followed a similar strategy for the S-PLUS survey and Walmsley et al. (2022) classified galaxies in the DECALS survey. In addition to observations, CNNs have also been extended to classify images from cosmological simulations

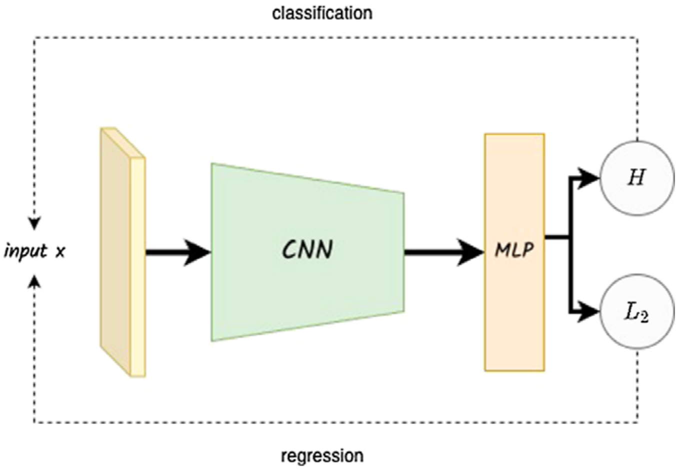

Figure 3. Schematic view of a simple *Vanilla* type Convolutional Neural Network, the most common approach for binary, multi-class classification and regression in extragalactic imaging. The input, which is typically an image is fed to a series of convolutional layers. The resulting embedding is used an input of a Multi-Layer Perceptron which outputs a float or array of floats. If the problem is a classification, the standard loss function is the cross-entropy (H), while if it is a regression the quadratic loss (L_2) is usually employed.

[a] https://www.kaggle.com/.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

in order to assess the realism of galaxy morphologies (Huertas-Company et al. 2019; Varma et al. 2022). In such applications, the CNN is trained on labelled observations and then applied to mock images. In addition to global morphology, Tadaki et al. (2020) used also a similar supervised CNN setting to classify spiral galaxies based on their resolved properties (i.e. type of spiral arms).

3.1.2. Radio galaxy morphology

In addition to the Optical and Near-Infrared domains, the radio astronomy community has also been very active in developing and testing deep learning techniques for the classification of radio galaxies. These efforts are motivated by the forthcoming arrival of new radio facilities such as SKA[b] that will change the landscape by detecting hundreds of thousands of new radio galaxies. Similarly to what happened in the optical, the availability of large datasets with labels has enabled the community to extensively test deep learning for classification. The Radio Galaxy Zoo project (Banfield et al. 2015) used indeed citizen science to determine the host galaxy of the radio emission and the radio morphology of ∼170 000 galaxies. This therefore constitutes an excellent database for applying neural networks. It highlights the importance of the preparatory work done by the community for accelerating the adoption of deep learning techniques.

The first work exploring CNNs for classification of radio galaxies is Aniyan & Thorat (2017). They use a simple sequential CNN (Figure 3) and conclude that an accuracy up to ∼95% can be achieved in classifying galaxies in three main classes—Fanaroff-Riley I (FRI), Fanaroff-Riley II (FRII) and bent-tailed galaxies. A number of works have followed. Alhassan, Taylor, & Vaccari (2018) also reported similar accuracy when using CNNs to classify compact and extended radio sources observed in the FIRST radio survey (see also Maslej-Krešňáková, El Bouchefry, & Butka 2021 for similar conclusions). Lukic et al. (2018) explores different network configurations and concludes that a three layer network is typically enough to reach more than 90% accuracy. Wu et al. (2019) explored Faster Region-based Convolutional Neutral Networks to detect and classify radio sources from the Radio Galaxy Zoo project.

3.1.3. Strong lenses

Deep Learning based supervised classification has been widely extended to other extragalactic classification tasks. An example of application which has significantly benefited from the advent of deep learning is the detection of strong gravitational lenses (e.g. Jacobs et al. 2017; Petrillo et al. 2017; Lanusse et al. 2018; Davies, Serjeant, & Bromley 2019; Schaefer et al. 2018; Metcalf et al. 2019; Petrillo et al. 2019; Jacobs et al. 2019; Li et al. 2020b; Huang et al. 2020). Gravitational lenses produce characteristic distortions of the light of background sources, caused by the presence of a foreground massive galaxy or cluster in the same line of sight. The analysis of strong lenses provides information about the total matter distribution of the foreground system. Strong lenses provide therefore a unique probe of the dark matter distribution in galaxies. The first step consists in identifying the lenses on large samples of galaxy images.

Similarly to what has been described for galaxy morphology, the usual method to identify lenses is through Convolutional Neural Networks as done for galaxy morphology (Figure 3). However, there are some specific issues related to strong lensing detection since the problem is severely unbalanced. The number densities of strong lenses are indeed several orders of magnitude smaller than the ones of regular galaxies. This poses two main problems. First, it is impossible to build a large enough training sample of observed lenses. Second, in order to be scientifically useful, the classifier needs to reach extremely high purity values because even a small contamination from negative examples provokes that the sample of lenses is dominated by false positives. The community has proposed two sorts of solutions to these problems which appear in most of the works. To cope with the lack of training examples, the CNNs are usually trained on simulations. The physics of strong lenses is sufficiently well known so that lenses can be simulated with some degree of realism (see Figure 4). This practice is rather common in astrophysical applications as we will describe in the forthcoming sections and especially in Section 7. It does not come free of biases though. The work by Lanusse et al. (2018) highlights the importance of using realistically complex simulations for training in order to limit the potential biases.

The problem with false positives is in general more difficult to solve. The usual solution consists in visually inspecting the strong lenses candidates to remove the false positives. In that context, deep learning helps to reduce by several orders of magnitude the amount of required visual inspections but does not completely get rid of them.

Despite these issues, deep learning has rapidly become the state-of-the-art technique to find strong lenses in large surveys. It has been successfully applied to multiple imaging surveys following very similar strategies as just described. The first work using CNNs for identifying lenses focused on the CFHTLS survey is Jacobs et al. (2017). Petrillo et al. (2017) and (2019) applied a similar strategy for KIDS galaxies and Jacobs et al. (2019) extended the approach to DES. Huang et al. (2020) focused on the DECALS datasets. Some other works focus more on simulations only in view of preparing future surveys. Lanusse et al. (2018) demonstrated the performance of CNNs to detect lenses in LSST images and Davies et al. (2019) focused on Euclid.

Overall, the conclusions are shared among all different works. Deep learning approaches are shown to improve more traditional techniques and therefore will likely be used on future surveys. In support for this, the work by Metcalf et al. (2019) shows the results of a strong lensing detection challenge in the framework of the Euclid survey where the five best algorithms were based on CNNs.

3.1.4. Open issues

In summary, deep learning techniques have rapidly replaced traditional approaches for classification of astronomical images to the point that it will most likely be the adopted approach to classify galaxies in forthcoming surveys across the electromagnetic spectrum. The main advantages are speed and accuracy. Convolutional Neural Networks in particular have demonstrated to be more accurate for these classification tasks than other feature-based automated methods as seen in other disciplines. Classification is also one of the less risky applications in the sense that it is—in most cases—very close to the application of deep learning in the computer vision community. Some open issues remain still.

Beyond vanilla CNNs Even though standard vanilla CNNs provide in general accurate results, several works have explored more complex configurations commonly used in computer vision such as ResNets (e.g. Zhu et al. 2019; Kalvankar, Pandit, &

[b] https://www.skatelescope.org/the-ska-project/.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

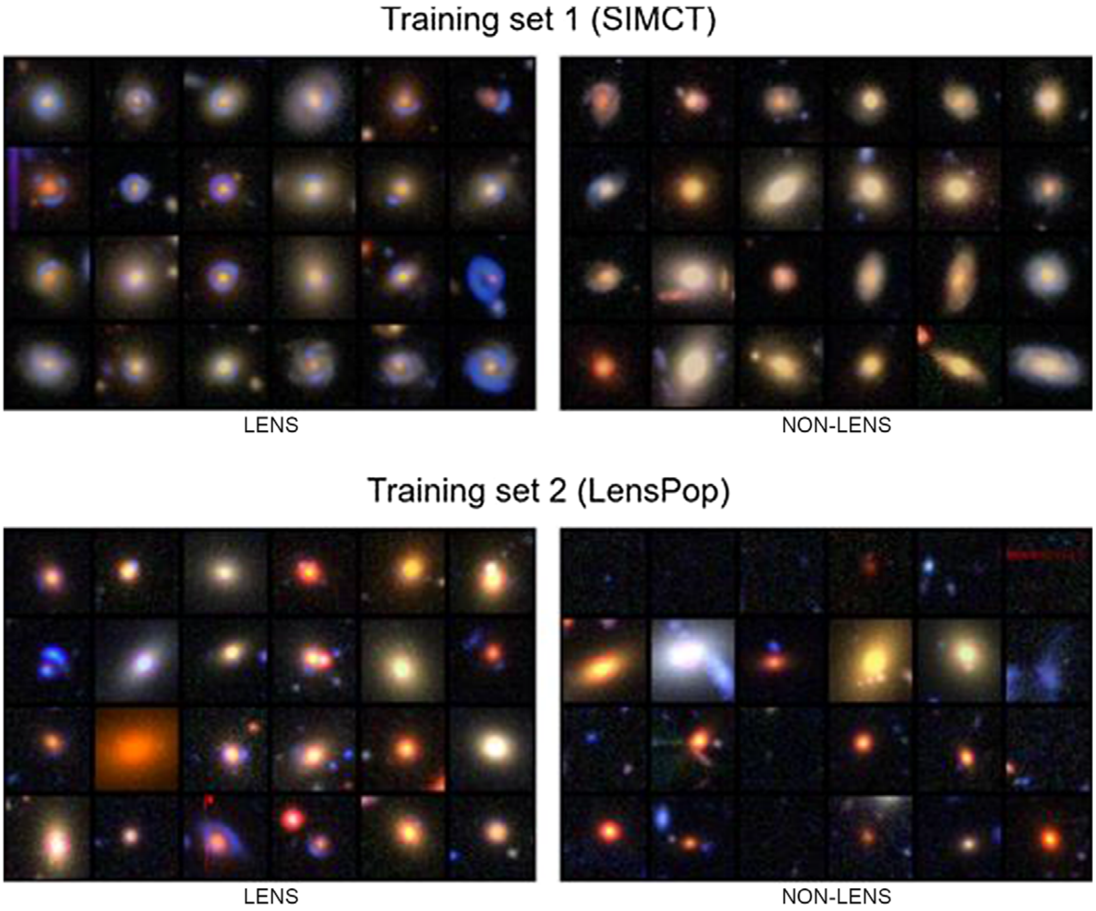

Figure 4. Example of two different simulated samples of strong lenses used for training a CNN. These simulations were used to detect strong lenses in the CFHTLS survey. Figure adapted from Jacobs et al. (2017)

Parwate 2020).Overall, the results are promising but do not significantly improve over approaches based on simpler CNNs. Some works have systematically compared different neural network architectures on the same dataset (e.g. Fielding, Nyirenda, & Vaccari 2021; Cavanagh, Bekki, & Groves 2021). The general conclusion is that, even if there are some differences in accuracy and in efficiency, all architectures fall in the same ballpark. A possible explanation for this is that astronomical images present in general less diversity than natural ones and hence relatively simple CNNs suffice to extract the relevant information. The works by Katebi et al. (2019) and Lukic et al. (2019) are among the few works in astronomy exploring the use of Capsule Networks (Sabour, Frosst, & E Hinton 2017). Capsule Networks are proposed as an alternative to CNNs that incorporates spatial information about the features present in the images. Very briefly, it uses a sort of *inverse rendering* to encode the presence of a given object in an image. It therefore encodes information, not only about the presence or not of a given object—which is what a CNN would do—but also information about where the object is and how it is oriented. We do not provide a detailed representation of the architectures given that this type of approach has only been marginally used. The conclusion of the work by Lukic et al. (2019) is that, in the case of radio galaxy classification, Capsule Networks perform less well than more standard CNNs, reaching only an accuracy of ~75% (Figure 5). One possible explanation for that is that Capsule Networks were initially thought to identify scenes which do not look realistic. For example a CNN would learn to recognise faces based on features like eyes or noses. However, they do not take into account

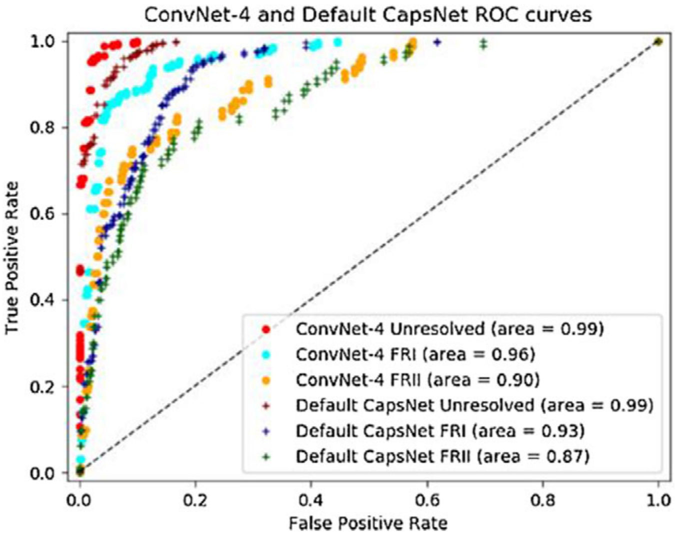

Figure 5. Comparison of Capsule Networks and CNNs applied to classify morphologies of radio galaxies. The ROC curves show the different performances. Capsule Networks do not offer a significant gain in this context. Figure adapted from Lukic et al. (2019)

the position of these features in the image. Capsule Network do, but this is not a common issue for astronomical imaging. This might be one of the reasons why Capsule Networks have not been very used in astronomy so far. Becker et al. (2021) did the exercise of systematically testing the performance of CNNs for radio galaxy classification using multiple performance metrics such as inference time, model complexity, computational complexity, and mean per class accuracy. They report three main types of architectures that perform best but they are all variations of sequential CNNs. In a recent work, Tang et al. (2021) explores the use of multi-branch CNNs to simultaneously learn from multiple survey inputs (NVSS and FIRST). Interestingly they confirm that including multi-domain information allows to reduce the number of miss classifications by ∼40%.

Labelled data A common bottleneck for deep learning based classification is the availability of labelled samples to train the supervised algorithms. In astronomy this is particularly delicate since the data properties (e.g. noise, resolution) change from one dataset to another so in theory the labelling process should be repeated for every new dataset. A number of works have addressed this issue with different approaches. The work by Walmsley et al. (2020) explores Active Learning as a way to reduce the amount of required examples for training. Active learning allows one to select the most informative examples for the model which are the ones showed to human classifiers. The work by Walmsley et al. (2020) is also the first to explore Bayesian deep learning as a way to both estimate uncertainties an also identify the most informative examples (Figure 6). Domínguez Sánchez et al. (2019) and Ghosh et al. (2020) explored transfer learning, which consists on refining the weights of a neural network trained on a similar labelled dataset to reduce the need of large training samples (Figure 7). They showed that the amount of labelled examples can be reduced by a factor of a few with this approach. The issue of the size of the training set that is needed is also investigated by Samudre et al. (2022). The authors explore whether reliable morphological classifications can be obtained with a small sample of 2 000 labelled images. They namely test transfer learning but also few-shot learning techniques based on twin networks. The conclusion is that even with small datasets, reliable classifications can be obtained using CNNs, with an adapted training strategy. The recent work by Walmsley et al. (2021) explores another version of transfer learning. They show that the features learned by the CNNs for a given task can be recycled to estimate other morphological properties. Vega-Ferrero et al. (2021) used instead a simulated training set built from an observational sample from SDSS to classify more distant galaxies from the Dark Energy Survey.

Despite some remaining issues, some of which are common to most of deep learning application—see Section 7 for a more detailed discussion—it is probably safe to argue that the community has accepted that deep learning will be employed to classify galaxies from future surveys such as LSST and Euclid.

3.1.5. Transient astronomy

The field of transient astronomy is about to change dramatically with the arrival of synoptic sky surveys such as the LSST survey by the Rubin Observatory, which will observe large areas of the sky with an unprecedented frequency to find variable and transient astronomical sources. The number of detections per night is expected to easily exceed several thousands. The community has seen in machine learning techniques and particularly in deep learning a promising way of classifying the detected objects and filtering the most potentially (unknown) interesting candidates (see the report by Ishida 2019). We will address the discovery of new types of transients in Section 5. We focus here one the supervised classification of variable sources.

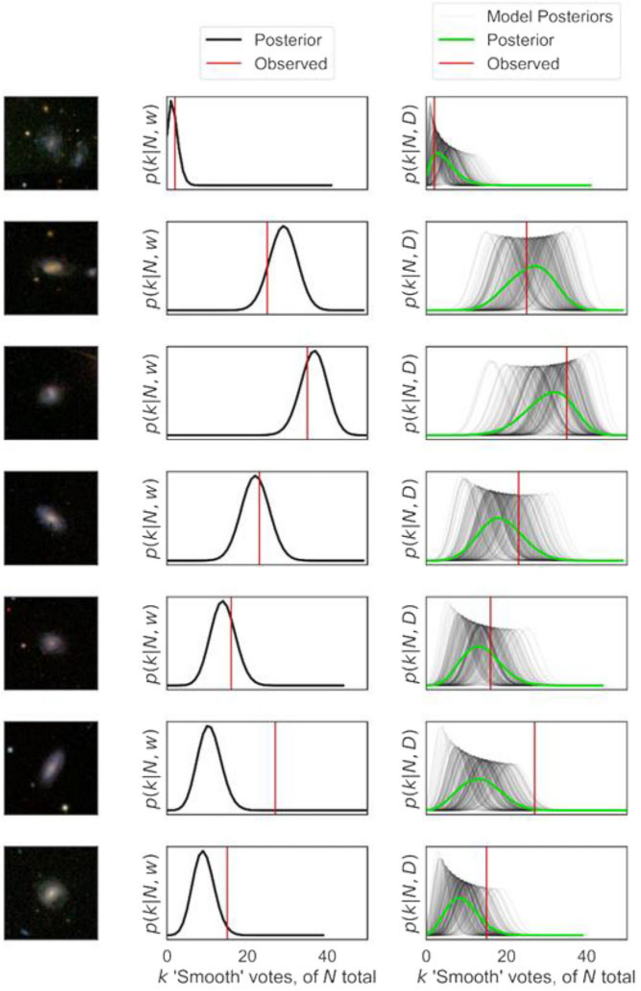

Figure 6. Example of posterior distributions of morphologies estimated from the votes of different classifiers. The leftmost column shows an image of a galaxy. The middle column shows the posterior predicted by a single network (black), while the right column shows the posterior averaged over 30 Monte Carlo dropout approximated networks. The red vertical lines indicate the ground truth value, which generally shows a good overlap with the posterior distribution. Figure adapted from Walmsley et al. (2020)

One key science topic for cosmology is the detection and characterisation of SuperNovae light curves. There are different types of Supernovae and not all are useful for the same purposes. For example, SNIa are used for cosmology. Rapidly identifying the type of object saves—among other things—telescope time. Although this is ideally done with spectroscopy, it is unfeasible to perform a spectroscopic follow-up of all the sources that will be detected. Therefore the community started as early as 2010 to prepare for this data deluge with the creation of simulated datasets such as the Supernovae Photometric Classification Challenge (SPCC) or the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC).[c] Because of these early

[c] https://plasticc.org/.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

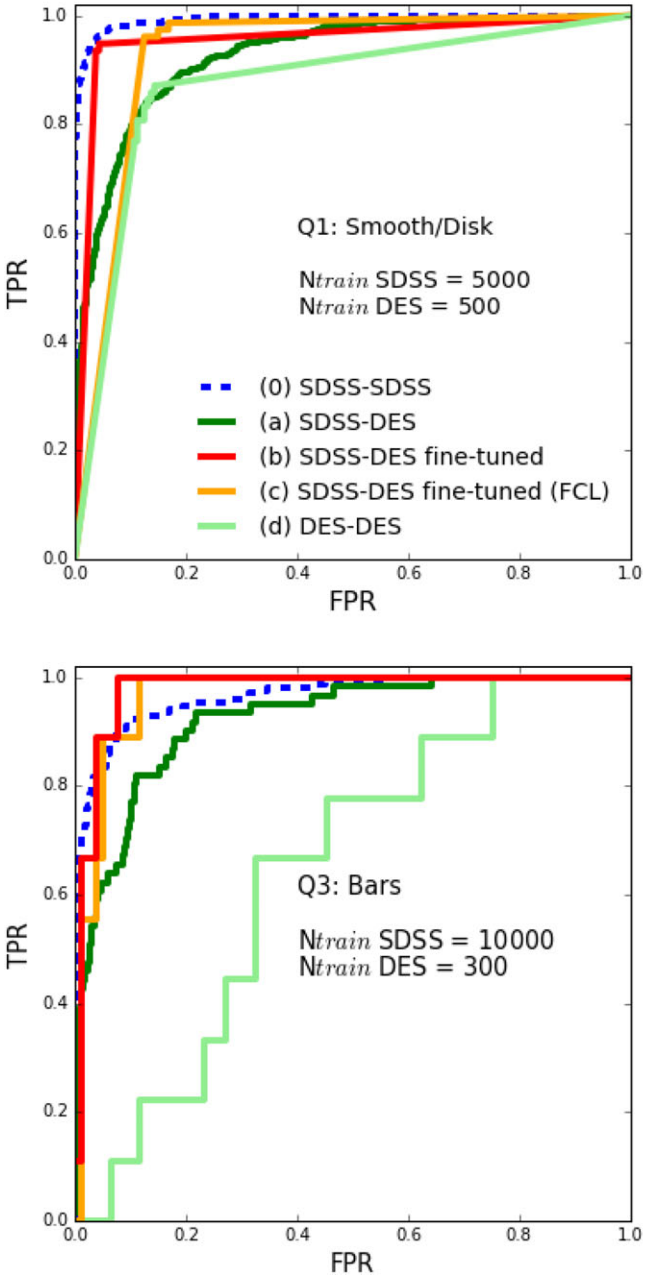

Figure 7. Example of Transfer Learning adapted to galaxy morphological classification exploring how to use a CNN trained on SDSS to the DES survey. Each panel shows a different morphological feature (smooth/disk, bars from top to bottom). The different lines show the ROC curves using different training models and test sets. The work concludes that when using a small subset of DES galaxies to refine the weights of a CNN trained on SDSS, the accuracy of the classification becomes optimal (red solid lines compared to blue dashed lines). Figure adapted from Domínguez Sánchez et al. (2019)

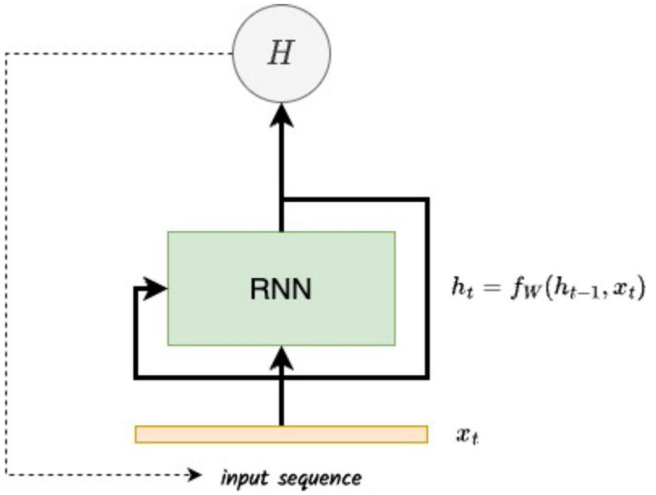

Figure 8. Schematic view of a Recursive Neural Network (RNN) which has been used for classification of light curves in several works. The photometric sequence is fed to a recursive block and trained with a cross-entropy loss. The RNN blocks keep a memory (h_t) of the previous time step which make them suitable for handling time sequences.

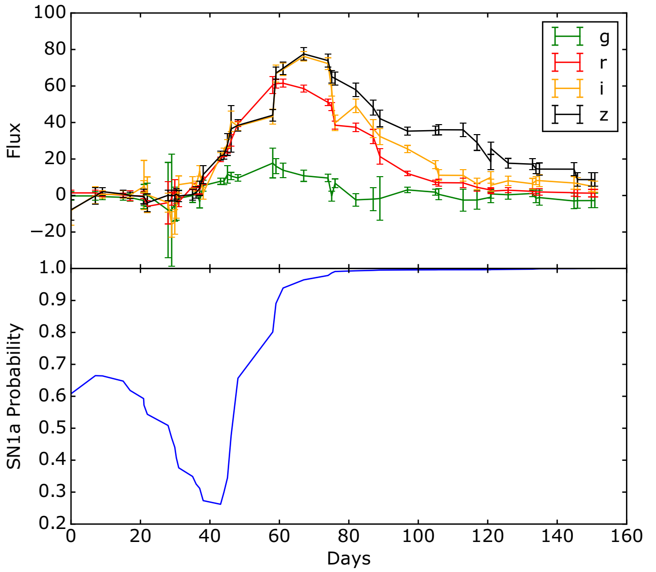

Figure 9. RNNs used for SN photometric light curve classification. The figure shows an example of light curve in five photometric bands (top panel) and the probability of classifications as a function of the time step (bottom panel). After roughly 50 d, the supernova type is classified with high confidence. Figure adapted from Charnock & Moss (2017)

efforts, there exists a consequent literature using *pre-deep learning* machine learning methods (e.g. SVMs, RFs and ANNs) to address the problem of SN light curve classification (e.g. Lochner et al. 2016; Villar et al. 2019; Hosseinzadeh et al. 2020; Vargas dos Santos, Quartin, & Reis 2020).

The first work to use deep learning for SN light curve classification is by Charnock & Moss (2017). They use to that purpose Recurrent Neural Networks (RNNs), which are a type of Neural Network architecture designed to handle sequences of variable length (see Figure 8 for a simple illustration). They are called recurrent because they keep a memory of the previous information in the sequence and use it to make the predictions. They are typically used for language modelling. The authors report an average accuracy above 90% for the classification of light curves in three types—supernovae types I, II and III—on the simulated sample from the Supernovae Photometric Classification Challenge (Figure 9). Despite the small training set of a hundred data points, RNNs achieved state-of-the-art results compared with a combination of template fits and boosted decision trees (Lochner et al. 2016). In addition of not requiring feature engineering, one advantage of RNNs is the ability to classify incomplete light curves. Moss (2018) also explores RNNs on the same simulated dataset.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

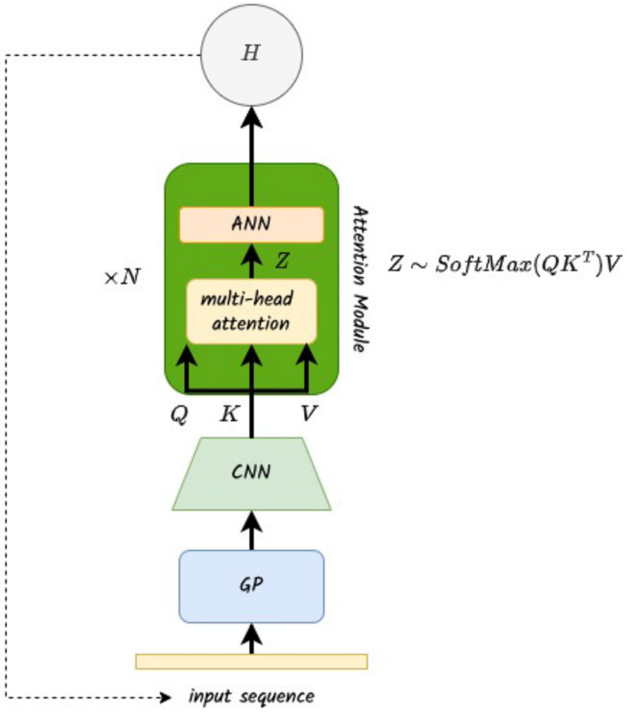

Figure 10. Example of typical transformer architecture for the classification of light curves. A Gaussian Process is first used to homogenise the time sampling and the resulting sequence is fed into a CNN for feature extraction. The attention modules are used to extract the final features which are used to classify.

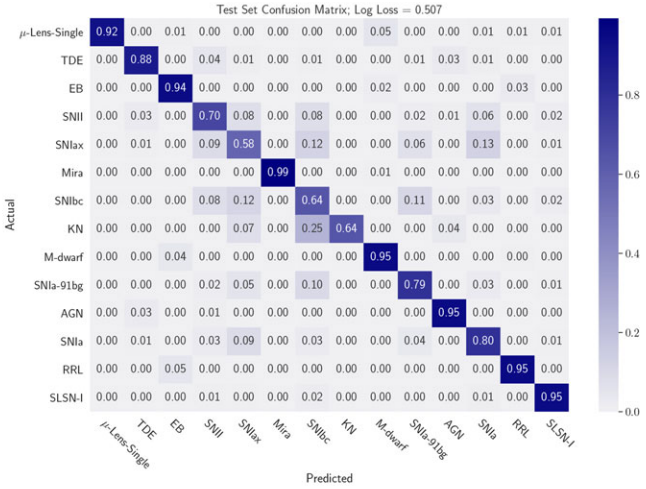

Figure 11. Transformer applied to SN photometric light curve classification. The figure shows a confusion matrix on the PLAsTiCC simulated dataset. Most of the classes are identified with accuracies larger than 90% as seen in the diagonal of the matrix. Figure adapted from Allam & McEwen (2021)

They propose some improvements such as a stronger data augmentation process to mitigate the effects of small samples and reach >95% accuracy. A similar conclusion is reached by Möller & de Boissière (2020) who also explored RNNs. They report a similar accuracy also on realistic simulations and confirm accuracies above 85% for incomplete light curves. See also Burhanudin et al. (2021) for similar conclusions. This latter work proposes handling imbalance with a focal loss function.

The processing of data in the form of sequences has experienced significant breakthroughs in the machine learning community over the past years. In particular, the so-called attention methods which identify the region of the sequences that contain the most relevant information have been demonstrated to be very powerful for sequence to sequence tasks such as translation and for classification of time series (Vaswani et al. 2017). This type of attention based architectures are commonly known as Transformers (see Figure 10). The application of Transformers to astronomy is still rather limited. However some works have already explored their performance for SN light curve classification and other types of transients. Allam & McEwen (2021) use a variation of the original Transformer architecture to classify photometric light curves from the PLAsTiCC simulated dataset. The authors demonstrate they achieve state-of-the-art accuracies. They claim the Transformer is able to deal with very unbalanced classes without need of augmentation, achieving the lowest logarithmic loss to date (Figure 11). However, as the authors emphasise, the comparison with other methods is not straightforward given that they are evaluated under different conditions. This highlights a general problem for the comparison of different works performing classification. Astronomy lacks in general of standardised datasets on which algorithms can be consistently tested (see Section 7 for a general discussion). Pimentel, Estevez, & Forster (2022) is a second work exploring the use of attention mechanisms. They apply their method to real data after a first training on simulations and a fine tuning step. They also conclude that the attention network outperforms classical approaches based on RFs and also RNNs especially for late-classification and early-classification of light curves.

In addition to purely photometric data, the classification of variable sources can also be performed on images. This is traditionally done by subtracting images from different epochs. Several works have explored the use of deep learning to work directly in the image sequences. Carrasco-Davis et al. (2019) used for example Recursive Convolutional Neural Networks (RCNNs) to identify real transients from artefacts. Gómez et al. (2020) extend the use of RCNNs to a multi-class classification of transients. They also use RCNNs to extract information from both temporal and spatial correlations.

In summary, deep learning approaches have naturally incorporated to other ML approaches to classify photometric light curves. In particular RNNs and more recently Transformers provide competitive results. However, for this particular task, deep learning does not seem to have dramatically improved previous techniques as it is the case for image classification. The work by Hložek et al. (2020) summarises the results of the PLAsTiCC challenge organised in the Kaggle Platform. It can clearly be seen that both *classical* and deep learning approaches provide competitive results.

3.1.6. Other classifications

Similar supervised approaches for other classification tasks have been tested over the past years, reaching also similar conclusions and facing similar challenges. Kim & Brunner (2017) used convolutional neural networks for separating stars from galaxies. Star-galaxy separation is a classical task in the analysis of deep surveys. As for other classification tasks, the advantage of CNNs is that they use the pixel-level information and do not rely on summary statistics. The general conclusion however is that CNNs offer a marginal gain over more ML feature-based approaches for this classification problem. Ono et al. (2021) used CNNs to distinguish between real Lyα emitters (LAEs) and contaminants using imaging data in six narrow band filters. Ackermann et al. (2018) made the first tests to classify galaxy mergers trained on Galaxy Zoo and reported significant improvements over state-of-the-art approaches (the case of galaxy mergers is extensively discussed in Subsection 4.2). Walmsley et al. (2019) and Tanoglidis, Ćiprijanović, & Drlica-Wagner (2021b) explored the use of CNNs for the classification of low surface brightness (LSB) structures in deep imaging surveys. The systematic exploration of the low surface brightness universe will be enabled by future surveys such as LSST or Euclid. Automatically identifying and classifying LSB structures is therefore a new challenge. The authors test a CNN model on the Dark Energy Survey data and report a ∼95% accuracy in separating artefacts from real LSB structures. Only a few works have explored this science case, probably due to the lack of proper labelled datasets.

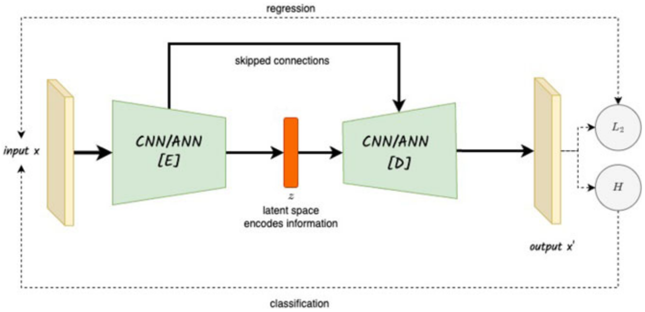

Figure 12. Schematic representation of a Unet type of architecture. This is typically used for addressing image to image problems (e.g. segmentation, galaxy properties in 2D). A first CNN (encoder) encodes information into a latent space (z), which is then fed to another CNN (decoder) which produces an image. One particularity of the Unet is the presence of skipped connections between the encoding and decoding parts which have been demonstrated to provide higher accuracies.

3.2. Segmentation, deblending and pixel-level classification

Object detection and deblending is another problem on which deep learning techniques have been extensively tested in these past years. The detection of sources to build catalogs with some measured properties is a first standard step in the processing of imaging from deep surveys. In image processing this typically fits into the field of image segmentation, which is precisely the task of identifying the positions and boundaries of different objects in an image. The type of segmentation can be semantic, if the objects belong to different classes, or instance if we aim at detecting objects of the same type.

Since the popularisation of CCDs, object detection in astrophysics is generally done through the software called SExtractor, which implements an advanced multi thresholding technique to detect and separate objects. Although it has been extensively used over the past years, the limitations become more obvious with the advent of deeper surveys in which the confusion between sources becomes very common. It is estimated that ∼80% of galaxies will be affected by some sort of overlapping or *blending*. Given that blending can severely affect the scientific conclusions, it is important to have reliable methods for detection and deblending (see Melchior et al. 2021 for a review on deblending).

Over the past years, there has been significant progress in the computer vision community on image segmentation by applying deep learning networks. Therefore, similarly to what happens for classification, deep learning segmentation techniques developed for general purpose computer vision applications are available. An out-of-the box implementation is thus expected to provide reasonable results. Astrophysical data has however some key properties which are not found in other types of images. The dynamic range is very large, typically spanning several orders of magnitude from the centres of the objects to the outskirts. Objects do not have clear edges. This is a fundamental difference with respect to natural imaging applications, which makes the segmentation task with neural networks more challenging.

A first very popular approach for object detection is the use of encoder-decoder networks. Unets (Ronneberger, Fischer, & Brox 2015), which incorporate skipped connections between the encoder and decoder branches, and have emerged as one of the state-of-the-art segmentation networks (see Figure 12). Originally designed for medical imaging, they have been commonly applied to astrophysics for detection over the past years. Boucaud et al. (2020) first applied a Unet to detect objects in image stamps and measure the photometry of overlapping sources. In this proof-of-concept work, it is shown that the measured photometry of pairs of galaxies is improved with respect to the standard SExtractor based approach (Figure 13).

However, this was tested in an idealised setting where the stamps only contained two images with one galaxy at the centre. Paillassa, Bertin, & Bouy (2020) explored a similar type of architecture, in a more realistic setting, to identify and classify artefacts in CCD images. In that case, classification and segmentation are performed simultaneously at the pixel level. The proposed approach successfully identifies multiple types of artefacts in an image (Figure 14).

A similar approach is explored by Hausen & Robertson (2020) combining this time object detection and morphological classification of galaxies at the pixel level. Using a sliding window, the Unet successfully classifies every pixel of the CANDELS survey in five morphological classes (Figure 15). Huertas-Company et al. (2020) used a similar type of architecture to study the resolved properties of galaxies by detecting giant star-forming clumps within high redshift galaxies in the CANDELS survey. Burke et al. (2019), Farias et al. (2020), Tanoglidis et al. (2021a) explored an alternative approach based on region based CNN architectures such as Mask RCNNs to perform similar tasks, that is, detection, deblending and classification.

Overall, these applications all show very promising results and clearly improve on more traditional methods both in terms of speed but also in accuracy, especially when combined with pixel-level classifications. In most cases, the application of out-of-the box architectures is enough to provide accurate results, once the input data are properly rescaled to limit the effect of the dynamic range. However, until now, and with the exception of the work by Hausen & Robertson (2020), the majority of the works have focused more on testing and demonstrating the performance of these deep learning based approaches. The application to real data to produce scientifically exploitable data products remains very moderate. These approaches suffer from similar limitations as the classification tasks, that is, training sets and uncertainty quantification (see Section 7). Finding suitable training sets is more challenging in this case as one need to both label and identify the boundaries of the objects. In astronomy, the definition of object boundaries strongly depends on the noise levels. The

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

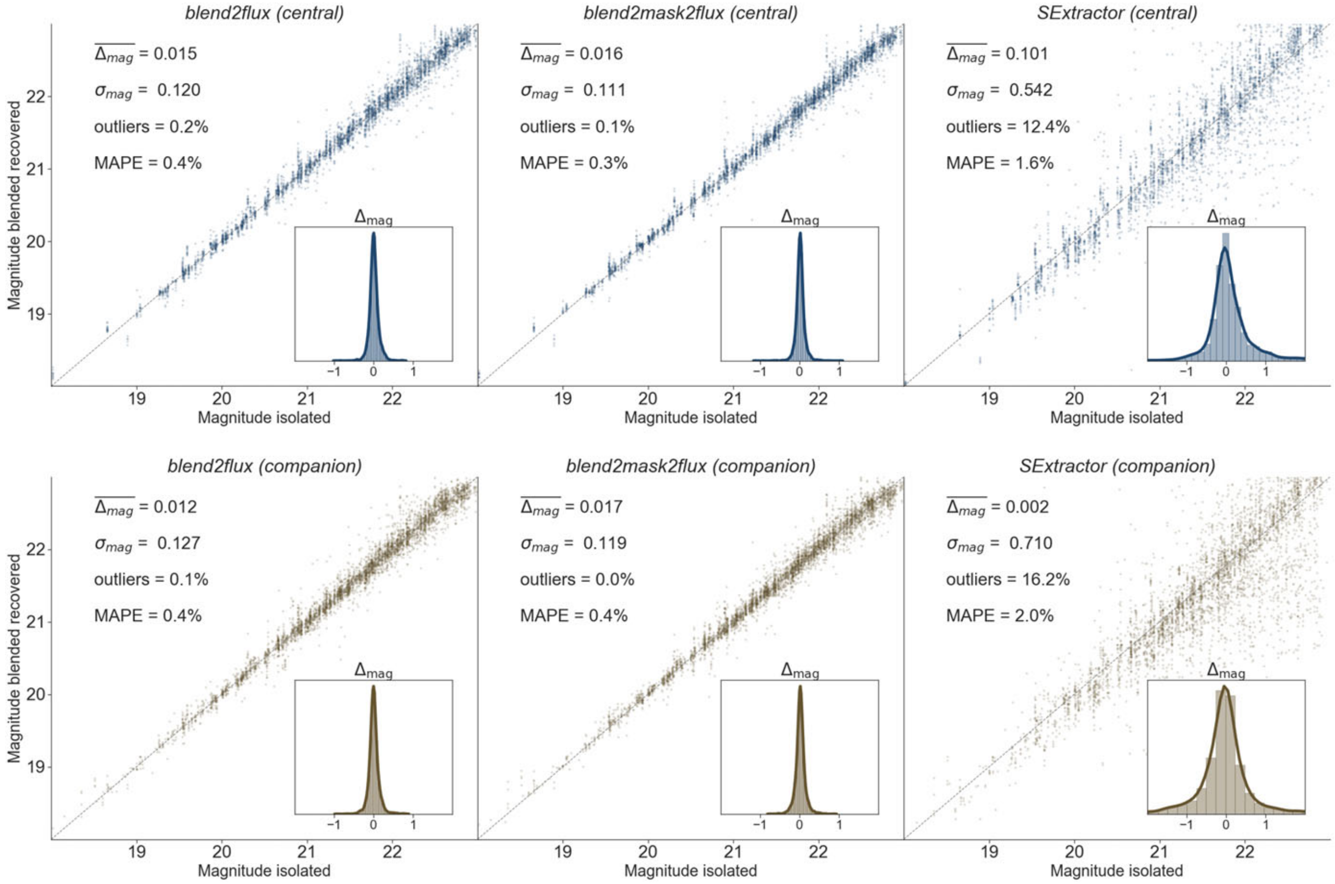

Figure 13. Comparison between the photometry measured with the Unet and with SExtractor. All panels show a comparison between input and recovered photometry. The top row shows the central galaxy in the stamp while the bottom row is the companion one. The two leftmost panels show two different deep learning architectures. The rightmost panel shows the results of the baseline SExtractor. Both dispersion and bias are improved with deep learning. Figure adapted from Boucaud et al. (2020)

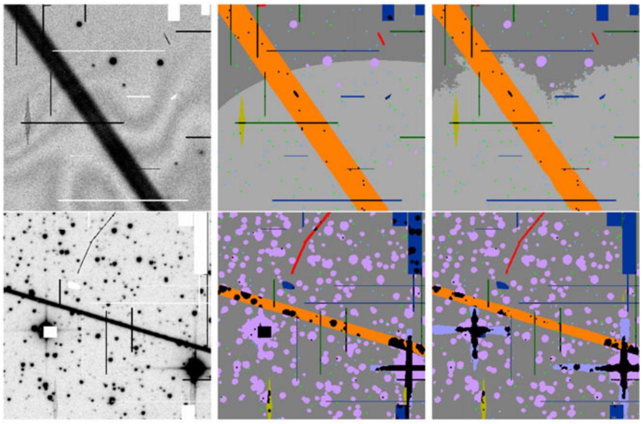

Figure 14. Unet used to segment different types of artefacts on CCD images. The leftmost figures show a field containing different artefacts. The middle panel shows the ground truth and the rightmost panels the predictions obtained by the Unet. Figure adapted from Paillassa et al. (2020)

adopted solutions so far are either training on simulations or relying on previous detections. None of them seems fully satisfactory. Simulations are usually too simplistic and the generalisation to data can induce unexpected behaviours. Using outputs from other software packages tends to propagate the existing biases. Possible solutions include the generation of more realistic training sets with generative models (e.g. Feder, Berger, & Stein 2020; Lanusse et al. 2021; Bretonnière et al. 2022). We will discuss this in more detail in Subsection 3.4. Recent works have also started to explore the implementation of uncertainties in the produced segmentation maps (Bretonnière, Boucaud, & Huertas-Company 2021) which seems a promising way to limit the impact of possible catastrophic failures when changing domain from simulations to observations. However, all these works remain at the proof-of-concept stage.

In addition to these segmentation based approaches, other groups have attempted to go a step forward by reconstructing the surface brightness profiles of overlapping galaxies. This implies moving from a classification to a regression problem, since the output of the networks is the flux at a given pixel. This task typically requires the use of generative models to learn the diversity of galaxy shapes in a data-driven way and then being able to generate likely solutions. Reiman & Göhre (2019) first explored the use of Generative Adversarial Networks (GANs) for this purpose (see Figure 16). In that case, the network input is a stamp of two overlapping galaxies and the output are two images containing each of the two galaxies separately. An adversarial loss is employed to ensure that the two produced galaxies are realistic and indistinguishable from observed galaxies (see Figure 17). Following a similar goal, Arcelin et al. (2021) used Variational Autoencoders (VAEs) to reconstruct the light distribution of blended galaxies applied to simulations of the LSST survey. Recently, Hausen & Robertson (2022) attempted an intermediate solution in between

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

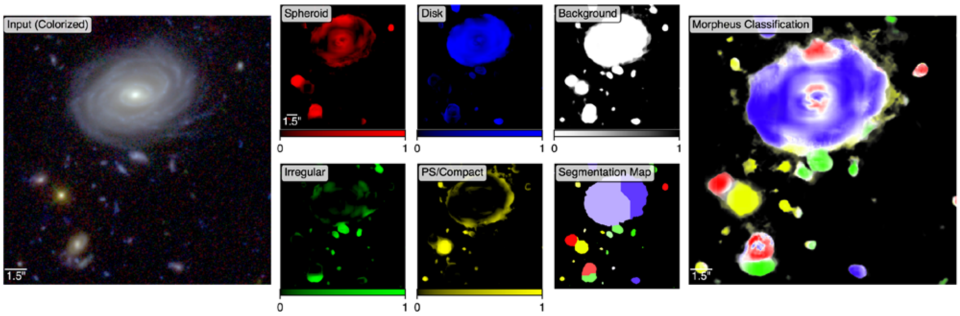

Figure 15. Example of pixel-level classification of a region in the CANDELS survey using a Unet. The leftmost image shows the original field. The six smaller images show different channels with different morphological types and the rightmost image is the combination of all channels with a pixel-level morphological classification. Figure adapted from Hausen & Robertson (2020)

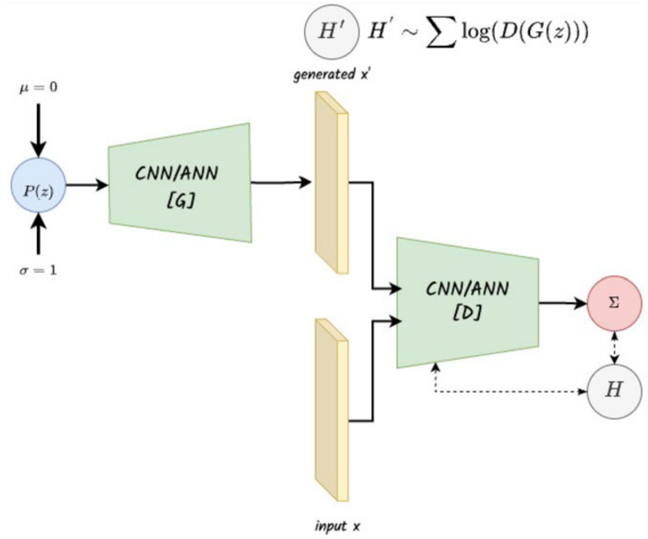

Figure 16. Schematic representation of a standard Generative Adversarial Network (GAN). It has been used as a generative model for deblending, identifying outliers as well as for accelerating cosmological simulations among others. A first CNN (generator) maps a random variable into an image, which is then fed to a second CNN (discriminator) together with real images to distinguish between both datasets. The generator and discriminator networks are trained alternatively.

full reconstruction and detection. They proposed a novel approach based on Partial-Attribution Instance Segmentation to estimate the fraction of fluxes in each of the galaxies from the blended system. This is interesting as it provides a solution specifically designed for the astrophysical problem, which remains rare in deep learning applications.

Works attempting this task are still at the exploration level as well, even though the results seem very promising (e.g. Figure 17). Reiman & Göhre (2019) used very simple simulations by just adding two SDSS images; Arcelin et al. (2021) employed analytical simulations. As all other deblending efforts, this approach suffers from finding a suitable training set which is close enough to observations without being too simplistic. Arcelin et al. (2021) showed that transfer learning from a network trained on simulations is a promising approach. More important, the statistical versus individual accuracy problem becomes more dramatic when generating images instead of masks. Generative Models produce realistic images in a statistical sense, that is, arising from the same probability density function than observations. However, on an individual basis, artefacts might appear on the generated images. These are in general very difficult to track down and can therefore induce significant biases. This individual versus statistical accuracy is an inherent problem to machine learning which needs to be taken into account when using ML predictions for scientific analysis.

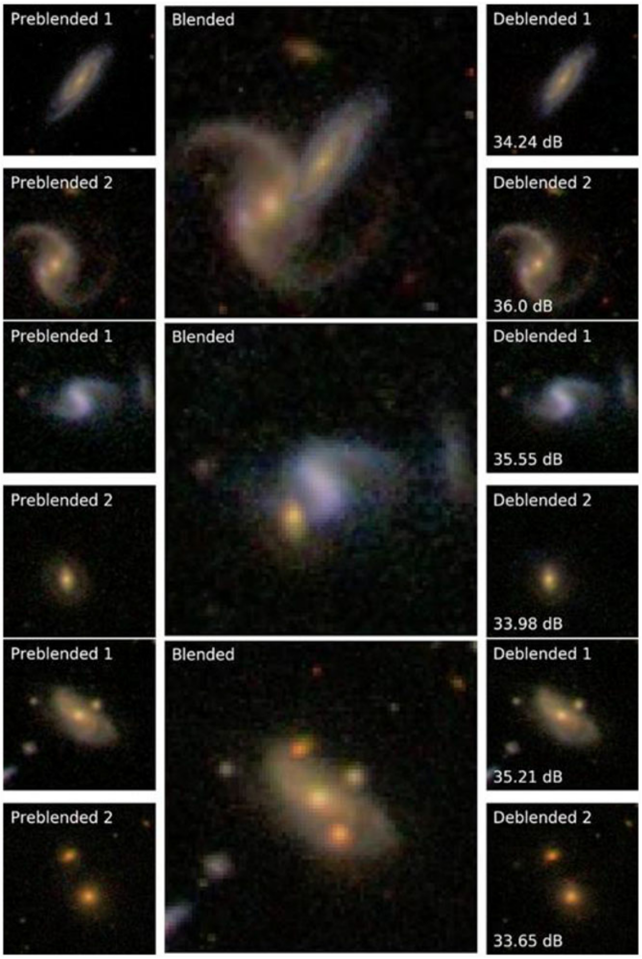

Figure 17. Example of galaxy deblending using GANs. The central images show three overlapping galaxies. On each side of the big images, the original (left) and reconstructed (right) images of the pair of galaxies are shown. Figure adapted from Reiman & Göhre (2019).

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

In order to limit the black-box effect, an interesting approach is proposed by Lanusse, Melchior, & Moolekamp (2019). The authors use an hybrid model that combines a physically motivated model with analytic expressions for known terms, with a data-driven prior for galaxy morphology learned with a generative model. In this approach, the output of the generative model is only used as a prior of the inverse problem, and therefore the impact of unexpected artefacts is reduced. The combination of physically motivated models with data-driven ones appears as an appealing solution which will likely become important in the future.

3.3. Improving data quality

Deep learning has also been explored to improve the quality of data, that is, denoising and deconvolution. Astronomical images are usually noisy and blurred by the effect of the Point Spread Function (PSF), which for ground based data includes both the telescope impulse response and the effect of the atmosphere. The processes of denoising and deconvolution aim therefore at recovering the information before degradation. This is typically a challenging inverse problem which needs significant regularisation. Data-driven approaches have emerged as alternative solutions to more classical deconvolution techniques. Schawinski et al. (2017) first explored the use of Generative Adversarial Networks (Figure 16) to deconvolve images from SDSS. They show that GANs can recover features even after degradation. This remains however a simple experiment since galaxies were previously degraded. Gan, Bekki, & Hashemizadeh (2021) built up on a similar idea using GANs to translate between ground and space based observations. Jia et al. (2021) proposes an improved solution based on two different GANs which reduces the need of large amount of paired images with and without degradation. Lauritsen et al. (2021) extended the idea to the sub millimetre regime by using Variational Autoencoders instead of GANs. Encoder-Decoder networks can also be used for denoising, and some works have explored this for astronomy. Vojtekova et al. (2021) showed that Unets can effectively increase the exposure time by a factor of two.

Generally speaking all the attempts show impressive results in solving the long standing problem of deconvolution. They remain however at the proof-of-concept stage and have not been applied for scientific analysis. Similarly to what happens with deblending, the main limitation is robustness. Generative Models such as GANs produce very realistic images but can also introduce artefacts which are statistically meaningful but not necessarily on an image per image basis. These artefacts can introduce uncontrolled biases in the scientific analysis.

3.4. Emulating astronomical images

In a number of applications, the ability of simulate survey data is very valuable, for instance to test and calibrate measurement pipelines. One of the difficulties faced in simulations of upcoming surveys such as LSST or Euclid is the relatively small set of deep and high-resolution imaging data (essentially limited to HST surveys such as COSMOS) that can be used to provide complex galaxy light profiles as inputs. This is one of the motivations behind the development of Deep Generative Models for galaxy morphology, which can be trained on the existing data and then generate significantly more examples realistic light profiles.

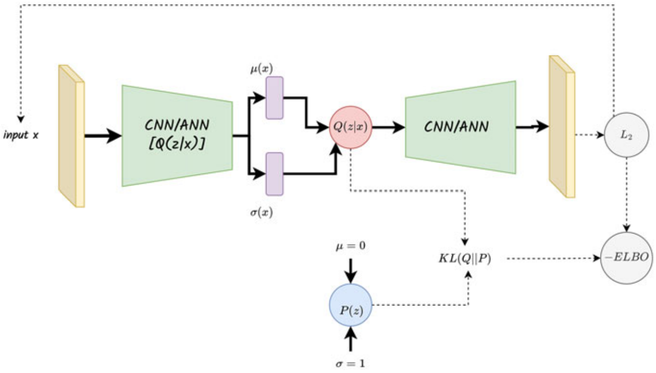

Figure 18. Illustration of a Variational Autoencoder (VAE). This generative model has been extensively used in astronomy for multiple emulation tasks. A first CNN maps the input into a distribution, which is then sampled to provide the input to a second CNN which reconstructs the input image. A regularisation term is added to ensure that the latent distribution behaves close to a prior distribution.

In one of the first works following that direction, Ravanbakhsh et al. (2017) demonstrated that training relatively simple GANs (Figure 16) and a Variational Autoencoders (see Figure 18) on HST COSMOS postage stamps successfully captured most of the relevant population-level parameters, such as size, magnitude, and ellipticity. In subsequent work, Lanusse et al. (2021) extended that model to explicitly account for the PSF, and proposed a hybrid Normalising Flow (see Figure 19) and VAE architecture which allowed to achieve diverse and high quality samples, while also making it possible to condition the light-profile generation on galaxy properties. Normalising Flows are a type of generative models which, as opposed to GANs or VAEs, allow the exact evaluation of the likelihood $p(y)$ and therefore, their weights can be directly learned by maximising the log likelihood of the training dataset. The idea is to construct a bijective mapping f such that $y = f(z)$ where z is a variable with a simple base density $p(z)$, typically a Normal distribution. As f is invertible one can evaluate the density y using the change of variable theorem, that is, simply inverting f and keeping track of the Jacobian of the transformation (see Figure 19). Bretonnière et al. (2022) used that generative model to create simulations of the Euclid VIS instrument on a 0.4 deg^2 field with complex galaxy morphologies and used those simulations to assess the magnitude limit at which the Euclid surveys (both deep and wide) will be able to resolve the internal morphological structure of galaxies.

One of the limitations of standard GANs and VAEs for the simulations of such galaxies is that it quickly becomes difficult to generate high quality samples on large stamps. To address these technical difficulties, Fussell & Moews (2019) developed for instance a StackGAN approach in which a first GAN is trained to generate a low-resolution image (e.g. 64 × 64), which is then up-sampled by a second GAN (e.g. to 128 × 128). Smith & Geach (2019) proposed to use a GAN variant, known as a Spatial-GAN (SGAN) to generate no longer only postage stamps, but entire fields of arbitrary size through a purely convolutional architecture. The authors demonstrated the ability to sample a 87 040 × 87 040 pixels image emulating the HST eXtreme Deep Field (XDF). More recently, Smith et al. (2022) explored applying a Denoising Score Matching approach (Song & Ermon 2019) and were abltexte to generate large high-resolution postage stamps of size 256 × 256 of remarkable quality.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

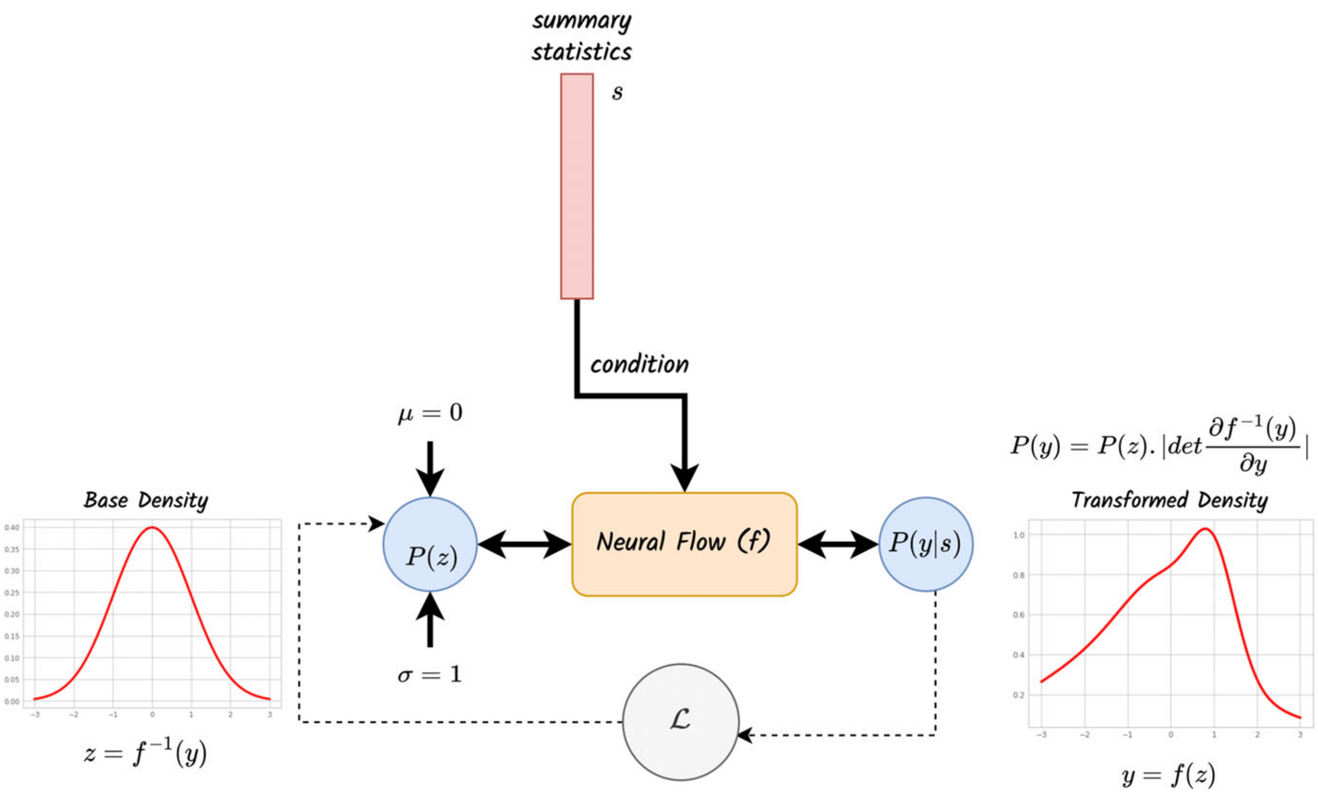

Figure 19. Illustration of a Normalising flow. This density estimator model is used to transform a simple distribution (typically a Normal distribution) into a more complex density using an invertible function *f* parametrised by a neural network which is optimised by maximising the log likelihood. The transformation can also be conditioned to an additional set of parameters (*s* in the figure).

Summary of deep learning for computer vision

- Deep learning has rapidly emerged as a solution for the classification of objects in large surveys. Galaxy morphology, strong lenses, and also light curves constitute the main applications. Deep learning based catalogs, especially for galaxy morphology, exist and are being used for scientific analysis.
- The most common approach for classification are supervised Convolutional Neural Networks with different degrees of complexity.
- Overall, CNNs achieve higher accuracy than previous approaches and are also faster.
- The lack of labelled training sets is a common bottleneck. Solutions involving Transfer Learning and/or the use of simulated training sets have been proposed. This implies some additional challenges which we discuss in Section 7
- False positives in the case of very unbalanced samples (e.g. for strong lensing) is also a commonly encountered problem. No satisfactory solution has been found to date apart from visual inspection.
- Standardised datasets to consistently compare the performances of different classification approaches are not common in astronomy which limits the possibility of comparing different approaches (see Section 7).
- Deep learning has also been explored for object detection and segmentation in images.
- The most popular approach for segmentation are convolutional encoder-decoder networks such as Unets, although more complex architectures have been also employed.
- Overall, the results are promising and tend to outperform pre-deep learning approaches for detection.
- Most of the works remain still at the proof-of-concept stage. Until now, there are no deep learning based catalogs in major deep imaging surveys. The robustness of such approaches is still a concern, especially for deblending. Physics informed models could be a solution to explore in the future.

4. Deep learning for inferring physical properties of galaxies

4.1. *Neural networks as fast emulators*

We now move to address the efforts done in the past years to estimate the physical properties of galaxies using deep learning. Compared to the previous section where we described computer vision tasks, these applications are typically more domain specific since they target physical quantities of galaxies from a regression point of view. The general approach followed by the community has been to test neural networks to emulate—replace—more specific tools, developed over the years, which are usually slow or not fast enough to deal with future big data surveys.

4.1.1. Photometric redshifts

All modern cosmological surveys require a more or less precise estimation of redshifts. When spectroscopy is not available, which is the case for most deep surveys, photometric redshift estimation is the standard way to proceed to measure distances of large numbers of galaxies using a combination of broad and narrow band photometry. Photometric redshift estimation is therefore a non-linear mapping between a set of photometric points and a real number measuring the redshift. The standard way to approach the problem is through the fitting of Spectral Energy Distributions generated from Stellar Population Models (e.g. Benítez 2000; Bolzonella, Miralles, & Pelló 2000). However, since it is—in theory—a well defined problem, it is among the most popular applications of deep learning supervised regression in astrophysics. The first attempts of estimating photometric redshifts with neural networks start well before the deep learning boom, in the early 2000s (Collister & Lahav 2004; Vanzella et al. 2004). These methods already relied on the idea of learning the mapping between photometry and redshift from data through a Multi-Layer Perceptron trained under a Mean Squared Error (MSE) loss. The only difference with a modern architecture would be in the depth of the model and the choice of activation function. Perhaps the most successful of these early neural methods for photometric redshifts, ANNz (Collister & Lahav 2004) has continued to evolve over time, with ANNz2 (Sadeh, Abdalla, & Lahav 2016) including some quantification of epistemic uncertainties through an ensemble of randomised estimators techniques reminiscent of modern deep ensembles.

Two significant evolutions of these methods appeared in recent years with the generalisation of deep learning: 1. probabilistic modelling of the redshift distribution to estimate posterior probabilities; 2. pixel-level models based on CNNs, thus going beyond photometric information and able to use morphology as well.

Probabilistic Modelling of Photometric Redshifts Going beyond a regression task, Bonnett (2015) proposed to use a neural network (still an Multi-Layer Perceptron—MLP), which for a given photometry would output a distribution in the form of a discretised probability density function (PDF). The model would then be trained with a standard cross-entropy loss to predict the redshift bin in which a given galaxy should fall, which in fact mathematically corresponds to estimating the posterior distribution of redshifts given photometry, under a prior given by the selection of the training set. This approach was subsequently reused to train other, more complex, neural networks for photometric redshifts (Pasquet-Itam & Pasquet 2018; Pasquet et al. 2019), but can potentially suffer from the discretisation needed to represent the PDF. Indeed, the network has no built-in notion that classes with adjacent indices actually correspond to adjacent bins.

Another approach to model distributions at the output of a neural network is to use a Mixture Density Network (MDN Bishop 1994). MDNs use an MLP to predict the parameters of a mixture of probability densities, and thus provide an analytic and continuous PDF model for a given input (see Figure 20). This approach was for instance proposed in D'Isanto & Polsterer (2018), where the neural network outputs are the mean, variance, and weights for a mixture of n one dimensional Gaussians. Overall, the general consensus, is that, when only using photometry as input, neural networks do not provide specially more accurate results than other template based approaches (see e.g. Schmidt et al. 2020). One important challenge is that training sets are generally biased as we will discuss in Subsection 7.3.

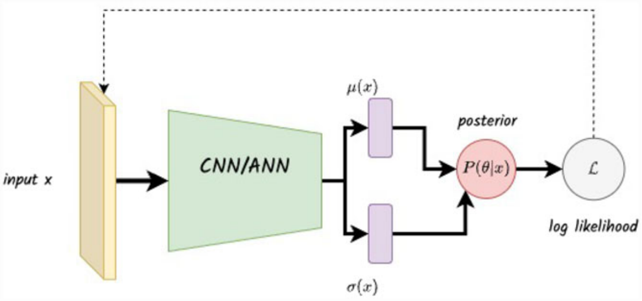

Figure 20. Representation of a Mixture Density Network (MDN). It is a generalisation of a standard CNN/ANN in which the output has been replaced by a probability distribution parametrised with some parameters (μ, σ in the figure). MDNs provide an estimation of the random (aleatoric) uncertainty. The loss function is the generalised log likelihood.

Convolutional Photometric Redshift Estimators The second major evolution of neural network-based photometric redshifts has of course been the introduction of CNNs to build a pixel-level model capable in principle of using the entire information content of a multi-band galaxy image to refine a redshift estimate. In the first instance of this approach, Hoyle (2016) used the state-of-the-art model at the time, AlexNet (Krizhevsky et al. 2012), to build a 5-layers deep CNN model taking as inputs a combination of *r*,*g*,*i*,*z* images of a given galaxy and trained to classify that galaxy into a discrete redshift bin. Interestingly, in this initial study performed over a set of 60 000 SDSS images, no significant improvements in redshift prediction accuracy were reported when compared to a more traditional photometric feature-based AdaBoost machine learning model. It would take a couple more years for a broader development of CNN-based methods, starting with D'Isanto & Polsterer (2018) which proposed to combine a simple 3-layers deep convolutional architecture with a mixture density output, but again reported only a mild improvement in terms of accuracy compared to a feature-based approach on an SDSS sample.

The benefits of a convolutional approach started to become clear with Pasquet et al. (2019), which used a much deeper convolutional network, comprised of one input convolution layer and 5 inception blocks, trained under a redshift bin classification loss. These inception blocks (Szegedy et al. 2014) essentially replace one convolutional layer by multiple parallel convolutional layers with different kernel sizes, the output of which are concatenated back into a single tensor at the end of the block. This study, based again on galaxies from the SDSS Main Galaxy Sample using *ugriz* images, illustrated in particular how the CNN is able to automatically make use of pixel-level data to extract information beyond colours, improving redshift estimation. In particular, the bottom row of Figure 21 shows the comparison between the photometric redshift bias of a standard colour-based k-NN photometric redshift estimate and the proposed CNN approach as a function of galaxy ellipticity (proxy for galaxy inclination) for star-forming galaxies. As can be seen, the colour-based approach exhibits a strong inclination-dependent bias caused by the various amounts of dust attenuation as a function of the viewing angle. The CNN shows however comparatively very little bias, indicating that the

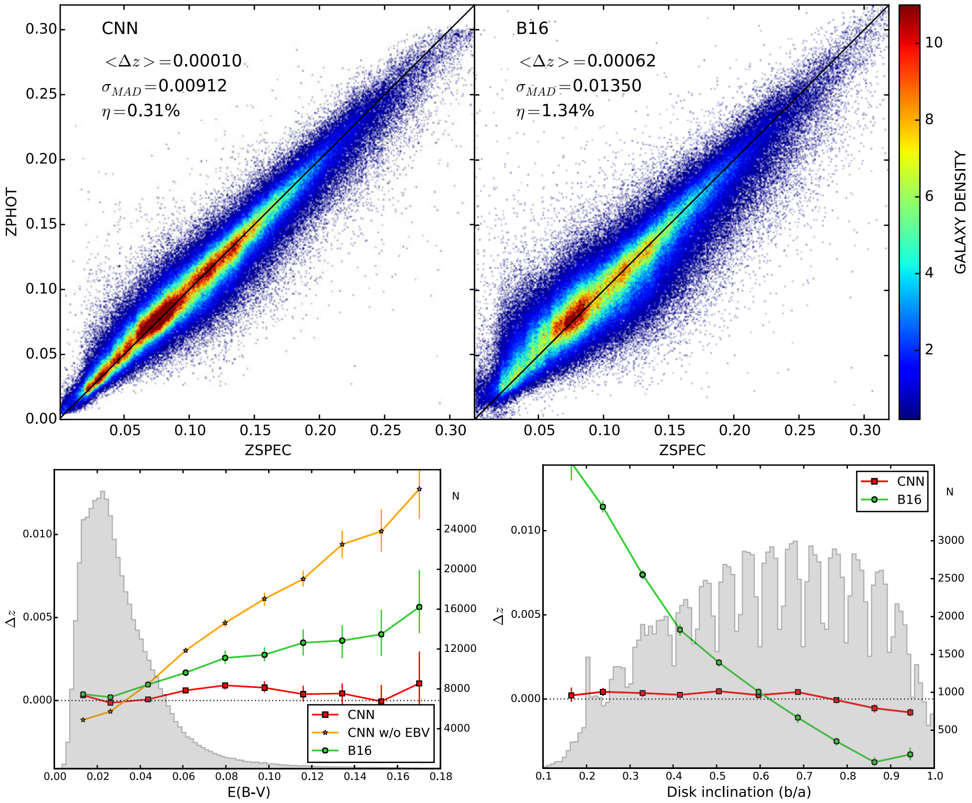

Figure 21. Comparison of photometric redshift performance between a deep CNN and a colour-based k-NN (B16) method reported in (Pasquet et al. 2019). The top row shows the predicted redshifts vs spectroscopic redshifts for the CNN (left) and the k-NN method (right). The distribution is noticeably tighter for the CNN with smaller overall bias, scatter, and outlier rate. The bottom row show the photometric redshift bias Δz for the two methods, as a function of extinction (left panel) and disk inclination (right panel) of galaxies classified as star-forming or starburst. Having access to the whole image information, the CNN is able to automatically account for the reddening induced by looking at galaxies with high inclination, whereas the k-NN method only using colour information is blind to this effect and exhibit a strong bias with inclination.

model is able to automatically estimate and account for inclination in its prediction by directly drawing that information from the postage stamps. This example illustrates the main advantage of a deep learning approach, it alleviates the need for handcrafted features, leaving it to the model to identify from the data itself the relevant information.

Pasquet et al. (2019) highlight however an important consideration when using a CNN approach. Whereas flux measurements can be standardised to account for varying noise and PSF, a CNN based only on the raw postage stamps, without additional information, is blind to these observing condition factors. The authors report for instance a noticeable bias as a function of seeing for the CNN, and mention the fact that this information could be provided to the model in future work to let the model compensate for these factors.

In a number of subsequent works (Menou 2019; Henghes et al. 2021; Zhou et al. 2021), it was proposed to extend this pure convolutional approach to photometric redshift estimation to hybrid models, combining an MLP branch tasked with processed photometric features (e.g. colours) and a CNN branch having access to the full image of the galaxy. It was found in these multiple studies that directly providing the model with highly informative features through the MLP branch improved the overall accuracy. Although all the relevant information is in principle already included in the images themselves, this approach reduces the amount of automatic feature extraction that the convolutional branch needs to perform. In all cases, the best results are achieved when both photometric features and images are provided jointly to the model.

Improving Scaling with Size of Training Set One particular aspect that may limit the applicability of these deep learning methods is the need for large training samples, and so in this case the need for large (expansive) spectroscopic samples to properly train these supervised methods. As a possible mitigation technique for this issue, Hayat et al. (2021) demonstrated that a self-supervised encoding provided by contrastive learning retained significant redshift information. We direct the interested reader to Section 5 for more details on contrastive learning. Here, the authors proposed a two-step approach, first training in complete self-supervision (without needing any spectroscopic redshifts) a 1d encoding of galaxies postage stamps. And in a second step, training in a supervised way a shallow MLP to predict redshifts on a small training set of available spectroscopic redshifts. They find two surprising results: 1. This fine-tuned self-supervised approach always outperforms a conventional supervised training, 2. The accuracy of self-supervised estimated redshifts scales very well into the low-data regime. They find for instance that their self-supervised approach using 20% of labelled data achieves similar accuracy to a fully supervised training using labels for the entire dataset.

With a different strategy for limiting the amount of data needed, Dey et al. (2021) proposed to replace a conventional

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

convolutional architecture by deep capsule networks. Compared to CNNs, capsule networks are robust to rotations and changes in viewpoints and thus provide a more natural representation for randomly oriented objects like galaxies (see also Section 3). The hope is therefore to not require as much training data if the model already provides the built-in invariances of the problem. In their proposed architecture, the capsule network generates a low dimensional encoding of the input image, which is then used to perform two tasks jointly: reconstructed in the input image with a CNN, and estimating the redshift of the galaxy with an MLP. In addition, their model also classifies, at the level of the capsule outputs, the morphology type of the galaxy (elliptical or spiral). The authors find that this approach leads to a better scaling with size of the training set than (Pasquet et al. 2019), especially at very small training set sizes, but the benefits are not as significant as the ones offered by the self-supervised approach of Hayat et al. (2021).

Another complementary approach to reduce the dependence on large spectroscopic datasets is to use transfer learning, to build a model on simulated data, and fine-tuning it on survey data. This approach was for instance explored in Eriksen et al. (2020) using a MDN trained on a combination of FSPS simulations and data from the PAU Survey. Using a pretraining on simulations was found to reduce the photo-z scatter by as much as 50% for faint galaxies.

Finally, the question of robustness and stability of these deep neural networks was investigated in Campagne (2020) which highlighted that inception models like the one proposed in Pasquet et al. (2019) can be highly sensitive to adversarial attacks. Although these attacks are unlikely to happen on astronomical data (see however Ćiprijanović et al. 2021a), this result underlines again the fact that these black-box methods are not as interpretable as more conventional approaches like template-fitting (interpretability is discussed in Section 7.3).

4.1.2. Galaxy structural parameters

In addition to classification, galaxy morphology can be also quantified with some parameters that define an analytical description of the surface brightness distribution of galaxies. The so-called Sersic models are defined by three quantities: the effective radius (r_e), the Sersic index (n) and the axis ratio (b/a). By combining these parameters with a normalisation factor to account for the different galaxy luminosities, one can describe most of the surface brightness distributions of galaxies. The standard way to measure these parameters is by fitting PSF convolved analytic models to the 2D surface brightness distributions of galaxies. The task can be formulated as a mapping between pixel values and real numbers, which characterise the galaxy shape. It is therefore well suited for a supervised regression problem, provided that a reliable training set is available. Given that the input data are galaxy images, Convolutional Neural Networks are the most common approach. Tuccillo et al. (2018) first used a CNN to estimate galaxy structural parameters. In this first work, the authors used a simple training set made of analytic profiles with noise added and demonstrated that CNNs achieve comparable or better performance than standard methods, with the key advantage of being a factor of ~50 faster. As previously said, computational efficiency is one of the main motivations behind these works aiming at emulating existing software. Tuccillo et al. (2018) also attempted to apply the trained CNNs to observed galaxies with HST. However, because the training set was too simple, the results did not appear to be satisfactory. In particular, the authors did not include foreground and background galaxies in the training set which constitutes an important difference with observations. The authors proposed a transfer learning step using measurements performed with standard approaches. Although the results quickly improve, the final results necessarily propagate the systematics of existing software, which cannot be improved by construction. In that respect, the main gain is speed. Ghosh et al. (2020) built on this and showed that with a transfer learning step, CNNs can provide reliable structural parameters for both low and high redshift galaxies. The authors estimate structural parameters for ~120 000 galaxies in the SDSS and CANDELS surveys.

More recently, Li et al. (2021) attempted a similar approach applied to ground-based observations. The training is still done on simulations but with realistic backgrounds. Additionally, the PSF is included as an input to the CNN, so that the networks can learn the effects of varying PSFs across the field of view. The authors show, that by including these improvements, the CNNs generalise well to observations without need of transfer learning and achieve comparable results to standard approaches, with the advantage of being ~3 times faster (see Figure 22). Other works have attempted to decompose the galaxies into bulges and disks. This is an equivalent problem but the number of parameters is increased by a factor of two. Grover et al. (2021) showed that CNNs can estimate the bulge-to-disk ratio—that is, luminosity ratio between the bulge and the disk components—of 20 000 galaxies in less than a minute. The main motivation is, once more, a gain in computational time. Tohill et al. (2021) explored the use of CNNs to estimate other morphological parameters of galaxies which quantify the distribution of light (i.e. Concentration-Asymmetry-Smoothness—CAS—system; Conselice 2003). The conclusion is very similar; neural networks accurately reproduce measurements compared with standard algorithms, but faster. Interestingly, they also show that using CNNs makes the measurements more robust in the low signal-to-noise regime, which is one of the main issues of the CAS system.

Overall, these approaches look very promising to deal with large amounts of photometric data such as the datasets that will be delivered by Euclid and the Rubin Observatory for example. The limitations are similar to other problems. The networks need to be trained on simulations by definition. The extrapolation to real observations is always complicated as one needs to make sure that the training set covers all the observed parameter space. As this is a common challenge, we discuss it in Section 7. This is particularly challenging for space based observations in which the enhanced spatial resolution increases the differences between the simulated datasets used for training and the observations. A possible solution is the inclusion of some sort of uncertainty estimation which could help identifying failures. This has been recently done by Ghosh et al. (2022) who showed that a combination of MDNs and Monte Carlo Dropout can provide well calibrated uncertainties of galaxy structural parameters. Aragon-Calvo & Carvajal (2020) explored a self-supervised approach to avoid using a fully supervised training based on simulations. However, the approach has not been applied so far to large samples of galaxies. In addition, since the inference time is very short, the bottleneck of this type of approaches is in the training. In current approaches, a specific training set needs to be built for every different application which is not an optimal solution.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

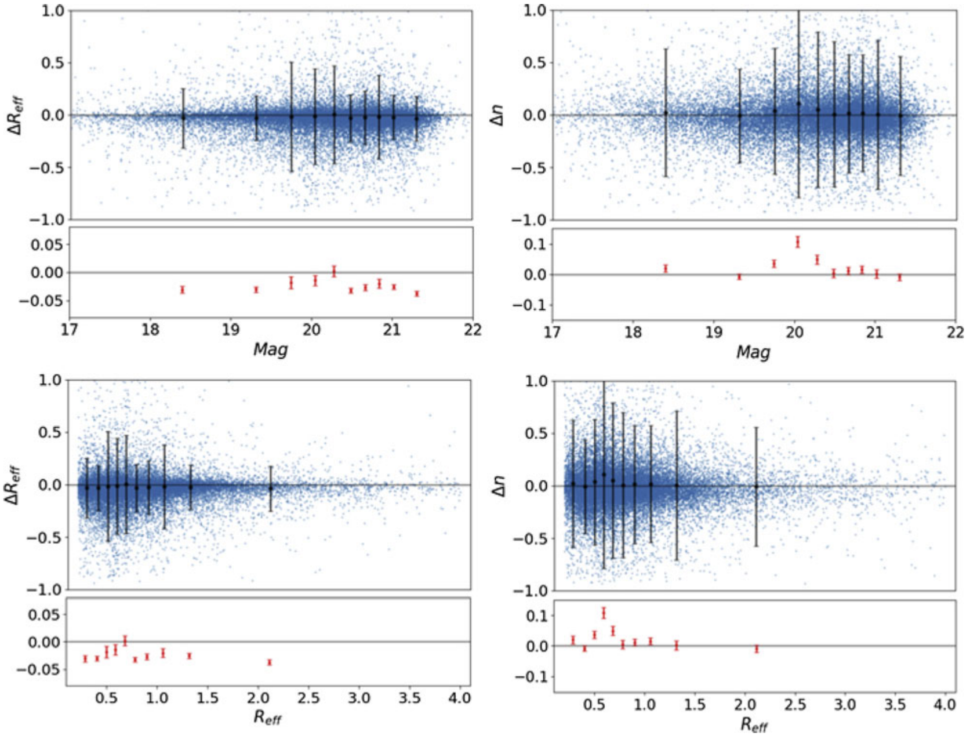

Figure 22. Convolutional Neural Network to measure galaxy structure. The left column shows the difference between true and estimated radius. The right column is the same for Sersic index. Top row shows the results as a function of magnitude while the bottom row is as a function of radius. Figure adapted from Li et al. (2021).

4.1.3. Stellar populations, star formation histories

A similar type of application of deep learning is to derive the properties of the stellar populations of large samples of galaxies. This is also a regression type of application, in which a mapping between the galaxy photometry and properties like the stellar mass, the metallicity or stellar age, is sought. As for the previous applications, it exists a standard approach based on the fitting of the Spectral Energy Distributions (SEDs). However, it is typically slow and not adapted to the large volumes of data that will be delivered by future surveys. Deep learning is thus used as an accelerator. Most of the published works so far follow a similar approach. A supervised Neural Network is trained for regression between photometric values and stellar population properties. For example, Surana et al. (2020) used fully connected Artificial Neural Networks applied to data from the GAMA survey to derive stellar masses, star formation rates and dust properties of galaxies. The training is performed on stellar population models and the results are compared to standard approaches. The conclusions are also very similar to other applications falling in the same category. Deep Learning performs similarly to standard methods, but a factor of a few faster. Similarly, Simet et al. (2019) used neural networks to estimate the stellar population properties of high redshift galaxies from the CANDELS survey. The training is in this case performed on semi-analytical models. The conclusion is that galaxy physical properties can be measured with neural networks at a competitive degree of accuracy and precision to template-fitting methods. It is worth noticing that Neural Networks are not the only approach to address this problem, although in this review we primarily focus on deep learning techniques. As this is essentially a mapping between two sets of real numbers, other Machine Learning techniques can be employed—Gilda et al. (2021), Davidzon et al. (2019) used for example Boosting Trees and SOMs respectively for a similar task.

In a recent work, Buck & Wolf (2021) pushed this idea further by trying to predict resolved stellar population properties instead of integrated quantities (Figure 23). In that case, the mapping is made between broad band photometric images of galaxies and 2D maps of stellar mass, metallicity and other stellar population properties. This is equivalent to a regression at the pixel level. The architecture for this type of work is en encoder-decoder Unet type of network as the ones used for segmentation (see Section 3 and Figure 12) but with a mean square error loss to work in regression mode.

In addition to stellar population properties at the time of observation, one can use the photometry of galaxies to infer the star formation histories (SFHs), that is, the star formation rate as a function of cosmic time. There are several established approaches using either parametric or non-parametric methods. However, the problem is known to be significantly degenerate and the star formation activity at early times is in general poorly constrained. Therefore the final estimation heavily relies on established priors. Lovell et al. (2019) first attempted to use CNNs in a supervised regression setting to estimate the SFHs of galaxies in the EAGLE cosmological simulation (Figure 24). One advantage of training on hydrodynamic simulations is that the prior is learned in a data-driven way by using fairly realistic distributions from the simulations. The authors of this first work show a reasonable reconstruction of the SFH and a decent robustness to domain changes. Qiu & Kang (2021) uses CNNs for the opposite task, that is, estimate the galaxy SED from the galaxy SFH taken from

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

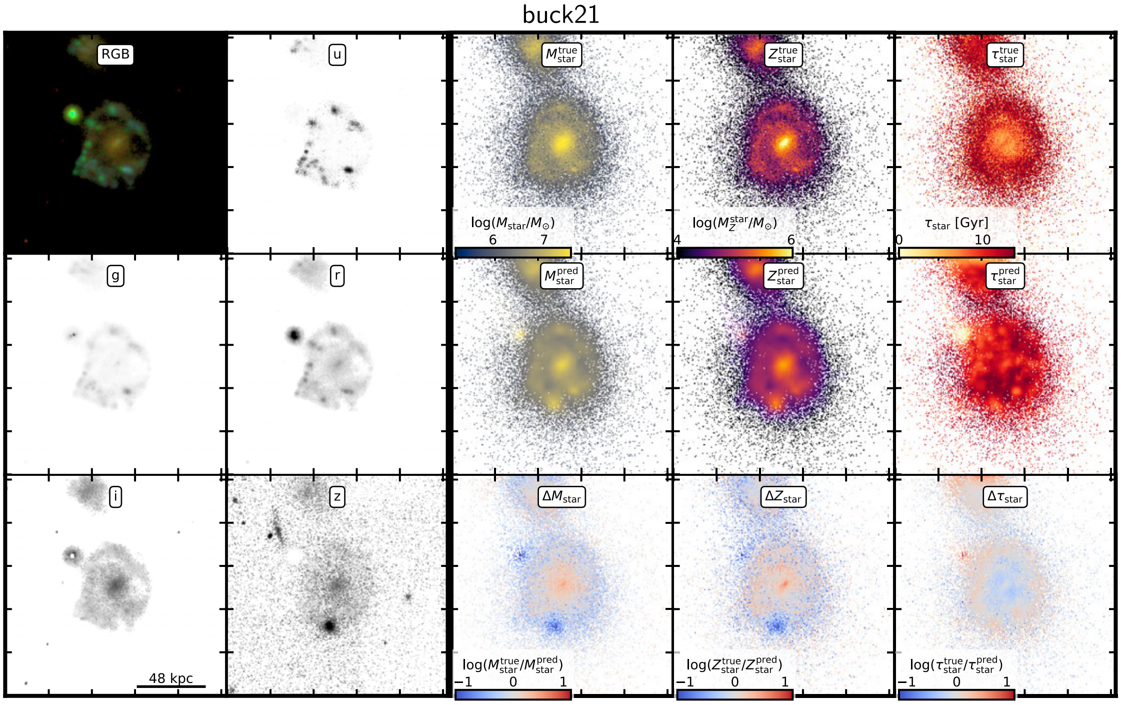
buck21
RGB
u
$M_{\rm star}^{\rm true}$
$Z_{\rm star}^{\rm true}$
$\tau_{\rm star}^{\rm true}$
$\log(M_{\rm star}/M_\odot)$
6 7
$\log(M_Z^{\rm star}/M_\odot)$
4 6
$\tau_{\rm star}$ [Gyr]
0 10
g
r
$M_{\rm star}^{\rm pred}$
$Z_{\rm star}^{\rm pred}$
$\tau_{\rm star}^{\rm pred}$
i
z
$\Delta M_{\rm star}$
$\Delta Z_{\rm star}$
$\Delta \tau_{\rm star}$
48 kpc
$\log(M_{\rm star}^{\rm true}/M_{\rm star}^{\rm pred})$
−1 0 1
$\log(Z_{\rm star}^{\rm true}/Z_{\rm star}^{\rm pred})$
−1 0 1
$\log(\tau_{\rm star}^{\rm true}/\tau_{\rm star}^{\rm pred})$
−1 0 1

Figure 23. CNN applied to estimate resolved stellar populations of galaxies. The input are images in different wavebands (2 left columns) and the output are different stellar population maps (3 right columns). In the stellar population maps, the top row is the ground truth, the middle one the prediction and the bottom row shows the residual. Figure adapted from Buck & Wolf (2021).

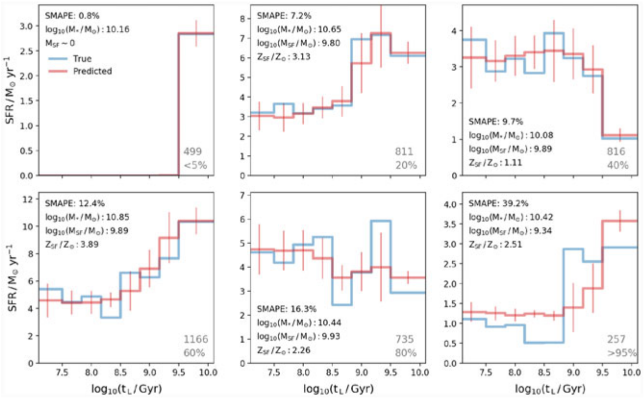

Figure 24. CNN applied to reconstruct the Star Formation Histories of galaxies. The input are photometric images and the output the Star Formation Rates in different bins of time. The different panels illustrate some examples of predictions (red lines) together with the ground truth from cosmological simulations (blue lines). Figure adapted from Lovell et al. (2019).

simulations. In this case deep learning acts as an emulator of radiative transfer codes.

As other applications of this kind, the results strongly rely on simulated training sets and on the implicit assumption that the simulations properly cover the observed parameter space. This is particularly critical here since the mocking from numerical simulation is usually done with existing stellar population models which are unavoidably a simplification of reality. Another important limitation is that, up to now, none of the published works on this front properly accounts for uncertainties, even though uncertainty estimation is far from being a solved issue with traditional methods. Both uncertainty estimation and domain shift are common challenges to many applications which are discussed in Section 7.

4.1.4. Strong lensing modelling

Another science case in which deep learning techniques have been extensively tested over the past years is the modelling of strong gravitational lenses—the formation of multiple images of distant sources due to the deflection of their light by the gravity of intervening structures. In Subsection 3.1 we have discussed efforts done to find these lenses on large datasets. The goal here is to characterise the lenses. This generally means quantifying image distortions caused by strong gravitational lensing and estimating the corresponding matter distribution of these structures (the 'gravitational lens'). Similarly to the previous applications in this category, there exists a method to perform this analysis, based on maximum likelihood modelling of observations. The process is however time consuming requiring complex dedicated software. Deep learning appears therefore as an appealing approach for accelerating the inference. The first work in exploring this is by Hezaveh et al. (2017). The authors test CNNs to estimate the lensing parameters from the images—Einstein radius, complex ellipticity, and the coordinates of the centre of the lens. They show that CNNs can recover the parameters with similar accuracy than standard methods but ten million times faster (Figure 25). An obvious caveat of the deep learning approach for inference is the lack of reliable uncertainties. Perreault Levasseur et al. (2017) is one of the first works exploring Bayesian Neural Networks to estimate uncertainties in the modelling of strong lenses, and in astrophysics in general. They use the technique of Monte Carlo dropout to approximate Bayesian Neural Networks (Gal & Ghahramani 2015; Charnock, Perreault-Levasseur, & Lanusse 2020) and show that, in

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

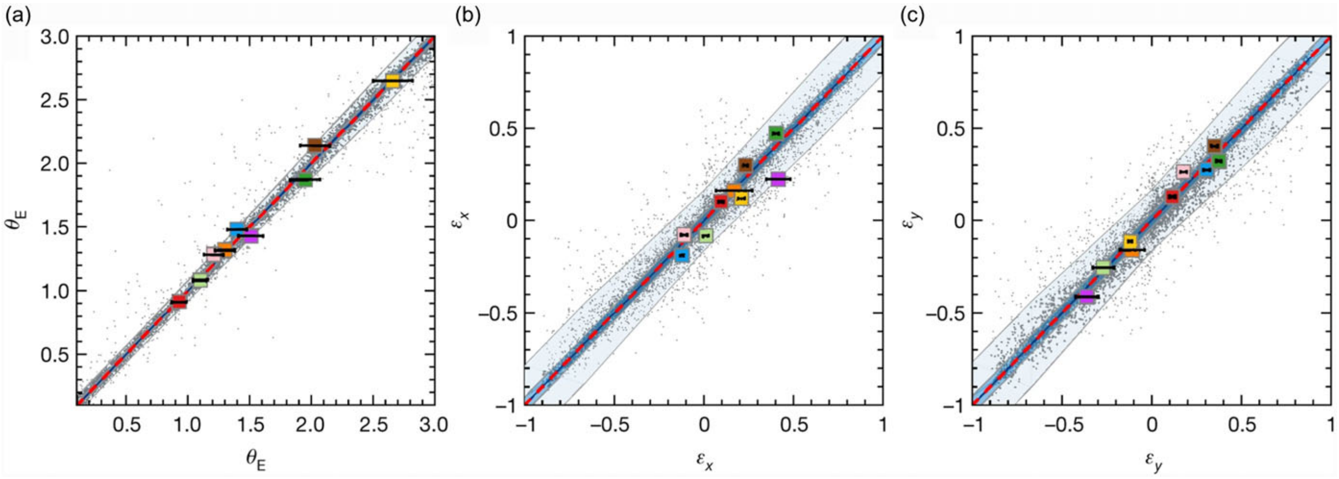

Figure 25. Estimation of strong lensing parameters with CNNs. Each panel shows a different parameter (ground truth versus CNN estimation). Figure from Hezaveh et al. (2017)

that particular case, it results in accurate and precise uncertainty estimates, significantly faster than Monte Carlo Markov Chains.

These two pioneer works have set the route for a fair amount of publications exploring the use of deep learning for lens modelling along similar lines. The most typical approach is the use of CNNs on images with an approximate Bayesian component to estimate uncertainties. For example, Madireddy et al. (2019) proposed a complete deep learning based pipeline including detection and classification of lenses followed by a modelling phase. Bom et al. (2019) apply Residual Neural Networks to simulated images of the Dark Energy Survey to predict Einstein Radius, lens velocity dispersion and lens redshift within ~10–15%. See also Schuldt et al. (2021) for similar conclusions on simulated Hubble Space Telescope and Hyper Suprime-Cam Survey images. Pearson, Li, & Dye (2019a) applied CNNs to simulated LSST and Euclid images. They find that including colour information results in a ~20% increase in accuracy compared to single band estimates. In a follow-up paper, Pearson et al. (2021) perform a systematic comparison between CNN-based estimation and conventional parametric modelling on increasingly realistic datasets going from smooth parametric profiles to real observations from the Hubble Ultra Deep Field. The main conclusion is that CNNs outperform traditional methods not only in terms of speed but also in accuracy by ~20%. However, the work also concludes that combining both approaches reduces further the errors. In addition, the use of CNN priors reduces the computational time of parametric modelling by a factor of ~2. Chianese et al. (2020) goes a step further by proposing a fully differentiable framework for lensing modelling. The pipelines combines a data-driven model based on a VAE for the source and a physical model for the lens which allows a forward modelling of the system to be compared with the observations. The main novelty, is that thanks to the differentiable programming framework, it becomes possible to compute the derivatives of the likelihood function with respect to both the parameters of the source and the lens, allowing for fast inference (Figure 26). Along similar lines, Morningstar et al. (2019) combines a physical model with a Recurrent Neural Network to iteratively reconstruct the lens model which is then fed to a CNN to estimate the lens parameters. Morningstar et al. (2018) applies the same methodology to analyse interferometric data. The modelling of lenses can be combined with a direct inference of cosmological parameters such as the Hubble constant (Park et al. 2021). We will address these applications in more detail in Subsection 6.2. Maresca, Dye, & Li (2021) proposed to use CNNs, not for reconstruction, but to identify unphysical models from parametric fitting.

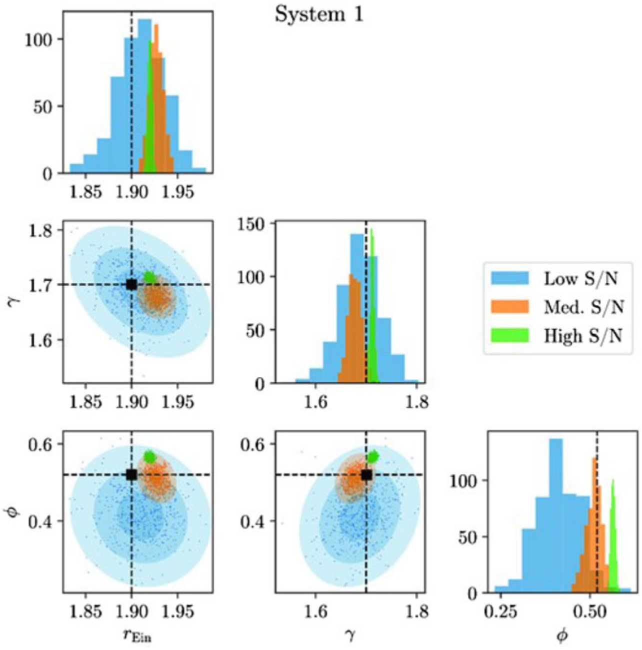

Figure 26. Lens modelling with a fully differentiable approach. The panels show the posterior distribution of different parameters of the model for different SNRs as labelled, together with the ground truth (black circle). Figure adapted from Chianese et al. (2020)

4.1.5. Other properties

Deep learning has also been applied to measure other galaxy properties, in addition to the major categories described in the previous subsections. Stark et al. (2018) used a Generative Model to measure the photometry of AGN hosts. The neural network is used in this case to separate the light of the quasar from the emission of the host galaxy. In that respect it is similar to a deblending problem discussed in Section 3. In line with works under the same category, the authors demonstrate that their approach is more than 40 times faster than standard model fitting approaches. Yao-Yu Lin et al. (2021) also addressed the issue of AGN quantification by inferring the mass directly without going through photometry. The input in this case are quasar light time series. The work shows that neural networks reach similar accuracy than traditional methods that use SDSS spectra which are more time consuming to obtain. Also within the time domain community, Stahl et al. (2020) developed a deep learning framework to measure the phase and the light curve shape of Type Ia supernova from an optical spectrum. Cabayol-Garcia et al. (2020) and Cabayol et al. (2021) explored CNNs to systematically measure photometry on the narrow band PAU survey. They find in this case that the deep learning approach improves the photometric accuracy by up to ∼40%.

4.2. Galaxy physical properties from simulations

In the previous section we have discussed applications where deep learning is used as an accelerator to replace existing methods. The main motivation of these works is therefore efficiency for dealing with large amounts of data. Because neural networks are universal approximators they can also be used to unveil new correlations between observables and physical quantities. In this case, deep learning is not replacing existing methods, but rather used as an exploration tool to unveil new patterns in the data which can be informative about underlying physical processes and/or about physical properties of galaxies from simulations.

4.2.1. Physical processes of galaxy formation

Deep Learning offers a new way of establishing correlations between image features and physical processes driving galaxy formation in general. The general procedure is that simulations are used to identify a physical process without ambiguity. For example, galaxy mergers are straightforward to identify galaxies in a simulation but challenging to find in observations. Mock observations can therefore be produced to train a CNN at identifying the physical process. The advantage is that the network is let free to identify the optimal features that characterise a physical process given the observables. This way of proceeding is partly new. It has been driven by both the emergence of deep learning techniques and large sets of numerical simulations.

Galaxy mergers A central example of this type of application which has caught the attention of many groups in the past years is the characterisation of galaxy mergers. Mergers of galaxies are arguably a key driver of stellar mass assembly in galaxies across cosmic time. Identifying and characterising the properties of large samples of mergers to assess their impact on diverse assembly processes has remained a key open issue in the field of galaxy formation for many years. The precise measurement of the merger rate—number of mergers per unit time and unit volume—is also an indirect probe of the cosmological model. Since galaxy mergers tend to disrupt the surface brightness distribution of galaxies because of the gravitational interaction, it is an old idea to use measurements of perturbations in the luminosity profiles of galaxies as indicators of a merger activity. A popular approach in the early 2000s was to measure some moments of light that are sensitive to asymmetries in the light distribution (e.g. Conselice 2003). The problem is that the link between these perturbations and the actual merger activity is loose. The observability timescale of a given feature such a tidal structure depends on the type of merger, the cosmic epoch and other properties, which makes it very difficult to establish a direct link between image features and merger status.

In that context, numerical simulations offer an attractive way of connecting measurements on images to a phase in the merger since the dynamical process of merger can be entirely tracked down in the simulation. Early efforts had indeed tried to establish some first order calibration using numerical simulations of galaxy pairs (e.g. Lotz et al. 2008). The work by Lotz et al. (2008) associated variations in the moments of light—concentration, Gini, asymmetry also known as CAS parameters—with the merger phases. However this was done manually and with a limited set of simulations. Snyder et al. (2019) improved on these early works by exploring Random Forests applied to moments of light on numerical simulations. More recently Whitney et al. (2021) used galaxies from the TNG simulation to calibrate the observability timescale of the so-called CAS parameters.

Deep learning offers however a new way of looking at this problem of detecting and characterising galaxy mergers by bypassing summary statistics and manually engineered features. Pearson et al. (2019b) first applied a CNN to the identification of galaxy mergers using a training set of mergers from the EAGLE simulation mimicking the SDSS observational properties. The key difference with previous works is that no manual features are extracted; the images of the different mergers are provided as input. In this first work, they found that CNNs did not achieve very high performance which is interpreted as a signature that the images used did not present significant perturbations. One possible reason for this poor performance is that mergers were selected in an non-homogeneous way over a large range of times. The importance of the training set was carefully analysed in the work by Bottrell et al. (2019). The authors explored how the performance of CNNs changed depending on the realism of the training set used for training. The main conclusion is that it is more important that images used for training reproduce the observational properties of the sample to be analysed, that is, PSF, noise, background sources, than using full radiative transfer to improve the conversion from stellar particles to light. Ferreira et al. (2020) followed up on this idea and used deep learning for the first time to compute the merger rate up to $z \sim 3$. They trained a supervised CNN to identify mergers labelled in the TN300 simulation and selected over a fixed time window of ∼1 Gyr. They showed that the CNNs can distinguish mergers from non-mergers based on multi-wavelength imaging with an accuracy of ∼90%. When applied to data from the CANDELS survey, their results reconcile measurements of the merger rate performed with pair counts and photometry (see Figure 27). These results confirm that the discrepancy between the two approaches was caused mainly because of calibration issues. Using deep learning on simulation allows one to properly calibrate the observability timescale based on the simulation metadata and therefore obtain a more reliable measurement of the merger rate. A similar approach was followed by Bickley et al. (2021), who trained a CNN to focus only on post-mergers identified in the TNG simulation. They confirm that deep learning techniques outperform moment based identifications of post-mergers and applied their model to the CFIS survey. However, the authors highlight a common problem with very unbalanced classification problems as this one. Since the number of post-merger galaxies is very small compared to the global population of

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

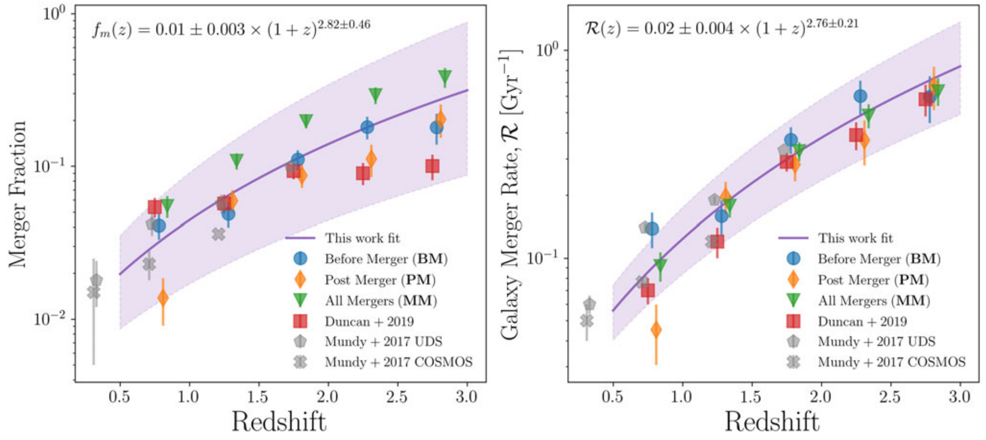

Figure 27. Supervised deep learning trained on simulations used to infer galaxy merger rates. The panels show the inferred merger fractions (left) and merger rates (right) with deep learning as compared to other independent techniques. Figure adapted from Ferreira et al. (2020)

galaxies, even a small fraction of false positives, strongly impacts the purity of the post-merger sample (see Subsection 3.1.3).

Because neural networks are very flexible, they can be easily used to combine the information of different types of inputs which is more difficult with other techniques. Bottrell et al. (2021) recently explored the combined information provided by photometry and kinematics to detect mergers. Both maps are fed into deep learning network which combines the information in an unsupervised way. Interestingly, the authors conclude that kinematics does not bring significant additional information.

Other works have attempted to go beyond the classification and use neural networks to regress on the properties of the mergers. Koppula et al. (2021) used a residual deep neural network to estimate the time to/from the first passage in a merger sequence. Using information from the Horizon-AGN simulation, the authors produce a large sample of images of major mergers at different stages. They show that, based on a single snapshot, the CNN is able to recover the position of the stamp in the merger sequence with an error of ~40 Myr. This is interesting because deep neural networks are used to provide temporal constraints on phases of galaxy formation, based on a single snapshot. A similar approach was followed by Cai et al. (2020). They trained a combination of Autoencoders and Variational Autoencoders to infer the properties of galaxy mergers. They conclude as well that with a single image, the dynamical status of the mergers can be inferred, bypassing dynamical modelling. Even more recently Eisert et al. (2022) used an invertible neural network to infer several mass assembly indicators of galaxies (i.e. mass of accreted stars, time since the last major merger) using a variety of observable quantities from the TNG simulation.

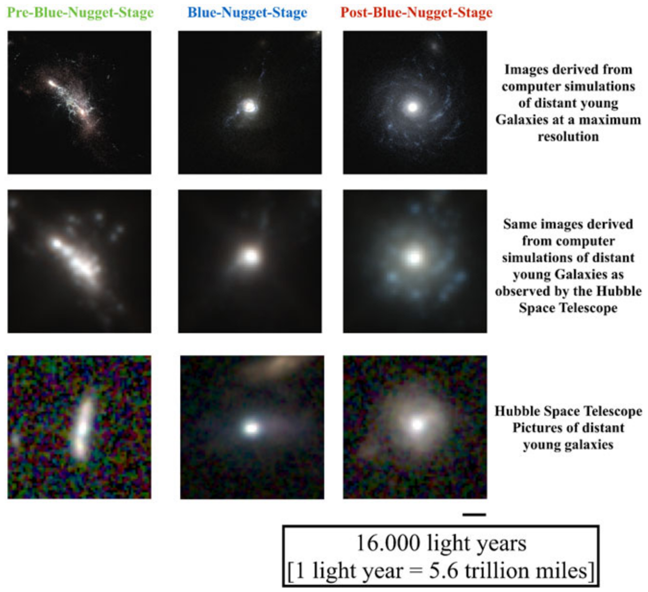
Pre-Blue-Nugget-Stage
Blue-Nugget-Stage
Post-Blue-Nugget-Stage
Images derived from computer simulations of distant young Galaxies at a maximum resolution
Same images derived from computer simulations of distant young Galaxies as observed by the Hubble Space Telescope
Hubble Space Telescope Pictures of distant young galaxies
16.000 light years
[1 light year = 5.6 trillion miles]

Figure 28. CNN applied to reconstruct to classify galaxies in different evolutionary stages defined by cosmological simulations. Each column shows a different phase of evolution as defined by the simulation. The top row is the high resolution from the simulation. The middle row shows the same galaxy with observational realism added. The bottom row shows real observed galaxies identified by the CNN as being in one of the three different phases. Figure adapted from Huertas-Company et al. (2018)

Other physical processes Huertas-Company et al. (2018) first applied this idea to the identification of a physical process called *compaction* or *blue nugget* phase. Several observational works suggested that diffuse star-forming galaxies become compact star-forming galaxies called ‘blue nuggets’ (BN) which subsequently quench (‘red nuggets’) following a sudden gas inflow towards their centre. The VELA zoom-in simulations (Ceverino et al. 2015) show also rapid gas inflows leading to central starbursts, and several mechanisms were identified that lead to this compaction phenomenon including major gas-rich mergers or disk instabilities often triggered by minor. The authors tested whether deep learning can detect the blue nugget phase of galaxy evolution purely identified in the simulation metadata. Mock HST images from the simulated galaxies were labelled according to their evolutionary stage from the simulation metadata (before, during, or after the BN phase), that is, the labelling is exclusively done based on physics. The result is that galaxies can be successfully classified into evolutionary stages without identifying specific features, that is, just using the pixel space (see Figure 28).

Diaz et al. (2019) applied a similar idea to the classification of formation mechanisms of lenticular galaxies. They identified different evolutionary tracks leading to the formation of S0 galaxies—isolated, tidal interaction in a group halo, and spiral-spiral merger—and trained a CNN to identify them based on stellar density and two-dimensional kinematic maps. They found that the CNNs are able to distinguish the different formation scenarios and conclude about the potential of deep learning to classify galaxies according to their evolutionary phases. Schawinski, Turp, & Zhang (2018) used a Generative Adversarial Network to try constraining the physical processes leading to the quenching of star formation in galaxies. Ginzburg et al. (2021) used a CNN trained on zoom-in cosmological simulations to infer the longevity of star-forming clumps in high redshift galaxies. Instead of relying on the conversion between photometry and age of the stellar populations, they defined two types of clumps—short and long lived—based on the information from the simulation and trained a supervised neural network in a binary classification mode to distinguish between the two types. This is yet another example where neural networks are used as universal approximators to find not obvious links between observables and physical properties.

Open issues Using deep learning trained on simulations to constrain the phases of galaxy formation is becoming increasingly popular in the community and the example of mergers clearly illustrates this. All these approaches suffer however from an obvious limitation, the so-called domain gap between observations and simulations. As this is a recurrent problem, more information is provided in Section 7. Since simulations do not perfectly reproduce observations, applying a trained network on observations will induce some biases. Moreover, since observations are generally not labelled at all, especially when trying to infer physical properties, it is impossible to evaluate the effect of the domain gap and the results need to be accepted blindly. As already mentioned in previous sections, including uncertainty quantification in the neural networks is an option to mitigate this effect. However the error induced by changes in the domains is in general difficult to capture by uncertainty quantification methods. Other options consist in performing the domain adaptation during training to ensure that the features learned by the neural networks are not specific to the simulation domain. There are different approaches to do so since it is a problem that exists in many fields of application (see e.g. Wang & Deng 2018). In extragalactic astronomy there has been limited exploration of these approaches. Ćiprijanović et al. (2021b) recently explored the impact of domain adaptation during training for the identification of galaxy mergers. They tested different domain adaptation techniques such as Maximum Mean Discrepancy and Domain Adversarial Neural Networks and concluded that including these leads to an increase of target domain classification accuracy of up to ∼20%. This is a promising result for future applications of neural networks trained on simulations. It is likely that future works will start to incorporate these approaches more often.

4.2.2. Dark matter

In a similar spirit of finding new correlations, deep learning has been increasingly used in the past years to estimate the dark matter masses and properties of galaxies and clusters, based on observable quantities.

Dynamical masses of clusters of galaxies The earliest applications focused on galaxy clusters, which are the largest gravitationally bound objects. This is because there exist alternative methods to measure dynamical masses of clusters. However standard approaches are based on simple scaling relations based on the virial theorem. The simplest approach is to use a power law relation between the dispersion of the line of sight velocities (LOS) and the cluster mass. This was indeed one of the first probes of dark matter by Zwicky (1933). Other classical approaches consist in using scaling relation between the X-ray luminosity of the gas and the mass of the cluster or the Sunyaev-Zeldovich effect. However, these scaling laws are all based on strong assumptions about the physical status of the cluster, the most important being that the system is in virial equilibrium. This entails some inherent biases in the estimated cluster masses.

Deep learning and machine learning in general, calibrated on simulations where dark matter properties are known, offer an interesting approach to look for additional correlations which could help in reducing the scatter in the dynamical mass estimates. Early efforts were done before the emergence of deep learning, especially in the works by Ntampaka et al. (2015, 2016). The authors explored Support Distribution Machine class of algorithms to predict cluster masses using LOS velocities and radial positions of cluster members. They reported an improvement of a factor of two (see also Armitage, Kay, & Barnes 2019 for similar conclusions).

Ntampaka et al. (2019) applied CNNs to cluster mass estimation in which is the first work using deep learning for this purpose. The training is performed with mock 2D X-ray images of Chandra observations. There is indeed a well known correlation between X-ray luminosity and cluster masses. They report a ∼10% smaller scatter than standard luminosity based methods, even without using any spectral information. Interpretability techniques based on attribution methods, reveal that the CNNs tend to ignore the cluster centres because they likely lead to more biased estimates. This is another example of CNNs used to find new correlations in the data and one of the few examples were basic attribution techniques provide meaningful information. Similarly, Yan et al. (2020) used a combination of feature maps (stellar mass, soft X-ray flux, bolometric X-ray flux, and the Compton y parameter) and reach comparable results. This work illustrates another advantage of deep learning over traditional approaches, which is the possibility of combining multiple observables in a transparent way.

Along the same lines, Ho et al. (2019) explored CNNs to estimate cluster masses. The training is based on LOS velocities and radial positions of galaxies in the cluster. They explored both 1D and 2D CNNs. They also report a factor of ∼2 improvement with respect to power law-based estimates. Interestingly CNNs also improve the results of more classical ML approaches explored in previous works.

Convolutional Neural Networks have also been explored on the third observable usually employed to infer cluster masses, the Sunyaev-Zeldovich effect. de Andres et al. (2021) used CNNs on mock maps of the Planck satellite from numerical simulations. The advantage of using deep learning is again that no assumptions on the symmetry of the cluster's gas distribution nor on the cluster physical state are made.

These previous works, although promising, remain at the exploratory level and suffer from the same limitations than other similar approaches. Namely, the results are restricted by the prior inferred from the simulations and they usually lack of uncertainty estimation with some exceptions. The works by Kodi Ramanah et al. (2020), Kodi Ramanah, Wojtak, & Arendse (2021) make a step forward to address some of these issues. In these works, the authors explore, for the first time, flow based neural networks (see Figure 19) trained of the phase space of cluster galaxies to infer cluster masses (Figure 29). The key addition of their approach is

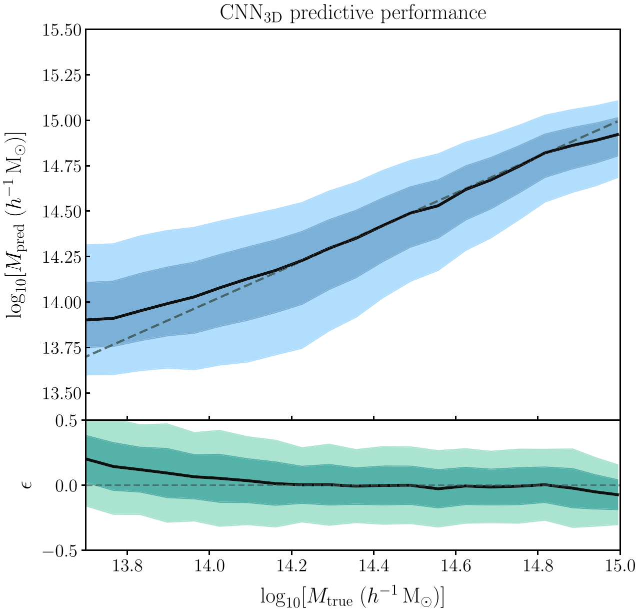

Figure 29. The figure shows the predicted dark matter mass as a function of the true mass along with uncertainties using a Normalising Flow. Figure adapted from Kodi Ramanah et al. (2021)

that the network provides therefore a full probability distribution of the cluster mass instead of a single point estimate, which can therefore be used to account for uncertainties. This is a key step forward towards an application of deep learning based approaches for estimation of cluster masses in large surveys. The authors claim a factor of 4 improvement compared to scaling relations based estimations. They then apply their model to a sample of observed clusters with well calibrated dynamical masses and show that the neural network provides both unbiased measurements and well calibrated uncertainties. Ho et al. (2021) also investigate the use of Bayesian CNNs to include uncertainty measurements. They show that BNNs recover well calibrated 68% and 90% confidence intervals in cluster mass to within 1% of their measured value.

Halos of galaxies Deep learning can be also extended to estimate dark matter halo masses of less massive galaxies than clusters. In low mass halos, there are also less galaxies and therefore it is more challenging to use LOS velocities. They do not present any X-ray emission either. The most standard way to proceed is by using Abundance Matching techniques. Abundance Matching main assumption is—with some variations—a monotonic relation between the stellar masses of galaxies and dark matter halos. By using dark matter halo mass functions from N-body simulations, one can then assign halo masses to galaxies. In this context, deep learning can be used to look for additional correlations between galaxy properties and dark matter beyond simple abundance matching assumptions. Calderon & Berlind (2019) explored several machine learning algorithms, including neural networks, for estimating the dark matter halo masses of galaxies in the SDSS. They used ML to explore how much information about halos is provided by several galaxy properties from synthetic catalogs. They conclude that including more physical properties is translated into a better accuracy than the one reached by abundance matching. A problem with this approach, acknowledged by the authors, is that secondary dependencies of halo masses on galaxy properties might be very model dependent. This can

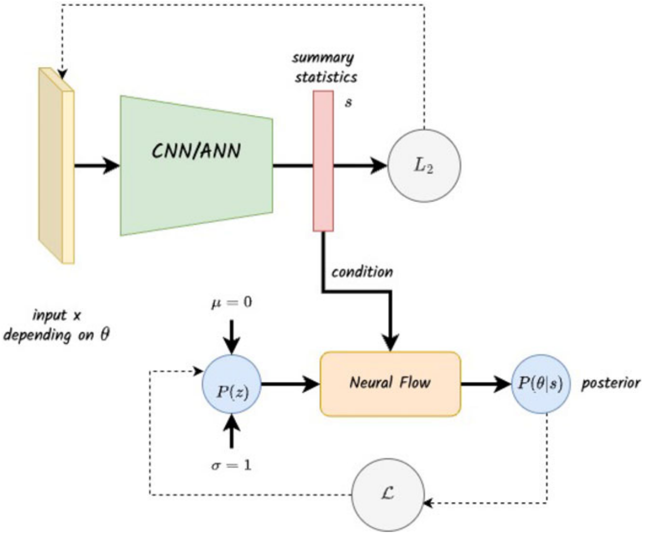

Figure 30. Illustration of a Neural Flow for estimating posterior distributions. This type of approach is starting to become common in simulation based inference approaches for estimating cluster masses or cosmological parameters. A first Neural Network with a standard L_2 loss is used to obtain some summary statistics of the data which are then use as a condition for a Neural Flow mapping a simple distribution into an approximation of the posterior.

therefore induce systematic biases in the inferred dark matter masses which do not exist in simpler approaches. Shao et al. (2021) investigated how sub-halo masses can be estimated using neural networks trained on a number of physical properties of galaxies (i.e. black hole mass, gas mass, stellar mass etc) from numerical simulations. They used the CAMELS simulation suite which is a series of numerical simulations performed with different codes and cosmologies, specially designed for ML. We will describe the simulations into more detail in Subsection 6.2. They found that sub-halo masses can be predicted accurately (~0.2 dex) at any redshift from simulations with different cosmologies, astrophysics models, subgrid physics, volumes, and resolutions. The authors argue that the neural networks might have found a universal relation, which turns out to be a generalised version of the virial theorem involving radius, velocity dispersion and maximum circular velocity. This is a good example of deep neural networks used to find hidden correlations which can be even translated into analytical expressions. We will discuss this further in Section 5. In a recent work, Villanueva-Domingo et al. (2021a) explored the use of Graph Neural Networks (GNNs) to estimate halo masses of galaxies. GNNs are a special type of neural networks that are built on graphs, and therefore allow one to account for the relations between neighbouring halos. The authors used the halos from cosmological simulations as nodes of the graphs and encoded the gravitational interaction between them in the edges of the graph. The nodes include the positions of the halos, the relative velocities, the stellar mass and the half-mass radius. They show that the model is able to estimate halo masses with a ~0.2 dex uncertainty (Figure 31). The model is also built to account for uncertainties and is shown to generalise reasonably well between different simulated datasets.

In a follow-up work (Villanueva-Domingo et al. 2021b), the authors apply the trained GNNs to measure the halo Way and Andromeda galaxies. They show that the inferred constraints are in good agreement with estimates from other traditional methods.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

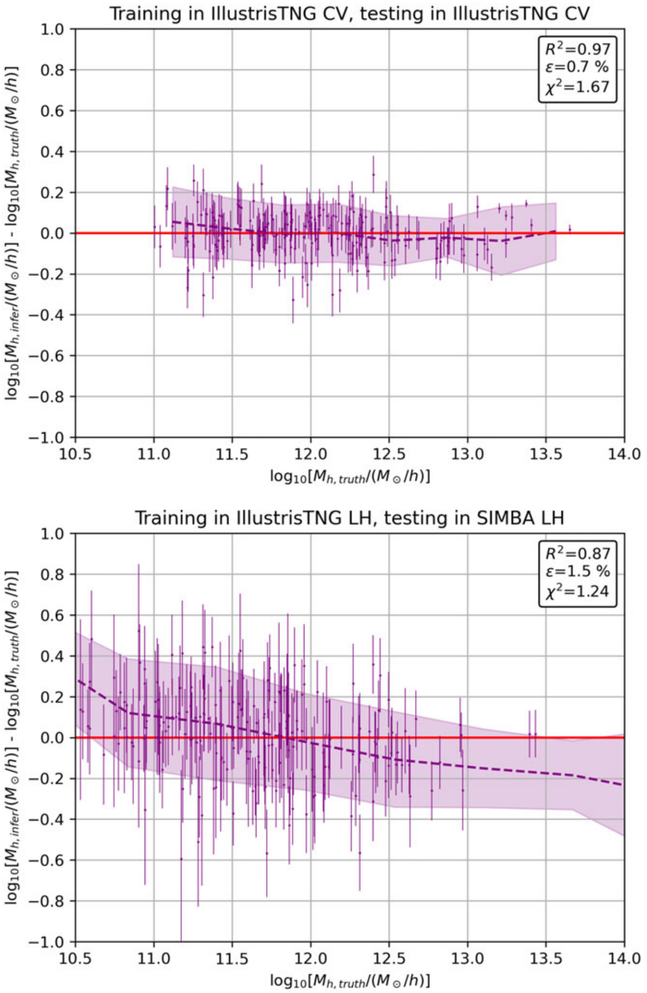

Figure 31. The top panel illustrates the accuracy obtained on simulations, when the training and testing is done on datasets coming from the same underlying cosmological simulation. The bottom right panel is when two different simulations are used for training and testing respectively. Figure adapted from Villanueva-Domingo et al. (2021a)

4.2.3. *Deep learning generated observations*

Some recent works have tried to push even further the ability of neural networks to establish not obvious non-linear mappings between domains to bypass telescope observations. Spectroscopic observations are more expensive in terms of observing time than imaging ones. However, spectroscopy is significantly richer in terms of available information about astrophysical processes. Some works have therefore explored if deep neural networks can infer spectra from images without the need for observing time. Wu & Peek (2020) showed that deep neural networks efficiently learn a mapping between galaxy imaging and spectra. They showed that SDSS spectra can be inferred from 5-band images with very high accuracy. The authors argue that this approach could be used as an exploration tool for future surveys such as LSST. Holwerda et al. (2021) followed up on this work by applying the same network to estimate the spectrum of an AGN and compare it to existing spectra from the literature. Although the neural network predicted spectrum is in overall good agreement with the observed ones there are some differences in the strength of emission lines. This suggests that the ML approach might be an interesting way of identifying specific types of objects such as AGN and/or as a first order exploration. It is difficult to use the inferred spectra to analyse the physical properties though. This is somehow expected since spectra contain more information than imaging, as previously stated. A similar experiment is performed by Hansen et al. (2020) who infer kinematic information of galaxies (velocity dispersion and rotation maps) from single band imaging.

Summary of Deep learning for inferring physical properties of galaxies

- Over the past years, deep learning has been tested as a tool to infer physical properties of galaxies in large surveys. The most common applications are: photometric redshift estimation, galaxy structure and stellar populations and strong lensing modelling.
- The main motivation for using deep learning for these tasks is computational speed. Deep learning is used here as a fast emulator for existing methods. Overall, the most common conclusion of these tests is that deep learning approaches achieve state-of-the-art performance, several orders of magnitude faster.
- The typical approach used is a supervised regression (C)NN trained on simulated datasets, for which the ground truth is known.
- A major challenge of these applications is robustness. Since the models are predominantly trained on simulations, the representativity of the training set is a major issue. Extrapolation with neural networks is in general problematic. Therefore, making sure that the training samples properly cover the inference dataset is a key challenge.
- Uncertainty quantification is also particularly important for this type of applications. Bayesian Neural Networks and Density estimators are among the most commonly employed solutions.
- As an extension to the derivation of physical properties, deep learning has also been explored as a tool to identify new correlations between observable and other physical properties of galaxies, which are generally not accessible with existing methods. Examples of this type of application include the inference of phases of galaxy evolution such as interactions, or the estimation of the dark matter content of galaxies.
- The main motivation is that neural networks are universal approximators. Deep learning is therefore employed to unveil hidden correlations using simulation based inference. By definition, the training is performed on (cosmological) simulations in which all the physical properties and evolutionary phases of galaxies are accessible.
- A key challenge for these type of approaches is robustness against the representativity of training sets and domain shifts. The effect of these issues is particularly dramatic here since cosmological simulations are known to be approximations to the observed universe. Moreover, in general, *pre-deep learning* approaches to compare with do not exist, as opposed to the previous type of application.
- Because deep learning is typically used *blindly* informed by simulations, interpretability becomes a key limitation. Solutions based on saliency maps or symbolic regression—when possible—have been explored. However these approaches still present a limited informative power. There is significant room for improvement in the future.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

5. Deep learning for discovery

In this section, we focus on efforts done by the community to use deep learning as a discovery tool. Applications typically include dimensionality reduction to visualise complex datasets and identify groups of objects, anomaly detection to automatically find potentially interesting objects in large datasets and some early efforts to automatically learn fundamental laws of physics.

5.1. *Visualisation of large datasets*

In addition of increasing in volume, datasets in astrophysics are becoming increasingly complex and of high dimensionality. Machine Learning can be employed to visualise datasets in a low dimension space to look for trends and correlations in the data, which otherwise are difficult to extract. It can also be used to identify classes of objects that share some properties which can help with the scientific analysis. These applications use typically unsupervised learning approaches, as opposed to what has been previously discussed. In unsupervised learning, data is unlabelled and therefore we seek a representation of the data instead of a mapping between data points and labels. We emphasise again that the present review focuses on deep learning applications and, therefore, we will not describe in detail works using other ML approaches for data visualisation. There exist however a large variety of techniques which do not involve neural networks and that have been applied to astronomy. For example, Baron & Ménard (2021) used graph representations to find structures in imaging and spectroscopic data. The works by Hocking et al. (2018) and Martin et al. (2020) also explore clustering coupled with graph representations to group images of galaxies that look similar. Self-Organising Maps have also been used to represent images of galaxies in the radio domain (see e.g. Galvin et al. 2020) and for spectral classification (e.g. Rahmani, Teimoorinia, & Barmby 2018). Other non-neural network-based dimensionality reduction techniques such as Principal Component Analysis (PCA), t-SNE (van der Maaten & Hinton 2008) or UMAP (McInnes, Healy, & Melville 2018) are also used in several works to explore data.

Deep learning offers several approaches for dimensionality reduction and visualisation. The most standard and widely used are Autoencoders, a particular type of encoder-decoder architectures (see Section 3 and Figure 12) in which the inputs and outputs are identical. The networks therefore simply learn how to reproduce the input data. However, since these architectures—which can be deterministic or probabilistic—typically present a bottleneck at the junction between the encoder and the decoder, they naturally represent the data in a low dimension space. Exploring the distribution in this bottleneck layer is useful to find structures in the data. Ma et al. (2019) used a Convolutional Autoencoder (CAE) to explore a sample of radio active galactic nuclei (AGNs). By feeding the CAE with images of radio AGNs with hosts of different morphologies, they showed that the network naturally clusters similar objects together in the bottleneck. In that particular work, the low dimensional representation is also used for a downstream supervised classification using the learned feature space as input for a supervised network. This is also a common application of the dimensionality reduction approach. When only a reduced subsample of labelled examples is available, the reduced dimensionality space can be used to train a supervised network with smaller training sets. A similar approach was followed by Cheng et al. (2020). They used a CAE to represent images of galaxies in a low dimension space with the goal of finding strong gravitational lenses. Since the samples of labelled lensed systems are typically small (see Subsection 3.1.3 for more details), unsupervised representation offers an alternative way to find lenses without labels. The authors perform a clustering step in the latent space learned by the neural network to automatically find groups of objects which similar properties. They find that the CAE based method successfully isolates ~60% of all lensing images in the training set. In Cheng et al. (2021a) they extend the same approach to the unsupervised exploration of galaxy morphology. Using a modified version of a Variational Autoencoder, they obtain an unsupervised representation of nearby galaxies from the SDSS survey. They then perform a hierarchical clustering in the latent space to conclude that the neural network representations share some properties with the classical Hubble sequence but provide a more meaningful representation, especially for ambiguous intermediate morphological types. See also the work by Spindler, Geach, & Smith (2021) for an application of VAEs to representation of galaxy morphology. A similar approach is presented in the work by Zhou et al. (2021). The authors apply a combination of CAE based representation with a multi-clustering model to study the morphologies of high redshift galaxies from the CANDELS survey. Portillo et al. (2020) also used the same type of approach involving a Variational Autoencoder to represent spectra of nearby galaxies. The authors projected SDSS spectra into a latent space of 6 dimensions and showed that the different types—that is, star-forming, quiescent—are naturally separated without labelling (Figure 32). Interestingly, the non-linear components of VAEs seem to enable a better separation than a simple PCA decomposition if the latent space remains of dimension lower than ~10. The conclusion is that dimensionality reduction with neural networks is a sensitive way of exploring data of high dimension. Notice however that they did not use convolutional layers.

In a recent work, Teimoorinia et al. (2021) follow a similar approach, but with additional layers of complexity, to explore Integral Field Unit (IFU) data from the Manga survey. In their approach, called DESOM, the authors propose to combine a convolutional Autoencoder and a SOM to represent spectra. The spectra are fed to a CAE and projected into a latent space of lower dimension. In the same training loop, the representations are used to train a SOM that further clusters similar objects. That way, all spectra of a galaxy can be passed into the machinery to obtain a *fingerprint* for every object which corresponds to the final projection into the SOM plane. The authors propose going a step forward by passing again the obtained *fingerprint* into the DESOM to obtain a single location for a galaxy based on the 2D distribution of all spectra belonging to the same galaxy.

Other works have also applied deep learning based dimensionality reduction to assess data quality. For example, Mesarcik et al. (2020) used an Autoencoder to explore radio data and identify possible technical failures in the observations. This illustrates another interesting use of deep learning dimensionality reduction techniques to quickly explore complex datasets.

Another deep learning approach for dimensionality reduction which is increasing in popularity in the recent years, is what is generally known as self-supervised learning through contrastive learning. As opposed to the Autoencoder approach, where the

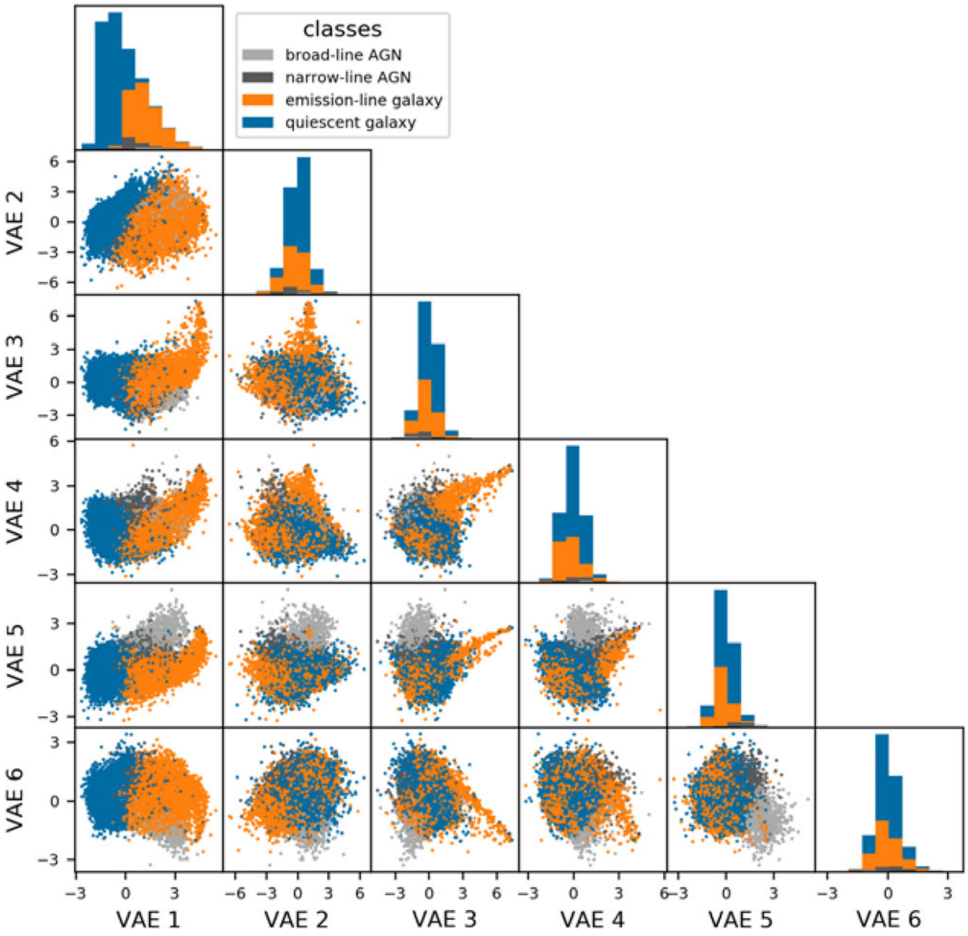

Figure 32. Variational Autoencoder for dimensionality reduction and data visualisation. The figure show how the different types of spectra (labelled with different colours) populate different projections of the latent space. Figure adapted from Portillo et al. (2020)

projection depends on the architecture, in contrastive learning, the computation of representations is more data oriented. The general idea is to apply some perturbations to the input data so that the networks learn to ignore those and cluster together data points coming from the same parent input data. This is obtained by what is called a contrastive loss term. We emphasise that it is not the goal of this review the technical details of the different deep learning approaches, but to review how they are being used in astronomy. We refer the reader to Chen et al. (2020) and references therein for more details (see Appendix A for references on the different deep learning methods mentioned in this review). Figure 33 shows a very schematic representation of a contrastive learning setting. The output is in essence similar to the one obtained with an Autoencoder—that is, a representation of data in a reduced latent space—but the underlying idea is significantly different. One of the key advantages of a contrastive approach is that the perturbations applied to the input data can be tuned for a science case and turn the representations independent to a known undesired effect. In astronomy, it can enable to mitigate the effects of instrumental or selection biases for example. Contrastive learning has only started to be applied in astrophysics relatively recently. The first work exploring self-supervised learning is by Hayat et al. (2021). The authors used an existing network to compute representations for multi-band SDSS images. Among the perturbations applied to the images, they included standard rotations and cropping, but also some adapted to astronomy, such as extinction. They showed that the contrastive learning model successfully clusters galaxies with similar morphological properties and therefore constitutes a promising way for data exploration in astrophysics (Figure 34).

Sarmiento et al. (2021) also applied contrastive learning to visualise data of nearby galaxies from the Manga survey. Instead of images, they used post-processed maps of stellar populations properties (metallicity, age) as well as stellar kinematic maps. They also show that the self-supervised learning setting is able to condense the information from this high-dimensional dataset into a subset of meaningful representations which contain information about the physical properties of galaxies. Interestingly, this is a case in which other simpler dimensionality reduction techniques such as PCA, or even Autoencoders, fail, given the large amount of instrumental effects present in the data. The authors show that more standard techniques tend to organise galaxies based on properties of the instrument (i.e. fibre size) instead of physical ones. Because the contrastive setting allows one to tune the augmentations to a specific problem, it can be trained so that the representations become independent of instrumental biases

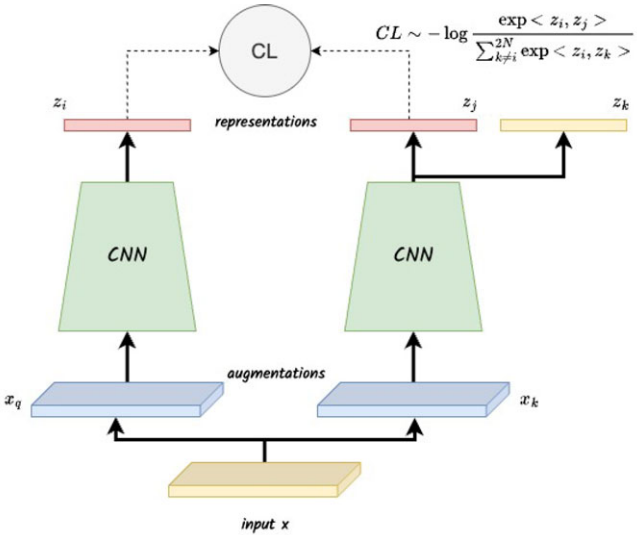
$CL \sim -\log \frac{\exp < z_i, z_j >}{\sum_{k \neq i}^{2N} \exp < z_i, z_k >}$
CL
z_i
representations
z_j
z_k
CNN
CNN
augmentations
x_q
x_k
input x

Figure 33. Illustration of a self-supervised contrastive learning architecture. Multiple random augmentations of the same image (positive pairs) are fed to two different CNNs which map them into a latent representation space. Also during training, pairs of completely different images (negative pairs) are also fed to the two CNNs. The contrastive loss is optimised to increase (decrease) the dot product of representations of positive (negative) pairs. Contrastive learning is starting to be used for dimensionality reduction and as a generalised feature extraction process for multiple downstream tasks such as galaxy classification or photometric redshift estimation.

(Figure 35). The inferred representations can be used for example to perform a clustering step and identify different classes of objects. Sarmiento et al. (2021) show that well-known types of galaxies appear naturally without any human supervision from a data-driven perspective.

In addition to visualisation, a common application of self-supervised representations is to use them as input for other downstream tasks. For example, the representations can be used for a subsequent supervised classification. The fact that objects have already been clustered together, helps to converge faster, which makes it especially appealing when a small amount of labelled data is available (see Section 3). Hayat et al. (2021) showed indeed that, by using the latent space for morphological classification of galaxies, they can reach a similar accuracy as with a fully supervised CNN but using >10 times less labelled data. In the follow-up work by Stein et al. (2021a), they also explore how the self-supervised representations can be used to find strong gravitational lenses reaching similar conclusions. Similarly to the work by Cheng et al. (2020), the representations are used with a small sample of lenses to train a simple linear classifier and find new strong lenses candidates.

5.2. Outlier detection

Anomaly or outlier detection is a fundamental aspect of discovery in physical sciences. A fair amount of new results in astrophysics have been triggered by serendipitous discoveries through the exploration of observational datasets. With the arrival of new big-data surveys, finding potentially interesting objects becomes increasingly difficult with purely human-based approaches. In the past year, unsupervised deep learning has been explored by several groups as a way to assist astronomers in the search of potentially interesting objects.

An anomaly or outlier is usually defined as a data point which properties deviate from the average properties of objects in the sample, under some metric. The visualisation techniques described in the previous subsection, which tend to cluster together data points with similar properties, can therefore be useful as well to identify deviant objects.

From a probabilistic point of view, an outlier can be also defined as an object which probability of observation under the probability density distribution of a data set is smaller than a given ϵ. In that context, modern generative models (e.g. VAEs, GANs), can approximate the probability density function $p(X)$ of a dataset X with increasing accuracy. Therefore, they can be employed to look for objects with a small probability of observation (see Chalapathy & Chawla 2019 for a generic review of deep learning techniques applied to anomaly detection).

5.2.1. Transient astronomy

The field of transient astronomy has been particularly active in this context. As previously summarised, the field is about to experience a data revolution. The forthcoming LSST survey will observe all the Southern Hemisphere sky every $\sim 2-3$ nights, producing an unprecedented real time *movie* of the night sky. The community expects to discover a significant amount of new types of variable objects using this dataset (Li et al. 2022). Therefore there have been over the past years a number of works exploring deep learning and machine learning in general, to identify anomalous light curves in preparation of LSST We emphasise again that deep learning is not the unique machine learning approach to identify anomalies. Malanchev et al. (2021) performed a comparison of several anomaly detection algorithms—isolation forests, on-class SVMs, Gaussian Mixture Models and Local Outlier Factor—to identify outliers in the Zwicky Transient Facility (ZTF). See also the work by Martínez-Galarza et al. (2021) which used decision trees and manifold learning. Pruzhinskaya et al. (2019) uses Isolation Forests as well on a set of features derived from the light curves using interpolation with Gaussian Processes. In the following, we will however focus on efforts relying on deep learning.

Villar et al. (2021b) used a Variational Recurrent Autoencoder (VRAE) network described in Villar et al. (2020) to identify anomalous light curves from the simulated PLAsTiCC dataset. The proposed methodology is based on three main steps involving three different ML algorithms. First the light curves are interpolated using Gaussian Processes (GPs). The resulting interpolations are then fed into a Variational Autoencoder. The temporal aspect is encoded by appending the time step to the elements representing the time series. The low dimension representation of the time series is finally passed through an Isolation Forest algorithm to assign an anomaly score. As opposed to well defined supervised problems, evaluating and comparing anomaly detection algorithms is always difficult since by definition the objective is not well defined. In that particular work, the authors quote a $\sim$95% purity in identifying light curves others than the ones generated by well-known types of objects. However, by definition the exercise is incomplete, since the sensitivity to unknown unknowns cannot be assessed. This work illustrates an interesting way of combining multiple ML approaches though. In particular, the introduction of a GP for preprocessing turns the model agnostic to the sampling frequency of the time series which is a very interesting feature for astronomical applications. A Variational

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

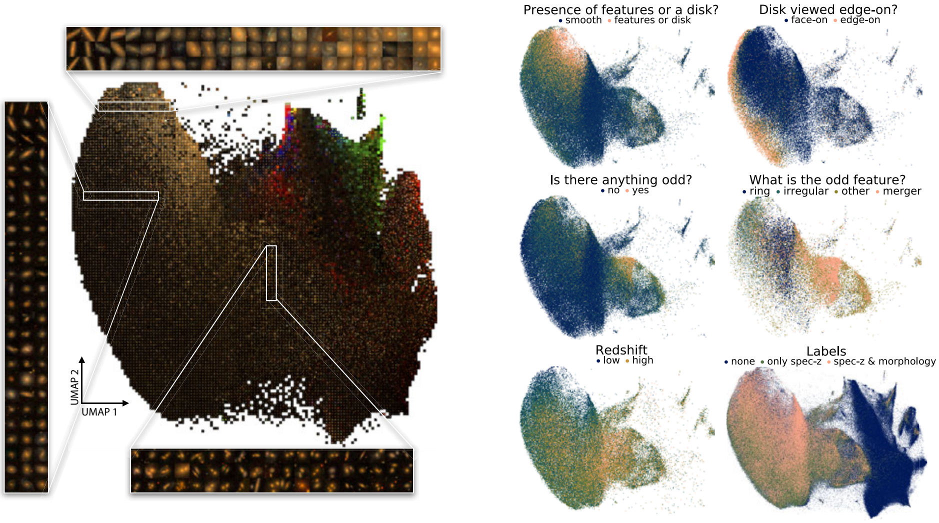

Figure 34. Self-supervised learning applied to multi-band SDSS images. The left panels shows a UMAP of the representations obtained with contrastive learning. The panels on the right show the same UMAP colour coded with different galaxy properties. Similar images are clustered together. Figure adapted from Hayat et al. (2021).

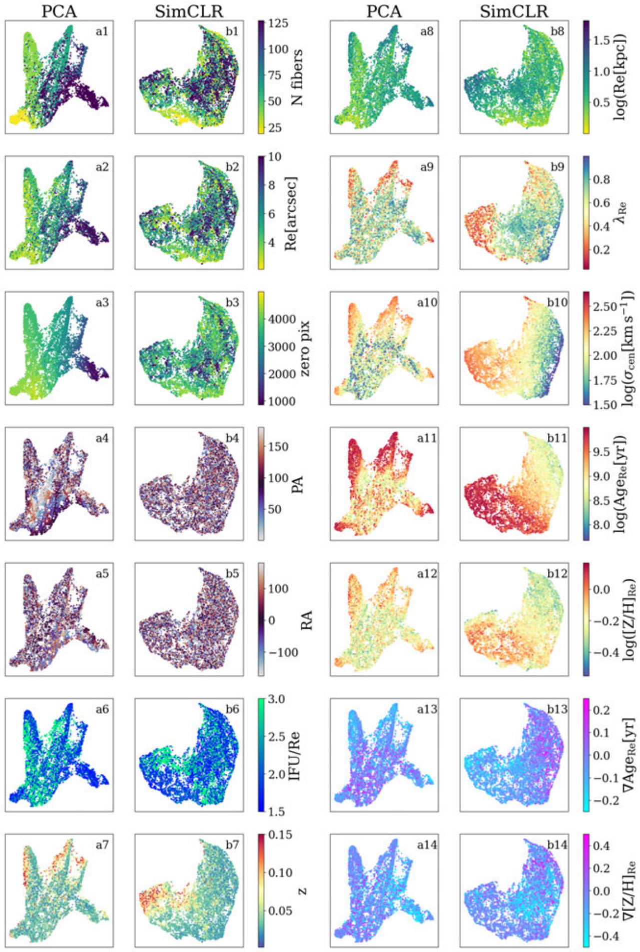

Figure 35. Representations of Manga maps using PCA and contrastive learning, and projected into a UMAP. The two leftmost columns show the plane colour coded with non-physical parameter (e.g. number of fibres in the IFU). The rightmost columns show the same maps colour coded with physical properties. Self-supervised representations cluster galaxies according to physical parameters while PCA focuses mostly on the number of fibres. Figure adapted from Sarmiento et al. (2021).

Recurrent Autoencoder is also used by Sánchez-Sáez et al. (2021) to identify changing-look Active Galactic Nuclei. The approach is analogous but it is applied to real observations from the ZTF survey. The VRAE is trained with light curves to obtain a representation in a low dimension space and then Isolation Forest is used to associate anomaly scores. Boone (2021) also employs a VAE to learn how to reconstruct light curves and then uses the latent space to assign anomaly scores to potentially deviant curves. The architecture used is slightly different than in the previous two works. Namely, they add a layer introducing some physical information about the light curve so that the setting can also be used for classification. However, the overall idea is analogous in essence.

Muthukrishna et al. (2021) explored a different approach. They trained instead an Autoregressive Generative Model to generate three known types of light curves (SNIa, SNII, SNIb). They try then to use the trained models to reconstruct other light curves and use as anomaly score the χ^2 difference between the input light curve and the reconstructed one. The underlying idea is that *common* light curves will be well reconstructed by the Neural Network if they have properly learned $p(X)$—the probability density function of the data distribution—while rare events will have larger reconstruction errors. Interestingly, they find that Autoregressive models used that way are not very efficient to identify outliers as compared for example to a Bayesian reconstruction of light curves. The explanation put forward is that neural networks are *too* efficient and are therefore able to generate, with descent accuracy, even light curves which were not part of the original dataset (Figure 36). This behaviour of AutoRegressive models has also been reported in the ML community (Ren et al. 2019). These models are indeed able to easily reconstruct less structured data than the data used for training, leading to small out-of-distribution probabilities. Some solutions have been suggested, which will be discussed in the following subsection.

Although these works are very recent and the community is still at an exploration phase, they confirm that outlier detection is in general a very complex task. Deep learning offers interesting options, especially because of the potential it has to accurately model the probability distribution of the data. However, there is

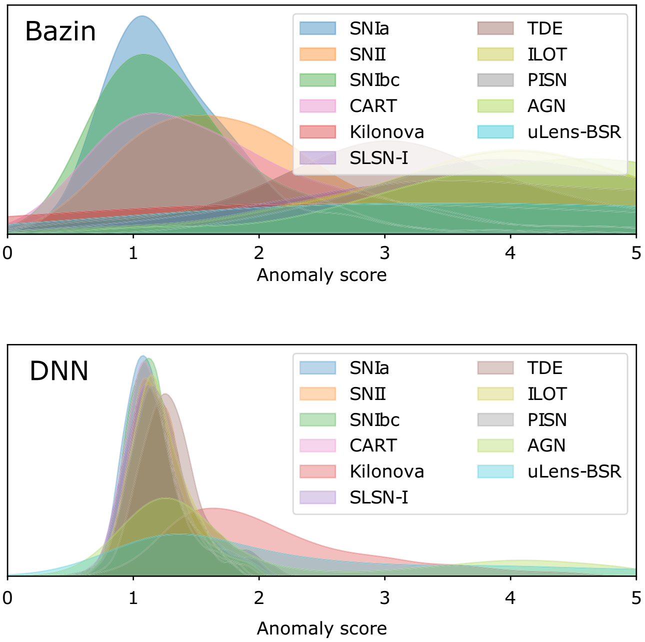

Figure 36. Anomaly scores for different types of light curves obtained with deep AutoRegressive Model (bottom panel) and with a Bayesian Reconstruction algorithm (top panel). Unknown light curves not used for training have larger anomaly scores when using the Bayesian method than with the Neural Network. Figure from Muthukrishna et al. (2021).

still no satisfactory solution available and all require some level of human interaction to identify the most interesting objects. This is eventually a structural issue, since the definition of *interesting* depends on the scientific case. Following these conclusions, the group of M. Lochner and collaborators have created dedicated tools to apply different outlier detection methods and enable a human inspection of potentially interesting candidates (see Lochner & Bassett 2021). One interesting feature this approach puts forward is the fact that another layer of ML is added to adapt the definition of *interesting objects* to each user.

5.2.2. *Imaging and spectroscopic outliers*

Anomalies can also be found in *static* data, that is, spectra or images of galaxies. This is again specially relevant in the context of future large imaging and spectroscopic surveys in which a manual inspection of all data points is prohibitively time consuming. Therefore the community has also explored several approaches to identify outliers in large imaging surveys. The underlying idea is analogous to what has been described for time series, that is, identifying objects which present some deviations from the general properties of the sample. Here again, there exist a large number of ML algorithms that can be employed; many of them not based on deep learning. For example, the recent work by Shamir (2021) uses a set of manually engineered features to find outlier candidates in HST images. Dimensionality reduction techniques such as Self-Organising Maps have also been explored as a way to identify anomalous spectra (Fustes et al. 2013). Baron & Poznanski (2017) used unsupervised Random Forests to isolate the rarest spectra in the SDSS by using individual fluxes as input. As done for the previous sections, we will focus here on deep learning applications.

Storey-Fisher et al. (2021) trained a Generative Adversarial Network using postage stamps of observed galaxies from the Hyper Suprime Cam (HSC) survey. They selected galaxies above an apparent magnitude limit and trained a Wassertsein Generative Adversarial Network (WGAN) to generate realistic images of galaxies. The underlying idea is that the model will learn how to accurately reproduce common galaxies but will fail when confronted to objects which appear with a small frequency in the training set. Once trained, WGANs do not provide an explicit latent space to sample. In order to associate anomaly scores to all galaxies, the authors perform an iterative search to identify the closest object that the WGAN can generate. They compute then an anomaly score based on a combination of the quadratic difference between the real and reconstructed image and an additional L2 difference of the features of the last layer of the critic network of the WGAN. They show that the framework is able to identify potentially interesting objects. However, a significant fraction of them are only image artefacts. The authors propose to add another dimensionality reduction layer with a CAE trained on the residual images (difference between the WGAN reconstruction and the original image). They show that, after this additional step, the different types of anomalies cluster together and a visual inspection is proposed to identify the most interesting anomalies (Figure 37). Interestingly, the work also compares the anomalies obtained with a less complex approach based on a CAE to reduce the dimension of the data. The WGAN is able to find more subtle anomalies because of the improved quality of the reconstruction. However, Tanaka et al. (2021) showed on the same dataset, that a Convolutional Autoencoder is also able to identify interesting anomalies. They quantify the performance of the anomaly detection algorithm on a set of known extreme emission lines galaxies and quasars. They report that ~60% of the objects belonging to these under represented classes are identified.

Margalef-Bentabol et al. (2020) use the same WGAN approach outlined in Storey-Fisher et al. (2021) but in a slightly different context. In this work, the anomaly detection setting is used to assess the realism of galaxies produced by cosmological simulations. In that context the WGAN is trained with mock images from simulations. The trained model is then confronted with real observations from the HST. The authors compare then the anomaly scores from both datasets and conclude that the neural network struggles to reconstruct some of the observed galaxies, meaning that they do not exist among the simulated galaxies. Zanisi et al. (2021) also explores whether anomaly detection approaches with deep learning can be used to compare galaxy images from cosmological simulations to observations. They use however a different approach based on the Autoregressive Model pixelCNN. Similar to GANs, pixelCNN is a generative model which can be use to learn the probability density function of data and generate new samples. However it provides an explicit expression of the likelihood function built in an Autoregressive fashion; that is, the values of a given pixel are determined based on the values of the previous ones. They apply the method to the comparison of SDSS and TNG galaxies. Similarly to what was reported by Muthukrishna et al. (2021) for time series, they find that the model can easily learn to generate simple objects and therefore very smooth galaxy profiles or even pure noisy images achieve high likelihoods of observation under the regressive model. To correct for this effect, the authors use instead the likelihood ratio presented in Ren et al. (2019) which forces the metric to become sensitive to the fine grained structure of galaxies. They can then show how the likelihood ratio metric is

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

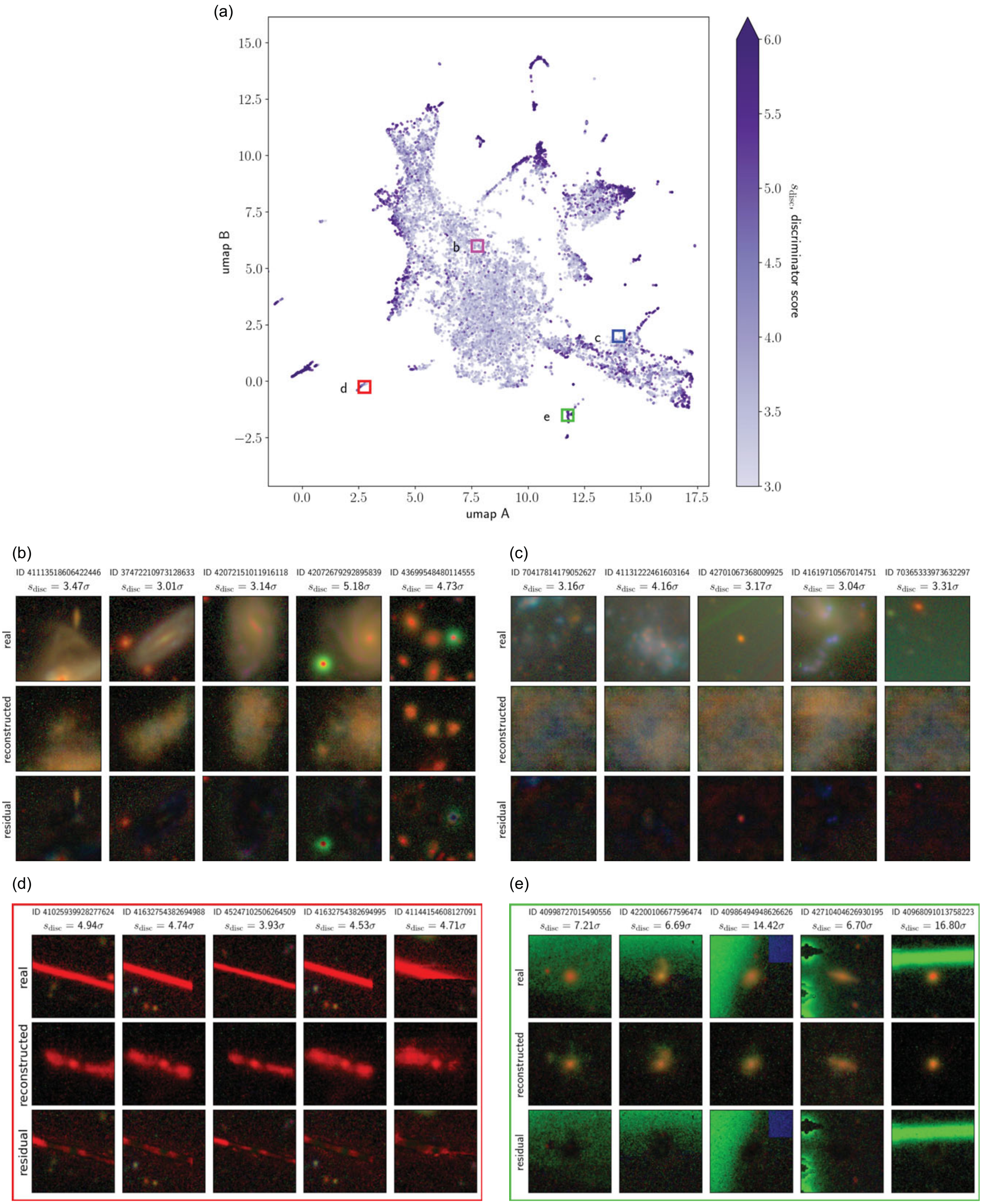

Figure 37. Example of anomalous objects identified with a combination of WGAN and a CAE applied to the HSC survey. The top panel shows the latent space from the image residuals of the WGAN reconstruction obtained with the CAE. The images below show examples of different regions of the parameter space. Figure from Storey-Fisher et al. (2021).

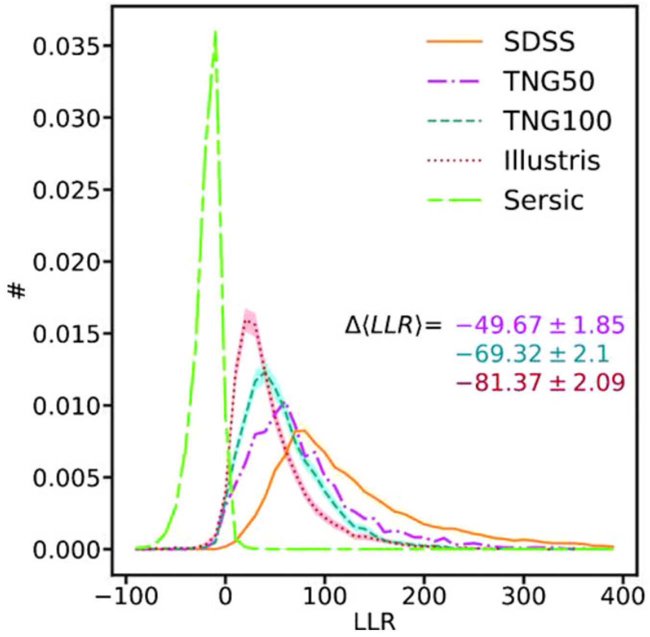

Figure 38. Distribution of likelihood ratios obtained with two pixelCNN networks of observations (SDSS) and different models as labelled. The closer the histograms from simulations are to the one from SDSS, the more realistic the simulation is. The unsupervised model is able to capture the improvement of simulations with time. Figure from Zanisi et al. (2021).

able to measure the improvement in the realism of cosmological simulations from the first Illustris model to the updated TNG one (Figure 38).

An additional way to identify outliers is through the representation space learned by contrastive learning, as one would do with a latent space from an Autoeconder. Stein et al. (2021b) explored this approach on images from the DESI survey and demonstrated the efficiency of self-supervised representations to identify outliers and perform similarity searches. See also the work by Walmsley et al. (2021) which uses the tools developed by Lochner & Bassett (2021) for similarity search and anomaly detection on the representation spaces.

5.3. *Discovery of physical laws*

Arguably one of the final goals of science is to find universal physical laws which can explain a broad set of observations. As said several times in the previous sections, deep neural networks offer an excellent predictive power but their interpretability is low as compared to model-driven approaches. Symbolic regression is the general term employed for the ensemble of techniques that aim at uncovering an analytical equation from data. They can be seen as a generalisation of polynomial regression to the space of all possible mathematical formulas that best predict the output variable taking as input the input variables. See Schmidt & Lipson (2009) for more details. One way to enable discovery with deep learning is therefore to apply symbolic regression techniques to the trained deep neural network model. This is usually very challenging given that neural network models are usually parametrised by a large number of parameters. There is one work in astrophysics attempting this by Cranmer et al. (2020). The authors train a GNN in a supervised manner to predict the properties of some dataset encouraging a sparse representation by the neural network. They then apply symbolic regression to the trained model (Figure 39). The authors show that they are able to recover for example some known Newtonian laws by predicting the movement of particles. More interestingly, they discover a new analytic formula which can predict the concentration of dark matter from the mass distribution of nearby cosmic structures. The formula is learned by applying symbolic regression to a GNN which learned the properties of a Dark Matter only simulation. In a follow-up work, Lemos et al. (2022) apply a similar approach to study orbital mechanics.

Summary of Deep learning for discovery

- Deep learning has been explored as a discovery tool. The main motivation is that future big data surveys are too large and too complex for efficient human-based exploration. Deep learning is therefore mainly used for visualisation and anomaly—outlier—detection. The transient astronomy community has been a particularly active field on this front in preparation for LSST.
- As opposed to previous applications, these applications rely on unsupervised deep learning. There is a variety of different approaches which have been tested: Autoencoders, Generative Models—GANs, VAEs, Autoregressive Flows. Self-supervised approaches using contrastive learning have also started to be used for this task.
- For visualisation, the usual approach is to use deep learning to obtain a low dimensional representation of the data which can be explored more easily.
- Anomaly detection implies learning a probabilistic description of the data and identifying objects with low likelihoods, that is, which can hardly reproduced by the trained models.
- A common result is that deep learning techniques correctly identify some anomalies, however the quantification of performance is challenging because by definition the problem is ill-posed. Some works find that complex deep learning networks might not be the most efficient way of detecting outliers because they are flexible enough to properly reproduce *simpler* data than the data used for training.
- In addition, filtering out interesting anomalies from artefacts remains an unsolved issue. Current solutions consists in providing anomalous candidates for further inspection.
- The issue of discovering physical laws from deep learning models has been just recently explored by applying symbolic regression methods on the trained models. It is difficult to generalise at this stage given the large amount of parameters of current deep learning models and the limited interpretability.

6. Deep learning for cosmology

In addition to galaxy formation, a key goal of modern deep surveys is to constrain cosmology. Deep learning is playing an increasingly large and promising role in at least two fronts: accelerating simulations—which are needed for efficient cosmological inference—and direct inference of cosmological parameters. This section is focused on these applications. We first review approaches aimed at producing simulations and then we move to the inference of cosmological parameters.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

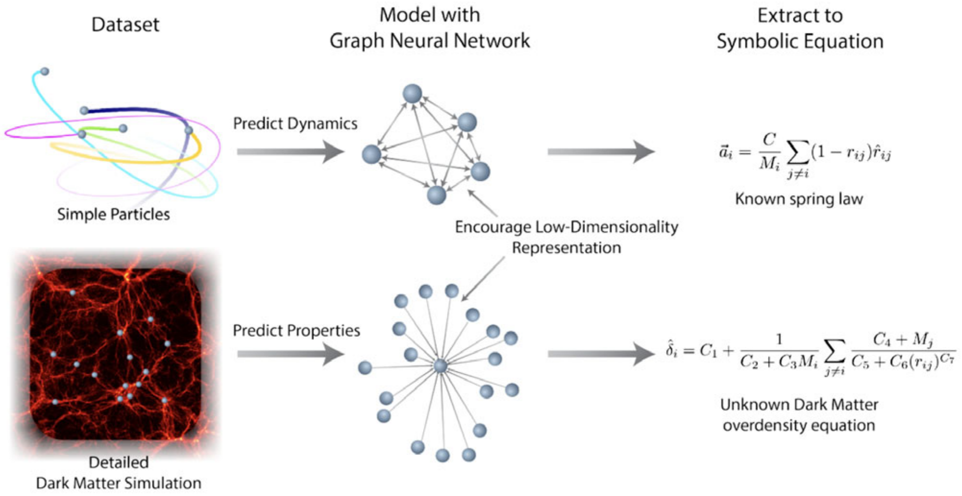
Dataset
Model with Graph Neural Network
Extract to Symbolic Equation
Simple Particles
Predict Dynamics
$\vec{a}_i = \frac{C}{M_i}\sum_{j\neq i}(1 - r_{ij})\hat{r}_{ij}$
Known spring law
Encourage Low-Dimensionality Representation
Predict Properties
$\hat{\delta}_i = C_1 + \frac{1}{C_2 + C_3 M_i}\sum_{j\neq i}\frac{C_4 + M_j}{C_5 + C_6 (r_{ij})^{C_7}}$
Unknown Dark Matter overdensity equation
Detailed Dark Matter Simulation

Figure 39. Cartoon illustrating the method to extract physical equation from a deep learning model applied to different datasets. Figure from Cranmer et al. (2020).

6.1. Accelerating simulations

Numerical simulations play an important role in our understanding of galaxy surveys, from shedding light into the physics of galaxy formation and evolution to the modelling of the large-scale structure of the Universe and its connection to galaxies. Fast and efficient simulations are needed to interpret the data. However, a major bottleneck is computational type. Ideally one would like to have high resolution and large volume simulations including both N-body and hydrodynamics. However, this is usually prohibitively time consuming and for that reason fast emulators are desirable. In the past years, deep learning has appeared as a promising solution, In this first subsection, we review how it has started to impact the field of cosmological simulations.

6.1.1. Learning N-body simulations

We begin by reviewing different strategies to either partially or entirely learn the physics of N-body simulations and to some extent complement a physical model to provide a significant acceleration compared to a full simulation. We will split these methods into two categories, depending on whether they act on and model matter density fields, or Lagrangian particle displacements.

Lagrangian displacement models N-body simulations typically model the matter distribution using tracer particles and evolving in time the position and velocity of these particles under the effect of gravitational forces. In this Lagrangian approach, the full output of a simulation can be seen as the displacement that each particle has undergone from its initial position on a regular lattice, along with its final velocity. The methods described here all aim at modelling this displacement field and therefore are not acting on 3D density fields, but on this displacement sampled at the initial particle positions on a regular grid.

Modelling residual displacements against fast simulation: In Dai, Feng, & Seljak (2018), the authors propose a physically motivated post-processing technique, dubbed Potential Gradient Descent (PGD), able to recover the small scales of fast Particle-Mesh (PM) simulations, and mimic the output of high-resolution N-body simulations, or even mimic the baryonic feedback from hydrodynamical simulations. The advantage of these fast PM simulations (such as FastPM Feng et al. 2016 or COLA Tassev, Zaldarriaga, & Eisenstein 2013) is that they can be run inexpensively on very large comoving volumes, but their lack of force resolution and their coarse time stepping limit their resolution, typically leading to a lack of power on small scales and inaccurate halo profiles. The method proposed in Dai et al. (2018) is to learn an additional displacement of the particles, moving them deeper into their local gravitational potential, which has the effect of sharpening the halo profiles. To compute this displacement, instead of using Convolutional Neural Networks, the authors use a physically motivated parameterisation in Fourier space, defined by an overall amplitude and a band-pass filter applied to the gravitational potential for a total of only 3 parameters, which respects the translational and rotational invariance of the problem. Training of the parameters of this model is done by either minimising the Mean Square Error (MSE) on the power spectrum or on the density field between a reference simulation and a fast simulation ran from the same initial conditions. With this simple, yet powerful, scheme the authors can emulate to within 5% accuracy the Illustris-3 simulation from only a 10 step FastPM simulation.

In one of the first works applying deep learning to N-body simulations, He et al. (2018) proposed a model based on a 3D convolutional U-Net (see Subsection 3.2) that learned to predict full nonlinear particle displacements given as an input analytic Zel'dovich Approximation (ZA) displacements (which corresponds to a single step of a FastPM algorithm). The 3D CNN takes as inputs a 3-channel 3D field providing this ZA displacement field sampled at the initial particle positions, and outputs the final displacement vector, still as a 3-channels 3D field. The model is then trained by Mean Squared Error loss on the reference displacement field provided by a FastPM simulation. An important realisation from that work was that a 3D CNN was able to accurately model this displacement field. Figure 40 illustrates the approximation error

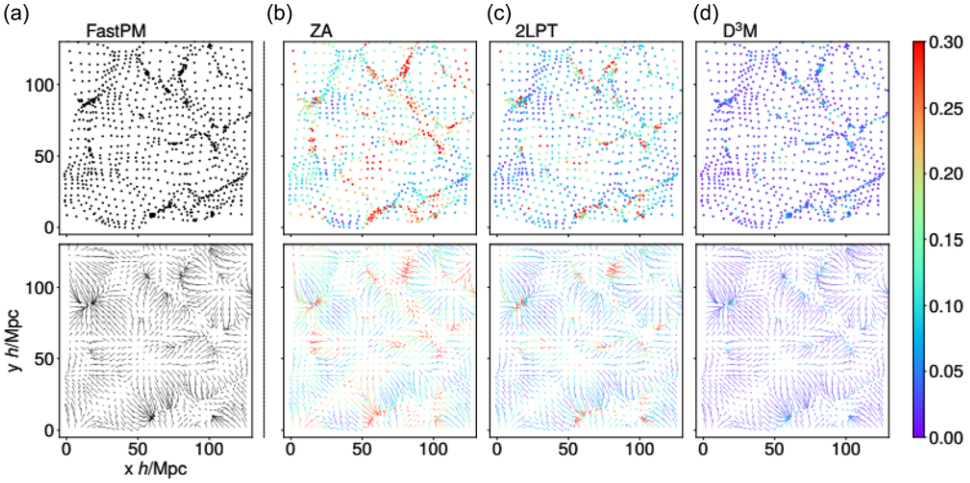

Figure 40. Illustration of learned displacement field in an N-body simulation from He et al. (2018). The first column is the reference simulation (FastPM), the second column shows a simple linear displacement (ZA), the third column shows a second order Lagrangian Perturbation Theory displacement (2LPT) and the last column shows the output of the 3D U-Net (D^3M). The top row shows final particle positions, the bottom row shows the final displacement field. The colour scale shows the error in position or displacement vectors between each approach and the reference FastPM simulation.

of this method (right column) compared to other fast approximations for the displacement field. It is found to be significantly more accurate than analytic solutions.

With a similar model, Giusarma et al. (2019) showed that it was possible to learn the residual displacements between a ΛCDM N-body simulation and a simulation with massive neutrinos. They used a modified version of D^3M to learn this residual displacement at $z = 0$ between the two sets of simulations, and found excellent results down to $k < 0.7h$ Mpc^{-1}.

Upsampling the displacement field from a low-resolution simulation: Recognising that upsampling the displacement field is equivalent to increasing the number of particles in a simulation, Li et al. (2020a) proposed a Super-Resolution technique based on a conditional GAN directly inspired from a StyleGAN2 (Karras et al. 2019) architecture. The generator takes as an input a low-resolution displacement field, and outputs an upsampled high-resolution displacement field. This approach achieves impressive results up to an upsampling factor of ×8 translating into a direct computational speedup of a factor ×1 000 in a setting where the goal would be to produce a $100h^{-1}$Mpc simulation with 512^3 particles. A visual illustration of this model is shown on Figure 41 where the rightmost panel is the output of the model. As a further extension of this approach Ni et al. (2021) trained a similar Lagrangian conditional GAN to model not only displacements but also velocities, yielding a full phase-space information that a real N-body simulation would produce. They further tested this model to demonstrate accurate matter power spectrum recovery up to within 5% up to $k \geq 10\ h^{-1}$ Mpc for an upsampling factor of ×8 of a 100 h^{-1} Mpc box, and validated the recovery of halo and sub-halo abundance.

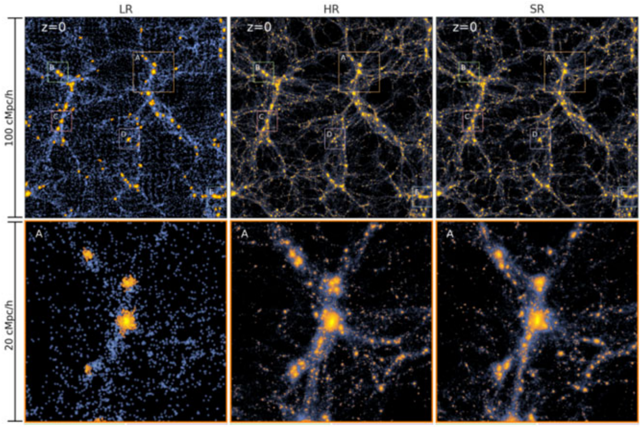

Figure 41. Illustration of N-body simulation super-resolution from Li et al. (2020a) showing from left to right, the Low-Resolution (LR) input, High-Resolution (HR) target, and Super-Resolution output of the model. The bottom row is a zoom-in on the region marked A.

Density field models The first method for simulation super-resolution, proposed by Ramanah et al. (2020), relied on a 3D Convolutional WGAN with a generator taking as inputs the density field from a low-resolution simulation run and a high-resolution set of initial conditions, and tasked with outputting a high-resolution final density field. The 3D convolutional discriminator compared the high-resolution density fields from the full simulation to the generator output. With this approach, the authors were able to upsample by a factor of 2 the resolution of the final density field, while reproducing faithfully a number of field properties including the power spectrum, the density contrast probability density function, and the bispectrum. This upsampling ratio represented a computational speedup of about ×11 for a 1 h^{-1}Mpc and 512^3 particles simulation. More recently, Schaurecker et al. (2021) proposed a similar model acting directly at the level of the density field, but only using the low-resolution final density field as an input (without needing the high-resolution initial conditions of Ramanah et al. 2020).

6.1.2. N-body emulation by deep generative modelling

All of the models from the previous section had the particularity of trying to model the residuals compared to a physical model instead of completely supplanting the physical simulation. In this section, we now cover the works that have taken the approach of trying to

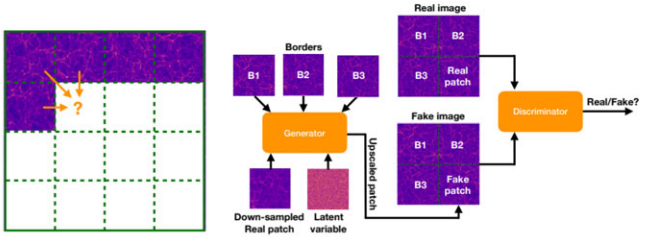

Figure 42. Sequential generation and upsampling strategy of the scalable GAN for N-body modelling presented in Perraudin et al. (2019a) scaling up to 256^3. The left illustration shows how the sequential generation of a large volume would proceed. The right plot illustrates the proposed architecture where the generator is conditioned on both neighbouring patches, and on the lower resolution patch, which at sampling time would be generated by a separate GAN trained on coarser volumes. Distribute under the Creative Commons CC BY licence (http://creativecommons.org/licenses/by/4.0/).

learn from scratch the entire simulation using a Deep Generative Model.

A number of works started to apply to the emulation of cosmological fields the new Deep Generative Models that were gaining in popularity at the time, and especially GANs. In the first instance of such an application, Mustafa et al. (2019) applied a conventional DCGAN to the modelling of weak-lensing convergence maps and demonstrated that the model was able to accurately reproduce a number of statistics of these fields, including their power spectra and Minkowski functionals. Shortly after, Rodriguez et al. (2018) presented a similar application of using a DCGAN to model slices of N-body simulations with results demonstrating that these models were able to capture most of the relevant statistics of the cosmic web.

These early works on generative modelling for cosmological fields were quickly confronted to the difficulty of building high quality models for very large or even 3D fields. Perraudin et al. (2019a) explored more in depth the limitations of simple DCGANs for generating large 3D N-body volumes and highlighted two key strategies enabling high quality results in this setting: 1. generating the field by patches, 2. using a multi-resolution approach based on a Laplacian pyramid. To generate N-body meshes of size 256^3, their proposed model uses 3 independent GANs that are trained on 3 increasingly high data resolution (32^3, 64^3, 256^3). The first model trained on the coarsest resolution is a conventional DCGAN while the other models are conditioned on lower resolution inputs. In addition, the GANs are made conditional on the neighbouring patches in the simulation, which allows at inference time to generate a large volume patch-by-patch in a sequential fashion, thus avoiding the need of storing the entire volume in GPU memory. Figure 42 illustrates the proposed strategy, which is shown in the paper to be able to recover both the power spectrum and peak counts to satisfying levels whereas a non-multiscale approach fails significantly.

Additional works have investigated other possible improvements to GANs for N-body simulations. In particular Feder et al. (2020) proposed modelling the latent space prior of a DCGAN with a heavy-tailed distribution instead of a Gaussian. In that work, using a Student's t-distribution is shown to improve the model's ability to capture the sampling variance of the peaks in the dark matter distribution, and overall improves power spectrum recovery on all scales.

These generative models of the matter distribution start to become useful when they are made conditional on some external parameters, for instance cosmological parameters or redshifts. The GAN model proposed in Feder et al. (2020) was for instance made conditional on redshift by simple concatenation of the conditional variable to the latent vector of the GAN, allowing the authors to generate volumes at intermediate redshifts, which would be useful to create lightcones. Perraudin et al. (2020) proposed a conditional GANs to produce 2D weak-lensing mass-maps conditioned on (σ_8, Ω_m) through a remapping of the latent vector of the GAN by a function that rescales the norm of that vector based on the conditional variable. More recently, Wing Hei Yiu, Fluri, & Kacprzak (2021) extended that work to the sphere, using a DeepSphere (Perraudin et al. 2019b) graph convolutional architecture, to emulate the KiDS-1000 survey footprint as a function of (σ_8, Ω_m).

While conditional GANs could be useful as emulators, they however cannot be directly used for cosmological inference due to the fact that GANs do not possess tractable likelihoods. One significantly different approach to generative modelling of the dark matter distribution proposed in Dai & Seljak (2022) relies on a Normalising Flow approach instead of a GAN to model explicitly the conditional distribution $p(x|\theta)$ where x is the dark matter distribution, and θ are cosmological parameters. Once trained, such a model can directly be used as the likelihood function of the high-dimensional data in a Markov-Chain Monte Carlo. In this paper, the authors introduce a Translation and Rotation Equivariant Normalising Flow (TRENF) model. It builds an n-d normalising flow based on learning filters and performs convolutions in Fourier space which impose by construction translation and rotation equivariance. The authors demonstrate that this approach accurately captures the high-dimensional likelihood of dark matter density fields and that it can be used not only for generating these fields, but also for inferring cosmological parameters.

6.1.3. Finding dark matter halos

In the pipeline needed to go from N-body dark matter simulations to observable galaxy distributions, a typically essential step is the identification of dark matter halos, which can then be populated with galaxies under a variety of techniques (e.g. HOD or SHAM). In this section, we review the various approaches which have been proposed to go from the dark matter density field to dark matter halos.

One of the first approaches to learn this connection was proposed in Modi, Feng, & Seljak (2018), and assumed a shallow neural network mapping between the local 3D dark matter density and 'halo mask' and 'halo mass' fields. The binary halo mask field was essentially used to model whether a given voxel actually contained a halo, while the halo mass field was predicting in each voxel a likely total halo mass. Given this model, a halo field could be recovered by multiplying these two outputs of the neural network. The actual neural network was based on a simple MLP taking as inputs a 3 × 3 × 3 voxel region of the dark matter density field itself, the field smoothed on a given scale, and the difference between the field smoothed on two difference scales. The authors find that the predicted halo mass field exhibits over a 95% correlation with the true field up to $k = 0.7h\text{Mpc}^{-1}$. Perhaps most interestingly, this model provided effectively a differentiable mapping between dark matter and halos, a differentiable halo finder, and the authors demonstrated that this mapping could be used in a reconstruction scheme to infer initial conditions

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

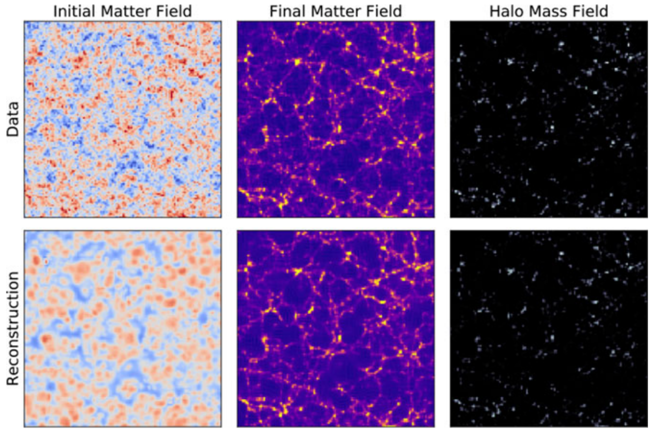

Figure 43. Illustration of an application of differentiable neural mapping between the dark matter density and dark matter halo fields from Modi et al. (2018). The top row shows the initial conditions dark matter field, final dark matter field (at $z = 0$), and the dark matter halo field obtained by a FoF halo finder. The bottom row shows the result of a reconstruction by gradient descent of these initial conditions, using a neural network to provide the mapping between the final density field and the halo field.

from a halo field by gradient descent through the neural network and a differentiable N-body simulation as illustrated on Figure 43.

Charnock et al. (2019) proposed a different approach to the same problem, building a probabilistic model based on a Mixture Density Network for the halo mass distribution in each voxel conditioned on the underlying dark matter density. The overall goal of the authors in that paper was to make the model minimalistically parametric while respecting the important physical properties of the problem, namely that the halo bias should be globally rotational and translational invariant. To reach this goal, instead of relying on standard CNNs, the authors proposed a model they refer to as the Neural Physical Engine (NPE), which applies a set of convolutional kernels parametrised to exhibit by design the desired symmetries. This leads to a reduced number of parameters compared to a similar 3D CNN. They apply this NPE on the dark matter density field, and use its output to condition a simple Gaussian MDN tasked with representing the local halo mass distribution. They demonstrate that a minimal model with only 17 parameters is able to accurately capture the halo mass distribution and its dependence on local environment, and further present an application where the model is used as part of a Bayesian reconstruction of initial conditions in a simulation from a given halo distribution.

Going in a deeper direction, a number of papers have looked into identifying dark matter halos, not from Eulerian space, but directly as Lagrangian patches at the level of the input density field of the simulation using a deep CNN model. Berger & Stein (2018) introduced the first deep CNN model to identify dark matter halos with this approach, relying on V-net model illustrated in Berger & Stein (2018). They used as a training set simulated initial density fields of size 128^3 with a 1 Mpc voxel resolution, along with a binary 128^3 segmented map indicating the position of Lagrangian patches identified as proto-halos by the Peak Patch semi-analytic code (Stein, Alvarez, & Bond 2019). The network is then trained with a binary classification loss to predict the presence or not of a halo in a given voxel. To build an actual halo catalog given a segmented volume outputted by the trained model, the authors then use a hierarchical Lagrangian halo-finding procedure that ultimately returns a list of halos. This entire procedure is found to lead to a halo mass function and power spectrum within 10% of the ground truth simulation. As a variation of this approach, Bernardini et al. (2019) proposed to replace a binary segmentation by a regression problem, where the target value corresponds to a normalised distance to the centre of the halo, leading to similar performance but with a smaller model.

In related work, Lucie-Smith et al. (2020), Etezad-Razavi et al. (2021) propose to use a 3D CNN to predict from a initial conditions centred on the location of halos, the mass of the collapsed halos at $z = 0$. They use these models to investigate the relative importance of various properties of the initial conditions for halo formation. Lucie-Smith et al. (2020) reports for instance that removing anisotropies in the initial conditions does not significantly affect the masses predicted by the model, hinting that initial shears may not be a significant factor in the halo formation process. Etezad-Razavi et al. (2021) reports that velocity information becomes more important to accurately predict masses at lower values of A_s.

6.1.4. Painting baryons on N-body simulations

As a way to bridge the gap between full hydrodynamical simulations and cheaper Dark Matter Only (DMO) simulations, a number of works have investigated the possibility of 'painting' baryons on top of DMO simulations.

A first category of papers proposed to model this mapping between dark matter density and baryonic fields in a probabilistic fashion to account for the inherent uncertainty. In the first work to attempt such modelling, Tröster et al. (2019) investigated the use of both conditional VAEs and conditional GANs to learn from 2D dark matter density slices a probabilistic map to 2D thermal Sunyaev-Zeldovich (tSZ) maps, which capture the electron pressure. They found excellent agreement with ground truth simulations, at different redshifts, indicating that this approach was very promising, but did report a tradeoff of GANs leading to more accurate results but being harder to train and less stable than VAEs.

More recently, Bernardini et al. (2021) explored the use of a conditional WGANs to learn a similar mapping to predict gas and H_I density on 2D maps. To achieve the generation of high-resolution and large maps, the authors adopt a multi-resolution strategy in which a U-Net generator outputs maps at 3 different resolutions that the critic compares to similarly downsampled versions of the target fields. This strategy allows them to successfully train the model on images of size 512^2, but after training this purely convolutional model can be applied on much larger fields. The authors report accurate H_I power spectra to within 10% accuracy up to the 10 kpc scales while being able to map simulation boxes of 100 h^{-1}Mpc on the side.

Extending these conditional generative approach to 3D fields, Horowitz et al. (2021) proposed a conditional VAE with a customised U-Net architecture. Contrary to a standard conditional VAE, this model features skip-connections between the conditional branch (the dark matter encoder) and the generative branch (the hydrodynamics decoder) which create a U-Net structure. It remains however probabilistic thanks to the variational bottleneck block which contains stochastic latent variables capable of capturing the aleatoric nature of the mapping. The resulting model inherits from the stability and robustness of VAEs but also benefits from this U-Net structure to enable high-resolution mapping.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Again, the model is kept strictly convolutional to make it insensitive to the size of the input field, so that it can be applied on larger volume than it is trained on. This model can also be made conditional on redshift and the authors demonstrate the possibility of generating lightcones with this approach.

A second class of papers propose similar but deterministic mappings, using 3D U-Nets trained to regress particular baryonic fields. Thiele et al. (2020) proposed a 3D U-Net to learn a similar mapping, although no longer probabilistic, between the 3D dark matter field and electron density, momentum, and pressure. This work reports two significant challenges in learning this mapping in 3D, one being the sparsity of interesting voxels (as in 3D most of the voxels are in empty regions), and the other being the high dynamic range of the fields to model. They address these challenges by biasing the loss function towards high density regions, and applying range-compression schemes. Overall they report better agreement with the reference simulations compared to semi-analytical models. In similar work, Wadekar et al. (2020) trained a U-Net to output H_I density maps and again reported better quality results than a standard HOD approach, while Harrington et al. (2021) used a U-Net to predict hydrodynamical fields (density, temperature, velocity) subsequently used to model Lyα fluxes with a physical prescription. Zhang et al. (2019) proposed a two-step approach to map the dark matter field to a 3D galaxy distribution, where a deep 3D CNN would predict a mask of the likely non-empty voxel, and a second CNN would regress the number of galaxies in these regions. They find that their CNN model is able to predict a galaxy distribution recovering the expected power spectrum to within the 10% level up to $k = 10h\,\mathrm{Mpc}^{-1}$ scale.

Finally, a singular approach was proposed in Dai & Seljak (2020) which extended the PGD method of Dai et al. (2018)—mentioned in the previous section—to parametrise a mapping between a dark matter only simulation and various fields accessible in hydrodynamical simulations using a combination of particle displacements and voxel-wise non-linearities. This method, dubbed Lagrangian Deep Learning, reused the same Fourier-based parametrisation to displace particles from a dark matter only simulation (e.g. FastPM), thus using a very small number of parameters (order 10) and providing translation and rotation equivariance, before painting them on a 3D mesh and applying a non-linear transform. The few parameters can be fitted by gradient descent on a single pair of dark matter only and hydrodynamical simulations. This scheme was successfully demonstrated to reproduce a range of maps from IllustrisTNG, including stellar mass distribution and electron momentum and pressure.

6.2. Deep learning for cosmological inference

In this section we will review in particular how emulators and Likelihood-Free Inference techniques are enabling the inference not only of cosmological parameters but also of cosmological fields.

6.2.1. Field-level cosmological constraints

Weak Gravitational Lensing It was realised early on that given access to simulations to act as a training set, neural networks can be used to extract cosmological information from high-dimensional data such as maps of weak gravitational lensing maps. The first example, presented in Schmelzle et al. (2017), demonstrated that a CNN-based classification model could discriminate between different discrete cosmological models, especially along the $\sigma_8 - \Omega_m$ degeneracy that conventional 2pt correlation functions are unable to resolve. Similar results in Peel et al. (2019), Merten et al. (2019) showed that a CNN classifier was able to distinguish between ΛCDM, modified gravity and massive neutrinos models from weak-lensing maps, with better discriminating power than more conventional higher-order statistics such as peak statistics. These results sparked a lot of interest into the potential use of CNNs to extract cosmological information from weak-lensing surveys, which resulted in a number of subsequent publications with the ultimate goal of yielding proper Bayesian posteriors on cosmological parameters.

Going beyond a classification task between discrete models, a second class of papers (Gupta et al. 2018; Fluri et al. 2018; Ribli et al. 2019b) built CNN regression models where the network is tasked with directly predicting (σ_8, Ω_m), either using a Maximum Absolute Error (MAE) loss (Gupta et al. 2018; Ribli et al. 2019b), or Gaussian-parameterised negative log likelihood loss (Fluri et al. 2018). It is important to note, as reported in all these papers, that the output of the network trained for regression will not be an unbiased estimator for the cosmological parameters, but should be interpreted as a low-dimensional summary statistic, which can then be used for inference in a second step, independent from the network training. To retrieve cosmological parameters, these papers assume a Gaussian likelihood on the output of the network and characterise the mean and covariance of that likelihood on a set of simulation, similarly to what is conventionally done for peak count statistics or other Higher-Order Statistics without an analytic likelihood. With this approach, all these papers reported the ability to extract more information than a 2pt function analysis, even on realistically noisy data, with Ribli et al. (2019b) reporting a factor of ∼2 smaller contours on (σ_8, Ω_m) in a Euclid or LSST setting.

Building on these promising results, the next phase of papers deployed these approaches to actual survey data. Fluri et al. (2019) followed a similar strategy to Fluri et al. (2018) and trained a ResNet model, on a suite of tomographic lensing simulations mimicking the KiDS-450 survey and spanning a range of $(\sigma_8, \Omega_m, A_{IA})$ values, where A_{IA} is the amplitude of the intrinsic galaxy alignment signal. This study found constraints broadly consistent with the fiducial 2pt function analysis of KiDS-450 (Hildebrandt et al. 2017) but when compared with an internal power-spectrum analysis yielded a 30% tighter posterior. In the most recent extension of that work, Fluri et al. (2022) performed a wCDM analysis of the KiDS-1000 weak-lensing maps including a large number of refinements. In particular, they used a spherical CNN architecture, DeepSphere (Perraudin et al. 2019b), in order to process spherical fields and extended their simulation suites to include a baryonic prescription, and left the dark energy equation of state parameter w_0 free to vary along with 5 other cosmological parameters. They find again broad agreement with KiDS-1000 wCDM, and internally consistent results with a power spectrum analysis, but with only a meagre 11% improvement on S8 constraints. The most likely reason for the limited constraining power of the CNN analysis comes from the low-resolution of maps used in the analysis (HEALPix nside=1024), but available non-Gaussian information content when baryonic systematics are taken into account is also an open question discussed below.

Jeffrey, Alsing, & Lanusse (2021) performed an analysis of the DES Science Verification (SV) data using a slightly different approach to previous works. Instead of training a CNN for regression, the authors introduced an information loss which explicitly trains the network to compress the input lensing maps into a low dimensional (asymptotically) sufficient statistic, that can further be used for inference with a Likelihood-Free Inference approach. More specifically, they used a Variational Mutual Information lower bound to train the model, which relies on using a Normalising Flow (NF) to approximate the posterior distribution on cosmological parameter from the low dimensional output of the convolutional compressor network, and training both models jointly as to minimise the negative log likelihood of the NF (Figure 30). Once trained under this information loss, the model can be applied to data, and a robust estimate of the posterior was achieved by Neural Likelihood Estimation using the pyDELFI package (Alsing et al. 2019). Compared to previous papers, this approach has asymptotic optimality guarantees, and does not rely on any Gaussian assumptions for the likelihood of the summary statistic. In this paper, the authors found consistent but tighter constraints with this approach compared to a power spectrum analysis, but the constraints remained very large due to the small size of the SV dataset.

One question remains unclear, however, regarding the amount of additional information deep learning can extract over the power spectrum or simpler higher-order statistics, when systematics like baryonic effects are taken into account. Lu, Haiman, & Matilla (2022) investigated this question using a simple baryonic correction model (BCM Aricò et al. 2020) for dark matter only simulations and trained a deep CNN to infer both cosmology and baryonic parameters from simulated lensing maps under a realistic HSC-like setting. The authors find that using a CNN instead of a power spectrum ($100 < \ell < 12\,000$) improves the constraining power on (Ω_m, σ_8) (in terms of 1-sigma area) by a factor of a few if the astrophysical parameters are kept fixed. However, the improvement degrades significantly when marginalising over astrophysical parameters. Indicating that there is some amount of non-Gaussian information left even after marginalising on baryons, but how sensitive are the resulting constraints to the specific baryonic model assumptions is uncertain.

With the success of these methods, several papers have attempted to introspect the CNN trained on weak-lensing maps, to try to identify or recognise what features of the data are being used to extract the cosmological information. Ribli et al. (2019a) proposed using at the first convolutional layer of the model a large 7×7 kernel, and recognised after training that this first layer was learning a kernel close to a Laplace operator and concluded that this operator allowed the network to be sensitive to the steepness of the peaks in convergence maps. Building on that insight, the authors handcrafted a new summary statistic based on histograms of Sobel-filtered lensing maps, which are sensitive to peak steepness. This simple statistic was found to outperform a deep CNN on noiseless data, while deteriorating in the presence of noise, but still outperforming conventional peak counts.

Using a different methodology, Matilla et al. (2020) performed a similar study using saliency methods to identify the features of the lensing maps relevant to the cosmological inference task. They found that in all cases, the most relevant pixels in the input maps were the ones with extreme values. In noiseless maps, regions with negative convergence accounted for the majority of the attribution, while on realistically noisy maps, the high value convergence regions (positive peaks) account for the majority of the attribution.

Large-Scale Structure Although not as actively researched, cosmological information can also be extracted from the galaxy distribution with deep learning. This was first illustrated by Ravanbakhsh et al. (2017) which used a 3D CNN to regress cosmological parameters (σ_8, Ω_m) from the 3D dark matter density in a suite of N-body simulations. Although not directly applicable to actual surveys, this work demonstrated that convolutional approaches where able to retrieve cosmological information from the 3D large-scale structure. Addressing the same problem, but with the computational aspects of training 3D CNNs on large-scale High Performance Computing (HPC) systems in mind, Mathuriya et al. (2018) presented a similar result on cubes of size 128^3, and showcasing distributed training on 2048 and up to 8192 CPU nodes on the NERSC Cori machine.

Going further in that direction, Ntampaka et al. (2019) presented a 3D CNN model acting this time on the 3D distribution of galaxies, with spectroscopic surveys in mind. This work relied on a suite of 40 dark matter simulations, populated with galaxies with a range of various HOD models (15 different models) as a way to marginalise over uncertainties on the galaxy-halo connection. Galaxies in a given comoving volume were painted on 3D slabs of size $550 \times 550 \times 220\ h^{-1}$Mpc to estimate a galaxy density field. A 3D CNN was then tasked with outputting (σ_8, Ω_m) and the model was trained by Maximum Absolute Error. The authors also proposed a variants on that model, with an MLP branch taking directly as an input the power spectrum of the volume, and combined or not with the CNN branch to provide the cosmological parameter estimates. The main takeaways of that paper were that the CNN was able to extract more information than the power spectrum alone, and that the model trained in this fashion generalised well to unseen HOD models.

Robustness to Baryonic Effects Because deep learning-based cosmological inference schemes remain opaque, one important question is how to robustify such an analysis to modelling uncertainties and systematics. Just like in modern 2 point function analyses, one of the most prominent questions is how to account for uncertainties in Baryonic physics.

Answering this question is one of the motivations for the Cosmology and Astrophysics with Machine-learning Simulations (CAMELS) suite (Villaescusa-Navarro et al. 2021b), a set of about 4000 simulations of $25(h^{-1}\text{Mpc})^3$ volumes which break down into 2000 dark matter only simulations, 1000 hydrodynamical simulations following the IllustrisTNG model, and 1000 hydrodynamical simulations using the SIMBA model (Davé et al. 2019). Each of these 2 different hydrodynamical sets of simulations varies not only cosmological parameters (Ω_m, σ_8) but also astrophysical parameters, namely (A_{SN1}, A_{SN2}) which regulate supernova feedback and (A_{AGN1}, A_{AGN2}) which parameterise AGN feedback. This suite of simulations therefore not only allows the study of the dependency between cosmological parameters and a given set of astrophysical systematics, but also can be used to check the robustness to an assumption of a particular baryonic feedback model (IllustrisTNG or SIMBA). On this dataset, the authors demonstrated that not only does the total matter distribution contain significant cosmological information accessible by deep neural networks, but also baryonic fields such as the H_I density (Villaescusa-Navarro et al. 2021a), and that even individual galaxies bare some

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

imprints of cosmological parameters (Villaescusa-Navarro et al. 2022). The question however is: can this information be retrieved under the uncertainties of the baryonic model?

In Villaescusa-Navarro et al. (2020), the authors illustrated on an analytically tractable toy model that a Neural Network can be trained to optimally marginalise over baryonic effects as long as the training data left the associated parameters free to vary according to a given prior. While promising, this result didn't necessarily imply that this implicit marginalisation would be robust to a change in baryonic model. To investigate precisely this question, Villaescusa-Navarro et al. (2020) trained a CNN on 2D projected density fields from the CAMELS dataset to regress (Ω_m, σ_8). They showed that models trained on IllustrisTNG lead to almost unbiased results on SIMBA and vice-versa, implying that in the process of learning a summary statistic that marginalises over baryonic effects, the neural networks are discarding the part of the signal affected by baryons, and are therefore no longer extremely sensitive to the details of the modelling of those effects. This result remains of course limited, but is very encouraging for the analysis of data.

6.2.2. Dark matter substructures from strong gravitational lensing

In previous sections we have described how deep learning has significantly impacted the detection and characterisation of strong gravitational lenses. Strong lensing systems can be used as well to constrain the substructure of dark matter on extended arcs which contains a wealth of information about the properties and distribution of dark matter on small scales and, consequently, about the underlying nature of the dark matter particle. The information can therefore be used to distinguish between various dark matter models—warm or cold dark matter for example.

However, probing this effect is challenging since the likelihood function for realistic simulations of population-level parameters is intractable. Alexander et al. (2020) followed a simple approach which consists in converting the inference problem into a classification of CNNs in classification mode to distinguish various types of dark matter models. They show they can reach AUC scores above 90%, for images with no substructure, spherical subhalos, and vortices on idealised simulations. Varma, Fairbairn, & Figueroa (2020) also used a CNN for multi-class classification in seven different categories corresponding to different lower mass cut-offs of the sub-halo mass function. They report being able to correctly identify the lower mass cut-off within an order of magnitude to better than ~90% accuracy.

Other works have attempted to go a step beyond by estimating the parameters describing the dark matter substructure in a regression mode using simulation based inference with deep learning. The first work exploring this is by Brehmer et al. (2019). The authors characterise substructure with a set of parameters and show, in a proof-of-concept application to simulated data, that neural networks can be trained to accurately estimate likelihood ratios associated to the dark matter substructure parameters (Figure 44). They conclude that ~100 strong lenses might be enough for characterising the abundance of substructure down to ~10%. Coogan, Karchev, & Weniger (2020) also used a likelihood-free approach to infer posterior distribution of the mass and positions of sub-halos in simulated systems. Vernardos, Tsagkatakis, & Pantazis (2020) applied a similar approach combining a simulator based on Gaussian Random Fields for the potential but combined with images of real galaxies for the lensed source and show they can also constrain the substructure parameters.

6.2.3. Reconstructing cosmological fields

Applications in cosmology also go beyond cosmological parameter estimation, and an increasingly large number of works explore applications of deep learning for inferring latent cosmological fields from observations.

Weak-Lensing Mass-Mapping One active research question in weak gravitational lensing is the reconstruction of the matter distribution that gives rise to the measured lensing effect, a task known as mass-mapping. This problem is made particularly difficult by the noisy nature of the observations (intrinsic galaxy ellipticities being much larger than the weak gravitational shear) and the need to invert a linear operator mapping shear to projected mass (also known as convergence) which becomes ill-posed in the presence of survey masks. This is therefore an instance of an ill-posed inverse problem, which does not have any unique solution, in the sense that different mass-maps can lead to a shear signal equally compatible with the data. For these problems, the hope of Deep Learning approaches is that they can learn, implicitly or explicitly, a prior on the signal to recover from training data, and use that prior to solve the inverse problem in an optimal fashion.

The first class of methods, explored in Shirasaki, Yoshida, & Ikeda (2019), Shirasaki et al. (2021), used a conditional adversarial network adapted from the pix2pix model (Isola et al. 2016) for image-to-image translation. In this approach, a first network with a U-Net structure is tasked with taking a noisy convergence obtained by a rough direct inversion of the shear field as an input, and outputting an estimate of the true convergence map. To train this denoiser, a second network is introduced to act as a discriminator, taking as an input both noisy and denoised convergence maps, coming either from the denoiser or from the training set, and outputting a probability between 0 and 1 of the denoised image being real. The model is then trained with a combination of a standard adversarial loss and an l1 loss between the recovered denoised mass-map and the truth from simulations. It is to be noted here that this adversarial model is not a generative model, the denoiser does not take random variables as an input and is therefore deterministic. Instead the adversarial loss can be understood as a learned similarity metric to compare recovered to true map. Training such a model typically requires a set of ray-traced lensing simulations, that are corrupted to include the same noise properties and masks as present in the data. Shirasaki et al. (2021) generated mock HSC observations including photometric redshift uncertainties, shape measurement uncertainties, realistic galaxy ellipticity noise and distribution on the sky, and actual HSC survey masks. On simulations the authors find that about 60% of the peaks identified on the denoised maps have significant clusters counterparts, against about 85% of positive matches on true maps, highlighting that the recovered map still correlate well with real structures, and the authors further show that the 1-point statistics of the recovered map shows stronger cosmological dependence than the noisy maps, hinting at interesting applications in constraining cosmological parameters. While this approach provides empirically good results, one drawback of this pix2pix training is that the recovered map does not have a clear Bayesian interpretation. As we will see below, subsequently developed techniques abandon this effective adversarial training

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

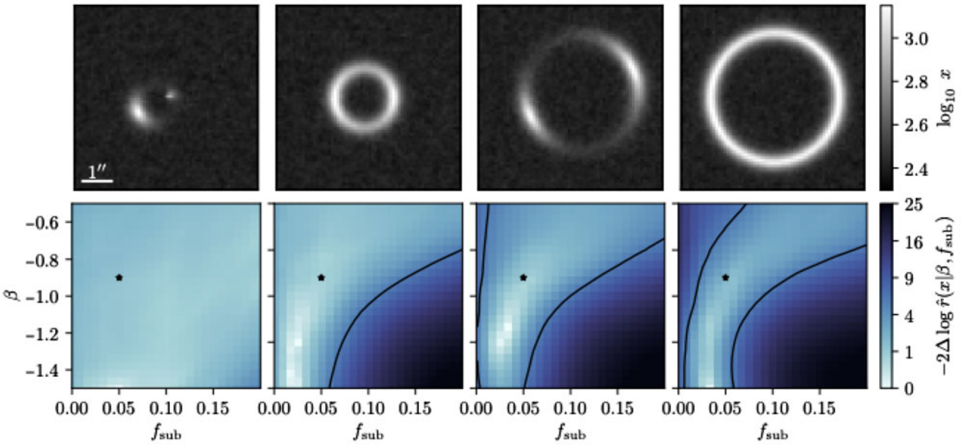

Figure 44. Illustration of lensed systems and the corresponding likelihood ratio maps estimated with simulation based inference and deep learning. The black crosses show the true values. Figure from Brehmer et al. (2019).

but gain a proper Bayesian interpretation of the output of the models.

In Jeffrey et al. (2020), the authors introduce a method, called `DeepMass`, using a similar Unet architecture but trained under a simple Mean Squared Error loss between true convergence map and output from the Unet. As highlighted by the authors, a regression model trained under an MSE implicitly learns to predict the mean of the posterior distribution of the target given the input. In the present case, the authors generate a suite of ray-tracing lensing simulations matching the DES Science Verification setting, the noiseless convergence maps from these simulations provide an implicit prior, while the simulated shear observations (including realistic noise and masks) provide an implicit likelihood. By training the model to reconstruct the true convergence given simulated shear data under an MSE loss, the model will therefore learn to output the mean posterior convergence map, under the implicit prior and implicit likelihood that are provided by the training set. In this DES SV setting, the authors demonstrate that this approach leads to an 11% improvement in MSE evaluated on simulations compared to a standard Wiener filter approach.

While Jeffrey et al. (2020) had the benefit of providing a Bayesian understanding of the network output, it did not provide any sort of uncertainty on the recovered map, which makes the interpretation and scientific exploitation of these results difficult. To overcome these limitations, Remy et al. (2022) introduced an approach allowing to sample from the full posterior distribution of the mass-mapping problem. They proposed to use a similar U-Net architecture, not to directly estimate the convergence map, but to learn from simulations a generative prior on convergence maps using a Denoising Score Matching technique (Song & Ermon 2019). With this approach, it can be shown that a neural network trained as a Gaussian denoiser under a simple MSE loss will actually learn the score of the data distribution, that is, the gradient of the log likelihood of the data. Once trained on simulations, this U-Net gives explicit access to the prior. The authors show that it is possible to combine this learned prior with an explicit data likelihood in an Hamiltonian Monte Carlo sampling procedure to sample from the full posterior distribution of the problem. Figure 45 illustrates on the bottom row posterior samples achieved with this method on a simulation of the HST/ACS COSMOS field, compared to the ground truth (top left). Most interestingly, it is shown that the mean of the posterior samples indeed converge to the same solution as the `DeepMass` estimate.

Initial Conditions Reconstructions One particularly interesting problem for the analysis of galaxy surveys is the reconstruction of the initial density field from the observed Large-Scale Structure. This can for instance be used to refine Baryonic Acoustic Oscillations (BAO) measurements (e.g. Schmittfull, Baldauf, & Zaldarriaga 2017), or part of a Bayesian forward modelling inference scheme (Seljak et al. 2017).

In a first example of applying deep learning to this problem, Mao et al. (2020) proposed a 3D CNN trained on N-body simulations to recover the initial density field at $z = 10$ given the final density field at $z = 0$ under a density weighted mean squared error loss. They find that their convolution model is capable of beating a standard linear reconstruction on scales smaller than $k \leq 0.2h\,\mathrm{Mpc}^{-1}$, but under performs on larger scales. Interestingly they find that their learned inversion can extrapolate to some extent to other cosmological parameters; a model trained on WMAP7 cosmology is capable of reconstructing initial conditions on WMAP5 simulations that lead to a slightly biased BAO signal, but still significantly different from the WMAP7 signal of the training data.

Going beyond a direct inversion method, Modi et al. (2021a) proposed an iterative reconstruction scheme based on Recurrent Inference Machines (RIM, Putzky & Welling 2017). This approach can be thought of as a learned iterative reconstruction algorithm. At each iteration a recurrent neural network proposes an update of the current reconstruction based on the knowledge of previous iterations and on the gradient of an explicit data likelihood term. In the absence of this neural network, the algorithm would result in a standard gradient descent scheme leading to a Maximum Likelihood Estimation of the initial conditions. By training the Neural Network to minimise at each iteration the Mean Squared Error between the current solution and the true initial conditions, the network will learn both an implicit prior, and a fast inference scheme to minimise the number of updates needed. The result will therefore be a fast convergence towards the mean posterior solution. Most interestingly, in order to compute this explicit likelihood, a differentiable forward model is needed, that is, in this case an N-body simulation. The authors make use of the

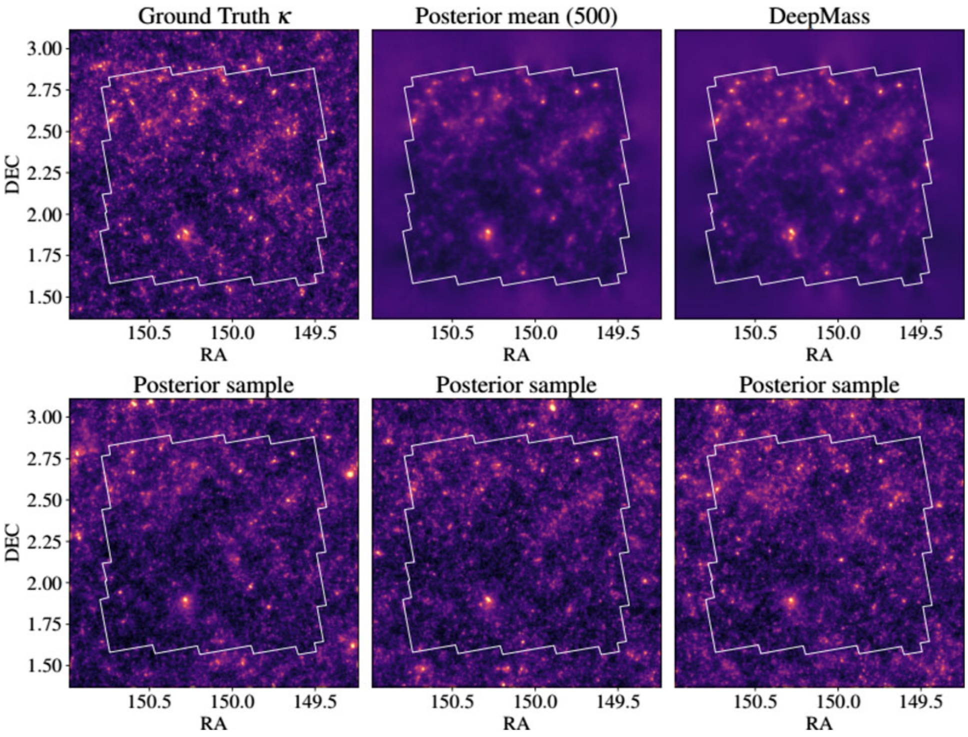

Figure 45. Illustration of weak-lensing mass-map reconstructions in a simulated COSMOS survey setting with the posterior sampling method of Remy et al. (2022) and the `DeepMass` direct posterior mean estimation method of Jeffrey et al. (2020). As can be seen, individual posterior samples (bottom row) are visually similar to a real convergence map (e.g. ground truth at the top left) but exhibit variability on structures not strongly constrained by data (e.g. outside of the survey region marked by the white contours). The top row illustrates that the `DeepMass` estimate indeed recovers the Bayesian posterior mean.

`FlowPM` TensorFlow-based fast N-body code (Modi, Lanusse, & Seljak 2021b) for this likelihood which becomes just a layer within a neural network (Figure 46). The authors propose to initialise the reconstruction at the standard linear reconstruction, and show that in only 10 iterations this method yields a better solution than an iterative reconstruction based on 400 iterations of an LBFGS minimiser.

Summary of Deep learning for Cosmology

1. Emulation

- The main motivation for using deep learning in simulations is to bypass some of the expensive computational steps needed to generate large volume and high-resolution simulations needed to model modern surveys. Applications have targeted in particular: emulating N-body simulations, enhancing the resolution of existing simulations (super-resolution), and learning mappings between the 3D dark matter distribution and dark matter halos or a range of hydrodynamical fields.
- Despite many impressive results, these methods have not yet been used for scientific applications. The main difficulty is that deep models can only be used within their training regimes (defined by specific training resolution, specific sets of cosmological parameters, or specific hydrodynamical run), and thus it is unclear whether they will bring concrete computational gains when the cost of the training sets are taken into account.

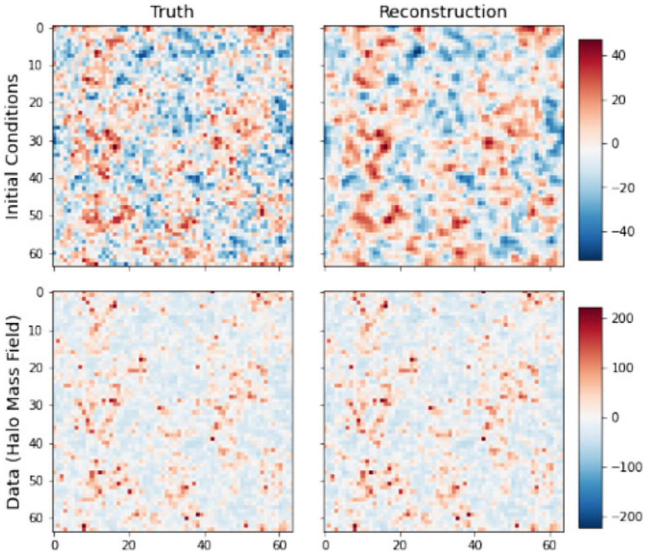

Figure 46. CosmicRIM initial conditions reconstruction technique of Modi et al. (2021a) based on a 3D LSTM recurrent neural network and includes explicitly at each iteration the gradients of the data likelihood. The plot shows a given ground truth initial condition field (top left) and associated final field (bottom left), along with the reconstructed initial conditions (top right) and reconstructed final field (bottom right).

- An alternative class of models, which so far as attracted limited attention, aims instead for a minimal set of parameters, building upon known symmetries and physical insight, greatly reducing the amount of training simulations needed

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Table 1. Overview of the different deep learning techniques used in the fields of galaxy formation and cosmology, divided by type of application (see text for details).

Application \ Model			CNNs	Enc.	Gene.	BNN	RNN	Trans.	GNN
1. Computer Vision	Classification	Morphology	🟦 ✓	🟥 ✓					
		Strong Lenses	🟦 ✓*	🟥 ✓*					
		Transients					🟦 ✓*)	🟦 ✓*)	
	Segmentation			🟦 ✓*	🟥 ✓*				
2. Galaxy Properties		Photoz	🟦 ✓			🟦 ✓			
		Structure	🟦 ✓*)						
		Stellar Populations	🟦 ✓*						
		Lensing	🟦 ✓*			🟦 ✓*			
		Physical Processes	🟦 ✓*						
		Dark Matter	🟦 ✓*			🟦 ✓*			🟦 ✓*
3. Discovery		Visualization	🟥 ✓	🟥 ✓	🟥 ✓				
		Outliers	🟥 ✓	🟥 ✓	🟥 ✓		🟥 ✓		
		Laws							🟥 ✓*
4. Cosmology		Emulation	🟦 ✓*	🟦 ✓*	🟦 ✓*	🟦 ✓*			
		Cosmological inference	🟦 ✓*		🟦 ✓*	🟦 ✓*			

CNNs: Standard classification and regression Convolutional Neural Networks including modern architectures such as ResNets. Enc: Encoder-Decoder networks and variants. Gene: Generative Models. BNNs: Bayesian Neural Networks; we also include Mixture Density Networks. RNNs: Recursive Neural Networks. Trans: Transformers. GNNs; Graph Neural Networks. A blue (red) background indicates supervised (unsupervised) learning. The star symbol highlights applications which require simulations to train the neural networks. The bracket after the star symbol indicates that the use of simulations is not always mandatory.

and even opening the possibility of inferring these parameters from the data itself.

2. Cosmological Inference

- Deep learning is opening a new way of comparing observations to theory: 1. it allows for the automatic extraction of cosmological information from high-dimensional data without requiring analytic summary statistics; 2. Neural Density Estimation makes it possible to perform Bayesian inference by leveraging numerical simulations.
- Although in theory a deep learning approach is statistically sound, it assumes that the simulators provide an accurate physical model of the observations. Any unaccounted for systematics may result in biases, which due to the black-box nature of deep neural networks are difficult to test/detect.
- A number of papers are starting to apply these methodology on data. In weak-lensing, the gains compared to a more standard power spectrum analysis have remained limited on current generation surveys when systematics are included and marginalised over in the analysis.
- A few applications have proposed to perform high-dimensional inference of cosmological fields (e.g. dark matter maps, reconstructing initial conditions). These works however do not yet attempt joint inference of fields and cosmological parameters.

7. Final thoughts: Assessing the present and future of deep learning for galaxy surveys

This final section is devoted to extract some indicators about the impact that deep learning has had in the analysis of galaxy surveys. We also attempt to highlight some of the key challenges these methods are facing, which in some cases might prevent or delay the general deployment of deep learning for scientific analysis. Some of these challenges have already been highlighted in the previous sections, but this section tries to extract the most commonly encountered

7.1. On the penetration of deep learning techniques in astronomy

We start by questioning what are the deep learning techniques most commonly used in astronomy and how efficiently the rapid progresses made in the ML community reach our community. In the previous sections, we have described different applications of deep learning making use of a variety of techniques. We summarise in Table 1, the broad type of neural network architectures used in the four categories of scientific applications we defined in this review. We have divided the neural network models in seven big groups. CNNs encapsulates all variants of convolutional neural networks, from Vanilla to more complex Residual Networks. The second group contains, generally speaking, image to image networks such as Encoder-Decoders, Autoencoders but also segmentation specific networks such as Mask RCNNs or YOLO.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

The third family of models are Generative Models which include Variational Autoencoders, Generative Adversarial Networks and also Autoregressive models. We then include Bayesian Neural Networks which allow for uncertainty quantification (Mixture Density Networks and Flow models are also included in this category), Recursive Neural Networks and Transformers mostly suited for sequences. The last group is made of Graph Neural Networks. The table first shows that applications in astronomy cover a wide range of deep learning techniques. Although standard CNNs are the most commonly used—probably because it is the most established approach and because imaging is the most common type of data—other more recent models are regularly applied to astronomical data. On the one side, this reflects the fact that astronomical data is rather diverse—including images, but also spectra, time sequences, simulations and observations. On the other hand, it suggests that the penetration of new ML techniques is efficient. State-of-the-art methods are rapidly applied to astronomy. This is likely a consequence of the fact that, even advanced ML methods are becoming increasingly easy to use for non-experts. It is almost straightforward to test a new technique on an astrophysical problem with current high level implementations. The downside is that, generally speaking, the methods are often applied *blindly* off the shelf, with little domain specific adaptation. Consequently, a feature that is still lacking in a fair amount of the applications of deep learning to astronomy is the inclusion of previous physical knowledge into the data-driven models. This can be done by adapting the loss functions or by modifying the neural network architectures to incorporate known symmetries (see work by Villar et al. 2021a; Bowles et al. 2021). It obviously requires deeper knowledge of the machine learning aspects which is something that will likely take more time.

Another interesting feature that emerges from Table 1 is that training on simulations is the most common approach in astronomy. Almost all supervised approaches rely at some stage on simulated data. It reflects that the samples of labelled data remain small and/or that the measurements in observations are noisy. Relying on simulations to train the models adds however an important element of uncertainty to all applications. Machine learning approaches are indeed very sensitive to domain shift issues. According to Table 1, almost all recent applications are affected by those at some extent. We will discuss this further in Subsection 7.3.

7.2. Measuring the impact of deep learning

We now move to measuring the impact of works using deep learning in the astronomical literature. We have seen in the introduction that the number of papers making use of neural networks has increased exponentially over the past half decade. In this subsection, we try to measure the impact of these works with some standard metrics. Figure 47 shows the evolution of the number of papers, number of citations and average citations per paper in the period 2015–2021. The publications are divided in the four different categories defined in this review, that is, computer vision, galaxy properties, discovery and cosmology. We have only included in the figure the works explored for this work. As a consequence, it is very likely that the figure is not complete and that the numbers presented are closer to a lower limit. However, it should provide a good overview of the general trends and represents a more controlled experiment than a purely automatic search. We also emphasise that the division in categories is a choice by the authors of this review. Therefore, there is some obvious overlap between the different types of applications.

Nevertheless, the figure reveals some interesting behaviours. We first confirm the global increasing trend of the number of papers using deep learning for galaxy surveys. Since 2015 there is a clear exponential increase. In 2021, there are at least 70 papers using deep learning in the context of galaxy surveys, while there were less than 5 in 2015. This is factor of ∼15 increase and implies more than a paper per week on average. If we look at the division per type of application, we see that computer vision type of applications (i.e. classification, segmentation) concentrate the largest fraction of publications. All the other remaining classes share similar fractions. However, there is clearly a decreasing trend of the relative importance of classification and segmentation applications. While the fraction of these papers was around 70–80% in 2015–2017, it is only of ∼20% for papers published in 2021. The trends seem to suggest a diversification of the applications of deep learning to astronomy moving from computer vision tasks—mainly classification and segmentation—to a large variety of different applications. The number of yearly works for data processing seems to flatten indeed after 2019, while other applications like data exploration (*discovery*) rapidly rise.

A similar behaviour is observed in terms of citations. In 2016, roughly 80% of the citations are for papers using deep learning for computer vision tasks. The fraction is steadily decreasing, but still remains close to ∼50% in 2021 even though the number of papers is only 20%. It probably reflects a delay between the publication time and the time papers start to be cited.

We attempt to remove this effect in Figure 48. We plot the number of papers per year as a function of the number of citations per paper and per year, averaged over the time elapsed since the first citation in a given group. That way, for papers focusing on computer vision, the time window is set to 6 years, while for the others, we consider 4 yr. We also show in the figure, for reference, the location of publications flagged with the keywords *galaxy evolution*, according to the NASA ADS search engine. Using this normalisation, we observe several interesting trends. Works using deep learning for galaxy surveys, represent roughly >5% of all works focusing on galaxies. This is not a large fraction but it is still remarkable given the relatively recent emergence of deep learning. They receive on average ∼1.5 times less citations per publication. It suggests that the impact of deep learning papers remains moderate as compared to the average. We might speculate about possible reasons. One possibility is that most of the works making use of deep learning are thought in preparation of future big data surveys which have not arrived yet (e.g. Euclid, LSST). The works are therefore more a demonstration of feasibility. It could be argued that existing surveys such as DES or SDSS for example are already good targets for data-driven science. This is certainly true, and as we have seen in this review, there are many works targeting these surveys. However, the sizes and dimensionality of current surveys still allow one to use relatively optimised codes and leave some room for some manual checks. This is also supported by the trends observed when the deep learning papers are divided by topic. The highest citation rate is measured for works focusing on simulation (i.e. emulation of cosmological simulations), which by definition do not require new observational data. Publications focused on unsupervised discovery, which strongly rely on new data being available, present the lowest impact, although they show a strong increase in numbers (Figure 47). It could also be that the technology is still too young,

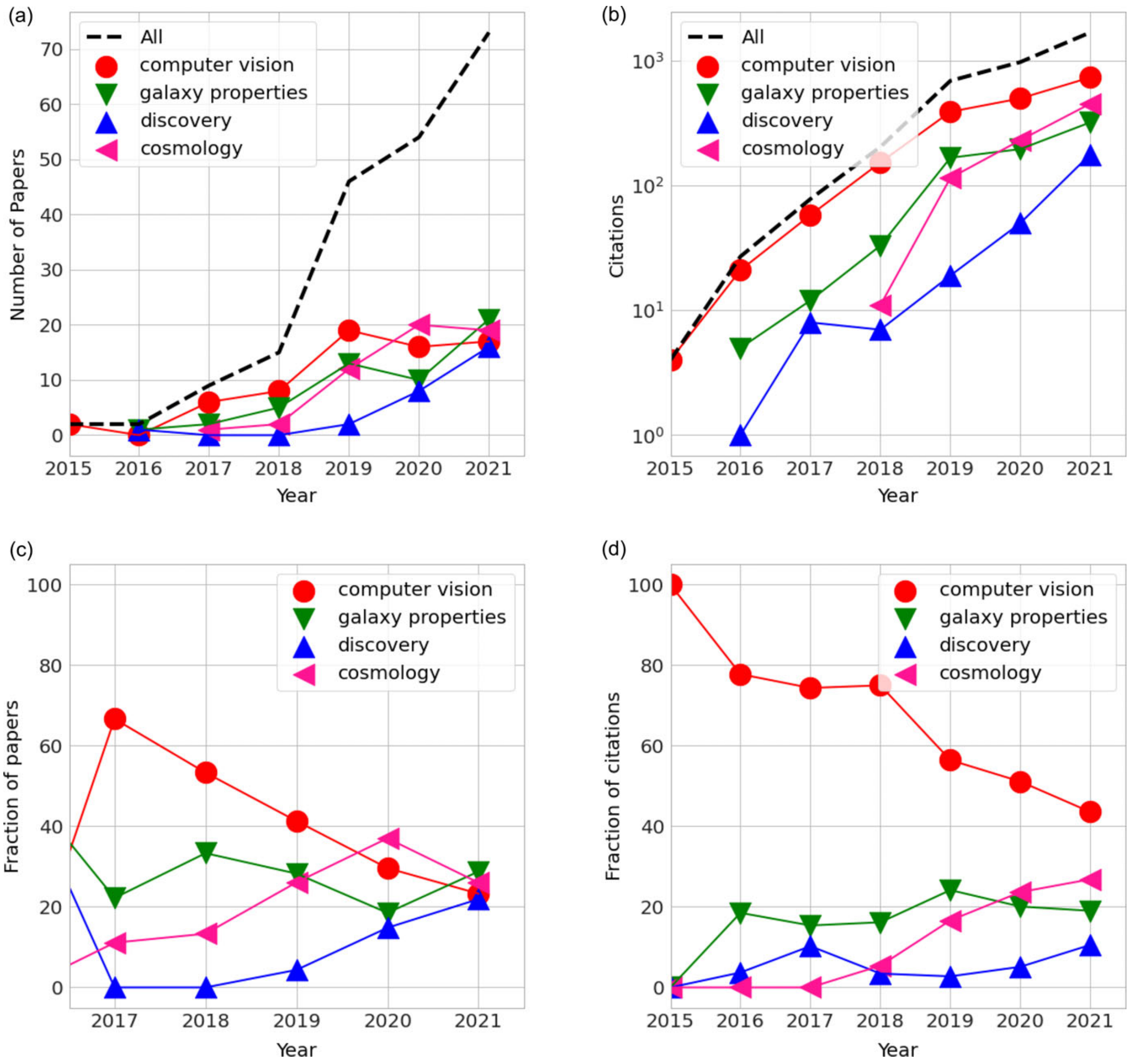

Figure 47. Impact of works using deep learning for galaxy surveys. Each symbol shows a different class of application as labelled (see text for details). The top left and right panels show the number of papers and number of citations as a function of time respectively. The bottom left and right panels show the fraction of papers and citations in each class of application.

and it is just a matter of time that the impact increases. The majority of the works are still at the proof-of-concept stage and have not reached the deployment stage. This could also mean eventually that there are some challenges/limitations which have not been deeply explored yet and that prevent these new methods to be fully adopted by the community. We explore these challenges more carefully in the following section.

7.3. Challenges

We list in Table 2 what we think are some of the major challenges that deep learning works face and which need to be addressed in the coming years by the community based on the works reported in this review. Some of these challenges are not specific to the astronomical community and can benefit from solutions arising from the field of Machine Learning. However, in some cases, the requirements are more strict in astronomy. The table also provides some possible solutions along with some list of—non-exhaustive—references which have explored these solutions.

7.3.1. Small (and biased) labelled datasets

A major challenge in applications of deep learning for astronomy is the lack of large enough labelled data sets to train supervised

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Table 2. Major challenges that deep learning works applied to astronomy might suffer and that will need to be addressed in the coming years. We also provide elements of solutions already being explored along with the corresponding references.

Challenge 1 Small (and biased) labelled datasets	
Solution 1.A Transfer Learning	Domínguez Sánchez et al. (2019), Samudre et al. (2022), Lukic et al. (2019)
Solution 1.B Simulated dataset	Jacobs et al. (2017), Vega-Ferrero et al. (2021)
Solution 1.C Self-supervised learning	Hayat et al. (2021)
Solution 1.D Active Learning and similar	Walmsley et al. (2020)
Challenge 2 Uncertainty	
Solution 2.A Bayesian approximations	Walmsley et al. (2020), Perreault Levasseur et al. (2017)
Solution 2.B Density Estimators	Kodi Ramanah et al. (2020)
Challenge 3 Interpretability	
Solution 3.A Saliency maps and similar	Huertas-Company et al. (2018), Bowles et al. (2021), Bhambra et al. (2022)
Solution 3.B Symbolic regression	Cranmer et al. (2020)
Solution 3.C Physics informed	Scaife & Porter (2021), Villar et al. (2021a), Charnock et al. (2019)
Challenge 4 Domain shift	
Solution 4.A Transfer Learning	Tuccillo et al. (2018), Domínguez Sánchez et al. (2019), Ghosh et al. (2020)
Solution 4.A Domain Adaptation	Ćiprijanović et al. (2021b)
Challenge 5 Benchmarking	
Solution 5.A Standardised datasets	PLAsTiCC, SKA data challenge, Galaxy Zoo

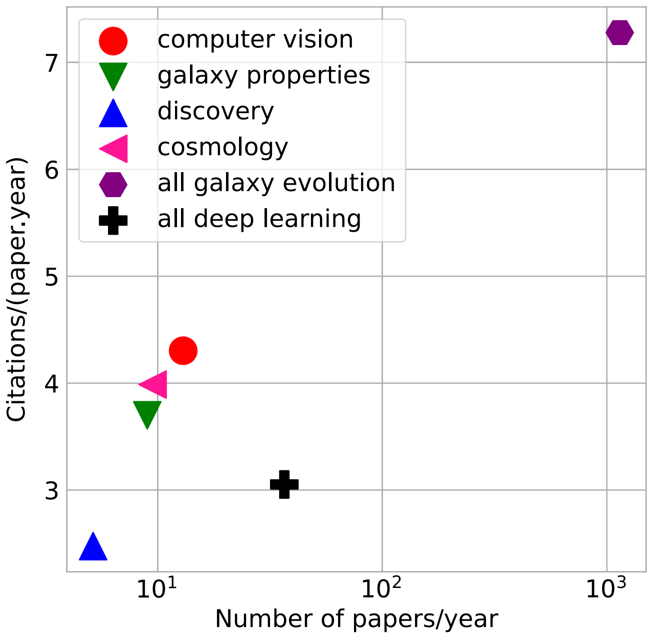

Figure 48. Number of citations normalised by the number of papers and years (from the first publication in that category) as a function of the number of papers per year. Each symbol shows a different category as defined in this work (see text for details). The ‘all galaxy evolution’ group includes an automatic search of all publications in the field of galaxy formation (with or without deep learning).

deep learning models. This issue has been highlighted multiple times during the description of works. We have seen in Section 3 that deep learning was first applied to classification of galaxy morphologies. This was mainly triggered by the fact that large samples of labelled images existed from previous citizen science efforts. A similar behaviour is observed in the radio domain. However, the vast majority of the data are unlabelled preventing the deployment of supervised deep learning methods. Although this issue is common to many disciplines, it is particularly critical in astronomy since the properties of the images change depending on the instrument used. Therefore, a sample of labelled images from one observational dataset cannot directly be exported to another survey, hampering the use of deep learning. We notice however, that large dataset are not always needed. Some works (e.g. Ackermann et al. 2018; Walmsley et al. 2019) have demonstrated that reliable results can be obtained with small training samples. This is probably a consequence of the relatively limited complexity of astronomical images compared to natural images.

Nevertheless, reducing the amount of labels is usually desirable for most supervised applications. There exist several types of solutions listed in Table 2. An obvious one is to manually label more data. It is however very time consuming and cannot be done for every new survey. Active learning approaches allow to label only the more informative examples for the deep learning model, hence reducing the needed time. Active learning has been particularly explored in the framework of galaxy morphology (e.g. Walmsley et al. 2020). Another fairly straightforward solution is Transfer Learning. This is usually done by training a model in a similar dataset to the target dataset for which labelled data exist. The neural network weights are then refined with a small sample of labelled examples from the target sample. Transfer learning has also been explored for galaxy morphology (e.g. Domínguez Sánchez et al. 2019). Other works use simulations to overcome the lack of labels. As seen in the previous section, using simulations is a fairly common approach in astronomy. This is probably because the complexity of astronomical objects—especially images—allows one to obtain quite realistic simulations with fairly simple approaches. This is the case, for instance, for the classification of strong lenses, which exclusively rely on simulated training sets. Although this is a very efficient approach, it also implies some potential issues related to the change of domain. This is a general challenge which we have grouped into Challenge 4 in Table 2. Finally, the ML community has recently started to explore the so-called self-supervised approaches to reduce the amount of labelled examples. The underlying idea is to first compute some meaningful representations of the data in an unsupervised way and then use the obtained representations to train a supervised network. Because the self-supervised step filters out non-informative features, the amount of needed examples for training is reduced. Hayat et al. (2021) and Sarmiento et al. (2021) have recently explored self-supervised learning for galaxy morphology, photometric redshifts and galaxy kinematics respectively. Contrastive self-supervised learning can also be potentially employed to reduce the gap between simulations and observations as well as Autoencoder representations.

In addition to the size of the labelled datasets, a major challenge is how representative are those. A common problem in astronomy is that the sample for which labels are available does not always overlap with the inference datasets. This typically could happen because it is easier to obtain labels for a sub population of objects. The extrapolation of the trained neural network results to a dataset which was not exactly used for training is usually a problem. This issue has become very obvious for photometric redshift estimation (see Section 4) in which ML approaches tend to fail for examples poorly represented in the training. The only solution adopted by the community has been to obtain more training data.

7.3.2. Uncertainty

Uncertainty quantification is a major challenge for applications of deep learning to physical sciences. This is a major difference with respect to standard computer vision applications on natural imaging, which usually do not require well calibrated uncertainties. Therefore, standard deep learning methods directly exported to astronomy do not quantify uncertainties and this is generally not acceptable for scientific applications. Some tasks such as classification can sometimes be accepted without precise uncertainty and rely on statistical measurements (i.e. ROC, Precision-Recall curves). However, if deep learning is intended to be used for accurate measurements of galaxy properties or to constrain models, they need to incorporate error measurements. We see a changing trend in the community. The early efforts did not include uncertainty estimation; however more and more works are introducing at least some quantification of errors. The community has explored several solutions, but the problem is far from being solved. A promising approach are Bayesian Neural Networks (BNNs) which aim at measuring posterior distributions from a Bayesian perspective. Several approaches have been proposed over the past years and the implementation has also become more straightforward, which favours the use by non-experts. For example BNNs are the common approach for the modelling of strong lenses for example (e.g. Perreault Levasseur et al. 2017). However, one need to keep in mind that BNNs compute an approximation of the true posterior distributions which sometimes might not be accurate enough (see review by Charnock et al. 2020). Density Estimator networks such as Regressive Flows or Mixture Density Networks are another approach to sample from complex posterior distributions. Although several works in astronomy pay careful attention to the quantification of uncertainties, they represent still a minority of the deep learning literature in astronomy. We believe this is a major challenge for the future deployment of deep learning in the analysis of deep surveys.

7.3.3. Interpretability

Related to the problem of uncertainty quantification is interpretability. By moving to a data-driven approach to data analysis, we unavoidably loose some control on what type of information is extracted and used by the neural network models. This effect is sometimes referred as the *black-box effect*. Deep learning models are in general opaque black boxes which perform complex non-linear mappings, difficult to unveil. Although this might be a problem for most applications, it is particularly worrying for scientific ones, and therefore constitutes a major challenge for the acceptance of deep learning by the astronomical community. For example, not properly understanding the information used can generate some biases, as demonstrated by Dhar & Shamir (2022) which show that deep learning based classifications are sensitive to the location of the galaxy in the sky. The field of interpretability of deep neural networks is even less developed than the one of uncertainty estimation and in general the techniques employed provide a limited amount of information. A common approach is to identify the regions of the input data that provide most of the information for the network decisions. These methods can be generally useful to identify biases—for example, the neural network model focuses on background noise—but are still far from providing any physical interpretation of what is being measured. Some works have looked at ways of enhancing the explainability (Huertas-Company et al. 2018; Bhambra, Joachimi, & Lahav 2022) but the amount of extracted information typically consists on the identification of pixels in the input images which contribute most to the decision. Although this is certainly valuable information to identify biases in particular, it does not provide a true explainability in terms of physical meaning. Interpreting the results is easier when the inputs are parameters instead of raw pixel values. In such cases, there exists the possibility of performing symbolic regressions to try to dig into the relations learned by the neural networks (e.g. Cranmer et al. 2020; Villaescusa-Navarro et al. 2022). An interesting research line to ease interpretability is the inclusion of prior physical constraints in the neural network model. Architectures that preserve known symmetries of the physical problem are for example an interesting way to keep a control on what the networks are extracting (e.g. Scaife & Porter 2021; Villar et al. 2021a; Bowles et al. 2021).

7.3.4. Domain shift

Another key challenge faced by deep learning applications to astronomy is related to the change of domains between the training and the inference steps. As highlighted in Table 1, a majority of the applications of deep learning to astronomy rely on simulations for training the models. This is typically justified because the availability of labelled samples is limited (see Subsection 7.3.1), because we aim at accessing information that is only available on simulations, for example, galaxy mergers, dark matter or information about the cosmological model or because the likelihood is intractable but we have an idea on how to simulate the data (see review by Cranmer, Brehmer, & Louppe 2019). For example, the work by Bottrell et al. (2019) examines very carefully the impact of using more or less realistic simulations for training. The community has explored several solution to mitigate the impact of training on simulations and apply to data. A simple approach is to use transfer learning. This is only possible when there exist some measurements in observations which can be used to fine tune the weights from the neural network model trained on simulations (e.g. Tuccillo et al. 2018). This is not always possible though, especially when we try to infer the parameters of a model and has also the problem of propagating the biases of any existing method previously applied to observations. Domain adaptation techniques are another alternative approach which attempt the make the features learned by the model agnostic to the differences between domains. As opposed to transfer learning, this is done during training so that no domain specific features are learned. Ćiprijanović et al. (2021a) have recently quantify the gain of such techniques for the identification of galaxy mergers. It remains however an open issue for the future.

7.3.5. Benchmarking and deployment

A final challenge which has not been discussed much in the literature so far is related to how the different approaches can be robustly compared. As we have thoroughly described in this review, the past years have witnessed an emergence of a large number of deep learning methods applied to a diversity of scientific topics. In many cases, the results are shown for a specific dataset, with a specific configuration, which makes it hard to compare with existing approaches. The ML community has been using since many years, what is called standardised datasets. These are common datasets which are publicly shared and on which any new approach is usually tested. This benchmarking approach has been an important channel for progress in the community. The astronomical one is not used to this type of approach and therefore, with some noticeable exceptions (e.g. PLAsTiCC, Galaxy Zoo for classification), we lack of a coherent way of comparing methods. We argue that this is an important aspect on which to work as a community to boost progress. Having standardised datasets on which test models can not only help comparing methods but also identify pitfalls and biases and therefore contribute to make the neural network models more robust. This is an important step towards a full deployment of these approaches into scientific pipelines.

8. Summary

This work reviews the use of modern deep learning techniques for the analysis of deep galaxy surveys. Although machine learning has been used in astronomy for several decades, the recent deep learning revolution as induced an unprecedented number of new works exploring the use of these novel techniques in astronomy. The purpose of this review is to assess how deep learning has been used in astronomy and what are the key achievements and challenges. We do not describe however the technical aspects of deep learning techniques.

We have divided deep learning applications in four broad categories defined by the type of application: 1—computer vision, 2—galaxy properties, 3—discovery and 4—cosmology. The first sections of these review (Sections 3–6) describe the most relevant works in each category. A summary of the main points for each type of application is included at the end of the corresponding section. The first category (Section 3) includes general computer vision applications such as classifications and object detection. The second (Section 4) is related to measure galaxy properties, deep learning acts as a fast emulator and as universal approximator. The third category (Section 5) illustrates all efforts related to visualisation and identification on new types of objects. The fourth group (Section 6) contains publications which use deep learning for cosmology. Namely we include two main applications: more efficient simulation and cosmological inference.

The last section (Section 7) focuses on extracting some lessons about the use of deep learning techniques in astronomy, on the impact they have had so far and on what are—in our humble opinion—the key challenges that will need to be addressed in the near future. We list below key take away messages from this analysis:

- The first work using deep learning in astronomy is from 2015. Since then, the number of works using deep learning for galaxy surveys has increased exponentially. There is factor of ∼15 increase between 2015 and 2021.
- The most common deep learning method used are sequential Convolutional Neural Networks with different degrees of complexities. However, there is a good variety of techniques which have been tested for astronomy including recent developments such as Transformers or self-supervised approaches. This reflects a *democratisation* of these techniques which are becoming increasingly easy to use. However, the methods are often applied with a limited amount of physically driven modifications. The combination of previous physical knowledge with data-driven models is still an open issue, even if it is a rapidly changing field.
- The majority of the works ($>50\%$) focus on what we call computer vision applications—which essentially include classification and segmentation. This is also the field in which deep learning has brought the most important breakthroughs. These are applications which are more prone to a direct import from the ML community. However, we measure a diversification of the applications which span a variety of topics such as the acceleration of cosmological simulations, the inference of galaxy properties or constraints on cosmology.
- The works using deep learning represent ∼5% of all works on galaxy formation which is remarkable. The receive however ∼3 citations per paper and per year on average. This is roughly ∼1.5 times less citations than publications on galaxies for example, although computer vision applications also perform better in this front. It suggests a moderate impact of deep learning so far which might be explained because most of the works are still at the exploratory stage.
- We have identified a set of 5 major challenges which frequently appear in deep learning applications and that we believe need to be addresses in the nearby future. 1—Small labelled datasets; 2—Uncertainty estimation; 3—Interpretability; 4—Domain shift and 5—Benchmarking.

A. Acronyms

We summarise in this appendix the acronyms used for designating types of deep learning methods, galaxy surveys as well as simulated datasets. For every method we also indicate a reference where more details can be obtained.

Machine Learning:

- **ANN:** Artificial Neural Network
- **ARF:** Auto Regressive Flow—Papamakarios, Pavlakou, & Murray (2017)
- **BNN:** Bayesian Neural Network—Charnock et al. (2020), Goan & Fookes (2020)
- **CAE:** Convolutional Autoencoder
- **CNN:** Convolutional Neural Network
- **DT:** Decision Tree
- **(W)GAN:** (Wasserstein) Generative Adversarial Network—Goodfellow et al. (2014), Arjovsky, Chintala, & Bottou (2017)
- **GNN:** Graph Neural Network
- **Mask R-CNN:** Mask Region Convolutional Neural Network—He et al. (2017)
- **MLP:** Multi-Layer Perceptron
- **MDN:** Mixture Density Network—Bishop (1994)
- **RF:** Random Forest
- **RNN:** Recursive Neural Network
- **SOM:** Self-Organising Map
- **SVM:** Support Vector Machines
- **VAE:** Variational Autoencoder—Pu et al. (2016)
- **YOLO:** You Only Look Once—Redmon et al. (2015)

Deep galaxy surveys where deep neural networks have been applied:

- **CANDELS:** Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey; Koekemoer et al. (2011)
- DECaLS: The Dark Energy Camera Legacy Survey; Dey et al. (2019)
- **DES:** The Dark Energy Survey; Dark Energy Survey Collaboration et al. (2016)
- **Euclid:** Laureijs et al. (2011)
- **HSC:** Hyper Suprime Cam; Aihara et al. (2018)
- MaNGA (SDSS IV): Mapping Nearby Galaxies at APO; Bundy et al. (2015)
- **Pan-STARRS:** Panoramic Survey Telescope and Rapid Response System; Chambers et al. (2016)
- **PAU:** Physics of the Accelerating Universe;
- **SDSS I / II Legacy Surveys:** Sloan Digital Sky Survey;
- **LSST:** Legacy Survey of Space and Time; Ivezić et al. (2019)
- **S-PLUS: Southern Photometric Local Universe Survey**; Mendes de Oliveira et al. (2019)
- **GAMA:** Galaxy and Mass Assembly; Driver et al. (2011)
- **ZTF:** Zwicky Transient Facility; Bellm (2014)

Simulated datasets used to train deep neural networks:
CAMELS: Villaescusa-Navarro et al. (2021b)
PLAsTiCC: Photometric LSST Astronomical Time-Series Classification Challenge
IllustrisTNG: Pillepich et al. (2018)
EAGLE: Schaye et al. (2015)
SIMBA: Davé et al. (2019)
VELA: Ceverino et al. (2015)
Horizon-AGN: Dubois et al. (2014)

Acknowledgements. MHC acknowledges financial support from the State Research Agency (AEI-MCINN) of the Spanish Ministry of Science and Innovation under the grants "Galaxy Evolution with Artificial Intelligence" with reference PGC2018-100852-A-I00, from the ACIISI, Consejería de Economía, Conocimiento y Empleo del Gobierno de Canarias and the European Regional Development Fund (ERDF) under grant with reference PROID2020010057, and from IAC project P/301802, financed by the Ministry of Science and Innovation, through the State Budget and by the Canary Islands Department of Economy, Knowledge and Employment, through the Regional Budget of the Autonomous Community.

References

Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K., & Turp, M. D. 2018, MNRAS, 479, 415
Aihara, H., et al. 2018, PASJ, 70, S4
Alexander, S., Gleyzer, S., McDonough, E., Toomey, M. W., & Usai E. 2020, ApJ, 893, 15
Alhassan, W., Taylor, A. R., & Vaccari, M. 2018, MNRAS, 480, 2085
Allam Jr., T., & McEwen, J. D. 2021, Technical report, Paying Attention to Astronomical Transients: Photometric Classification with the Time-Series Transformer, https://ui.adsabs.harvard.edu/abs/2021arXiv210506178A.
Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440
Andreon, S., Gargiulo, G., Longo, G., Tagliaferri, R., & Capuano, N. 2000, MNRAS, 319, 700
Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20
Aragon-Calvo, M. A., & Carvajal, J. C. 2020, MNRAS, 498, 3713
Arcelin, B., Doux, C., Aubourg, E., Roucelle, C., & Collaboration, L. D. E. S. 2021, MNRAS, 500, 531
Aricò, G., Angulo, R. E., Hernández-Monteagudo, C., Contreras, S., Zennaro, M., Pellejero-Ibañez, M., & Rosas-Guevara, Y. 2020, MNRAS, 495, 4800
Arjovsky, M., Chintala, S., & Bottou, L. 2017, Technical report, Wasserstein GAN, https://ui.adsabs.harvard.edu/abs/2017arXiv170107875A.
Armitage, T. J., Kay, S. T., & Barnes, D. J. 2019, MNRAS, 484, 1526
Ball, N. M., & Brunner, R. J. 2010, IJMPD, 19, 1049
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., & Brunner, R. J. 2004, MNRAS, 348, 1038
Banerji, M., Abdalla, F. B., Lahav, O., & Lin, H. 2008, MNRAS, 386, 1219
Banfield, J. K., et al. 2015, MNRAS, 453, 2326
Baron, D. 2019, Technical report, Machine Learning in Astronomy: a practical overview, https://ui.adsabs.harvard.edu/abs/2019arXiv190407248B.
Baron, D., & Ménard, B. 2021, ApJ, 916, 91
Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530
Bazell, D., & Peng, Y. 1998, ApJS, 116, 47
Becker, B., Vaccari, M., Prescott, M., & Grobler, T. 2021, MNRAS, 503, 1828
Bellm, E. 2014, The Zwicky Transient Facility. eprint: arXiv:1410.8185, https://ui.adsabs.harvard.edu/abs/2014htu.conf...27B
Benítez, N. 2000, ApJ, 536, 571
Berger, P., & Stein, G. 2018, MNRAS, 482, 2861
Bernardini, M., Feldmann, R., Anglés-Alcázar, D., Boylan-Kolchin, M., Bullock, J., Mayer, L., & Stadel, J. 2021, 10.1093/mnras/stab3088

Bernardini, M., Mayer, L., Reed, D., & Feldmann, R. 2019, 10.1093/mnras/staa1911

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Bhambra, P., Joachimi, B., & Lahav, O. 2022, MNRAS, 511, 5032

Bickley, R. W., et al. 2021, MNRAS, 504, 372

Bishop, C. M. 1994, Mixture Density Networks

Bluck, A. F. L., Maiolino, R., Brownson, S., Conselice, C. J., Ellison, S. L., Piotrowska, J. M., & Thorp, M. D. 2022, Technical report, The quenching of galaxies, bulges, and disks since cosmic noon: A machine learning approach for identifying causality in astronomical data, https://ui.adsabs.harvard.edu/abs/2022arXiv220107814B.

Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476

Bom, C., Poh, J., Nord, B., Blanco-Valentin, M., & Dias, L. 2019, Technical report, Deep Learning in Wide-field Surveys: Fast Analysis of Strong Lenses in Ground-based Cosmic Experiments, https://ui.adsabs.harvard.edu/abs/2019arXiv191106341B.

Bom, C. R., et al. 2021, MNRAS, 507, 1937

Bonnett, C. 2015, MNRAS, 449, 1043

Boone, K. 2021, AJ, 162, 275

Bottrell, C., et al. 2019, MNRAS, 490, 5390

Bottrell, C., Hani, M. H., Teimoorinia, H., Patton, D. R., & Ellison, S. L. 2021, MNRAS

Boucaud, A., et al. 2020, MNRAS, 491, 2481

Bowles, M., Bromley, M., Allen, M., & Scaife, A. 2021, arXiv e-prints, p. arXiv:2111.04742

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., & Cranmer, K. 2019, ApJ, 886, 49

Bretonnière, H., Boucaud, A., & Huertas-Company, M. 2021, arXiv e-prints, p. arXiv:2111.15455

Bretonnière, H., et al. 2022, A&A, 657, A90

Buck, T., & Wolf, S. 2021, Technical report, Predicting resolved galaxy properties from photometric images using convolutional neural networks, https://ui.adsabs.harvard.edu/abs/2021arXiv211101154B.

Bundy, K., et al. 2015, ApJ, 798, 7

Burhanudin, U. F., et al. 2021, MNRAS, 505, 4345

Burke, C. J., Aleo, P. D., Chen, Y.-C., Liu, X., Peterson, J. R., Sembroski, G. H., & Lin, J. Y.-Y. 2019, MNRAS, 490, 3952

Cabayol-Garcia, L., et al. 2020, MNRAS, 491, 5392

Cabayol, L., et al. 2021, MNRAS, 506, 4048

Cai, M. X., Bédorf, J., Saletore, V. A., Codreanu, V., Podareanu, D., Chaibi, A., & Qian, P. X. 2020, Technical report, DeepGalaxy: Deducing the Properties of Galaxy Mergers from Images Using Deep Neural Networks, https://ui.adsabs.harvard.edu/abs/2020arXiv201011630C.

Calderon, V. F., & Berlind, A. A. 2019, MNRAS, 490, 2367

Campagne, J.-E. 2020, Technical report, Adversarial training applied to Convolutional Neural Network for photometric redshift predictions, https://ui.adsabs.harvard.edu/abs/2020arXiv200210154C.

Carrasco-Davis, R., et al. 2019, PASP, 131, 108006

Cavanagh, M. K., Bekki, K., & Groves, B. A. 2021, MNRAS, 506, 659

Ceverino, D., Dekel, A., Tweed, D., & Primack, J. 2015, MNRAS, 447, 3291

Chalapathy, R., & Chawla, S. 2019, Technical report, Deep Learning for Anomaly Detection: A Survey, https://ui.adsabs.harvard.edu/abs/2019arXiv190103407C.

Chambers, K. C., et al. 2016, Technical report, The Pan-STARRS1 Surveys, https://ui.adsabs.harvard.edu/abs/2016arXiv161205560C.

Charnock, T., Lavaux, G., Wandelt, B. D., Boruah, S. S., Jasche, J., & Hudson, M. J. 2019, 10.1093/mnras/staa682

Charnock, T., & Moss, A. 2017, ApJ, 837, L28

Charnock, T., Perreault-Levasseur, L., & Lanusse, F. 2020, Technical report, Bayesian Neural Networks, https://ui.adsabs.harvard.edu/abs/2020arXiv200601490C.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, Technical report, A Simple Framework for Contrastive Learning of Visual Representations, https://ui.adsabs.harvard.edu/abs/2020arXiv200205709C.

Cheng, T.-Y., Huertas-Company, M., Conselice, C. J., Aragón-Salamanca, A., Robertson, B. E., & Ramachandra, N. 2021a, MNRAS, 503, 4446

Cheng, T.-Y., et al. 2021b, MNRAS, 507, 4425

Cheng, T.-Y., Li, N., Conselice, C. J., Aragón-Salamanca, A., Dye, S., & Metcalf, R. B. 2020, MNRAS, 494, 3750

Chianese, M., Coogan, A., Hofma, P., Otten, S., & Weniger, C. 2020, MNRAS, 496, 381

Ćiprijanović, A., et al. 2021a, Technical report, DeepAdversaries: Examining the Robustness of Deep Learning Models for Galaxy Morphology Classification, https://ui.adsabs.harvard.edu/abs/2021arXiv211214299C.

Ćiprijanović, A., et al. 2021b, MNRAS, 506, 677

Cohen, S. H., Windhorst, R. A., Odewahn, S. C., Chiarenza, C. A., & Driver, S. P. 2003, AJ, 125, 1762

Collister, A. A., & Lahav, O. 2004, PASP, 116, 345

Conselice, C. J. 2003, ApJS, 147, 1

Coogan, A., Karchev, K., & Weniger, C. 2020, Technical report, Targeted Likelihood-Free Inference of Dark Matter Substructure in Strongly-Lensed Galaxies, https://ui.adsabs.harvard.edu/abs/2020arXiv201007032C.

Cranmer, K., Brehmer, J., & Louppe, G. 2019, Technical report, The frontier of simulation-based inference, https://ui.adsabs.harvard.edu/abs/2019arXiv191101429C.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., & Ho, S. 2020, Technical report, Discovering Symbolic Models from Deep Learning with Inductive Biases, https://ui.adsabs.harvard.edu/abs/2020arXiv200611287C.

D'Abrusco, R., Staiano, A., Longo, G., Brescia, M., Paolillo, M., De Filippis, E., & Tagliaferri, R. 2007, ApJ, 663, 752

D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111

Dai, B., Feng, Y., & Seljak, U. 2018, JCAP, 2018, 009

Dai, B., & Seljak, U. 2020, 10.1073/pnas.2020324118

Dai, B., & Seljak, U. 2022, Technical report, Translation and Rotation Equivariant Normalizing Flow (TRENF) for Optimal Cosmological Analysis, https://ui.adsabs.harvard.edu/abs/2022arXiv220205282D.

Dark Energy Survey Collaboration, et al. 2016, MNRAS, 460, 1270

Davé, R., Anglés-Alcázar, D., Narayanan, D., Li, Q., Rafieferantsoa, M. H., & Appleby, S. 2019, MNRAS, 486, 2827

Davidzon, I., et al. 2019, MNRAS, 489, 4817

Davies, A., Serjeant, S., & Bromley, J. M. 2019, MNRAS, 487, 5263

de Andres, D., et al. 2021, Technical report, Mass Estimation of Planck Galaxy Clusters using Deep Learning, https://ui.adsabs.harvard.edu/abs/2021arXiv211101933D.

Dey, A., et al. 2019, AJ, 157, 168

Dey, B., Andrews, B. H., Newman, J. A., Mao, Y.-Y., Rau, M. M., & Zhou, R. 2021, Technical report, Photometric Redshifts from SDSS Images with an Interpretable Deep Capsule Network, https://ui.adsabs.harvard.edu/abs/2021arXiv211203939D.

Dhar, S., & Shamir, L. 2022, Technical report, Systematic biases when using deep neural networks for annotating large catalogs of astronomical images, https://ui.adsabs.harvard.edu/abs/2022arXiv220103131D.

Diaz, J. D., Bekki, K., Forbes, D. A., Couch, W. J., Drinkwater, M. J., & Deeley, S. 2019, MNRAS, 486, 4845

Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441

Domnguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661

Domnguez Sánchez, H., et al. 2019, MNRAS, 484, 93

Driver, S. P., et al. 2011, MNRAS, 413, 971

Dubois, Y., et al. 2014, MNRAS, 444, 1453

Eisert, L., Pillepich, A., Nelson, D., Klessen, R. S., Huertas-Company, M., & Rodriguez-Gomez, V. 2022, Technical report, ERGO-ML I: Inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks, https://ui.adsabs.harvard.edu/abs/2022arXiv220206967E.

Eriksen, M., et al. 2020, MNRAS, 497, 4565

Etezad-Razavi, S., Abbasgholinejad, E., Sotoudeh, M.-H., Hassani, F., Raeisi, S., & Baghram, S. 2021

Farias, H., Ortiz, D., Damke, G., Jaque Arancibia, M., & Solar, M. 2020, A&C, 33, 100420

Feder, R. M., Berger, P., & Stein, G. 2020, PhyRvD, 102, 103504

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, MNRAS, 463, 2273

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Ferreira, L., Conselice, C. J., Duncan, K., Cheng, T.-Y., Griffiths, A., & Whitney, A. 2020, ApJ, 895, 115

Fielding, E., Nyirenda, C. N., & Vaccari, M. 2021, Technical report, A Comparison of Deep Learning Architectures for Optical Galaxy Morphology Classification, https://ui.adsabs.harvard.edu/abs/2021arXiv211104353F.

Fluri, J., Kacprzak, T., Lucchi, A., Refregier, A., Amara, A., & Hofmann, T. 2018, PhyRvD, 98, 123518

Fluri, J., Kacprzak, T., Lucchi, A., Refregier, A., Amara, A., Hofmann, T., & Schneider, A. 2019, PhyRvD, 100, 063514

Fluri, J., Kacprzak, T., Lucchi, A., Schneider, A., Refregier, A., & Hofmann, T. 2022, Technical report, A Full $w$CDM Analysis of KiDS-1000 Weak Lensing Maps using Deep Learning, https://ui.adsabs.harvard.edu/abs/2022arXiv220107771F.

Fussell, L., & Moews, B. 2019, MNRAS, 485, 3203

Fustes, D., Manteiga, M., Dafonte, C., Arcay, B., Ulla, A., Smith, K., Borrachero, R., & Sordo, R. 2013, A&A, 559, A7

Gal, Y., & Ghahramani, Z. 2015, Technical report, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, https://ui.adsabs.harvard.edu/abs/2015arXiv150602142G.

Galvin, T. J., et al. 2020, MNRAS, 497, 2730

Gan, F. K., Bekki, K., & Hashemizadeh, A. 2021, Technical report, SeeingGAN: Galactic image deblurring with deep learning for better morphological classification of galaxies, https://ui.adsabs.harvard.edu/abs/2021arXiv210309711G.

Gao, D., Zhang, Y.-X., & Zhao, Y.-H. 2008, MNRAS, 386, 1417

Ghosh, A., Urry, C. M., Wang, Z., Schawinski, K., Turp, D., & Powell, M. C. 2020, ApJ, 895, 112

Ghosh, A., et al. 2022, arXiv e-prints, p. arXiv:2207.05107

Gilda, S., de Mathelin, A., Bellstedt, S., & Richard, G. 2021, Technical report, Unsupervised Domain Adaptation for Constraining Star Formation Histories, https://ui.adsabs.harvard.edu//abs/2021arXiv211214072G.

Ginzburg, O., Huertas-Company, M., Dekel, A., Mandelker, N., Snyder, G., Ceverino, D., & Primack, J. 2021, MNRAS, 501, 730

Giusarma, E., Hurtado, M. R., Villaescusa-Navarro, F., He, S., Ho, S., & Hahn, C. 2019

Goan, E., & Fookes, C. 2020, Technical report, Bayesian Neural Networks: An Introduction and Survey, https://ui.adsabs.harvard.edu/abs/2020arXiv200612024G.

Goddard, H., & Shamir, L. 2020, ApJS, 251, 28

Gómez, C., Neira, M., Hoyos, M. H., Arbeláez, P., & Forero-Romero, J. E. 2020, MNRAS, 499, 3130

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. 2014, Technical report, Generative Adversarial Networks, https://ui.adsabs.harvard.edu/abs/2014arXiv1406.2661G.

Grover, H., Bait, O., Wadadekar, Y., & Mishra, P. K. 2021, MNRAS, 506, 3313

Gupta, A., Zorrilla Matilla, J. M., Hsu, D., & Haiman, Z. 2018, Physical Review D, 97, 103515

Hansen, S., Conselice, C. J., Fraser-McKelvie, A., & Ferreira, L. 2020, RNAAS, 4, 185

Harrington, P., Mustafa, M., Dornfest, M., Horowitz, B., & Lukic, Z. 2021

Hausen, R., & Robertson, B. E. 2020, ApJS, 248, 20

Hausen, R., & Robertson, B. 2022, Technical report, Partial-Attribution Instance Segmentation for Astronomical Source Detection and Deblending, https://ui.adsabs.harvard.edu/abs/2022arXiv220104714H.

Hayat, M. A., Stein, G., Harrington, P., Lukic, Z., & Mustafa, M. 2021, ApJ, 911, L33

He, K., Gkioxari, G., Dollár, P., & Girshick, R. 2017, Technical report, Mask R-CNN, https://ui.adsabs.harvard.edu/abs/2017arXiv170306870H.

He, S., Li, Y., Feng, Y., Ho, S., Ravanbakhsh, S., Chen, W., & Poczos, B. 2018, PNAS, 116, 13825

Henghes, B., Pettitt, C., Thiyagalingam, J., Hey, T., & Lahav, O. 2021, Technical report, Investigating Deep Learning Methods for Obtaining Photometric Redshift Estimations from Images, https://ui.adsabs.harvard.edu/abs/2021arXiv210902503H.

Hezaveh, Y. D., Perreault Levasseur, L., & Marshall, P. J. 2017, Natur, 548, 555

Hildebrandt, H., et al. 2017, MNRAS, 465, 1454

Hložek, R., et al. 2020, Technical report, Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC), https://ui.adsabs.harvard.edu/abs/2020arXiv201212392H.

Ho, M., Farahi, A., Rau, M. M., & Trac, H. 2021, ApJ, 908, 204

Ho, M., Rau, M. M., Ntampaka, M., Farahi, A., Trac, H., & Poczos, B. 2019, ApJ, 887, 25

Hocking, A., Geach, J. E., Sun, Y., & Davey, N. 2018, MNRAS, 473, 1108

Holwerda, B. W., et al. 2021, ApJ, 914, 142

Horowitz, B., Dornfest, M., Lukic, Z., & Harrington, P. 2021

Hosseinzadeh, G., et al. 2020, ApJ, 905, 93

Hoyle, B. 2016, arXiv:1504.07255 [astro-ph, physics:physics]

Huang, X., et al. 2020, ApJ, 894, 78

Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., & Sánchez Almeida, J. 2011, A&A, 525, A157

Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971

Huertas-Company, M., et al. 2015, ApJS, 221, 8

Huertas-Company, M., et al. 2018, ApJ, 858, 114

Huertas-Company, M., et al. 2019, MNRAS, 489, 1859

Huertas-Company, M., et al. 2020, MNRAS, 499, 814

Ishida, E. E. O. 2019, NatAs, 3, 680

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. 2016, Technical report, Image-to-Image Translation with Conditional Adversarial Networks, https://ui.adsabs.harvard.edu/abs/2016arXiv161107004I.

Ivezić, Z., et al. 2019, ApJ, 873, 111

Jacobs, C., Glazebrook, K., Collett, T., More, A., & McCarthy, C. 2017, MNRAS, 471, 167

Jacobs, C., et al. 2019, MNRAS, 484, 5330

Jeffrey, N., Alsing, J., & Lanusse, F. 2021, MNRAS, 501, 954

Jeffrey, N., Lanusse, F., Lahav, O., & Starck, J.-L. 2020, MNRAS, 492, 5023

Jia, P., Ning, R., Sun, R., Yang, X., & Cai, D. 2021, MNRAS, 501, 291

Kalvankar, S., Pandit, H., & Parwate, P. 2020, Technical report, Galaxy Morphology Classification using EfficientNet Architectures, https://ui.adsabs.harvard.edu/abs/2020arXiv200813611K.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. 2019, Technical report, Analyzing and Improving the Image Quality of StyleGAN, https://ui.adsabs.harvard.edu/abs/2019arXiv191204958K.

Katebi, R., Zhou, Y., Chornock, R., & Bunescu, R. 2019, MNRAS, 486, 1539

Kim, E. J., & Brunner, R. J. 2017, MNRAS, 464, 4463

Kodi Ramanah,D., Wojtak, R., Ansari, Z., Gall, C., & Hjorth, J. 2020, MNRAS, 499, 1985

Kodi Ramanah, D., Wojtak, R., & Arendse, N. 2021, MNRAS, 501, 4080

Koekemoer, A. M., et al. 2011, ApJS, 197, 36

Koppula, S., et al. 2021, Technical report, A Deep Learning Approach for Characterizing Major Galaxy Mergers, https://ui.adsabs.harvard.edu/abs/2021arXiv210205182K.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in Advances in Neural Information Processing Systems (Curran Associates, Inc.), https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Lahav, O., et al. 1995, Sci, 267, 859

Lahav, O., Naim, A., Sodré Jr., L., & Storrie-Lombardi, M. C. 1996, MNRAS, 283, 207

Lanusse, F., Ma, Q., Li, N., Collett, T. E., Li, C.-L., Ravanbakhsh, S., Mandelbaum, R., & Póczos, B. 2018, MNRAS, 473, 3895

Lanusse, F., Mandelbaum, R., Ravanbakhsh, S., Li, C.-L., Freeman, P., & Póczos, B. 2021, MNRAS, 504, 5543

Lanusse, F., Melchior, P., & Moolekamp, F. 2019, arXiv e-prints, p. arXiv:1912.03980

Laureijs, R., et al. 2011, Technical report, Euclid Definition Study Report, https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L.

Lauritsen, L., Dickinson, H., Bromley, J., Serjeant, S., Lim, C.-F., Gao, Z.-K., & Wang, W.-H. 2021, MNRAS, 507, 1546

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., & Battaglia, P. 2022, Technical report, Rediscovering orbital mechanics with machine learning, https://ui.adsabs.harvard.edu/abs/2022arXiv220202306L.

Li, L.-L., Zhang, Y.-X., Zhao, Y.-H., & Yang, D.-W. 2007, ChJAA, 7, 448

Li, Y., Ni, Y., Croft, R. A. C., Matteo, T. D., Bird, S., & Feng, Y. 2020a, PNAS 118, e2022038118

Li, R., et al. 2020b, ApJ, 899, 30

Li, R., Napolitano, N. R., Roy, N., Tortora, C., La Barbera, F., Sonnenfeld, A., Qiu, C., & Liu, S. 2021, Technical report, GAlaxy Light profile convolutional neural NETworks (GaLNets). I. fast and accurate structural parameters for billion galaxy samples, https://ui.adsabs.harvard.edu/abs/2021arXiv211105434L.

Li, X., Ragosta, F., Clarkson, W. I., & Bianco, F. B. 2022, ApJS, 258, 2

Lintott, C. J., et al. 2008, MNRAS, 389, 1179

Lochner, M., & Bassett, B. A. 2021, A&C, 36, 100481

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31

Lotz, J. M., Jonsson, P., Cox, T. J., & Primack, J. R. 2008, MNRAS, 391, 1137

Lovell, C. C., Acquaviva, V., Thomas, P. A., Iyer, K. G., Gawiser, E., & Wilkins, S. M. 2019, MNRAS, 490, 5503

Lu, T., Haiman, Z., & Matilla, J. M. Z. 2022, arXiv:2109.11060 [astro-ph]

Lucie-Smith, L., Peiris, H. V., Pontzen, A., Nord, B., & Thiyagalingam, J. 2020

Lukic, V., Brüggen, M., Banfield, J. K., Wong, O. I., Rudnick, L., Norris, R. P., & Simmons, B. 2018, MNRAS, 476, 246

Lukic, V., Brüggen, M., Mingo, B., Croston, J. H., Kasieczka, G., & Best, P. N. 2019, MNRAS, 487, 1729

Ma, Z., et al. 2019, ApJS, 240, 34

Madgwick, D. S. 2003, MNRAS, 338, 197

Madireddy, S., Li, N., Ramachandra, N., Butler, J., Balaprakash, P., Habib, S., & Heitmann, K. 2019, Technical report, A Modular Deep Learning Pipeline for Galaxy-Scale Strong Gravitational Lens Detection and Modeling, https://ui.adsabs.harvard.edu/abs/2019arXiv191103867M.

Malanchev, K. L., et al. 2021, MNRAS, 502, 5147

Mao, T.-X., Wang, J., Li, B., Cai, Y.-C., Falck, B., Neyrinck, M., & Szalay, A. 2020, MNRAS, 501, 1499

Maresca, J., Dye, S., & Li, N. 2021, MNRAS, 503, 2229

Margalef-Bentabol, B., Huertas-Company, M., Charnock, T., Margalef-Bentabol, C., Bernardi, M., Dubois, Y., Storey-Fisher, K., & Zanisi, L. 2020, MNRAS, 496, 2346

Martin, G., Kaviraj, S., Hocking, A., Read, S. C., & Geach, J. E. 2020, MNRAS, 491, 1408

Martínez-Galarza, J. R., Bianco, F. B., Crake, D., Tirumala, K., Mahabal, A. A., Graham, M. J., & Giles, D. 2021, MNRAS, 508, 5734

Maslej-Krešňáková, V., El Bouchefry, K., & Butka, P. 2021, MNRAS, 505, 1464

Mathuriya, A., et al. 2018, Technical report, CosmoFlow: Using Deep Learning to Learn the Universe at Scale, https://ui.adsabs.harvard.edu/abs/2018arXiv180804728M.

Matilla, J. M. Z., Sharma, M., Hsu, D., & Haiman, Z. 2020, PhRvD, 102, 123506

McInnes, L., Healy, J., & Melville, J. 2018, arXiv e-prints, p. arXiv:1802.03426

Melchior, P., Joseph, R., Sanchez, J., MacCrann, N., & Gruen, D. 2021, NatRvPh, 3, 712

Mendes de Oliveira, C., et al. 2019, MNRAS, 489, 241

Menou, K. 2019, MNRAS, 489, 4802

Merten, J., Giocoli, C., Baldi, M., Meneghetti, M., Peel, A., Lalande, F., Starck, J.-L., & Pettorino, V. 2019, MNRAS, 487, 104

Mesarcik, M., Boonstra, A.-J., Meijer, C., Jansen, W., Ranguelova, E., & van Nieuwpoort, R. V. 2020, MNRAS, 496, 1517

Metcalf, R. B., et al. 2019, A&A, 625, A119

Miller, A. S., & Coe, M. J. 1996, MNRAS, 279, 293

Modi, C., Feng, Y., & Seljak, U. 2018, 10.1088/1475-7516/2018/10/028

Modi, C., Lanusse, F., Seljak, U., Spergel, D. N., & Perreault-Levasseur, L. 2021a

Modi, C., Lanusse, F., & Seljak, U. 2021b, A&C, 37, 100505

Morningstar, W. R., Hezaveh, Y. D., Perreault Levasseur, L., Blandford, R. D., Marshall, P. J., Putzky, P., & Wechsler, R. H. 2018, Technical report, Analyzing interferometric observations of strong gravitational lenses with recurrent and convolutional neural networks, https://ui.adsabs.harvard.edu/abs/2018arXiv180800011M.

Morningstar, W. R., et al. 2019, ApJ, 883, 14

Moss, A. 2018, Technical report, Improved Photometric Classification of Supernovae using Deep Learning, https://ui.adsabs.harvard.edu/abs/2018arXiv181006441M.

Mustafa, M., Bard, D., Bhimji, W., Lukic, Z., Al-Rfou, R., & Kratochvil, J. M. 2019, CAC, 6, 1

Muthukrishna, D., Mandel, K. S., Lochner, M., Webb, S., & Narayan, G. 2021, Technical report, Real-time Detection of Anomalies in Multivariate Time Series of Astronomical Data, https://ui.adsabs.harvard.edu/abs/2021arXiv211208415M.

Möller, A., & de Boissière, T. 2020, MNRAS, 491, 4277

Naim, A., Ratnatunga, K. U., & Griffiths, R. E. 1997, arXiv e-prints, pp astro–ph/9704012

Ni, Y., Li, Y., Lachance, P., Croft, R. A. C., Matteo, T. D., Bird, S., & Feng, Y. 2021, 10.1093/mnras/stab2113

Ntampaka, M., Trac, H., Sutherland, D. J., Battaglia, N., Póczos, B., & Schneider, J. 2015, ApJ, 803, 50

Ntampaka, M., Trac, H., Sutherland, D. J., Fromenteau, S., Póczos, B., & Schneider, J. 2016, ApJ, 831, 135

Ntampaka, M., et al. 2019, ApJ, 876, 82

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., & Zumach, W. A. 1992, AJ, 103, 318

Odewahn, S. C., Windhorst, R. A., Driver, S. P., & Keel, W. C. 1996, ApJ, 472, L13

Ono, Y., et al. 2021, ApJ, 911, 78

Paillassa, M., Bertin, E., & Bouy, H. 2020, A&A, 634, A48

Papamakarios, G., Pavlakou, T., & Murray, I. 2017, Technical report, Masked Autoregressive Flow for Density Estimation, https://ui.adsabs.harvard.edu/abs/2017arXiv170507057P.

Park, J. W., Wagner-Carena, S., Birrer, S., Marshall, P. J., Lin, J. Y.-Y., Roodman, A., & LSST Dark Energy Science Collaboration 2021, ApJ, 910, 39

Pasquet-Itam, J., & Pasquet, J. 2018, A&A, 611, A97

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26

Pearson, J., Li, N., & Dye, S. 2019a, MNRAS, 488, 991

Pearson, J., Maresca, J., Li, N., & Dye, S. 2021, MNRAS, 505, 4362

Pearson, W. J., Wang, L., Trayford, J. W., Petrillo, C. E., & van der Tak, F. F. S. 2019b, A&A, 626, A49

Peel, A., Lalande, F., Starck, J.-L., Pettorino, V., Merten, J., Giocoli, C., Meneghetti, M., & Baldi, M. 2019, PhRvD, 100, 023508

Perraudin, N., Defferrard, M., Kacprzak, T., & Sgier, R. 2019b, A&C, 27, 130

Perraudin, N., Marcon, S., Lucchi, A., & Kacprzak, T. 2020

Perraudin, N., Srivastava, A., Lucchi, A., Kacprzak, T., Hofmann, T., & Réfrégier, A. 2019a

Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJ, 850, L7

Petrillo, C. E., et al. 2017, MNRAS, 472, 1129

Petrillo, C. E., et al. 2019, MNRAS, 484, 3879

Pillepich, A., et al. 2018, MNRAS, 473, 4077

Pimentel, O., Estevez, P. A., & Forster, F. 2022, Technical report, Deep Attention-Based Supernovae Classification of Multi-Band Light-Curves, https://ui.adsabs.harvard.edu/abs/2022arXiv220108482P.

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, AJ, 160, 45

Pruzhinskaya, M. V., Malanchev, K. L., Kornilov, M. V., Ishida, E. E. O., Mondon, F., Volnova, A. A., & Korolev, V. S. 2019, MNRAS, 489, 3591

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. 2016, Technical report, Variational Autoencoder for Deep Learning of Images, Labels and Captions, https://ui.adsabs.harvard.edu/abs/2016arXiv160908976P.

Putzky, P., & Welling, M. 2017, Technical report, Recurrent Inference Machines for Solving Inverse Problems, https://ui.adsabs.harvard.edu/abs/2017arXiv170604008P.

Qin, D.-M., Guo, P., Hu, Z.-Y., & Zhao, Y.-H. 2003, CJAA, 3, 277

Qiu, Y., & Kang, X. 2021, Technical report, Starduster: A multi-wavelength SED model based on radiative transfer simulations and deep learning, https://ui.adsabs.harvard.edu/abs/2021arXiv211214434Q.

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Rahmani, S., Teimoorinia, H., & Barmby, P. 2018, MNRAS, 478, 4416

Ramanah, D. K., Charnock, T., Villaescusa-Navarro, F., & Wandelt, B. D. 2020, 10.1093/mnras/staa1428

Ravanbakhsh, S., Oliva, J., Fromenteau, S., Price, L. C., Ho, S., Schneider, J., & Poczos, B. 2017

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2015, Technical report, You Only Look Once: Unified, Real-Time Object Detection, https://ui.adsabs.harvard.edu/abs/2015arXiv150602640R.

Reiman, D. M., & Göhre, B. E. 2019, MNRAS, 485, 2617

Remy, B., Lanusse, F., Jeffrey, N., Liu, J., Starck, J.-L., Osato, K., & Schrabback, T. 2022, Technical report, Probabilistic Mass Mapping with Neural Score Estimation, https://ui.adsabs.harvard.edu/abs/2022arXiv220105561R.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., & Lakshminarayanan, B. 2019, Technical report, Likelihood Ratios for Out-of-Distribution Detection, https://ui.adsabs.harvard.edu/abs/2019arXiv190602845R.

Ribli, D., Pataki, B. A., & Csabai, I. 2019a, NatAs, 3, 93

Ribli, D., Pataki, B. A., Matilla, J. M. Z., Hsu, D., Haiman, Z., & Csabai, I. 2019b, MNRAS, 490, 1843

Rodriguez, A. C., Kacprzak, T., Lucchi, A., Amara, A., Sgier, R., Fluri, J., Hofmann, T., & Réfrégier, A. 2018, CAC, 5, 4

Ronneberger, O., Fischer, P., & Brox, T. 2015, Technical report, U-Net: Convolutional Networks for Biomedical Image Segmentation, https://ui.adsabs.harvard.edu/abs/2015arXiv150504597R.

Sabour, S., Frosst, N., & E Hinton, G. 2017, Technical report, Dynamic Routing Between Capsules, https://ui.adsabs.harvard.edu/abs/2017arXiv171009829S.

Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502

Samudre, A., George, L. T., Bansal, M., & Wadadekar, Y. 2022, MNRAS, 509, 2269

Sánchez-Sáez, P., et al. 2021, AJ, 162, 206

Sarmiento, R., Huertas-Company, M., Knapen, J. H., Sánchez, S. F., Domínguez Sánchez, H., Drory, N., & Falcón-Barroso, J. 2021, ApJ, 921, 177

Scaife, A. M. M., & Porter, F. 2021, MNRAS, 503, 2369

Schaefer, C., Geiger, M., Kuntzer, T., & Kneib, J. P. 2018, A&A, 611, A2

Schaurecker, D., Li, Y., Tinker, J., Ho, S., & Refregier, A. 2021

Schawinski, K., Turp, M. D., & Zhang, C. 2018, A&A, 616, L16

Schawinski, K., Zhang, C., Zhang, H., Fowler, L., & Santhanam, G. K. 2017, MNRAS, 467, L110

Schaye, J., et al. 2015, MNRAS, 446, 521

Schmelzle, J., Lucchi, A., Kacprzak, T., Amara, A., Sgier, R., Réfrégier, A., & Hofmann, T. 2017, Technical report, Cosmological model discrimination with Deep Learning, https://ui.adsabs.harvard.edu/abs/2017arXiv170705167S.

Schmidt, M., & Lipson, H. 2009, Sci, 324, 81

Schmidt, S. J., et al. 2020, MNRAS, 499, 1587

Schmittfull, M., Baldauf, T., & Zaldarriaga, M. 2017, PhRvD, 96, 023505

Schuldt, S., Suyu, S. H., Meinhardt, T., Leal-Taixé, L., Cañameras, R., Taubenberger, S., & Halkola, A. 2021, A&A, 646, A126

Seljak, U., Aslanyan, G., Feng, Y., & Modi, C. 2017, JCAP, 2017, 009

Shamir, L. 2021, MNRAS, 501, 5229

Shao, H., et al. 2021, Technical report, Finding universal relations in subhalo properties with artificial intelligence, https://ui.adsabs.harvard.edu/abs/2021arXiv210904484S.

Shirasaki, M., Moriwaki, K., Oogi, T., Yoshida, N., Ikeda, S., & Nishimichi, T. 2021, MNRAS, 504, 1825

Shirasaki, M., Yoshida, N., & Ikeda, S. 2019, PhRvD, 100, 043527

Simet, M., Chartab, N., Lu, Y., & Mobasher, B. 2019, Technical report, Comparison of Observed Galaxy Properties with Semianalytic Model Predictions using Machine Learning, https://ui.adsabs.harvard.edu/abs/2019arXiv190508996S.

Smith, M. J., & Geach, J. E. 2019, MNRAS, 490, 4985

Smith, M. J., Geach, J. E., Jackson, R. A., Arora, N., Stone, C., & Courteau, S. 2022, MNRAS, 511, 1808

Snyder, G. F., Rodriguez-Gomez, V., Lotz, J. M., Torrey, P., Quirk, A. C. N., Hernquist, L., Vogelsberger, M., & Freeman, P. E. 2019, MNRAS, 486, 3702

Song, Y., & Ermon, S. 2019, Technical report, Generative Modeling by Estimating Gradients of the Data Distribution, https://ui.adsabs.harvard.edu/abs/2019arXiv190705600S.

Spiekermann, G. 1992, AJ, 103, 2102

Spindler, A., Geach, J. E., & Smith, M. J. 2021, MNRAS, 502, 985

Stahl, B. E., Martínez-Palomera, J., Zheng, W., de Jaeger, T., Filippenko, A. V., & Bloom, J. S. 2020, MNRAS, 496, 3553

Stark, D., et al. 2018, MNRAS, 477, 2513

Stein, G., Alvarez, M. A., & Bond, J. R. 2019, MNRAS, 483, 2236

Stein, G., Blaum, J., Harrington, P., Medan, T., & Lukic, Z. 2021a, Technical report, Mining for strong gravitational lenses with self-supervised learning, https://ui.adsabs.harvard.edu/abs/2021arXiv211000023S.

Stein, G., Harrington, P., Blaum, J., Medan, T., & Lukic, Z. 2021b, Technical report, Self-supervised similarity search for large scientific datasets, https://ui.adsabs.harvard.edu/abs/2021arXiv211013151S.

Storey-Fisher, K., Huertas-Company, M., Ramachandra, N., Lanusse, F., Leauthaud, A., Luo, Y., Huang, S., & Prochaska, J. X. 2021, MNRAS, 508, 2946

Storrie-Lombardi, M. C., Lahav, O., Sodre Jr., L., & Storrie-Lombardi, L. J. 1992, MNRAS, 259, 8P

Surana, S., Wadadekar, Y., Bait, O., & Bhosale, H. 2020, MNRAS, 493, 4808

Szegedy, C., et al. 2014, Technical report, Going Deeper with Convolutions, https://ui.adsabs.harvard.edu/abs/2014arXiv1409.4842S.

Tadaki, K.-I., Iye, M., Fukumoto, H., Hayashi, M., Rusu, C. E., Shimakawa, R., & Tosaki, T. 2020, MNRAS, 496, 4276

Tanaka, T. S., Shimakawa, R., Shimasaku, K., Toba, Y., Kashikawa, N., Tanaka, M., & Inoue, A. K. 2021, PASJ

Tang, H., Scaife, A. M. M., Wong, O. I., & Shabala, S. S. 2021, MNRAS

Tanoglidis, D., et al. 2021a, arXiv e-prints, p. arXiv:2109.08246

Tanoglidis, D., Ćiprijanović, A., & Drlica-Wagner, A. 2021b, A&C, 35, 100469

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, JCAP, 2013, 036

Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., & Bluck, A. F. L. 2021, Technical report, Mapping the Diversity of Galaxy Spectra with Deep Unsupervised Machine Learning, https://ui.adsabs.harvard.edu/abs/2021arXiv211203425T.

Thiele, L., Villaescusa-Navarro, F., Spergel, D. N., Nelson, D., & Pillepich, A. 2020, 10.3847/1538-4357/abb80f

Tohill, C., Ferreira, L., Conselice, C. J., Bamford, S. P., & Ferrari, F. 2021, ApJ, 916, 4

Tröster, T., Ferguson, C., Harnois-Déraps, J., & McCarthy, I. G. 2019, MNRAS, 487, L24

Tuccillo, D., Huertas-Company, M., Decencière, E., Velasco-Forero, S., Domínguez Sánchez, H., & Dimauro, P. 2018, MNRAS, 475, 894

van der Maaten, L., & Hinton, G. 2008, JMLR, 9, 2579

Vanzella, E., et al. 2004, A&A, 423, 761

Vargas dos Santos, M., Quartin, M., & Reis, R. R. R. 2020, MNRAS, 497, 2974

Varma, S., Fairbairn, M., & Figueroa, J. 2020, Technical report, Dark Matter Subhalos, Strong Lensing and Machine Learning, https://ui.adsabs.harvard.edu/abs/2020arXiv200505353V.

Varma, S., et al. 2022, MNRAS, 509, 2654

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. 2017, Technical report, Attention Is All You Need, https://ui.adsabs.harvard.edu/abs/2017arXiv170603762V.

Vega-Ferrero, J., et al. 2021, MNRAS, 506, 1927

Vernardos, G., Tsagkatakis, G., & Pantazis, Y. 2020, MNRAS, 499, 5641

Villaescusa-Navarro, F., Wandelt, B. D., Anglés-Alcázar, D., Genel, S., Zorrilla Mantilla, J. M., Ho, S., & Spergel, D. N. 2020, Technical report, Neural networks as optimal estimators to marginalize over baryonic effects, https://ui.adsabs.harvard.edu/abs/2020arXiv201105992V.

Villaescusa-Navarro, F., et al. 2021a

Villaescusa-Navarro, F., et al. 2021b, ApJ, 915, 71

Villaescusa-Navarro, F., et al. 2022

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press

Villanueva-Domingo, P., et al. 2021a, Technical report, Inferring halo masses with Graph Neural Networks, https://ui.adsabs.harvard.edu/abs/2021arXiv211108683V.

Villanueva-Domingo, P., et al. 2021b, Technical report, Weighing the Milky Way and Andromeda with Artificial Intelligence, https://ui.adsabs.harvard.edu/abs/2021arXiv211114874V.

Villar, V. A., et al. 2019, ApJ, 884, 83

Villar, V. A., et al. 2020, ApJ, 905, 94

Villar, S., Hogg, D. W., Storey-Fisher, K., Yao, W., & Blum-Smith, B. 2021a, Technical report, Scalars are universal: Equivariant machine learning, structured like classical physics, https://ui.adsabs.harvard.edu/abs/2021arXiv210606610V.

Villar, V. A., Cranmer, M., Berger, E., Contardo, G., Ho, S., Hosseinzadeh, G., & Lin, J. Y.-Y. 2021b, ApJS, 255, 24

Vojtekova, A., Lieu, M., Valtchanov, I., Altieri, B., Old, L., Chen, Q., & Hroch, F. 2021, MNRAS, 503, 3204

Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2020, ApJ, 916, 42

Walmsley, M., Ferguson, A. M. N., Mann, R. G., & Lintott, C. J. 2019, MNRAS, 483, 2968

Walmsley, M., et al. 2020, MNRAS, 491, 1554

Walmsley, M., et al. 2021, arXiv e-prints, p. arXiv:2110.12735

Walmsley, M., et al. 2022, MNRAS, 509, 3966

Wang, M., & Deng, W. 2018, Technical report, Deep Visual Domain Adaptation: A Survey, https://ui.adsabs.harvard.edu/abs/2018arXiv180203601W.

Weir, N., Fayyad, U. M., & Djorgovski, S. 1995, AJ, 109, 2401

White, R. L. et al. 2000, ApJS, 126, 133

Whitney, A., Ferreira, L., Conselice, C. J., & Duncan, K. 2021, ApJ, 919, 139

Wing Hei Yiu, T., Fluri, J., & Kacprzak, T. 2021, Technical report, A tomographic spherical mass map emulator of the KiDS-1000 survey using conditional generative adversarial networks, https://ui.adsabs.harvard.edu/abs/2021arXiv211212741W.

Wu, J. F., & Peek, J. E. G. 2020, Technical report, Predicting galaxy spectra from images with hybrid convolutional neural networks, https://ui.adsabs.harvard.edu/abs/2020arXiv200912318W.

Wu, C., et al. 2019, MNRAS, 482, 1211

Yan, Z., Mead, A. J., Van Waerbeke, L., Hinshaw, G., & McCarthy, I. G. 2020, MNRAS, 499, 3445

Yao-Yu Lin, J., Pandya, S., Pratap, D., Liu, X., Carrasco Kind, M., & Kindratenko, V. 2021, Technical report, AGNet: Weighing Black Holes with Deep Learning, https://ui.adsabs.harvard.edu/abs/2021arXiv210807749Y.

Zanisi, L., et al. 2021, MNRAS, 501, 4359

Zhang, X., Wang, Y., Zhang, W., Sun, Y., He, S., Contardo, G., Villaescusa-Navarro, F., & Ho, S. 2019

Zhou, C. C., Gu, Y. Z., Fang, G. W., & Lin, Z. S. 2021, Technical report, Automatic morphological classification of galaxies: convolutional autoencoder and bagging-based multiclustering model, https://ui.adsabs.harvard.edu/abs/2021arXiv211213957Z.

Zhu, X.-P., Dai, J.-M., Bian, C.-J., Chen, Y., Chen, S., & Hu, C. 2019, Ap&SS, 364, 55

Zwicky, F. 1933, HPA, 6, 110

https://doi.org/10.1017/pasa.2022.55 Published online by Cambridge University Press