MASTER THESIS

From Galaxies Spectra to Stellar Mass using Neural Networks on the EAGLE Cosmological Simulations

Author
E.G. van Weenen

Supervisors
Prof.dr. Joop Schaye
Dr. Camila A. Correa
Dr. James W. Trayford

LEIDEN UNIVERSITY

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science
in
Astronomy & Data Science

July 1, 2019

Abstract

The spectral energy distribution of a galaxy is a probe for its star formation activity. Artificial neural networks are used to study the relationship between modelled *ugriz* broad-band fluxes of the Sloan Digital Sky Survey (SDSS) and the stellar mass, determined using galaxies of the EAGLE (Evolution and Assembly of GaLaxies and their Environment) `RefL0100N1504` simulation at redshift $z \sim 0.1$ (Schaye et al., 2015; Trayford et al., 2015). The hyperparameters of the network are optimised using Tree-structured Parzen Estimators with 5-fold cross-validation, after which the acquired relation is applied to galaxies of a separate test set. For EAGLE galaxies and galaxies of the (Chang et al., 2015) SDSS catalogue, following a uniform logarithmic stellar mass distribution, an accuracy of $R^2 = 0.99$ and $R^2 = 0.94$ is obtained respectively, indicating a discrepancy in the photometry-stellar mass relation of EAGLE and SDSS galaxies. Attempts to diminish this discrepancy, by adding noise to the flux and stellar mass of EAGLE galaxies, or by using stellar masses of the Brinchmann et al. (2004) catalogue, do not show improvements, suggesting that differences in the neural network's predictions for EAGLE and SDSS galaxies are caused by intrinsic differences of the data. Analysing the performance of centrals and satellites separately, slightly lower prediction errors can be obtained on centrals. Finally, the importance of individual magnitudes and colours towards a prediction is analysed with the SHapley Additive ExPlanations (SHAP) framework, and a bottom-up sequential search is performed to find the magnitudes and colours that obtain the lowest prediction errors.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

The formation and evolution of galaxies is a topic of broad interest in cosmology. Physical processes occurring in galaxies such as the formation of stars leave their imprint on the electromagnetic spectrum of a galaxy. Young and massive stars emit strongly in the UV and have their peak around the blue wavelength, the optical and near-infrared is dominated by the contribution of old and non-massive stars and the mid-infrared luminosity is mostly due to dust heated by the emission of young stars (Kennicutt, 1998; Kennicutt and Evans, 2012). Therefore the spectral energy distribution of a galaxy should encode for many fundamental properties of the stellar population, such as the star formation history (SFH), stellar mass, metallicity and the initial mass function. Understanding the star formation history of galaxies allows us to constrain theories about the evolution of large scale structure in the Universe.

The field of galaxy evolution has rapidly evolved over the past twenty years. Multi-spectral imaging surveys, such as the *Sloan Digital Sky Survey* (SDSS), have collectively observed millions of nearby galaxies, the order of 10000 galaxies up to a redshift of $z \sim 6$ and thousands of galaxies up to $z \sim 8$ (Madau and Dickinson, 2014). Through wide coverage of the electromagnetic spectrum, detailed spectral energy distributions have been constructed and intrinsic properties have been estimated for a large number of galaxies (Somerville and Davé, 2015), thereby allowing the cosmic star formation history to be tightly constrained (e.g. Hopkins, 2018).

Additionally, methods of computational fluid dynamics have allowed for detailed simulations of the formation and evolution of large scale structure in the Universe. These simulations provide fundamental insights into the processes involved in galaxy formation and evolution. Examples of such hydrodynamical simulations are Illustris (Vogelsberger et al., 2014), the EAGLE project (Schaye et al., 2015), MUFASA (Davé et al., 2016) and the BAHAMAS project (McCarthy et al., 2016). The EAGLE (*Evolution and Assembly of GaLaxies and their Environments*) project is a suite of numerical hydrodynamic simulations following the evolution of structure in the Universe in volumes of 25 to 100 comoving Mpc (cMpc) in a universe following a Λ Cold Dark Matter (CDM) cosmology, with a resolution sufficient to marginally resolve the Jeans scale in the warm interstellar medium (ISM) (Schaye et al., 2015). Even though the complex underlying physics of galaxies in hydrodynamical simulations is limited by the resolution of the simulations (Schaye et al., 2010), the properties of galaxies produced in the EAGLE simulations correspond relatively well to observations.

From both observational studies of galaxy evolution and simulations, a picture arises of a Universe that was more active in the past. The star formation rate of the Universe averaged over a volume – the star formation rate density (SFRD) – has increased since the Big Bang, peaked 5 Gyr ago ($z \leq 1$) and decreased to the value it is today (e.g. Lilly et al., 1996; Madau et al., 1998; Heavens et al., 2004), as can be observed in Figure 1. Before the peak SFRD ($z > 2$), approximately 25% of the present day stellar mass had formed (Madau and Dickinson, 2014). At earlier times, the increase in the cosmic star formation rate can be explained by efficient cooling and abundant quantities of star-forming gas in dark matter halos (Hernquist and Springel, 2003). Cosmic star formation is limited at later times because of inefficient cooling due to the Hubble expansion (Hernquist and Springel, 2003). Several models of the cosmic star formation rate exist, one of which is a log-normal function of the SFRD in time (Gladders et al., 2013). Even though the log-normal function provides an excellent fit of the cosmic SFRD, as can be observed in Figure 1, it does not supply any physical explanation of the star formation history.

The cosmic SFRD describes the evolution of the star forming properties of the Universe

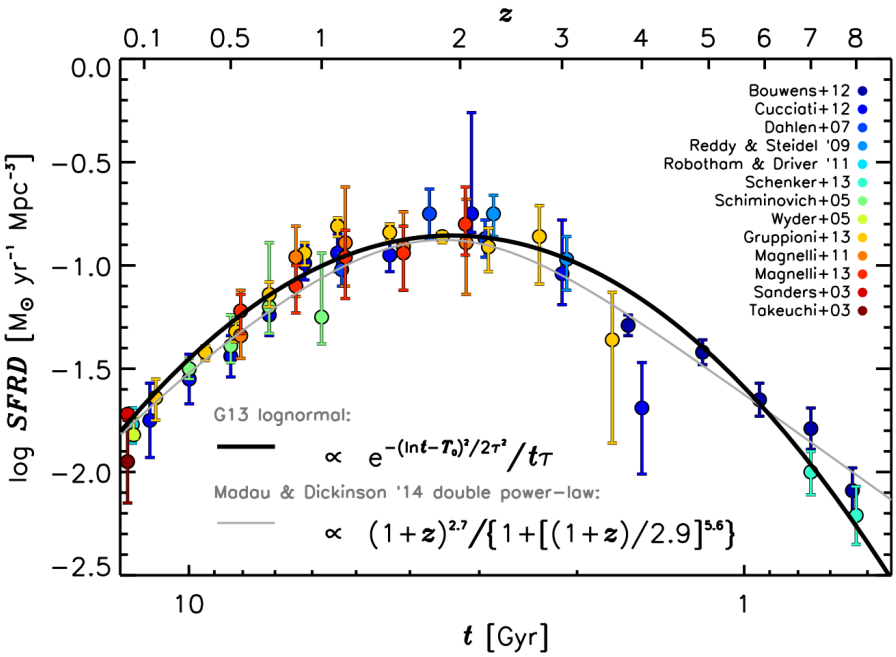
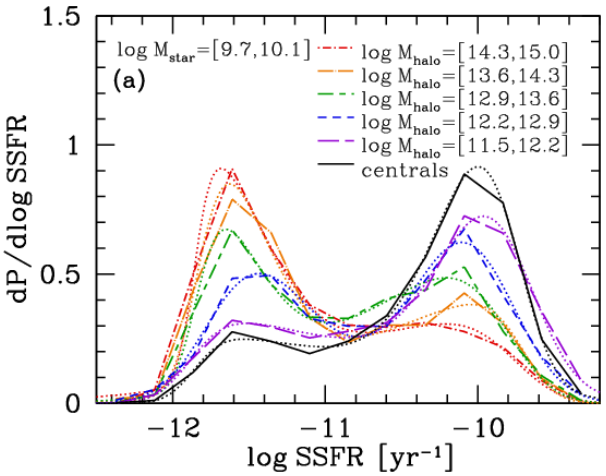


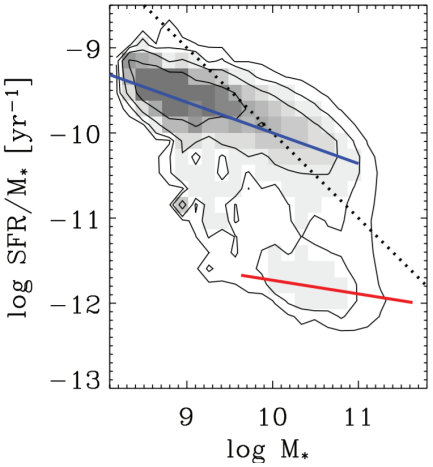
Figure 1: The cosmic star formation rate density (SFRD) can be described by a log-normal function. The coloured scatters indicate the SFRD at a certain redshift from studies collected by Madau and Dickinson (2014). The black line illustrates the log-normal fit from Gladders et al. (2013) (G13) and the grey line follows the double power law fit from Madau and Dickinson (2014). [Source: Abramson et al. (2016)]

adequately, but not of the individual galaxies that it is comprised of. The star formation history of individual galaxies consists of multiple brief star formation episodes and is therefore much more stochastic. Observations demonstrate that massive galaxies form earlier and more rapidly than low-mass galaxies, a process also known as galaxy mass assembly "downsizing" (e.g. Cowie et al., 1996; Cimatti et al., 2006; Noeske et al., 2007; Fontanot et al., 2009), which indicates a distinction in the star formation history of low- and high-mass galaxies. The star formation rate, colour, surface mass density and concentration of galaxies of constant mass has been observed to be "bimodal". Figure 2a shows the bimodally distributed specific star formation rate (sSFR = SFR/M_*) of galaxies observed in SDSS. The adjacent figure, Figure 2b, shows a diagram of the star formation rate versus stellar mass (or colour versus luminosity) in which two distributions can be identified: a distribution of red elliptical galaxies – the *red sequence*– that have their star formation extinguished (*quenched*) with predominantly old stellar populations; and a distribution of blue disc galaxies – the *blue cloud* – with high star formation rate and a relatively young stellar population that show a correlation between star formation rate and stellar mass (e.g. Brinchmann et al., 2004; Kauffmann et al., 2003). The number of galaxies in the red sequence has been increasing since $z \sim 2$ and the number of galaxies in the blue cloud has stayed relatively constant (e.g. Muzzin et al., 2013). This implies that an increasing number of galaxies has their star formation quenched.

The process that makes a galaxy transition from the blue cloud towards the red sequence is not fully understood. Several feedback mechanisms have been proposed as a possible explanation for quenching in galaxies. These feedback processes often suppress cooling and star formation and can be subdivided into two types: preventive feedback, a process that stops gas from accreting into the interstellar medium (ISM), and ejective feedback, a process in which gas is removed from the ISM after it has been accreted (Somerville and Davé, 2015). Two



(a) Distribution of the specific star formation rate (sSFR) of galaxies from SDSS DR7. The solid black line indicates central galaxies and the dotted coloured lines indicate satellite galaxies. A bimodality is observed in this distribution with a break at $s\mathrm{SFR} \approx 10^{-11}\mathrm{yr}^{-1}$. [Source: Wetzel et al. (2012)]



(b) Bimodal distribution of galaxies from SDSS DR4 as a function of specific star formation rate $\log \mathrm{SFR}/M_*$ and stellar mass M_*. The red line indicates the red sequence and the blue line indicates the blue cloud. [Source: Schiminovich et al. (2007)]

Figure 2: Bimodal distributions of galaxies from SDSS

common feedback processes are feedback from Active Galactic Nuclei (AGN) and supernova (SN) feedback. In the process of AGN feedback, energy and momentum produced by the AGN couples with gas inside galaxies. AGN can heat gas (thermal feedback), cause winds that eject gas (kinetic feedback) and change the ionisation state of gas (radiative feedback) (Somerville and Davé, 2015; McNamara and Nulsen, 2007; Fabian, 2012). Supernova explosions release energy into the ISM and perhaps even remove gas from galaxies (White and Rees, 1978; Dekel and Silk, 1986; White and Frenk, 1991), thereby modifying cooling processes and star formation.

When feedback processes are excluded, the theoretical number of galaxies formed of a certain mass or luminosity does not correspond to the number of galaxies observed. On the low-mass end, supernovae feedback is suggested to be the suppressor (Silk and Mamon, 2012) and on the high-mass end, AGN feedback could be the cause of the decrease in star formation (e.g. Khochfar and Ostriker, 2008; Johansson et al., 2009; Birnboim and Dekel, 2011). Other possible feedback mechanisms include harassment (frequent high speed galaxy encounters) (Moore et al., 1995) and strangulation (a cutoff of the supply of cold gas to the galaxy) (Peng et al., 2015), processes mainly associated with satellite galaxies that are not in the centre of the dark matter halo.

Even though many processes that trigger or quench star formation in galaxies have been proposed, the exact stellar evolution of galaxies is far from understood. Several methods can provide information about the star formation rate and stellar mass of galaxies, both probes for the star formation activity of a galaxy.

The star formation rate can be retrieved through observational tracers, i.e. monochromatic fluxes tracing the star formation rate of specific stellar populations. The UV stellar tracers represent young stellar populations, the optical ($\sim$ 0.4–0.8 μm) and near-IR ($\sim$ 0.8–3 μm) traces old and non-massive stars, and in the mid-IR ($\sim$ 3–70 μm) the luminosity is mostly due to dust heated by the emission of young stars (Kennicutt, 1998; Kennicutt and Evans, 2012).

Measuring the star formation rate through observational tracers does have its limitations. Observational tracers are sensitive to different types of galactic populations and timescales (Stensbo-smidt et al., 2017). The correlation between the star formation rate and broad-band photometric properties of galaxies cause difficulties in the derivation of the star formation rate from optical and far-IR tracers (e.g. Rafelski et al., 2016; Pearson et al., 2018) and inconsistencies occur between various SFR indicators (Davies et al., 2016). Furthermore, estimators in the UV need to be calibrated for the presence of dust (e.g. Calzetti and Kinney, 1994).

Another method for gathering information about the star formation rate is spectral energy distribution (SED) fitting. The SED of a galaxy contains information about fundamental properties of stellar populations such as the star formation history. Galaxy spectra can be created by taking the sum of the spectra of simple stellar populations (Tinsley, 1972; Searle et al., 1973; Larson and Tinsley, 1978). Within SED fitting, the observed galaxy spectrum is compared to template spectra generated by stellar population synthesis models (Walcher et al., 2011; Conroy, 2013). The physical properties of the template that provides the best match are adopted (e.g. Charlot et al., 2002). The star formation history is often parametrised as a falling exponential function modulated by the timescale (τ-models). Even though this parametrisation increases computational speed and simplicity of the fitting procedure, it imposes strong priors on the star formation rate and stellar mass derived from it (Carnall et al., 2019). In high-redshift galaxies, these parametrisations of the SFH cause underestimation of star formation rates and overestimation of ages (Lee et al., 2009). Therefore, often no functional form of the SFH is assumed (non-parametric SED fitting). The advantage of SED fitting is that it allows us to retrieve the full star formation history, stellar mass, and many other physical properties of galaxies. However, the star formation histories that are retrieved are often a poor representation of the true star formation history of a galaxy, due to the fact that not all possible spectra can be generated. Specifically, the generation of one spectrum involves a large number of free parameters that can be tweaked and is therefore computationally limited (Smith and Hayward, 2015). Additionally, the age-dust-metallicity-degeneracy introduces difficulties in estimating ages and thus also in measuring star formation histories, requiring high-quality data (Papovich et al., 2001; Conroy, 2013). Furthermore, even in non-parametric SED fitting, the choice of prior introduces a bias in the physical properties derived (Leja et al., 2019).

A relatively new method that can be used in astronomy is machine learning (ML). Machine learning algorithms allow computers to perform specific tasks without explicit instructions, thereby learning from the data itself rather than using pre-specified models. Machine learning has offered solutions to a wide variety of problems in various fields, examples of which include the recognition of handwritten digits (LeNet: Lecun et al., 1998), the recognition of objects in images (ImageNet: Deng et al., 2009), facial recognition (DeepFace: Taigman et al., 2014), automated language translation (Google Neural Machine Translation: Wu et al., 2016) and even computers exceeding superhuman levels of play in games (DeepBlue: Campbell et al., 2002; AlphaGo: Silver et al., 2016; AlphaZero Silver et al., 2017). In general, these types of problems can be subdivided into three categories: supervised learning, unsupervised learning and reinforcement learning. Supervised learning involves learning the relationship between variables while receiving feedback and it concerns classification and regression problems. Unsupervised learning entails learning from data without receiving feedback and can therefore be used in feature analysis, dimensionality reduction and clustering. Reinforcement learning is mainly used when computers need to take an action in an environment that maximizes the reward such as in robotics and games. Various machine learning algorithms and techniques

have been developed to solve these problems, including artificial neural networks (ANN) and random forests.

Within astronomy, machine learning algorithms have been proven successful in various situations. These techniques have been used in large astronomical surveys, for example in the search (or classification) of specific stellar objects (e.g. Marchetti et al., 2017; Pashchenko et al., 2018; Viquar et al., 2018), the estimation of photometric redshifts of galaxies (e.g. Collister and Lahav, 2004; Geach, 2012; Carrasco Kind and Brunner, 2013; Bilicki et al., 2014, 2016) and in galaxy (morphology) classification (e.g. Aghanim et al., 2015; Geach, 2012; Huertas-Company et al., 2015; Krakowski et al., 2016; Domíngues Sánchez et al., 2018; Siudek et al., 2018). In astronomical simulations, machine learning is used to connect galaxy and halo properties. For example, Kamdar et al. (2016), Agarwal et al. (2018) and Lucie-Smith et al. (2018) use machine learning to predict final halo and galaxy physical properties from initial dark matter halo properties in simulations. Machine learning has also been used to find a relationship between the spectral energy distribution of galaxies and their star formation history. Stensbo-smidt et al. (2017) use k-Nearest Neighbours regression to estimate star formation rates and photometric redshifts from SDSS *ugriz* photometry. With the same aim and using the same data, Delli Veneri et al. (2019) apply a Multi-Layer Perceptron algorithm (MLP) and Random Forests (RF) and focus on the selection of the most important SDSS magnitudes. Bonjean et al. (2019) use WISE luminosities in the near-IR and spectroscopic redshifts for their estimation of SDSS spectra-extracted star formation rates and stellar masses. Lovell et al. (2019) estimate star formation rates and stellar masses from the spectral energy distributions of galaxies of the EAGLE simulations at various redshifts.

In this thesis, we combine the success of machine learning in connecting galaxy and halo properties from simulations and in connecting star formation histories to photometry. We determine the relationship between the SDSS magnitudes of galaxies at redshift $z \sim 0.1$ of the EAGLE simulations (Trayford et al., 2017) and their stellar mass using supervised machine learning. As the stellar mass is equal to the integral over the star formation history up until the redshift it is observed (minus the mass loss) it is a less stochastic measure of the stars formed in a galaxy. Using machine learning to recover the stellar mass of galaxies at redshift $z \sim 0.1$ serves as an initial probe of recovering the star formation history of galaxies with machine learning. Simulations are used specifically to learn this relationship because they have the advantage of tracking properties of galaxies throughout their entire cosmic history, something that is not possible in observations. By learning the relation between photometric magnitudes and the SFH of individual galaxies directly, it should be possible to obtain unbiased star formation histories. This method is contrary to SED fitting, in which the resemblance between the SEDs of galaxies is maximised to obtain the star formation history, and for which biases still exist (Carnall et al., 2019), even when no functional form for the SFH is assumed (Leja et al., 2019). In this thesis, an artificial neural network (ANN) is used, efficient in learning complex relations without making any prior assumptions about the underlying structure of the data. An overview of the data can be found in Chapter 2 and the neural network and its optimisation is described in Chapters 3 and 4. We apply our models to galaxies observed with SDSS and their estimated stellar masses and test their accuracy (Chapter 5). Furthermore, we distinguish between central and satellite galaxies (Chapter 6) and analyse the importance of the individual magnitudes and colours for the prediction of the stellar mass (Chapter 7). In this way, we provide new insights into the process of galaxy evolution and compare it with existing methods that deduce star formation histories from SEDs.

2 Data

This thesis investigates the theoretical relation between fluxes and stellar mass of galaxies from the EAGLE simulations, and applies the trained model to broad-band fluxes and stellar mass observed in SDSS. The EAGLE simulations and SDSS observations are briefly described in Sections 2.1 and 2.2. The conversion of the data into an applicable format for the neural network is described in Section 2.3 and the conversion of the predicted values of the neural network to the original format is described in Section 2.4.

2.1 EAGLE

The EAGLE simulations are a suite of hydrodynamical simulations that can be used to explore properties of galaxy evolution. The EAGLE simulations adopt a flat ΛCDM cosmology with cosmological parameters from Planck Collaboration (2014): $\Omega_\Lambda = 0.693$, $\Omega_m = 0.307$, $\Omega_b = 0.048$, $\sigma_8 = 0.8288$, $n_s = 0.9611$ and $H_0 = 67.77$ km s^{-1}Mpc^{-1}. The simulations are run with the parallel N-body Tree-PM smoothed particle hydrodynamics (SPH) code GADGET 3 (based on the code of Springel, 2005), that computes hydrodynamics and gravitational forces on particles that represent a collisionless fluid. For processes that occur on a scale too small for the EAGLE simulations to resolve, *subgrid* recipes are adopted, to mimic their effect on galactic scales (Somerville and Davé, 2015). Subgrid models are implemented for the processes of radiative cooling, star formation, stellar mass loss and metal enrichment, energy feedback from star formation, gas accretion onto, and mergers of, supermassive black holes and AGN feedback. These subgrid models depend merely on properties of the local ISM. Star formation is implemented on a probability basis. Gas particles are eligible for star formation once they cross a pressure threshold dependent on metallicity. The eligible gas particles are assigned a probability of forming stars. The gas particles selected for star formation are then converted to star particles, transferring their metallicity as well, whereby each star particle represents a simple stellar population with a Chabrier (2003) initial mass function (IMF) and a given metallicity and age. Stellar (supernova) feedback and AGN feedback are implemented through stochastic heating (Schaye et al., 2015; Furlong et al., 2015). For each resolution, the parameters of the subgrid models are calibrated to observations of the galaxy stellar mass function (GSMF) at redshift $z \sim 0$ and the simulations show good weak convergence (Schaye et al., 2015). The galaxy properties such as stellar mass, colour and morphology are calculated within a three-dimensional 30 physical kpc (pkpc) radius with respect to the subhalo's centre of gravitational potential, resembling a two-dimensional Petrosian aperture (Li and White, 2009; Schaye et al., 2015). Halos are identified using the friends-of-friends (FoF; Davis et al., 1985) algorithm on dark matter particles. Gas and star particles are assigned to the halo of the dark matter particle they are closest to. Within halos, overdense regions - *subhalos* - are identified with the SUBFIND algorithm (Springel et al., 2001; Dolag et al., 2009). The structure with the lowest gravitational potential within a halo is identified as the *central* galaxy of the halo; all other galaxies in the halo are *satellites*.

The creation of dust attenuated broad-band fluxes of galaxies of the EAGLE database is described in Trayford et al. (2017) and we will shortly provide an overview. As each galaxy consists of star particles that represent a simple stellar population, the spectral energy distribution of a galaxy can be approximated by a superposition of simple stellar populations for which the SED is known. The population synthesis model GALAXEV (Bruzual and Charlot, 2003) is used to model the SED of each star particle, interpolating in age and stellar metallicity. However, the SED retrieved from this method is not the SED that would normally be

observed from a galaxy. Interstellar dust blocks the light at various wavelengths in galaxies, a process called *dust attenuation*, which plays an important role in modelling the observed light from a galaxy (Trayford et al., 2015). The impact of dust on the observed light of a galaxy depends on the morphology, the mixture of stellar ages present and the orientation of the galaxy. Unfortunately, dust is not a phase in the EAGLE simulations and therefore dust attenuation models have to be adopted. The Monte Carlo radiative transfer code SKIRT (Baes et al., 2003, 2011; Camps and Baes, 2015) tracks the radiative transfer of monochromatic photon packets in their trajection from the galaxy to a hypothetical detector. Dust in birth clouds of young stellar populations is modelled with MAPPINGS-III (Groves et al., 2008) and dust in the diffuse ISM is modelled by tracing the cold metal-rich gas in galaxies. The spectra are integrated for three galaxy orientations: edge-on, face-on and random. The final fluxes are calculated by convolving the integrated spectra with SDSS *ugriz* filters.

The EAGLE simulations consist of simulations with a range of resolutions and box sizes. The simulation used in this work is the `RefL0100N1504`, the simulation with the largest number of dark matter and baryonic particles (1504^3) and largest box size ($(100\ \mathrm{cMpc})^3$). The prefix denotes the subgrid model that was used, in this case `Ref` denotes the reference model. This simulation has a resolution of $1.2 \cdot 10^6 M_\odot$ and consists of 8072 galaxies at redshift $z \sim 0.1$ for which the broad-band fluxes have been modelled. From this simulation, we select the galaxies at snapshot 27 corresponding to redshift $z = 0.10063854$. We distinguish between centrals and satellites, and extract their stellar mass M_* and their modelled fluxes corrected for dust attenuation in the SDSS u, g, r, i and z bands integrated for random galaxy orientations. The corresponding query from the EAGLE database can be found in Table 14 of Appendix C.

2.2 SDSS

The galaxies of the Sloan Digital Sky Survey (SDSS) (York et al., 2000) provide an additional test set for the neural network. SDSS is a broad-band photometric and spectroscopic survey conducted with the Sloan Foundation 2.5m Telescope at Apache Point Observatory. We use the extinction corrected *ugriz* fluxes of Data Release 7 (DR7) (Abazajian et al., 2009) cross-matched with the stellar masses of Chang et al. (2015). The stellar masses are estimated from the SDSS spectroscopic galaxy sample and *WISE* 3-22 μm photometry using the SED modelling approach MAGPHYS (Da Cunha et al., 2008) that accounts for dust attenuation and emission. The 50th percentile (median) value is adopted as the estimate of the stellar mass and this catalogue contains the stellar masses and fluxes of 858 365 galaxies. The SED fitting method adopts a Chabrier (2003) Galactic disc IMF and the cosmology adopted is $(\Omega_m, \Omega_\Lambda, h) = (0.30, 0.70, 0.70)$.

The stellar masses created by Brinchmann et al. (2004), that assume the same cosmological parameters and a Kroupa (2001) IMF, are used as an alternative SDSS test set for comparison with the stellar masses from Chang et al. (2015). In case no SDSS stellar mass catalogue is defined in this thesis, we refer to the stellar masses from Chang et al. (2015).

2.3 Preprocessing

The EAGLE and SDSS data from the catalogues can not be used immediately in the neural network. The data sets first have to be preprocessed, whereby the data is converted into a format that leads to faster convergence towards a finding a solution for the relation between photometry and star formation history, or even convergence at all. The preprocessing steps are performed in the order in which they are listed.

Redshift selection The first step in converting the data from galaxies into an applicable format for the neural network is performing a selection on the redshift of these galaxies. From the EAGLE database, only the galaxies at snapshot 27 are selected, equivalent to a redshift of $z_{\text{ref}} = 0.10063854$. To ensure that a corresponding sample of SDSS galaxies is used, the SDSS galaxies within a range of $(1 \pm f) \cdot z_{\text{ref}}$ are selected,

$$z_{\text{ref}}(1 - f) < z < z_{\text{ref}}(1 + f)\,, \tag{1}$$

with fraction $f = 5 \cdot 10^{-3}$. This results in a sample of ~ 4500 SDSS galaxies.

Conversion to linear scale The EAGLE and SDSS fluxes are exponentially distributed. To convert the input to a linear scale we convert the fluxes to apparent AB magnitudes:

$$m_{AB} = -2.5 \log_{10} \left(\frac{F_\nu}{3631 \text{ Jy}} \right) . \tag{2}$$

Consecutively to remove any distance bias the apparent magnitudes are linearly scaled to absolute magnitudes.

$$M = m - 5 \log_{10} \frac{d_L}{10 \text{ pc}}\,, \tag{3}$$

with d_L the luminosity distance. The luminosity distance is calculated as follows,

$$d_L(z) = (1 + z) d_c(z)\,, \tag{4}$$

$$= (1 + z) c \int_0^z \frac{dz'}{H}\,, \tag{5}$$

with d_c the comoving distance and H the Hubble parameter,

$$H(z) = H_0 \sqrt{\Omega_{r,0}(1 + z)^4 + \Omega_{m,0}(1 + z)^3 + \Omega_{k,0}(1 + z)^2 + \Omega_{\Lambda,0}}\,. \tag{6}$$

Recall that the following cosmological parameters are used for EAGLE (Planck Collaboration, 2014) and SDSS:

	EAGLE	**SDSS**
$\Omega_{\Lambda,0}$	0.693	0.70
$\Omega_{m,0}$	0.307	0.30
$\Omega_{r,0}$	0	0
$\Omega_{m,0}$	0	0
h	0.6777	0.70

Table 1: Cosmological parameters used for EAGLE and SDSS data

The output feature stellar mass M_* is exponentially distributed as well. This feature is transformed to a linear scale by taking the logarithm of the stellar mass.

Remove missing values All galaxies from the SDSS data sets with stellar mass and flux that is missing or that is equal to zero or minus one are removed, resulting in an SDSS data set of 4494 galaxies.

Three neural network architectures The magnitudes calculated in the previous step are used to calculate the necessary colours. In our experiments three neural network architectures are defined that are differentiated by their input features:

- `nocolours` in which only the *ugriz* magnitudes of galaxies are used

$$\boldsymbol{x}^{(\text{nocolours})} = \begin{pmatrix} u & g & r & i & z \end{pmatrix}^\top ; \tag{7}$$

- `subsetcolours` in which the *ugriz* magnitudes and the basic colour combinations of these magnitudes (*u*-*g*, *g*-*r*, *r*-*i* and *i*-*z*) are used

$$\boldsymbol{x}^{(\text{subsetcolours})} = \begin{pmatrix} u & g & r & i & z & u\text{-}g & g\text{-}r & r\text{-}i & i\text{-}z \end{pmatrix}^\top ; \tag{8}$$

- `allcolours` in which the *ugriz* magnitudes and all colour combinations of these magnitudes are used

$$\boldsymbol{x}^{(\text{allcolours})} = \begin{pmatrix} u & g & r & i & z & u\text{-}g & u\text{-}r & u\text{-}i & u\text{-}z & g\text{-}r & g\text{-}i & g\text{-}z & r\text{-}i & r\text{-}z & i\text{-}z \end{pmatrix}^\top . \tag{9}$$

The architectures of the `nocolours`, `subsetcolours` and `allcolours` neural networks therefore consists of 5, 9 and 15 input nodes respectively. In all three neural networks, *ugriz* magnitudes are always used as input nodes.

Normalisation Once all features are linearly distributed, a MinMax normalisation is performed on both the input and output features. Each feature of the data is scaled linearly to fall within the range of $min = -1$ and $max = 1$. The j^{th} feature (e.g. u magnitude) of the EAGLE and SDSS data is then scaled as follows,

$$x_j := scale \cdot (x_j - \min_j x_j) + min \,, \tag{10}$$

where the variable *scale*,

$$scale = \frac{max - min}{\max_j x_j - \min_j x_j} \,, \tag{11}$$

is calculated using only the EAGLE data and is then used as a scale for both the EAGLE and SDSS data to ensure that both data sets are scaled equally.

Shuffle & split In the next step the galaxies of the EAGLE and SDSS data set are shuffled to ensure that the neural network receives galaxies with a variety of stellar masses and therefore a variety of input feature values during each epoch of training. If the neural network is shown a wide variety of input feature values, it will learn the characteristics of the relation faster (LeCun et al., 1998).

Successively, the EAGLE data is split into a training set and a test, which comprise 80% and 20% of the EAGLE data set respectively. The proportion of the training and test set is a trade-off between having enough data for the network to learn the relation between input and output features, and between estimating accurately the network's performance on unseen data.

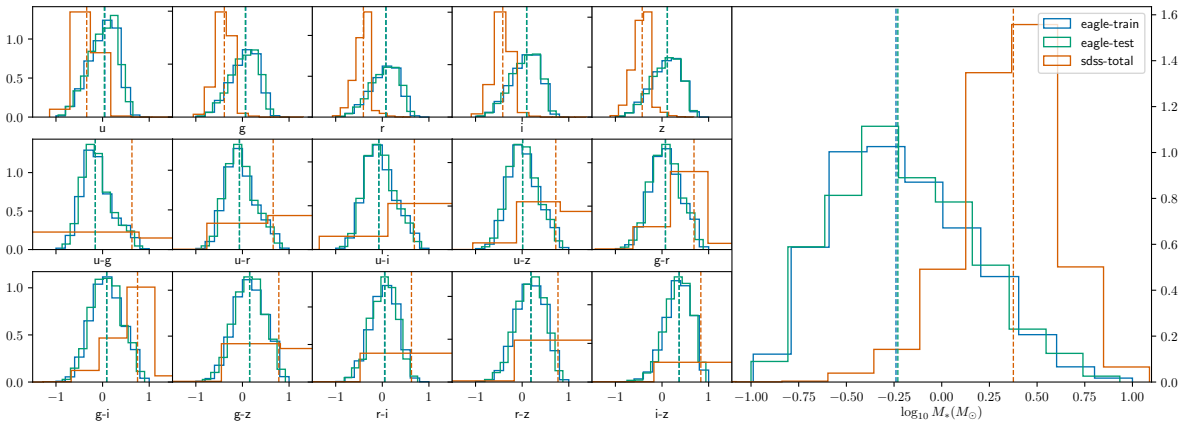
(optional) Sampling After the previous preprocessing steps, the data is distributed as shown in Figure 3a. The EAGLE galaxies are distributed between -1 and 1, as expected. Nonetheless, the SDSS data is not limited to this minimum and maximum, as the *scale* of the MinMax normalisation performed is calculated using the minimum and the maximum of the EAGLE galaxy features. Therefore, the SDSS galaxy features show a wider distribution. However, in order for the SDSS data set to function as a second test set, it is preferable that the features of SDSS galaxies follow the same distribution as those of the EAGLE galaxies. Therefore, two additional data sets are created. The preprocessing procedure remains the same for these sampled data sets, except for the shuffle & split step which is performed after the current sampling step.

Random sampling The first data set is created with the objective of having the same number of galaxies in the EAGLE test set as in the SDSS data set, since the evaluation should not be biased by the number of galaxies in the sample. Because the test set only comprises 20% of the EAGLE data set, the complete EAGLE data set should consist of five times as many galaxies as the SDSS data set. This data set, the *random* sampling set, is created by randomly selecting (sampling) $N = 125$ galaxies of the SDSS data set and $5N$ galaxies of the EAGLE data set. The final distribution of features can be observed in Figure 3b. Note that the medians of EAGLE and SDSS will not be altered by the random sampling.

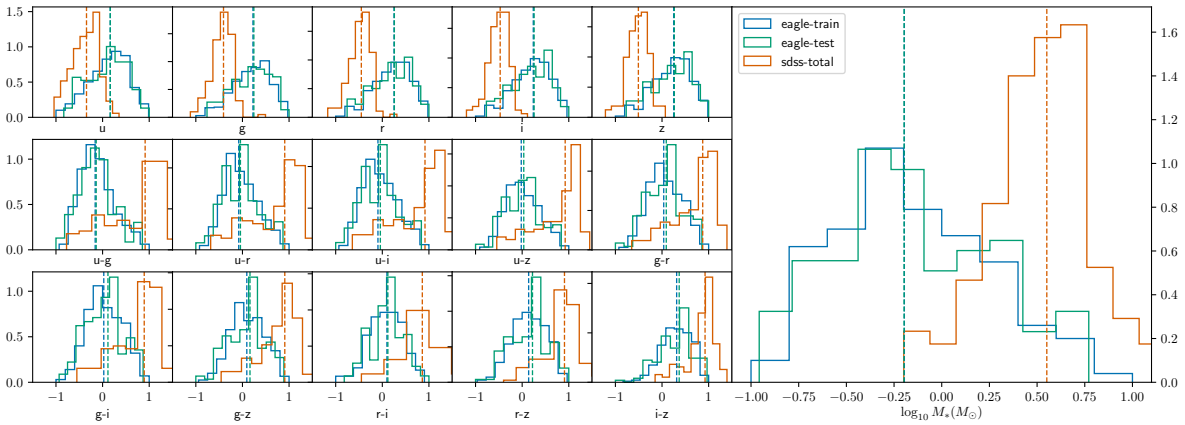
Uniform sampling Ideally, for us to make a fair comparison on the performance of the neural network on the EAGLE test set and the SDSS set, the features of both data sets should follow the same distribution. This is attempted by creating a second data set, with a uniform distribution of the stellar mass. This uniform distribution is created by dividing the stellar mass of the EAGLE and SDSS galaxies into $b = 10$ bins of size 0.3 dex where the EAGLE and SDSS histograms share the same edges. A minimum bin-count of $c = 100$ for the EAGLE galaxies and $c/5$ for the SDSS galaxies is imposed and all bins with a count underneath either the EAGLE or the SDSS threshold are disregarded. Of each remaining bin c EAGLE galaxies and $c/5$ SDSS galaxies are randomly selected to form the data set called the *uniform* sampling set. In order to compare two different sampling methods, we randomly select $5N = 625$ EAGLE galaxies and $N = 125$ SDSS galaxies from the uniform sampling set to ensure both *random* and *uniform* data sets have the same size. The data sets are then shuffled and split into a training and test set as described in the previous paragraph. By randomly splitting the data set into a training and test after the uniform sampling procedure we cannot guarantee that the stellar mass distributions is entirely uniform, but this approach should lead to a close enough uniform distribution. The distribution of features of this data set can be found in Figure 3c. The medians of the features of the EAGLE and SDSS data sets lie much closer now.

(optional) Adding noise In this thesis we explore the possibility of adding noise to the photometry and stellar masses of EAGLE galaxies. This step in the preprocessing procedure is not by default included and it will be mentioned if it is.

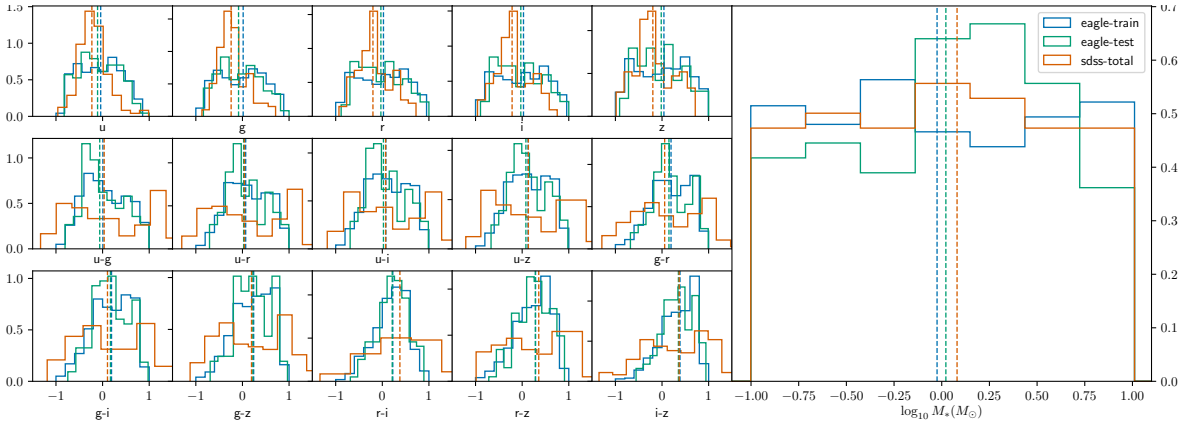
Photometry noise Noise is added to the broad-band fluxes of an individual EAGLE galaxy by randomly drawing the flux from a Gaussian distribution around the original flux and with a standard deviation determined as follows. The logarithmic SDSS flux is divided into 20 bins, and the bins that count less than 25 galaxies are disregarded. For each bin the average flux error of the SDSS galaxies in that bin is calculated. All EAGLE galaxies in this



(a) Distribution of features of 8072 EAGLE galaxies and 4494 SDSS galaxies with *no galaxy sampling* after preprocessing. There is a large difference in the medians of the EAGLE and SDSS distributions.



(b) Distribution of features of 625 EAGLE galaxies and 125 SDSS galaxies after *random sampling* of galaxies.



(c) Distribution of features of 625 EAGLE galaxies and 125 SDSS galaxies after *uniform stellar mass sampling* of galaxies. The medians of the distributions of EAGLE and SDSS galaxy features lie much closer than if no uniform sampling procedure would have been applied.

Figure 3: Distribution of features of the EAGLE `RefL0100N1504` and SDSS galaxies at redshift $z \sim 0.1$ after preprocessing. The top figure (3a) shows the feature distributions if no sampling procedure is applied after preprocessing. The bottom two figures show the distributions after *random* sampling (Fig. 3b) and after *uniform* sampling (Fig. 3c). The three rows on the LHS of each subfigure show the distribution of magnitudes and colours and the large figure on the RHS of each subfigure shows the distribution of stellar mass. The EAGLE training set is shown in blue, the EAGLE test set is shown in green and the SDSS data set is shown in orange. The vertical dashed lines show the median of distributions. The EAGLE galaxy features follow a distribution between -1 and 1, but the SDSS galaxies do not because the *scale* of the MinMax normalisation is fitted to the EAGLE data.

bin are appointed the average SDSS flux error of galaxies in that bin, which implies that the EAGLE galaxies corresponding to an SDSS logarithmic flux bin with less than 25 galaxies are removed from the sample. The standard deviation of the Gaussian distribution that the EAGLE flux is drawn from is then equal to half the appointed error. This random sampling occurs for all galaxies and all fluxes individually. After the sampling procedure, all preprocessing steps continue from paragraph *Conversion to linear scale*.

Stellar mass noise The noisy stellar masses for EAGLE galaxies are retrieved in a similar manner. The noisy stellar masses are randomly drawn from a Gaussian distribution around the logarithmic stellar mass, with the following standard deviation. The logarithmic stellar mass of SDSS galaxies is divided into 300 bins (the bins with less than 5 galaxies are disregarded), and for each bin the average mass standard deviation is calculated of the galaxies in that bin. Here the standard deviation is equal to a half times the logarithmic stellar mass at the 84^{th} percentile minus the logarithmic stellar mass at the 16^{th} percentile. The mass bins are smaller here (~ 0.010 dex) to ensure that they are larger than the error in stellar mass.

2.4 Postprocessing

After using the neural network to predict the stellar mass from our preprocessed photometry, the predicted stellar masses and the original magnitudes, colours and stellar masses have to be scaled to their original values. The MinMax normalisation performed in the preprocessing procedure is inversed to retrieve these values,

$$x_j := \frac{x_j - min}{scale} + \min_j x'_j \,, \tag{12}$$

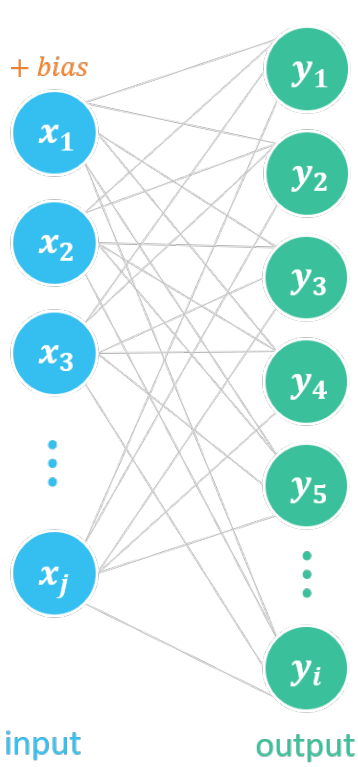
where $\min_j x'_j$ is the minimum value of the non-normalised jth feature determined in Eq. (10).

3 Neural networks

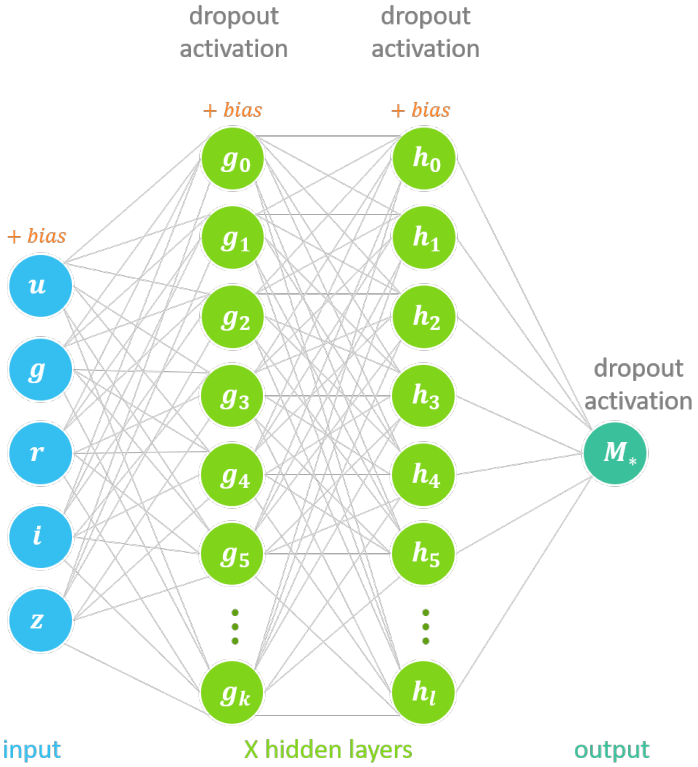
In this thesis the method used to model the relation between the flux and stellar mass of galaxies is the supervised machine learning method neural networks. The following sections provide a broad introduction on the topic of neural networks and describe the neural network used.

3.1 Introduction

Artificial neural networks are computer systems in which interconnected nodes, designed after the neurons in the human brain, transfer signals to one another (McCulloch and Pitts, 1943). The nodes carry a value and are connected by weights, indicating the importance of the signal. This form of machine learning can amongst others be used to learn the relationship between two variables. In the single-layer perceptron (Rosenblatt, 1957, 1958) the nodes are aligned in layers and the nodes of subsequent layers are fully interconnected. An illustration of such a network can be found in Figure 4a.



(a) Illustration of a simple single-layer perceptron with a j-dimensional input layer, an i-dimensional output layer and no hidden layers. All nodes are fully connected.



(b) Illustration of a basic multi-layer perceptron used in this thesis. The input nodes correspond to the SDSS *ugriz* magnitudes, and the output node corresponds to the stellar mass M_* of the galaxy. The input layer and hidden layers carry a bias node and the hidden layers and output layer carry an activation function and dropout. The number of hidden layers, the number of nodes per hidden layer, the dropout percentages and the activation functions are not defined here. All nodes are fully connected.

Figure 4: Illustrations of a simple neural network without hidden layers, and of a neural network with hidden layers. The neural network with hidden layers has 5 input nodes corresponding to the 5 SDSS bands and one output node corresponding to the stellar mass of a galaxy.

The nodes in the first layer represent the *input* feature(s) $\boldsymbol{x}$ and the nodes in the final layer represent the *output* $\boldsymbol{y}$ as a function of $\boldsymbol{x}$. Each layer carries an activation function $a^{(n)}$ that maps values of nodes that can range from negative to positive infinity to a value between 0 or -1 and 1, thereby introducing non-linearity in the relation between the input and the output. In this work, we use the *linear*, *sigmoid*, hyperbolic tangent *tanh* and Rectified Linear Unit *ReLU* (Hahnloser et al., 2000; Glorot et al., 2011) activation functions,

$$\text{linear} \qquad f(x) = x\,, \tag{13}$$

$$\text{sigmoid} \qquad f(x) = \frac{1}{1+e^{-x}}\,, \tag{14}$$

$$\tanh \qquad f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\,, \tag{15}$$

$$\text{ReLU} \qquad f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}\,. \tag{16}$$

The relationship between the input nodes and output nodes is fully determined by the weights such that the predicted i-dimensional output $\hat{\boldsymbol{y}}$ of an single layer perceptron with j-dimensional input $\boldsymbol{x}$ can be written as,

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y_1} \\ y_2 \\ \vdots \\ y_i \end{pmatrix} = a\left(\begin{pmatrix} W_{11} & \dots & W_{1j} \\ W_{21} & \dots & W_{2j} \\ \vdots & \ddots & \vdots \\ W_{i1} & \dots & W_{ij} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \end{pmatrix} \right) = a(\boldsymbol{W} \cdot \boldsymbol{x})\,, \tag{17}$$

where matrix $\boldsymbol{W}$ with indices W_{ij} indicates the weights between x_j and y_i and a denotes the activation function. Therefore, the i^{th} feature of the predicted output can be written as,

$$\hat{y_i} = a\left(\sum_j W_{ij} x_j \right)\,. \tag{18}$$

Extra layers of nodes – *hidden layers* – can be added in between the input and output layer adding extra complexity to the network. Such artificial neural networks are often called multi-layer perceptrons (hereafter MLP; Minsky and Papert, 1969) and are also used in this work. In this thesis the input nodes of the network correspond to the SDSS colours and magnitudes of a galaxy and the output node of the network corresponds to the stellar mass M_*. An illustration of such a network can be found in Figure 4b where only the five SDSS magnitudes are shown as inputs.

3.2 Optimisation

To find the weights of the artificial neural network that best describe the relationship between the input and output variables the difference between the predicted output $\hat{\boldsymbol{y}}$ and the target or true output $\boldsymbol{y}$ is minimised, a process called *training* the network. Generally during training, many samples of the data of which the true output is known (in our case EAGLE galaxies with modelled SDSS photometry and known stellar mass) are propagated through the network. In each iteration, the predicted output is calculated using the current weights in the network. Consecutively, in a single layer perceptron, these weights are updated with the *loss function*

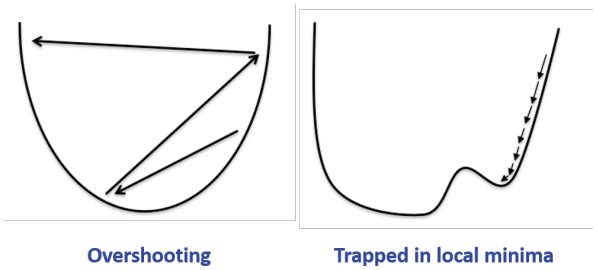


Figure 5: Visualisation of the loss function during training. If the learning rate during training is chosen too large, the minimum of the loss function will never be reached and will oscillate and diverge – a process called *overshooting*. If the learning rate is chosen too small, the algorithm might strand in a local minimum of the loss function.

$E(\boldsymbol{y}, \hat{\boldsymbol{y}})$, a measure of the error between the predicted and true output, multiplied by a *learning rate* η,

$$\boldsymbol{W} := \boldsymbol{W} - \eta \cdot E(\boldsymbol{y}, \hat{\boldsymbol{y}}) \cdot \boldsymbol{x}^\top . \tag{19}$$

The learning rate indicates the step size taken to update the weights and should be well tuned. If the learning rate is chosen too large, the minimum of the loss function will never be reached and will oscillate and diverge – a process called *overshooting*. If the learning rate is chosen too small the algorithm may be trapped in a local minimum of the error function, illustrated in Figure 5.

For neural networks with hidden layers, optimisation is more complicated. In MLPs the optimal weights are obtained with *backpropagation* whereby updating the weights occurs backwards through the network (Linnainmaa, 1970; Werbos, 1974; Parker, 1985; LeCun, 1985; Rumelhart and McClelland, 1986). Today, many optimisation algorithms are adapted forms of the *gradient descent* algorithm (e.g. Kelley, 1960; Bryson, 1961). In this algorithm, weights are updated proportionally to the gradient of the loss function $E(\boldsymbol{W})$ evaluated at the current point,

$$\boldsymbol{W} := \boldsymbol{W} - \eta \frac{\partial E}{\partial \boldsymbol{W}} . \tag{20}$$

The gradient of the loss function towards one weight W_{ij}, assuming a loss function that is some kind of error function $E \propto \sum_i (y_i - \hat{y}_i)$, is proportional to the following,

$$\frac{\partial E}{\partial W_{ij}} \propto \frac{\partial}{\partial W_{ij}} \sum_i (y_i - \hat{y}_i) \propto \frac{\partial}{\partial W_{ij}} \sum_i \left(y_i - a \left(\sum_j W_{ij} x_j \right) \right) . \tag{21}$$

This equation indicates that the activation functions chosen in the (hidden) layers play a role when updating the weights.

The number of samples that is propagated through the network (and that the error is calculated upon before updating the weights) is the *batch size*. Each time the weights are updated counts as one *epoch*. The training process takes place for a fixed number of epochs or until some form of convergence is reached.

In the adapted forms of gradient descent, the learning rate is adapted based on previous weight updates (*momentum*; Wiegerinck et al., 1994), preventing oscillations and ensuring smoother convergence towards the minimum of the loss function. The optimisers explored in this thesis that use this method are *stochastic gradient descent* (sgd), *RMSprop* (Hinton et al., 2012), *Adagrad* (Duchi et al., 2011), *Adadelta* (Zeiler, 2012), *Adam* (Kingma and Lei Ba, 2015;

Reddi et al., 2018), *Adamax* (Kingma and Lei Ba, 2015) and *Nadam* (Dahl et al., 2013; Dozat, 2016). In this work, *Adam* is most often used and its update rule is the following,

$$W_{ij}^{(t+1)} = W_{ij}^{(t)} - \eta \frac{\frac{m_{ij}^{(t)}}{1-\beta_1}}{\sqrt{\frac{v_{ij}^{(t)}}{1-\beta_2}} + \epsilon}\,, \tag{22}$$

where ϵ avoids division by zero and where β_1 and β_2 denote the decay rate of the first order moment m_{ij} and the second order moment v_{ij} of the gradient $g^{(t)} = \frac{\partial E\left(W_{ij}^{(t)}\right)}{\partial W_{ij}}$, respectively,

$$m_{ij}^{(t)} = \beta_1 m_{ij}^{(t-1)} + (1-\beta_1) g^{(t)}\,, \tag{23}$$

$$v_{ij}^{(t)} = \beta_2 v_{ij}^{(t-1)} + (1-\beta_2) g^{(t)} \odot g^{(t)}\,, \tag{24}$$

initialised at $m_{ij}^{(0)} = 0$ and $v_{ij}^{(0)} = 0$. In the above equation, $g^{(t)} \odot g^{(t)}$ denotes the element-wise multiplication of the gradient. In this thesis the default values for all optimiser-parameters of the `keras`[1] package are used. These values are listed in Table 15 of Appendix D.

Fast convergence towards the minimum of the loss function can also be attained by adapting the data that is propagated through the network. The format suggested by LeCun et al. (1998) is features of the data (in this thesis the photometry of a galaxy) following a distribution with zero mean and equal covariances among features (e.g. u or g-band flux), preferably uncorrelated. The distribution of input features should be centred around zero, because strictly positive or strictly negative inputs will lead to reciprocal weight updates of the same sign, meaning that the updates of all weights of the weight matrix $\boldsymbol{W}$ share the same sign (LeCun et al., 1998). The process of updating the weights is in this case very inefficient as the weights oscillate towards a solution. It can also introduce a directional bias in the learning process. Secondly, the input variables should have the same covariances as this will balance out the weights (LeCun et al., 1998).

3.3 Evaluation

The performance of the network can be assessed through various measures. These measures indicate by how much the predicted values by the network correspond to the true values. The error measures are calculated on the normalised data before the postprocessing procedure is applied. In this thesis the following measures are used. The *mean squared error* (MSE),

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,, \tag{25}$$

is the sum of the squared errors divided by the number of samples n that the error is calculated upon. The *mean absolute error* (MAE) indicates the average absolute error of the model and can be calculated with the following equation,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,. \tag{26}$$

[1] https://keras.io/

The mean error μ indicates the average error of the model and is calculated as follows,

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,. \tag{27}$$

It can also be viewed as the mean of the probability distribution of the errors $y_j - \hat{y_j}$ of the model and is therefore expected to be approximate to zero for an unbiased model. The variance error σ^2 is the variance of the probability distribution of the model's error and can be calculated with the following equation,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i - \mu)^2\,, \tag{28}$$

where μ is defined in Eq. (27).

A statistic often used in regression is the coefficient of determination R^2. The R^2 statistic indicates the fraction of the variance of the true output that can be explained by the model. The total sum of squares ($TSS = \sum_i (y_i - \bar{y})$ with $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$) is the total variance of the output. The residual sum of squares ($RSS = \sum_i (y_i - \hat{y_i})$) is the amount of variance in the output that is not explained by the model. The R^2 statistic is therefore equal to the variability explained by the model $TSS - RSS$ divided by the total variance of the true output TSS,

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i (y_i - \bar{y})}\,, \tag{29}$$

with $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ the mean of the target data. If the R^2 statistic equals 1, all variability of the target output is explained by the model, defining a good model. If the R^2 statistic is near 0, the model does not explain the variability observed. The R^2 statistic is therefore an indication of the goodness of the fit.

The R^2 statistic will increase when more inputs are added to the network, even though this extra information might not always be meaningful. The adjusted coefficient of determination $\bar{R^2}$ attempts to counter this effect by penalising extra input variables. It is equal to,

$$\bar{R^2} = 1 - \frac{RSS}{TSS}\frac{n-1}{n-j-1} = 1 - (1 - R^2)\frac{n-1}{n-j-1} = R^2 - \frac{j}{n-j-1}(1 - R^2)\,, \tag{30}$$

with j the number of input variables and n the sample size (Henri, 1961).

During training, the performance of the network is monitored with these measures. If we train with an infinitely complex network, or for a disproportionally large number of epochs, it is probable that almost the exact weights are found that map the input to the true output of the sample that is trained upon. However, this often implies that the trained network will not correctly predict the output on a new unseen sample of the data with the same probability distribution (James et al., 2013). This process is called overfitting and can be prevented with various actions. The foremost effort is dividing the data set into a training set and a test set, which are in our case 80 % and 20% of the complete data set of galaxies respectively. The weights are then updated using the training set and after the training process the performance of the network can be tested on the test set. Even during training the error on the test set can be monitored. Another measure against overfitting is randomly dropping a percentage of the connections between nodes during training (*dropout*; Srivastava et al., 2014).

4 Hyperparameter optimisation

In order to find the relation between the photometry and stellar mass of EAGLE galaxies it is necessary to find the parameters governing the learning process that allow for the best possible results. These parameters are called *hyperparameters* and are involved with the architecture and optimisation of the weights of the neural network. Hyperparameters can be the number of nodes in each hidden layer, the activation function of each layer and the optimisation algorithm to use. The following sections describe the hyperparameter optimisation methods used, the results obtained with these methods and the final architecture used in subsequent experiments.

4.1 Validation

Decisions about hyperparameters should be made based on the performance of the network on a validation set. This validation set comprises a fraction of the training set, not destined for training, and is independent of the test set. After tuning the hyperparameters of the network with the validation set, the trained network can be tested on the separate test set to provide an unbiased evaluation of the model (James et al., 2013). To ensure that we do not include bias in our decisions, all decisions about the architecture of the network are based on the evaluation of the validation set, rather than the separate test set.

The validation method used in this thesis is K-fold cross-validation (Mosteller and Tukey, 1968; Stone, 1974). In K-fold cross-validation the training set is divided into K random separate sub-samples – *folds*. Each fold is used as a validation set, while the remaining $K-1$ folds are used for training. The cross-validated loss computed is the average loss of the K folds. In this thesis we use $K = 5$ folds. Figure 6 shows an an illustration of the division of the data set. The advantage of cross-validation is that it removes effects caused by the randomness involved in the selection of the training and validation set (James et al., 2013).

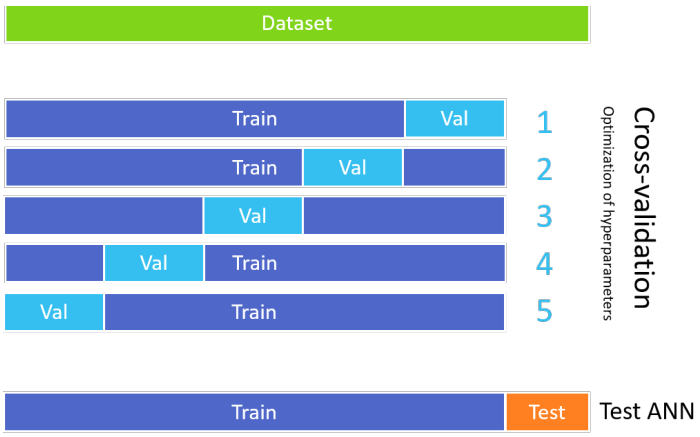


Figure 6: Illustration of the division of the EAGLE data set in this thesis. 80% of the complete EAGLE data set is randomly assigned to the training set and the remaining 20% is assigned to the test set. In this thesis $K = 5$ fold cross-validation is performed which implies that the previously defined training set is randomly divided into 5 independent samples. Each sample functions as a validation set while the remaining 4 samples are used for training. After 5 iterations, the average error is calculated and this error can be used for decisions about optimisation of the hyperparameters. Once the hyperparameters have been tuned, the ANN is trained using the complete training set and can be tested using the test set.

4.2 Hyperparameter-space and optimisation methods

In order to discuss all possible hyperparameter-optimisation methods, it is important to first understand the notion of *hyperparameter-space*. Each free hyperparameter such as the activation function of a certain layer, the number of nodes in a certain layer and the dropout in a certain layer adds an extra dimension to the hyperparameter-space. A neural network with all hyperparameters fixed, except for the number of nodes in the first and second hidden layer, would have a two-dimensional hyperparameter-space. If each hyperparameter would have three options, e.g. 10, 20 or 30 nodes in the first hidden layer and 20, 30 and 40 nodes in the second hidden layer, the hyperparameter-space would have a size of 3×3. A simple solution is to tune the hyperparameters by hand, but this method risks not exploring the entire hyperparameter-space and requires expertise. In a world with unlimited computation time, it would be possible to explore all possible values for all hyperparameters (a method called *grid search*) and select the hyperparameters that achieve the best results. However, as each hyperparameter adds an extra dimension to the hyperparameter-space, the hyperparameter-space grows exponentially and this option is often not computationally feasible. It is therefore important that the hyperparameter-space is reduced as much as possible by making educated guesses about which hyperparameters can already be determined.

There are various algorithms designed to approximate the optimal hyperparameters in a computationally feasible way. A slight improvement on grid search in terms of computational expense is *random search*. Random search randomly explores a percentage of the hyperparameter-space, assuming that the best hyperparameters found are in the vicinity of the optimal hyperparameters and therefore these model settings retrieve equally good results (Bergstra and Bengio, 2012). However, both random search and grid search generally explore a large proportion of the hyperparameters that retrieve high errors and are thus not worth exploring. Bayesian optimisation tries to circumvent the hyperparameters with high errors by using the results of previous hyperparameter evaluations. We define the hyperparameters as h and the score of the network (e.g. MSE) returned given those hyperparameters as $s = f(h)$. To prevent computational expense by minimising and thus evaluating the function $f(h)$, Bayesian optimisation builds a surrogate function, $p_M(s \mid h)$, for this objective function. This surrogate function equals the probability of score s given certain hyperparameters h of a model M. The surrogate function is used to determine which hyperparameters achieve the best score. The selection criterion that determines which hyperparameters achieve the best score is the Expected Improvement (EI), defined as,

$$EI_{s^*}(h) = \int_{-\infty}^{\infty} \max\left[(s^* - s), 0\right] \cdot p_M(s \mid h)\, ds\,. \tag{31}$$

The Expected Improvement is the expectation that a score $s = f(h)$ exceeds a threshold s^*, the best score approached so far (Bergstra et al., 2011). The pseudocode for a Bayesian optimisation algorithm would resemble the following:

Algorithm 1 Bayesian Optimisation - adapted form of Frazier (2018)

Place a prior on f
Observe $s = f(h)$ for various sets of hyperparameters h
for $n = 1, 2, ..., N$ **do**
  Update the posterior probability distribution $p(s \mid h)$
  Maximise the Expected Improvement using the new posterior distribution $h_n = \arg\max EI_{s^*}(h)$
  Observe $s_n = f(h_n)$
Return h with the smallest posterior mean

In *Gaussian Processes*, the surrogate function $p(s \mid h)$ itself is estimated. *Tree-structured Parzen Estimators* (TPE), a form of Bayesian optimisation and the approach used in this thesis, uses Bayes' Rule,

$$p(s \mid h) = \frac{p(h \mid s) \cdot p(s)}{p(h)}, \tag{32}$$

to estimate the surrogate function. This algorithm begins with a random search evaluating hyperparameters for a few iterations to create observations of the score s given certain hyperparameters h. The observations are consecutively divided into two groups: the group $g(h)$, with a score above or equal to a threshold s^*, and the group $l(h)$, with a score below the threshold. These two groups define the complete distribution $p(h \mid s)$,

$$p(h \mid s) = \begin{cases} l(h) & \text{if } s < s^* \\ g(h) & \text{if } s \geq s^* \end{cases}. \tag{33}$$

In the TPE algorithm, the threshold s^* is defined as $p(s < s^*) = \gamma$, thereby introducing a quantile factor γ,

$$p(h) = \int p(h \mid s)\, p(s)\, ds = \gamma l(h) + (1-\gamma) g(h). \tag{34}$$

Using this information, the Expected Improvement becomes the following:

$$\begin{aligned}
EI_{s^*}(h) &= \int_{-\infty}^{\infty} \max\left[(s^* - s), 0\right] \cdot p(s \mid h)\, ds, \\
&= \int_{-\infty}^{s^*} (s^* - s)\, p(s \mid h)\, ds, \\
&= \int_{-\infty}^{s^*} (s^* - s) \frac{p(h \mid s) \cdot p(s)}{p(h)} ds, && \text{[Bayes' theorem]} \\
&= l(h) \int_{-\infty}^{s^*} (s^* - s) \frac{p(s)}{p(h)} ds, && \text{[Eq. (33)]} \\
&= \frac{l(h)}{\gamma l(h) + (1-\gamma) g(h)} \int_{-\infty}^{s^*} (s^* - s)\, p(s)\, p(h)\, ds, && \text{[Eq. (34)]} \\
&\propto \left(\gamma + \frac{g(h)}{l(h)}(1-\gamma)\right)^{-1},
\end{aligned}$$

as shown in Bergstra et al. (2011). From this equation it is apparent that the Expected Improvement is maximised when the candidate h is more likely to be in $l(x)$ (the group with the score below the threshold and therefore best score) than in $g(x)$. Further details about the implementation of the TPE algorithm will be discussed in Section 4.6.

4.3 Fixed hyperparameters

A number of hyperparameters are not optimised in this thesis, either because they are robust to changes of hyperparameters or because their optimal values have been well researched. One such hyperparameter is the initialisation of weights, for which the Glorot and Bengio (2010) normal initialiser is used, drawing weights from a normal distribution with zero mean and standard deviation $\sigma = \sqrt{\frac{2}{i+j}}$ with i and j the dimensions of the output and input features respectively. For the sake of simplicity, no regularisation is used. The learning rate is not optimised either, because the learning rate depends on the type of optimiser used and this

Hyperparameter	Description	Options
`h_nodes0`	# nodes in 1st hidden layer	*10, 20, 30, 40, 50*
`h_nodes1`	# nodes in 2nd hidden layer	*0, 10, 20, 30, 40, 50*
`h_nodes2`	# nodes in 3rd hidden layer	*0, 10, 20, 30, 40, 50*
`activation0`	activation function on 1st hidden layer	*sigmoid, tanh, ReLU, linear*
`activation1`	activation function on 2nd hidden layer	*sigmoid, tanh, ReLU, linear*
`activation2`	activation function on 3rd hidden layer	*sigmoid, tanh, ReLU, linear*
`dropout0`	dropout fraction after the 1st hidden layer	*0, 0.2, 0.4, 0.7*
`dropout1`	dropout fraction after the 2nd hidden layer	*0, 0.2, 0.4, 0.7*
`dropout2`	dropout fraction after the 3rd hidden layer	*0, 0.2, 0.4, 0.7*
`optimiser`	optimisation algorithm	*sgd, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam*
`activation_out`	activation function in the output layer	*linear*
`loss function`	function optimised during training	*MSE*
`epochs`	number of training iterations	*15*
`batch size`	number of samples evaluated in each epoch	*128*

Table 2: Hyperparameters used in the 'subset' grid search. The hyperparameters that are fixed and not optimised are highlighted blue. The dashed horizontal lines divide the hyperparameters into the subsets used for this grid search. The green highlighted options denote the best hyperparameters found. The underlined options denote the hyperparameters assumed when the best hyperparameters have not been determined. Note that we require at least one hidden layer in the neural network.

would transform our hyperparameter-space into a tree structure. The learning rates used can be found in Table 15 of Appendix D.

4.4 Manual tuning

The hyperparameters `batch size` and number of `epochs` are manually tuned based on the improvement in the loss function and R^2 statistic. The point at which we stop training is determined by estimating the point at which no significant improvement in the loss function can be observed. This generally occurs after *15 epochs*. The number of samples to evaluate during each epoch is often determined by compromising between a good estimate of the gradient and computation time. If the entire training set of galaxies would be evaluated in each epoch, we would obtain the best possible estimate of the gradient and observe a smooth loss curve but this would be computationally expensive. If only one sample would be propagated through the network in one epoch, we would observe a stochastic loss curve but this would be computationally inexpensive (LeCun et al., 1998). The equilibrium between computation time and a smooth loss curve is found at a *batch-size of 128*. Note that the hyperparameters `number of input and output nodes` are not discussed, as they are determined by the number of features of the data itself. Furthermore, since our problem is a regression problem it is common to use a linear output activation function.

4.5 Subset Grid Search

To refrain from implementing prior assumptions on the architecture of the network, a broad exploratory grid search on the hyperparameters of the network is performed. The hyperparameters explored in this grid search can be found in Table 2. The hyperparameters that are not optimised are highlighted blue in this table. The full hyperparameter-space has a size of $5 \times 6 \times 6 \times 4^3 \times 4^3 \times 7 \approx 5 \cdot 10^6$ and is therefore too large to perform a normal grid search upon. To get a sense of which hyperparameter values should be used in our network, we separate the hyperparameters into four subsets as indicated by the dashed lines in the table, and perform

	EAGLE CV	EAGLE test	SDSS
MSE	$1.478{\cdot}10^{-3}$	$1.368{\cdot}10^{-3}$	$7.985{\cdot}10^{-3}$
R^2	0.988	0.989	0.874
$\bar{R}^2$	0.988	0.989	0.874
μ	$1.698{\cdot}10^{-3}$	$-2.410{\cdot}10^{-3}$	$7.885{\cdot}10^{-2}$
σ^2	$1.471{\cdot}10^{-3}$	$1.362{\cdot}10^{-3}$	$1.768{\cdot}10^{-3}$

Table 3: Evaluation of the `nocolours` network with the hyperparameters found by the 'subset' grid search hyperparameter optimisation. The five error measures as described in Section 3.3 are used to evaluate the network. During grid search the network is evaluated with 5-fold cross-validation (CV) on the EAGLE training set, consecutively the network is evaluated on the EAGLE test set and on the SDSS data set. The input features used are those corresponding to the `nocolours` architecture.

an exploratory grid search on these subsets. The best hyperparameters found are highlighted in green in the table.

When performing a grid search on a subset of the hyperparameters, assumptions have to be made for the values of the remaining hyperparameters that are not included in the current subset grid search. If a subset grid search has already been performed on these hyperparameters we assume the best values found. Else, we assume the underlined values from the table. The subset grid search is performed in the order of descending expected importance for the performance of the neural network:

1. `h_nodes` - the number of nodes in the hidden layers
2. `activation` - the activation functions of the hidden layers
3. `dropout` - the dropout of the hidden layers
4. `optimiser` - the optimisation algorithm used

By performing the subset grid search in this particular order we try to explore the parts of the hyperparameter-space that will retrieve the best results. However, please note that this approach might lead to possible biases in the hyperparameters that we choose. Therefore we only use these hyperparameters as a guideline and as a comparison with other approaches.

The exploratory grid search is performed with 5-fold cross-validation on the EAGLE training data, with the `nocolours` neural network. The EAGLE data set does not include any sampling procedure or added noise. The best hyperparameters are identified as those with the lowest *MSE* and are highlighted green in Table 2. The consecutive evaluation of the neural network with these hyperparameters on the EAGLE test and SDSS data can be found in Table 3. What can be observed is that the network already performs well, as an R^2 close to 1 is obtained. However, the performance on the SDSS set can be improved.

The figures illustrating the performance of the various hyperparameters can be found in Figure 9, containing the three subsets `h_nodes`, `activation` and `dropout`, and Figure 8, containing the subset `optimiser`. Figure 9 shows two-dimensional slices of the three-dimensional subset hyperparameter-space explored. These slices allow us to observe how well two combinations of hyperparameters perform, instead of viewing the performance of a singular performance. Unfortunately, it is difficult to show a three-dimensional view of the performance.

For the number of nodes in the hidden layer, a second and third hidden layer are in general preferred because the option of *no hidden layer* has on average a high mean squared error. Furthermore a large number of nodes is preferred over a small number of nodes, but this is again dependent on the combination of the three layers. This finding makes sense as an increasing number of nodes will often lead to asymptotically increasingly better predictions.

In terms of the activation function the *sigmoid* activation function generally retrieves high errors. The updates of the weights are influenced by the derivative of the activation function, as can be seen in Eq. (21). The low performance of the sigmoid function can be explained by

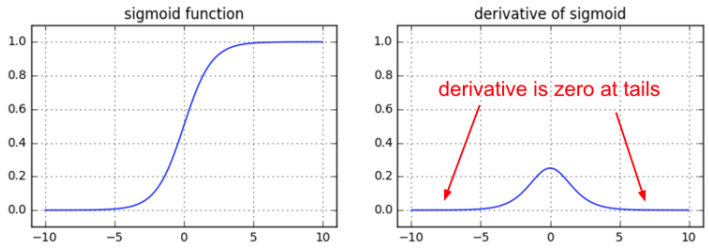


Figure 7: Illustration of the vanishing gradients problem with a sigmoid activation function. The figure on the LHS shows a sigmoid. The figure on the RHS shows the derivative of the sigmoid. The derivative is low at the center but even lower at the tails of the function. At the tails the derivative of the sigmoid approaches zero resulting in extremely low weight updates.

the fact that the sigmoid function is relatively flat compared to the other activation functions used, and therefore has an overall low gradient, resulting in relatively low weight updates. Furthermore if the weights are initialised in a range that is relatively large, the derivatives at a large range of the function will be approximate to zero and these weights will not be updated, an illustration of which can be found in Figure 7. This problem is also known as the *vanishing gradients problem*. This vanishing gradient problem can occur in the hyperbolic tangent function as well. However, no clear preference is observed for these other three activation functions. Regarding the hyperparameter subset dropout it is apparent that the lower the dropout the better the performance of the neural network on the test set. This observation makes sense as dropout is used as a measure against overfitting and randomly disregarding a percentage of the connections increases the prediction errors. Finally, the optimiser that obtains the lowest MSE is *Adamax*, closely followed by *Adam*. This observation corresponds with the research of Kingma and Lei Ba (2015).

An analysis of the loss function (the function that is optimised during training) with the hyperparameters from the subset grid search quickly shows that the mean absolute error MAE is a better loss function than the mean squared error MSE. A possible explanation for this observation is that the MSE squares the errors and therefore might perform better on outliers and worse on the average samples.

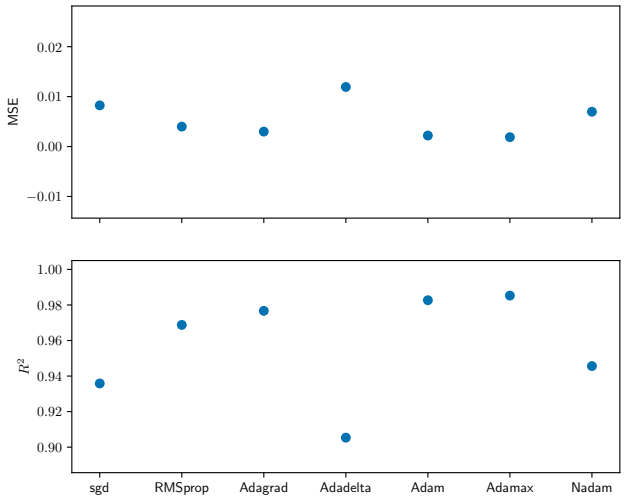
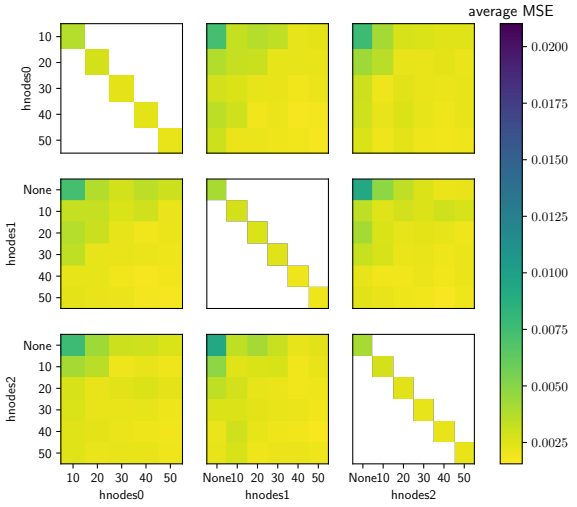
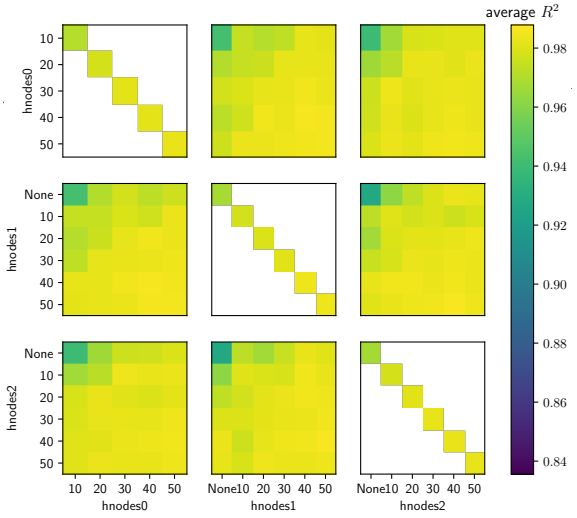


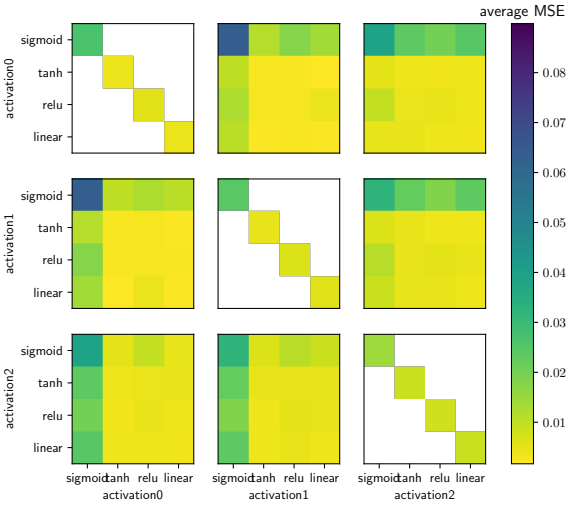
Figure 8: Results of the grid search performed with the `nocolours` neural network on the subset `optimiser` to explore the optimal hyperparameters. The trained network is evaluated with the measure MSE and the R^2 statistic. The optimiser *Adamax* shows the best performance, closely followed by *Adam*.



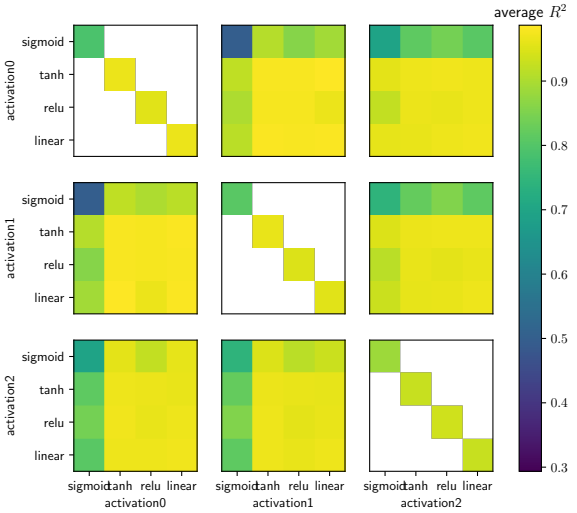
(a) `h_nodes`: nodes per hidden layer



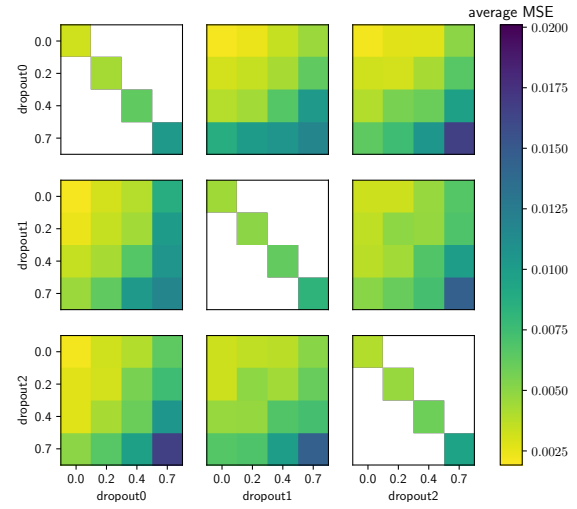
(b) `h_nodes`: nodes per hidden layer



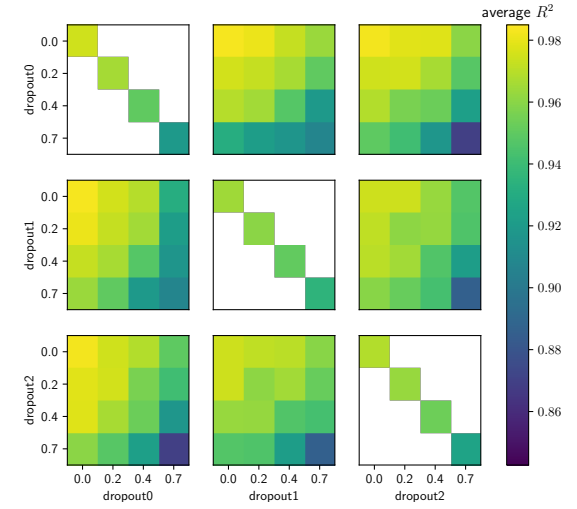
(c) `activation`: activation function for each hidden layer



(d) `activation`: activation function for each hidden layer



(e) `dropout`: dropout fraction after each hidden layer



(f) `dropout`: dropout fraction after each hidden layer

Figure 9: Results of the 'subset' grid search performed with the `nocolours` neural network to explore the optimal hyperparameters. The subsets explored are `hnodes` (the number of hidden nodes per hidden layer; Fig. 9a and 9b), `activation` (the activation function in each hidden layer; Fig. 9c and 9d) and `dropout` (the dropout fraction after each hidden layer; Fig 9e and 9f). The suffixes *0*, *1* and *2* indicate the hyperparameter of the first, second and third hidden layer respectively. For each subset of hyperparameters, the trained network is evaluated with the measure MSE and the R^2 statistic (results shown on the LHS and RHS respectively). The figures are two-dimensional slices of the three-dimensional subsets of hyperparameter-space explored, where we averaged over the remaining dimension. Note that each subfigure (consisting of nine smaller figures) is symmetric over the diagonal axis; the nine smaller figures individually are not.

4.6 Tree-structured Parzen Estimators

The Tree-structured Parzen Estimators (TPE) algorithm is performed with the `hyperas` package[2]. The hyperparameter-space for this algorithm is adapted with respect to the hyperparameter-space of the subset grid search. Because the subset grid search obtains an optimal number of nodes of 50 for two hidden layers, an extra option of 60 nodes is added. Due to high errors obtained in the subset grid search, the sigmoid function is removed from the hyperparameter-space and the dropout is completely ignored. The resulting hyperparameter-space is shown in Table 4. The hyperparameters that stay fixed in this approach are shown in the blue bottom rows of the table.

The TPE algorithm is performed for the three architectures `nocolours`, `subsetcolours` and `allcolours`, using 5-fold cross-validation on the EAGLE training set. This data set is not sampled and no noise is added to it. The algorithm uses $N = 100$ iterations to update the surrogate function $p(s \mid h)$ and the best hyperparameters are identified using the lowest cross-validated MAE over all epochs. The best hyperparameters found with the TPE algorithm for each architecture `nocolours`, `subsetcolours` and `allcolours` can be found in Table 5.

It can be observed that a large number of nodes in one of the hidden layers is always preferred. Nevertheless, the additional option of 60 nodes added to the hyperparameter-space is never selected as the best value. Moreover, a combination of a hyperbolic tangent and linear functions is favoured as activation functions in the hidden layers. Last of all, the *Adam* optimiser is selected as the best optimiser. The choice of optimiser is not surprising as it has been demonstrated by Kingma and Lei Ba (2015) that Adam and *Adamax* are the best performing optimisers. The choice of optimiser according to the subset grid search is Adamax, but the hyperparameter `optimiser` was optimised only with the specific architecture found in previous `hnodes`, `activation` and `dropout` subset grid searches, where the performance of the Adamax optimiser is closely followed by the performance of the Adam optimiser, as can be observed in Figure 8.

4.6.1 Evaluation of final architectures

The hyperparameters obtained with the TPE algorithm are used in all subsequent analyses. When referring to the neural networks `nocolours`, `subsetcolours` or `allcolours` the hyper-

[2]https://github.com/maxpumperla/hyperas

Hyperparameter	Description	Options
`h_nodes0`	# nodes in 1st hidden layer	*10, 20, 30, 40, 50, 60*
`h_nodes1`	# nodes in 2nd hidden layer	*0, 10, 20, 30, 40, 50, 60*
`h_nodes2`	# nodes in 3rd hidden layer	*0, 10, 20, 30, 40, 50, 60*
`activation0`	activation function on 1st hidden layer	*tanh, ReLU, linear*
`activation1`	activation function on 2nd hidden layer	*tanh, ReLU, linear*
`activation2`	activation function on 3rd hidden layer	*tanh, ReLU, linear*
`optimiser`	optimisation algorithm	*sgd, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam*
`activation_out`	activation function in the output layer	*linear*
`loss function`	function optimised during training	*MAE*
`epochs`	number of training iterations	*15*
`batch size`	number of samples evaluated in each epoch	*128*

Table 4: Hyperparameter-space of the TPE algorithm. The hyperparameters that are fixed and not optimised are highlighted blue. Note that at least one hidden layer is required in the network.

Hyperparameter	nocolours	subsetcolours	allcolours
h_nodes0	40	50	30
h_nodes1	50	50	30
h_nodes2	30	20	50
activation0	linear	ReLU	linear
activation1	tanh	tanh	tanh
activation2	tanh	linear	ReLU
optimiser	Adam	Adam	Adam
activation_out		*linear*	
loss function		*MAE*	
epochs		*15*	
batch size		*128*	

Table 5: Final optimised hyperparameters of the three neural networks `nocolours`, `subsetcolours` and `allcolours` determined with the TPE algorithm. All hyperparameters underneath the dashed line are determined through different methods and are the same for all three architectures.

Input	Hyperparameters	MAE	R^2	$\bar{R}^2$	μ	σ^2
nocolours	nocolours	$2.24{\cdot}10^{-2}$	0.994	0.994	$-5.07{\cdot}10^{-4}$	$7.87{\cdot}10^{-4}$
nocolours	subsetcolours	$2.22{\cdot}10^{-2}$	0.994	0.994	$3.71{\cdot}10^{-3}$	$7.19{\cdot}10^{-4}$
nocolours	allcolours	$2.21{\cdot}10^{-2}$	0.994	0.994	$3.20{\cdot}10^{-3}$	$7.49{\cdot}10^{-4}$
subsetcolours	nocolours	$1.84{\cdot}10^{-2}$	0.996	0.996	$8.14{\cdot}10^{-4}$	$5.38{\cdot}10^{-4}$
subsetcolours	subsetcolours	$1.80{\cdot}10^{-2}$	0.996	0.996	$8.84{\cdot}10^{-4}$	$5.10{\cdot}10^{-4}$
subsetcolours	allcolours	$-3.76{\cdot}10^{-2}$	0.996	0.996	$-3.76{\cdot}10^{-4}$	$4.94{\cdot}10^{-4}$
allcolours	nocolours	$1.81{\cdot}10^{-2}$	0.996	0.996	$-2.29{\cdot}10^{-3}$	$5.19{\cdot}10^{-4}$
allcolours	subsetcolours	$1.80{\cdot}10^{-2}$	0.996	0.996	$1.00{\cdot}10^{-3}$	$5.12{\cdot}10^{-4}$
allcolours	allcolours	$1.78{\cdot}10^{-2}$	0.996	0.996	$-9.73{\cdot}10^{-4}$	$5.00{\cdot}10^{-4}$

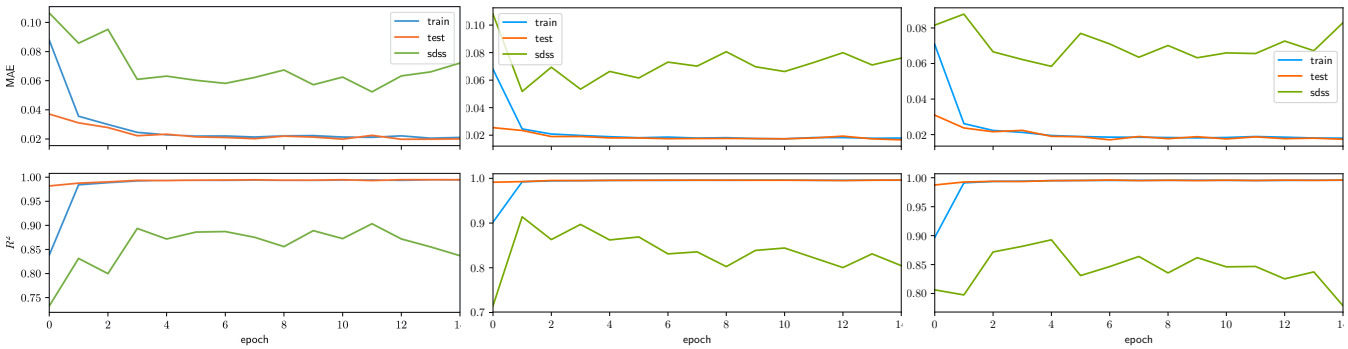
Table 6: Evaluation of the three different neural networks `nocolours`, `subsetcolours` and `allcolours` on different input subsets *nocolours*, *subsetcolours* and *allcolours*. The neural networks are assessed using 5-fold cross-validation on the EAGLE training data with the measures described in Section 3.3. The features that the neural network's hyperparameters are optimised upon are highlighted pink.

parameters in Table 5 will be assumed. After the identification of the optimal hyperparameters with TPE, the three neural networks are trained with the EAGLE training set and consecutively evaluated with galaxies of the EAGLE test set and the SDSS set. The EAGLE galaxies and SDSS galaxies are not sampled and no noise is added to the data. This assessment can be observed in Figures 10, 11 and 12. Figure 10 shows the evolution of the loss function MAE and monitored R^2 during training. The network's weights are optimised with the EAGLE training set and evaluated with the EAGLE test set and SDSS data set after each epoch of training. These figures illustrate that the error of the training set is a smooth curve, revealing that the right hyperparameters are chosen, specifically those involved in training (such as the optimiser, batch size and learning rate). Additionally, the training and test set errors follow each other closely, indicating that no overfitting takes place. The error on the SDSS data set is much higher however, suggesting a discrepancy exists between the EAGLE and SDSS data set. Further discussion on this topic can be read in Chapter 5.

In Figure 11, the true versus predicted stellar mass of all galaxies of the EAGLE test set and SDSS set is shown. Ideally all galaxies lie on the diagonal dashed line. The EAGLE galaxies are narrowly distributed on this diagonal line and the stellar mass prediction by all three neural networks obtains low errors ($MAE \sim 2{\cdot}10^{-2}$ and $R^2 \sim 0.99$). The prediction error is defined as the difference between the predicted stellar mass and true stellar mass of galaxies. Figure 12 shows the kernel density estimated distribution of the prediction error as a function of the true stellar mass of galaxies. The distributions of the prediction errors by the `subsetcolours` and `allcolours` neural networks on galaxies of the EAGLE test set are centred

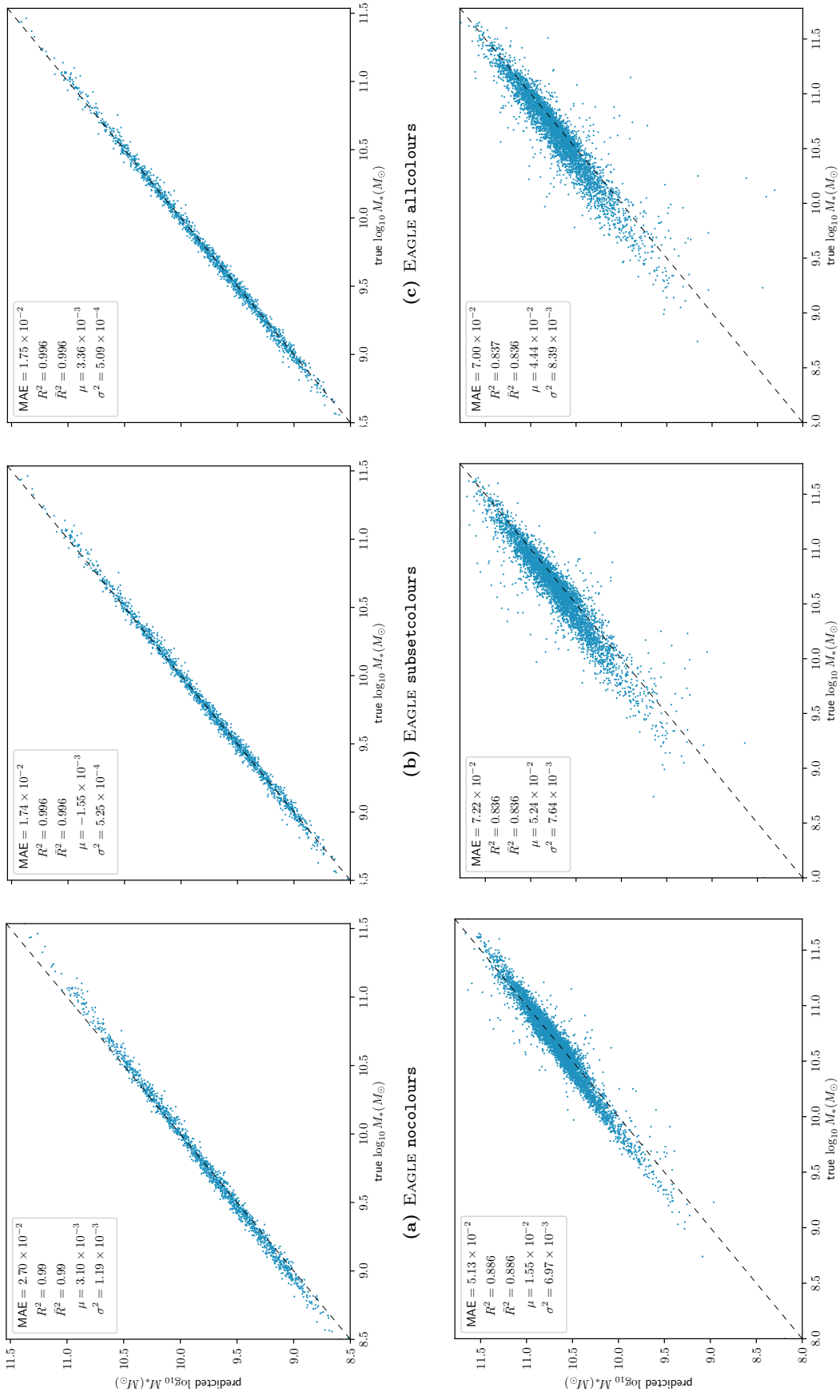
around zero and thus show promising results. The `nocolours` network is centred around zero error as well for EAGLE galaxies with average stellar mass. However, the predicted stellar mass by the `nocolours` neural network is too high for galaxies with small stellar masses, and too low for galaxies with large stellar masses.

The `nocolours` neural network differs from the `subsetcolours` and `allcolours` network in two ways: it uses different input features of the data (namely those described in Section 2.3, e.g. magnitudes and colours) and the networks are defined by different hyperparameters (i.e. the activation functions and the number of nodes in the hidden layers). Therefore a short analysis is performed, combining all three neural networks with their optimised hyperparameters and all three input feature combinations. The results of this analysis can be found in Table 6. From this table it is apparent that there is negligible difference in the performance of the three neural networks for the same input features evaluated in terms of MAE, R^2, $\bar{R}^2$ and σ^2: the different architectures of the three neural networks obtain relatively similar results. The difference seems to occur when different input features are used. This implies that the features that are used in the neural network are an important determinant of the performance of the network. The exception is the mean of the error μ, that seems to vary more among different architectures, possibly caused by the nature of the measure. When μ varies orders of magnitudes, the lowest μ for each input can be observed for the network that was optimised with this specific input. Therefore it is unsurprising that other network-input combinations obtain higher μ. From this experiment, it seems that adding colours to the input features increases the performance of the network. Further assessment of the importance of individual features will be further assessed in Chapter 7.



(a) `nocolours` **(b)** `subsetcolours` **(c)** `allcolours`

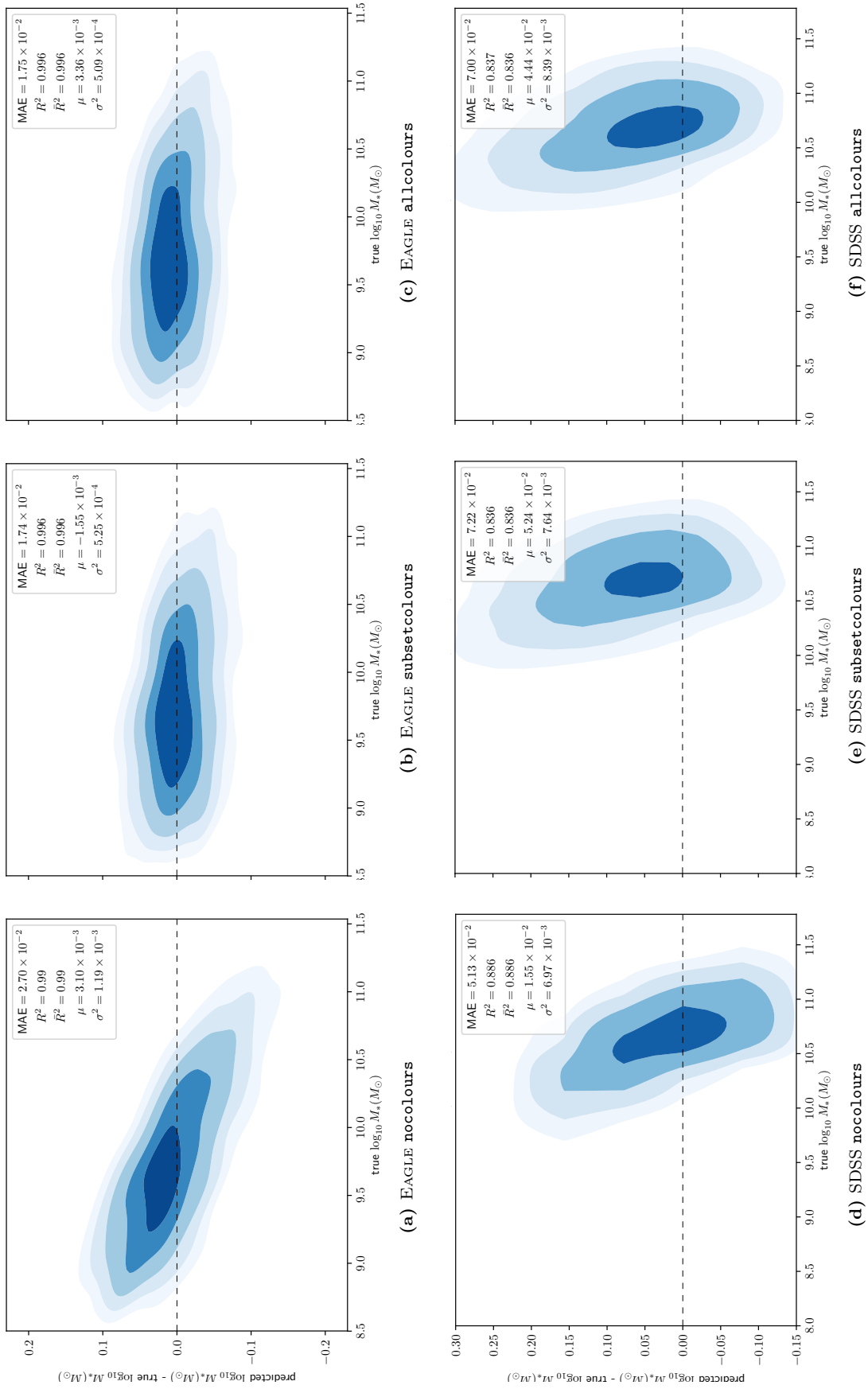
Figure 10: Evolution of the loss-function MAE and the monitored R^2 statistic during training. The measures of the EAGLE training set, test set and SDSS data set are indicated by the blue, orange and green line respectively. The EAGLE training set is used for optimising the weights of the neural network (training) and the EAGLE test set and SDSS data set are tested during each epoch of training with the current weights. The errors of the training and test set follow each other closely during the training process. The error of the SDSS data set however does not.



(a) EAGLE `nocolours` **(b)** EAGLE `subsetcolours` **(c)** EAGLE `allcolours`

(d) SDSS `nocolours` **(e)** SDSS `subsetcolours` **(f)** SDSS `allcolours`

Figure 11: Evaluation of optimal hyperparameters found with the TPE algorithm for the three networks `nocolours`, `subsetcolours` and `allcolours`. The networks are tested on EAGLE test galaxies (11a, 11b and 11c) and SDSS galaxies (11d, 11e and 11f), both not sampled and containing no noise. The figures show the true logarithmic stellar mass of each galaxy versus the predicted logarithmic stellar mass by the network. The error measures are shown in the top left corner. When the prediction is perfect all galaxies lie exactly on the diagonal dashed line.



(a) EAGLE `nocolours` (b) EAGLE `subsetcolours` (c) EAGLE `allcolours`

(d) SDSS `nocolours` (e) SDSS `subsetcolours` (f) SDSS `allcolours`

Figure 12: Evaluation of optimal hyperparameters found with the TPE algorithm for the three networks `nocolours`, `subsetcolours` and `allcolours`. The networks are tested on EAGLE test galaxies (12a, 12b and 12c) and SDSS galaxies (12d, 12e and 12f), both not sampled and containing no noise. The figures show the kernel density estimated (KDE) distribution of the prediction error of the stellar mass as a function of the true stellar mass. The error measures are shown in the top left corner. A good prediction is identified with a small variance in the predicted error and a distribution centred on the horizontal dashed line.

5 Comparison EAGLE and SDSS

In this chapter, the three optimised neural networks are evaluated with SDSS galaxies. To maximise the performance on the SDSS data set, both the EAGLE data (that the network is trained and tested upon) and the SDSS data (that serves as an additional test set) are manipulated to avoid that a discrepancy in performance is caused by differences in the data. We ensure that both the EAGLE test set and the SDSS set consist of an equal amount of galaxies (Section 5.1) and that their stellar masses follow the same uniform distribution (Section 5.2). We ensure that the same stellar mass range is used in all experiments, thereby defining a benchmark to compare the subsequent experiments to (Section 5.2.1). Additionally we add noise to the EAGLE data (Section 5.3) and analyse if differences in performance might be caused by different stellar mass modelling methods, through evaluation of the stellar masses of Brinchmann et al. (2004) (Section 5.4). Finally an experiment is performed in which the two data sets are interchanged: the SDSS set is used as the training set and the EAGLE data set is used as an additional test set (Section 5.5).

5.1 Size restriction (random sampling)

In order to properly compare the performance of the neural networks on two data sets, it is important that these data sets resemble each other. A first step in this direction is ensuring that the number of galaxies propagated through the network is equal for both data sets, as described by *random sampling* in Section 2.3. The training set then consists of $4 \cdot 125 = 500$ EAGLE galaxies and the EAGLE test set and SDSS data set comprise 125 galaxies each. Their feature distributions can be observed in Figure 3b. The results of training and cross-validating the three neural networks on this size-limited EAGLE training set can be observed in Table 7, showing the five cross-validated error measures as described in Section 3.3 for each of the three neural networks. After cross-validation, the networks are trained again on the entire size-limited EAGLE training set, and the evolution of the MAE and R^2 statistic during training evaluated on the EAGLE training, test and SDSS set can be observed in Figure 26 of Appendix D. The evaluation of these trained neural networks on the size-limited EAGLE and SDSS test set can be observed in Figures 13a to 13f. These figures show the kernel density estimated distribution of the prediction error as a function of the original stellar mass. The five error measures as defined in Section 3.3 are shown in the top-right corner of the figures. Figures 13a, 13b and 13c show the evaluation of the three networks `nocolours`, `subsetcolours` and `allcolours` on 125 galaxies of the EAGLE test set and Figures 13d, 13e and 13f show the evaluation of the three networks on 125 galaxies of the SDSS data set. If the distribution is centred on the horizontal dashed line (a prediction error of zero) for all stellar masses and has small variance, the network performs well on the data set evaluated.

What can be immediately seen from Table 7 is that restrictions on the size of the data set (*sampling*) lead to higher errors. This is straightforward as propagating more data through the network during training will allow for a better learning process. For the EAGLE test set, the error distributions are centred roughly around a prediction error of zero, with higher variance towards the low-mass end. It is interesting to note that the performance of the networks `subsetcolours` and `allcolours` (containing colour information) on the SDSS set has significantly improved even though less data is used in the training process. The error distributions of both networks evaluated on the SDSS set are centred around zero (low μ) and the best result obtained on the SDSS data is $MAE = 5.75{\cdot}10^{-2}$ and $R^2 = 0.921$. Nonetheless, for low masses an inclination towards predicted masses that are too high (and vice versa)

still exists, therefore it is worth exploring other methods that might diminish the discrepancy between the EAGLE test set and SDSS set.

5.2 Stellar mass distribution restriction (uniform sampling)

As mentioned in Section 2.3 and observed in Figures 3a and 3b, the medians of the feature distributions of EAGLE and SDSS show large differences. To ensure that the discrepancy in the evaluation of the neural network on the EAGLE test set and SDSS data set is not caused by these differences, we aim to diminish the difference in medians by enforcing both data sets to follow the same *uniform* stellar mass distribution. This does imply however that the resulting stellar mass range is smaller than the range of previous experiments, which might therefore lead to higher errors. After the uniform stellar mass sampling procedure, the resulting data set consists of 500 EAGLE training galaxies, 125 EAGLE test galaxies and 125 SDSS galaxies – the same number of galaxies as in the *random* sampling experiment described above. The results of training and cross-validating the three networks on the EAGLE training set with uniform mass distribution can be observed in Table 7, showing the five cross-validated error measures as described in Section 3.3 for each of the three neural networks. After 5-fold cross-validation the neural networks are trained again on the EAGLE training set with uniform stellar mass distribution. The evolution of the MAE and R^2 statistic during the training process evaluated on the EAGLE training, test and SDSS set with uniform stellar mass distribution can be observed in Figure 27 of Appendix D. The results of the evaluation of the three networks with the EAGLE test set and SDSS set with uniform stellar mass distribution can be observed in Figures 14a to 14f, showing the kernel density estimated distribution of the prediction error as a function of the original stellar mass. The five error measures as defined in Section 3.3 are shown in the top-right corner of the figures. Figures 14a, 14b and 14c show the evaluation of the three networks `nocolours`, `subsetcolours` and `allcolours` on 125 galaxies of the EAGLE test set with a uniform stellar mass distribution and Figures 14d, 14e and 14f show

		`nocolours`	`subsetcolours`	`allcolours`
no sampling	MAE	$2.14{\cdot}10^{-2}$	$1.75{\cdot}10^{-2}$	$1.75{\cdot}10^{-2}$
	R^2	0.994	0.996	0.996
	$\bar{R}^2$	0.994	0.996	0.996
	μ	$6.78{\cdot}10^{-4}$	$1.83{\cdot}10^{-3}$	$-2.14{\cdot}10^{-4}$
	σ^2	$7.35{\cdot}10^{-4}$	$4.96{\cdot}10^{-4}$	$4.94{\cdot}10^{-4}$
random	MAE	$4.52{\cdot}10^{-2}$	$3.15{\cdot}10^{-2}$	$2.84{\cdot}10^{-2}$
	R^2	0.980	0.990	0.991
	$\bar{R}^2$	0.978	0.989	0.989
	μ	$7.73{\cdot}10^{-3}$	$5.17{\cdot}10^{-3}$	$1.56{\cdot}10^{-3}$
	σ^2	$2.95{\cdot}10^{-3}$	$1.41{\cdot}10^{-3}$	$1.33{\cdot}10^{-3}$
uniform	MAE	$5.22{\cdot}10^{-2}$	$3.93{\cdot}10^{-2}$	$3.66{\cdot}10^{-2}$
	R^2	0.987	0.993	0.994
	$\bar{R}^2$	0.986	0.992	0.993
	μ	$-2.19{\cdot}10^{-3}$	$-4.17{\cdot}10^{-3}$	$4.72{\cdot}10^{-3}$
	σ^2	$4.12{\cdot}10^{-3}$	$2.25{\cdot}10^{-3}$	$2.04{\cdot}10^{-3}$

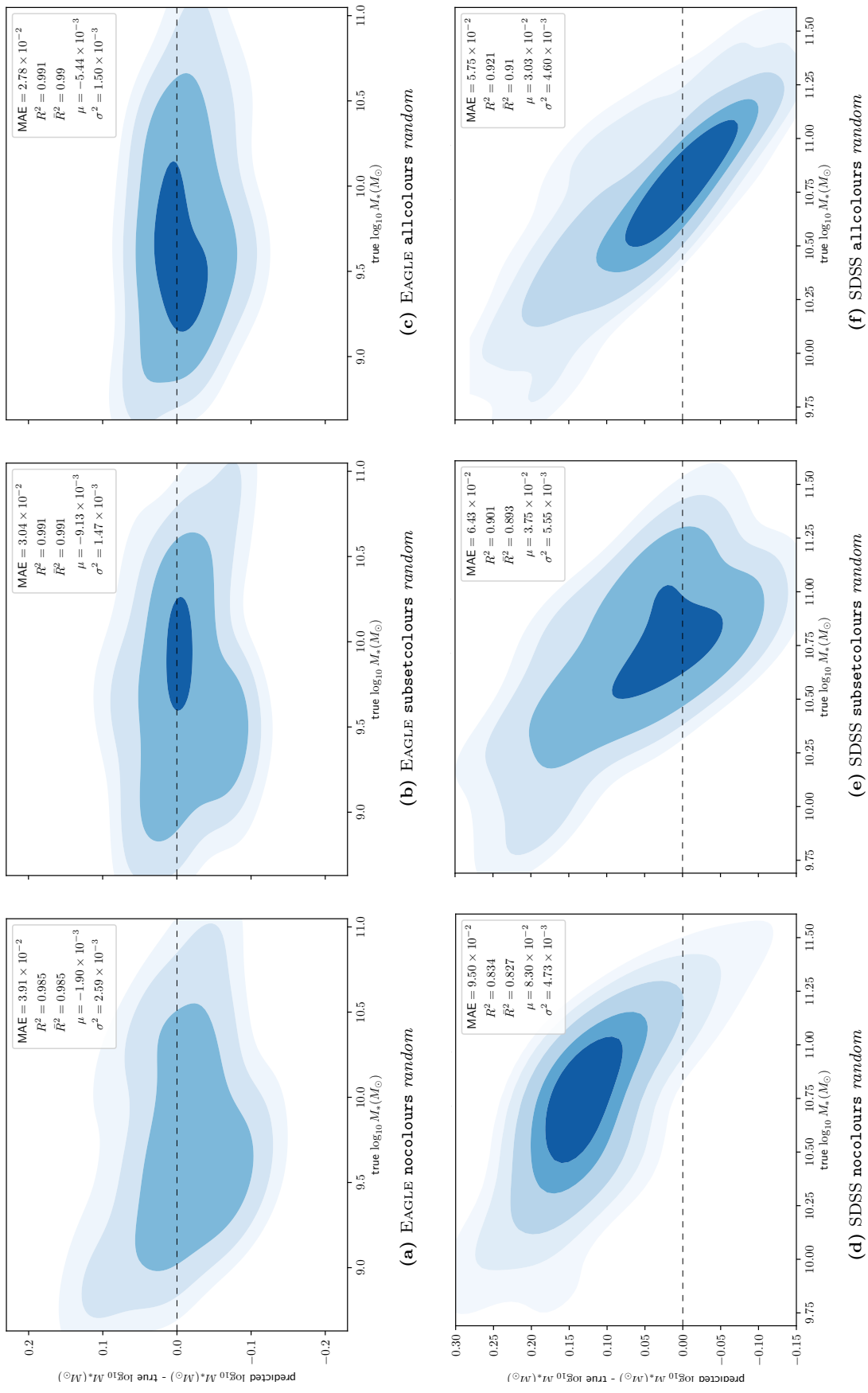
Table 7: Cross-validation of the three neural networks (`nocolours`, `subsetcolours`, `allcolours`) on the complete EAGLE training set (*no sampling*), 500 *random*ly sampled and 500 *uniform*ly sampled galaxies of the EAGLE training set, as described in Section 2.3. The five error measures evaluated are described in Section 3.3.

the evaluation of the three networks on 125 SDSS galaxies following the same uniform stellar mass distribution. If the error distribution is centred around the horizontal dashed line (a prediction error of zero) for all stellar masses and has small variance, the network performs well on the data set evaluated.

From Figure 14 it is apparent that the error distribution for EAGLE galaxies following a uniform stellar mass distribution is centred around zero for all masses in all three neural networks, now with a slightly higher variance towards the high-mass end. The three neural networks show worse performance on the SDSS galaxies however, of which the error distributions are centred ~ 0.2 dex away from the dashed line indicating zero error. The exception to this discrepancy is the evaluation of the `allcolours` neural network on SDSS galaxies, for which the error distribution is centred around zero towards the high-mass end.

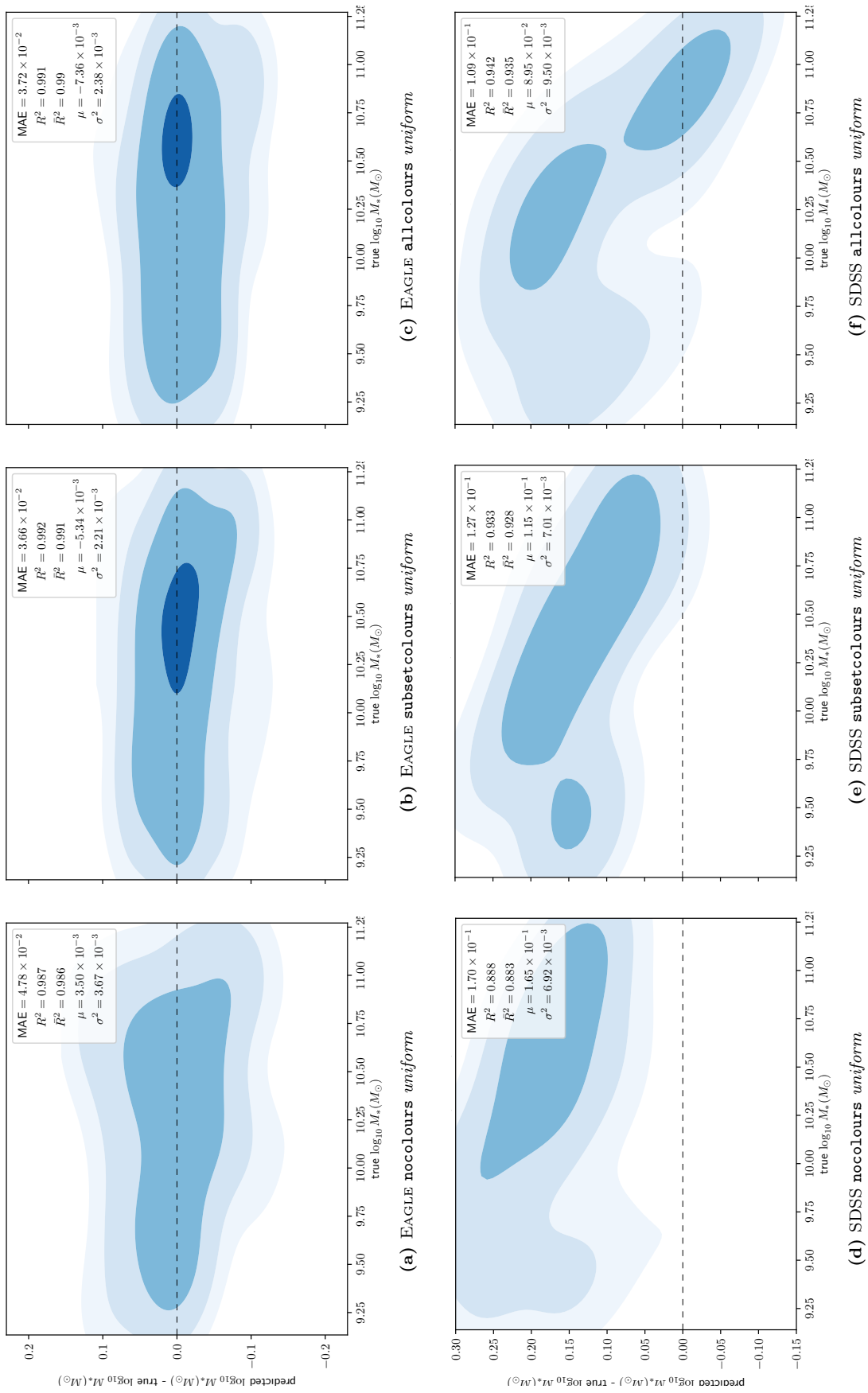
It is apparent that the cross-validated MAE increases with respect to cross-validation on the *random*ly sampled data described in the section above for all of the three networks. The R^2 statistic is expected to decrease for an increasing MAE, however, the R^2 increases as well. Note that a decrease in performance is expected with respect to the *random* sampling as the data that the network is trained and tested upon has a smaller stellar mass range. These differences are quite small however hence for the EAGLE data no clear preference for *random* or *uniform* sampling can be concluded from these results.

The same trend is observed for the MAE and R^2 evaluated on the SDSS set, for which the differences are slightly larger. The best results on the SDSS set are obtained with the `allcolours` network of $MAE = 1.09{\cdot}10^{-1}$ and $R^2 = 0.942$. The increase in MAE with respect to random sampling might again be due to a smaller stellar mass range used in the current experiment. Even though in some cases, the results of the random data sampling method might be better, in the continuing experiments the *uniform* stellar mass sampling method is applied to ensure that no bias towards certain masses is put into the model.



(a) EAGLE `nocolours` *random*
(b) EAGLE `subsetcolours` *random*
(c) EAGLE `allcolours` *random*
(d) SDSS `nocolours` *random*
(e) SDSS `subsetcolours` *random*
(f) SDSS `allcolours` *random*

Figure 13: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` on 125 **randomly sampled** EAGLE galaxies (13a, 13b and 13c) and 125 randomly sampled SDSS galaxies (13d, 13e and 13f) after training on 500 randomly sampled EAGLE training galaxies. The figures show the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass. The error measures are shown in the top right corner.



(a) EAGLE `nocolours` *uniform*

(b) EAGLE `subsetcolours` *uniform*

(c) EAGLE `allcolours` *uniform*

(d) SDSS `nocolours` *uniform*

(e) SDSS `subsetcolours` *uniform*

(f) SDSS `allcolours` *uniform*

Figure 14: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` on 125 EAGLE galaxies (14a, 14b and 14c) and 125 SDSS galaxies (14d, 14e and 14f) that are sampled to follow a **uniform stellar mass distribution** after training on 500 uniformly sampled EAGLE training galaxies. The figures show the KDE distribution of the prediction error of the true stellar mass as a function of the true stellar mass. The error measures are shown in the top right corner.

5.2.1 Benchmark for future experiments

In future experiments of this chapter the EAGLE and SDSS data are manipulated to research whether this might reduce the discrepancy between the EAGLE and SDSS data. The manipulation of either data sets results in a smaller overlap in the mass range of the two. To be able to compare the results of these future experiments with the experiment of this section, in which the data is uniformly sampled but not further manipulated, we reduce the mass bin range of the data and define a *benchmark* experiment to ensure that approximately the same uniform stellar mass range is used in all experiments. This benchmark experiment follows the exact same steps as the *uniform* mass sampling experiment of Section 5.2, except for the mass-bin range used to uniformly sample the data. Where a stellar mass bin range of $\log(9.12 - 11.35)M_{\odot}$ is used in previous experiments, for the benchmark and for consecutive experiments a mass-bin range of $\log(9.44 - 11.02)M_{\odot}$ is used. We specifically state mass-*bin* range as the restrictions are on the mass-bins that we randomly select galaxies from. Therefore, due to random selections, slight differences in mass ranges among different experiments might still occur. For a detailed description of the uniform stellar mass sampling procedure, please refer to Section 2.3.

The results of training and cross-validating the three networks on 500 galaxies of the EAGLE training set, following a *uniform* stellar mass distribution and a smaller mass-range, can be observed in Table 8. This table shows the five cross-validated error measures as described in Section 3.3 for each of the three neural networks. After 5-fold cross-validation, the neural networks are trained again on the same training set and evaluated on the EAGLE test set and SDSS data set, also following a *uniform* stellar mass distribution and a smaller mass-range. The evaluation can be observed in Figure 15, showing the kernel density estimated distribution of the prediction error as a function of the true stellar mass. The same discrepancy between the evaluation of the networks on the EAGLE and SDSS data can be observed, where the errors on SDSS show a much higher μ than for EAGLE and an inclination of low errors for high stellar mass and high errors for low stellar mass can be observed. The results of this experiment are used as a *benchmark* that future experiments can be compared to.

	`nocolours`	`subsetcolours`	`allcolours`
MAE	$6.57{\cdot}10^{-2}$	$5.04{\cdot}10^{-2}$	$4.95{\cdot}10^{-2}$
R^2	0.980	0.988	0.988
$\bar{R}^2$	0.978	0.987	0.985
μ	$1.90{\cdot}10^{-3}$	$-4.42{\cdot}10^{-4}$	$-3.02{\cdot}10^{-3}$
σ^2	$6.40{\cdot}10^{-3}$	$3.91{\cdot}10^{-3}$	$4.06{\cdot}10^{-3}$

Table 8: Cross-validation of the three neural networks (`nocolours`, `subsetcolours`, `allcolours`) on 500 *uniform*ly sampled **benchmark** galaxies of the EAGLE training set. The five error measures evaluated are described in Section 3.3.

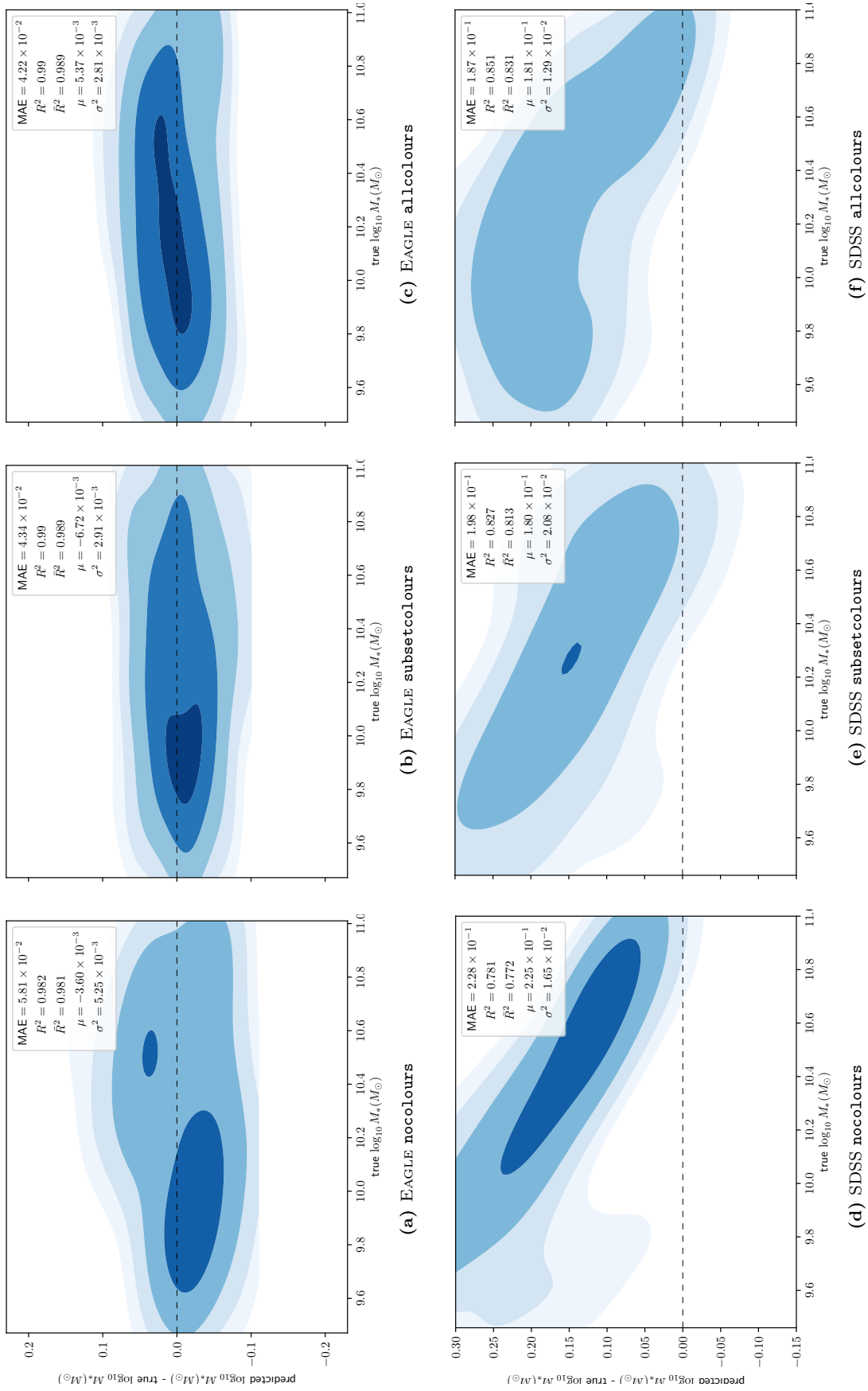
	`nocolours`	`subsetcolours`	`allcolours`
MAE	$7.50{\cdot}10^{-2}$	$5.91{\cdot}10^{-2}$	$5.77{\cdot}10^{-2}$
R^2	0.973	0.982	0.983
$\bar{R}^2$	0.971	0.980	0.980
μ	$1.24{\cdot}10^{-3}$	$4.14{\cdot}10^{-3}$	$1.33{\cdot}10^{-3}$
σ^2	$8.83{\cdot}10^{-3}$	$5.76{\cdot}10^{-3}$	$5.38{\cdot}10^{-3}$

(a) Noisy photometry

	`nocolours`	`subsetcolours`	`allcolours`
MAE	$1.72{\cdot}10^{-1}$	$1.62{\cdot}10^{-1}$	$1.62{\cdot}10^{-1}$
R^2	0.849	0.862	0.865
$\bar{R}^2$	0.841	0.843	0.841
μ	$-1.84{\cdot}10^{-3}$	$-4.16{\cdot}10^{-3}$	$-6.08{\cdot}10^{-4}$
σ^2	$4.64{\cdot}10^{-2}$	$4.23{\cdot}10^{-2}$	$4.19{\cdot}10^{-2}$

(b) Noisy stellar masses

Table 9: Cross-validation of the three neural networks (`nocolours`, `subsetcolours`, `allcolours`) on 500 *uniform*ly sampled galaxies of the EAGLE training set. To the photometry (9a) and the stellar mass (9b) **noise is added** as described in Section 2.3. The five error measures evaluated are described in Section 3.3.



(a) EAGLE `nocolours` **(b)** EAGLE `subsetcolours` **(c)** EAGLE `allcolours`

(d) SDSS `nocolours` **(e)** SDSS `subsetcolours` **(f)** SDSS `allcolours`

Figure 15: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` on 125 EAGLE galaxies (15a, 15b and 15c) and 125 SDSS galaxies (15d, 15e and 15f) following a **uniform** stellar mass distribution that serve as our **benchmark** in future experiments. The figures show the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass. The error measures are shown in the top right corner.

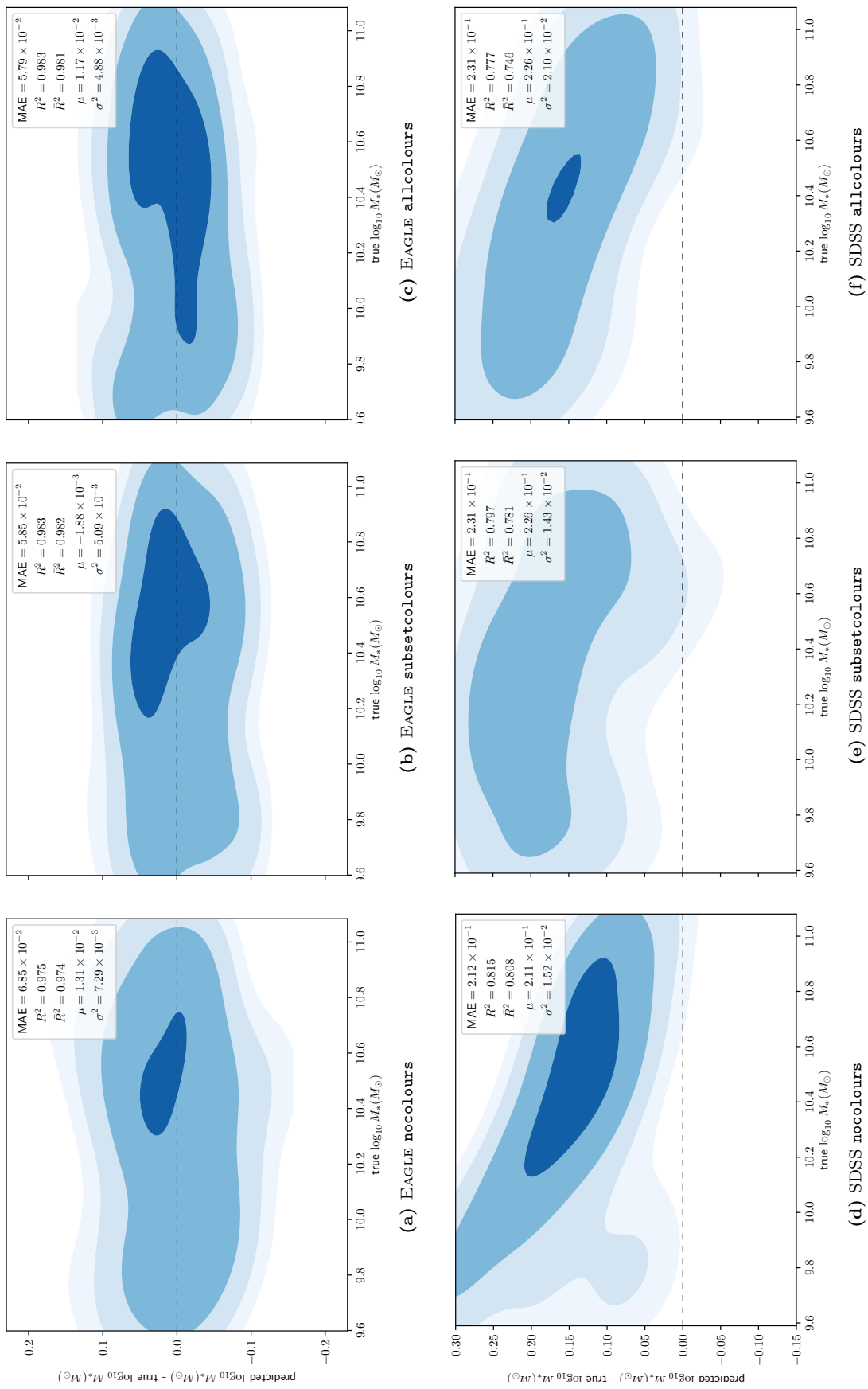
5.3 Adding noise

The observed SDSS flux and stellar mass of galaxies are values sampled from a normal distribution around their actual true value and with a standard deviation equal to the measurement error. We try to mimic these effects by adding noise to the EAGLE data in the same manner, as described in Section 2.3, hoping that this can reduce the difference in μ between EAGLE and SDSS.

The three neural networks `nocolours`, `subsetcolours` and `allcolours` are 5-fold cross-validated on the EAGLE training set with noisy photometry (Table 9a), and on the EAGLE training set with noisy stellar mass (Table 9b), showing the five cross-validated error measures as described in Section 3.3 for each of the three neural networks. After cross-validation, the networks are trained again on the noisy EAGLE training sets. The evaluation of the three neural networks trained on EAGLE galaxies with noisy photometry and noisy stellar mass can be observed in Figure 16 and Figure 17, respectively. The top-half of the figures show the evaluation on noisy EAGLE galaxies, and the bottom-half of the figures show the evaluation on SDSS galaxies. The figures show the kernel density estimated distribution of the prediction error of the stellar mass as a function of the true stellar mass. The five error measures as described in Section 3.3 evaluated on the data set in the figure are shown in the top-right corner of the figure. A neural network performs well on the data set evaluated if the error distribution for that data set is centred around the horizontal dashed line for all stellar masses, and if the distribution has low variance.

For the neural networks trained on EAGLE galaxies with photometry noise (Figure 16), the error distribution of the EAGLE test set is centred around zero (Figures 16a to 16c), but shows higher variance than in the benchmark experiment of Section 5.2.1. As observed in Table 9a, the cross-validation errors with respect to the benchmark are higher as well. The same observation can be made for the neural networks trained on EAGLE galaxies with stellar mass noise (Figure 17 and Table 9b), with even higher differences in error measures with respect to the benchmark and with significantly higher variance on EAGLE data. In Figures 17a to 17c, an inclination in the prediction of stellar mass can be noticed, where lower mass galaxies tend to have their stellar mass predicted too high and higher mass galaxies tend to have their stellar mass predicted too low. These results indicate that the relation between photometry and stellar mass becomes less clear and harder to predict when noise is added to the data, which is in line with expectations.

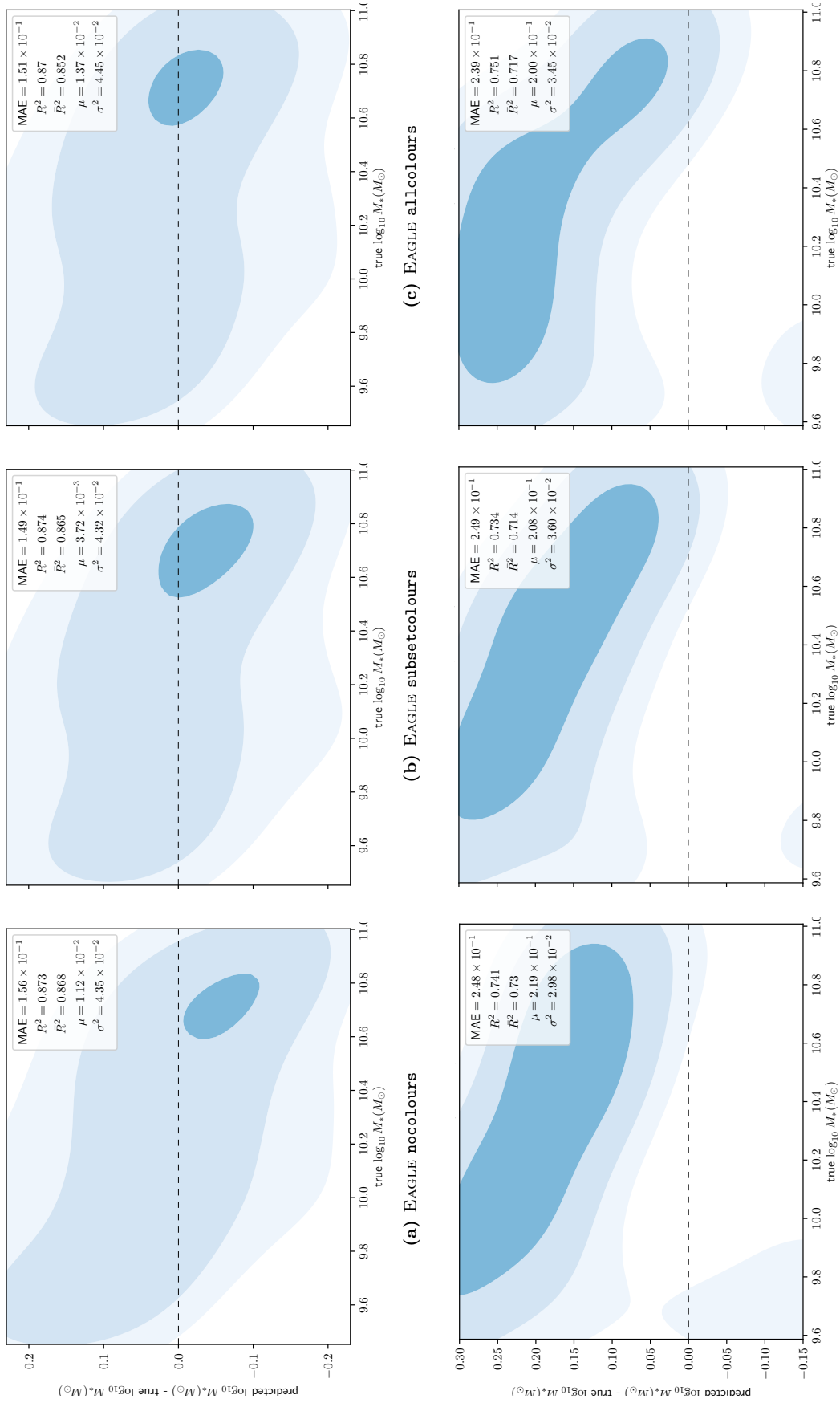
Significantly higher errors are measured in the evaluation of the networks on SDSS galaxies with respect to the benchmark. For both networks trained on EAGLE galaxies with photometry noise (Figures 16d to 16f) and on EAGLE galaxies with stellar mass noise (Figures 17d to 17f), the error distribution of the SDSS galaxies is ~ 0.2 dex from the horizontal dashed line that indicates zero prediction error. Furthermore, SDSS galaxies with lower mass are more inclined to have their stellar mass predicted too high than galaxies with higher mass. The higher errors on the evaluation of the SDSS galaxies with respect to the benchmark are possibly a consequence of the observation that the relation between photometry and stellar mass is more difficult to predict on EAGLE galaxies with noisy features. Therefore, by adding noise to the EAGLE galaxies, the discrepancy between EAGLE and SDSS galaxies with respect to the benchmark does not improve.



(a) EAGLE `nocolours` (b) EAGLE `subsetcolours` (c) EAGLE `allcolours`

(d) SDSS `nocolours` (e) SDSS `subsetcolours` (f) SDSS `allcolours`

Figure 16: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` trained and evaluated with EAGLE galaxies to which **photometry noise is added** and evaluated with SDSS galaxies, both following a *uniform* stellar mass distribution. The figures show the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass The error measures are shown in the top right corner.



(a) EAGLE `nocolours` (b) EAGLE `subsetcolours` (c) EAGLE `allcolours`

(d) SDSS `nocolours` (e) SDSS `subsetcolours` (f) SDSS `allcolours`

Figure 17: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` trained and evaluated with EAGLE galaxies to which **stellar mass noise is added** and evaluated with SDSS galaxies, whereby both data sets follow a *uniform* stellar mass distribution. The figures show the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass The error measures are shown in the top right corner.

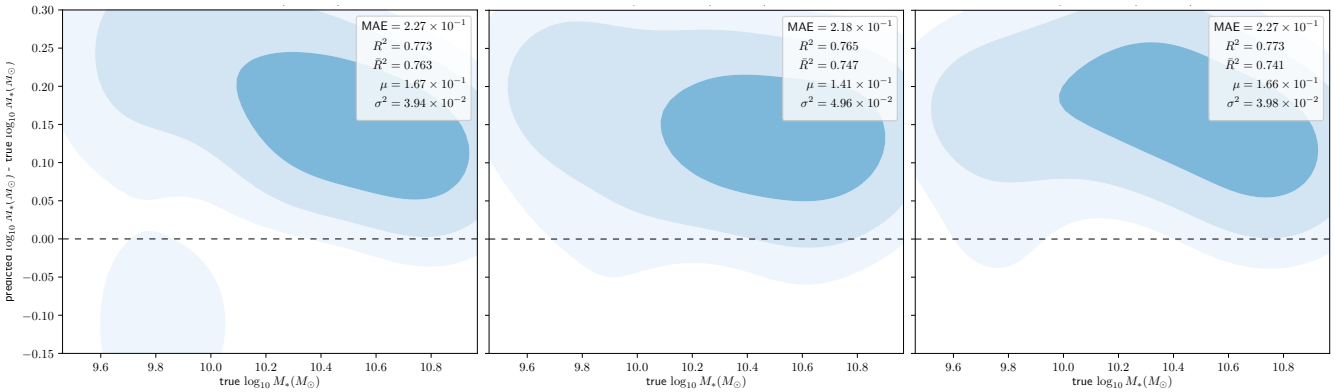
5.4 Comparison with Brinchmann et al. (2004)

The stellar masses of SDSS used in the previous experiments were those of Chang et al. (2015). Here we compare those stellar masses with the stellar masses obtained by Brinchmann et al. (2004) to investigate if this catalogue will improve our predictions. The difference between the Chang et al. (2015) and Brinchmann et al. (2004) catalogues is the addition of *WISE* photometry, an updated Galactic extinction correction and different dust attenuation laws.

Both methods assume fairly similar initial mass functions (Kroupa (2001) and Chabrier (2003)) and use the same cosmological parameters. The same preprocessing steps are performed on the Brinchmann et al. (2004) data set as described in Section 2.3 (without noise) and approximately the same *uniform* stellar mass distribution is enforced on the Brinchmann et al. (2004) data set as in our benchmark experiment, described in Section 5.2.1. Cross-validation and evaluation with the three neural networks on the EAGLE data set accomplishes similar results to those of Section 5.2.1, therefore only the evaluation with the Brinchmann et al. (2004) catalogue is shown.

The evaluation of the three trained neural networks on the (Brinchmann et al., 2004) catalogue can be observed in Figure 18, showing the kernel density estimated distributions of the prediction error as a function of the true stellar mass. Comparing the performance of the network on the Brinchmann et al. (2004) catalogue and the Chang et al. (2015) catalogue (Figures 15d-15f), a higher variance σ^2 for the Brinchmann et al. (2004) catalogue is visible. For the `nocolours` neural network, no significant differences between the MAE and R^2 statistic can be noticed. Once colours are introduced in the neural networks however, considerable differences do seem to occur. For the `subsetcolours` and `allcolours` neural networks, relatively large differences for the R^2 and σ^2 can be observed, which is also clearly visible from the figures. This result suggests that there is a substantial difference in the colour-mass relation of the two catalogues. For an in-depth comparison of the two catalogues, we refer to Chang et al. (2015).

Unfortunately, using a different SDSS stellar mass catalogue does not solve our issues of the discrepancy between EAGLE and SDSS. In subsequent experiments we will continue adopting the Chang et al. (2015) catalogue as our SDSS catalogue.



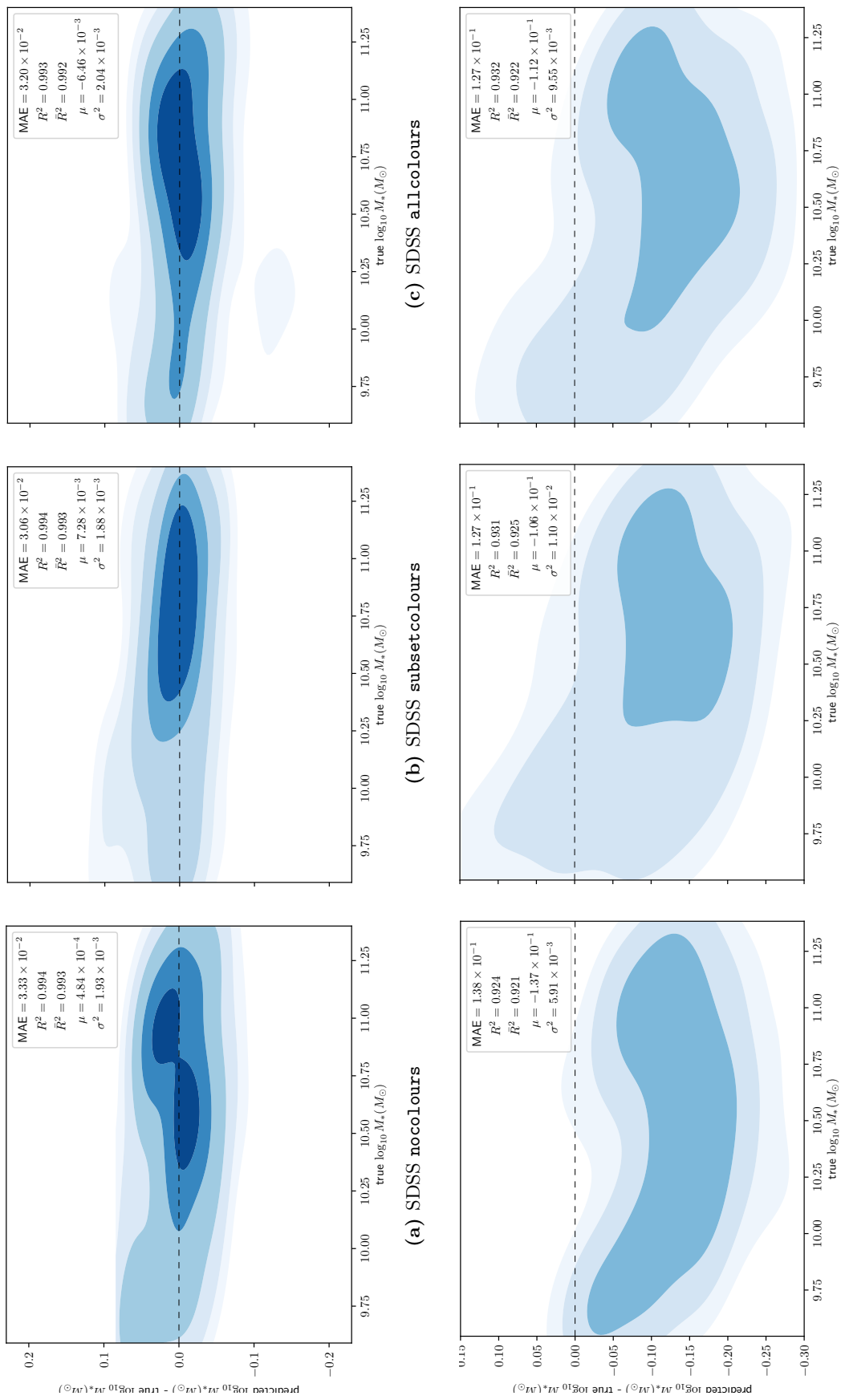
(a) SDSS `nocolours` **(b)** SDSS `subsetcolours` **(c)** SDSS `allcolours`

Figure 18: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` on 125 SDSS galaxies of the **Brinchmann et al. (2004) catalogue** (18a, 18b and 18c) following a uniform stellar mass distribution. The figures show the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass. The error measures are shown in the top right corner.

5.5 Training on SDSS

In the final experiment of this chapter the EAGLE and SDSS data sets are interchanged: the SDSS (Chang et al., 2015) data set is used as the training set and the EAGLE data set is used as an additional test set. In this way we research whether discrepancies between EAGLE and SDSS are caused by the incapability of the neural network to reproduce the photometry - stellar mass relation of the SDSS galaxies. In this experiment a uniform stellar mass range is used, but this stellar mass range is not the same as that of the benchmark therefore no qualitative comparison can be made between the two.

The results of the experiment can be observed in Figure 19, showing the kernel density estimated distributions of the prediction error as a function of true stellar mass for a neural network trained on SDSS galaxies and evaluated with SDSS and EAGLE galaxies. The neural network trained on the SDSS data obtains very low errors and is thus able to learn the relation between photometry and stellar mass well. Comparing these results with the evaluation of the EAGLE galaxies the same discrepancy between EAGLE and SDSS can be observed, where the EAGLE galaxies are repeatedly predicted to have lower stellar masses than the SDSS galaxies. This result suggests that the discrepancy between the EAGLE and SDSS galaxies are caused by differences that exist in the data itself, even after all the previous measures taken against these differences.



(a) SDSS `nocolours` (b) SDSS `subsetcolours` (c) SDSS `allcolours`

(d) EAGLE `nocolours` (e) EAGLE `subsetcolours` (f) EAGLE `allcolours`

Figure 19: Evaluation of neural networks `nocolours`, `subsetcolours` and `allcolours` trained on SDSS galaxies and evaluated on SDSS galaxies (19a, 19b and 19c) and on EAGLE galaxies (19d, 19e and 19f). The figures show the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass. The error measures are shown in the top right corner.

6 Centrals & Satellites

In this chapter the performance of the three neural networks on EAGLE central and satellite galaxies is analysed. Within EAGLE, centrals and satellites go through different evolutionary channels and have different quenching mechanisms, hence their mass accretion processes are different. Therefore it is interesting to analyse if there is a different relation between the photometry and stellar mass of centrals and satellites.

In this experiment we distinguish between three categories: *centrals (c)*, *satellites (s)* and *centrals + satellites (c+s)*. In the *c+s* category, no central or satellite constraint is placed on the galaxies, to ensure that same ratio of centrals to satellites is present in the training and test set as in the complete EAGLE data set. For each category, the same number of galaxies in each stellar mass bin is randomly selected from the complete EAGLE data set, to ensure that each category follows the same uniform stellar mass distribution (as described in Section 2.3). A total of 500 galaxies in each category is selected to serve as three separate training sets per category. In the same manner, 125 galaxies of each category are randomly selected that serve as three test sets. In this way we obtain the following 6 data sets with the number of galaxies of the data set in between brackets: *c+s* training set (500), centrals training set (500), satellite training set (500), *c+s* test set (125), *centrals* test set (125) and *satellites* test set (125). We ensure that there is no overlap between galaxies in the training and test set. Furthermore, 125 SDSS galaxies are selected to follow the same uniform stellar mass distribution, serving as an additional test set. The feature distributions of these data sets are shown in Figure 20. All EAGLE data sets follow roughly the same feature distributions for the same uniform stellar mass distribution thus indicating no directly observable difference in the photometry-stellar mass relation of centrals and satellites. The features of SDSS follow a different distribution from those of EAGLE .

Three experiments are performed in which we train on one of the three training sets and evaluate the network on all three test sets. We exclusively show the results of the `allcolours` neural network, as this network has demonstrated the lowest evaluation errors in the previous experiments and the result of the current analysis is similar for all three neural networks. The results of these experiments are shown in Table 10 and Figures 21, 22 and 23.

From Table 10a it is apparent that the three training and test sets obtain fairly similar results, with only small differences among them. Comparing the cross-validated results in this table, we observe that the best cross-validated result is obtained with the centrals training set, with an MAE of $4.75{\cdot}10^{-2}$. Comparing the error measures on the three *test* sets, the central galaxies also obtain the lowest evaluation errors in each training experiment. Additionally, except for the satellite training experiment in Table 10c, the satellite galaxies achieve the highest errors among the test sets. What is interesting to note from Table 10c, is that even when the network is trained on the satellite training set, the centrals test set obtains lower errors than the satellites test set.

If the network is trained on both centrals and satellites, the centrals and centrals + satellites test sets follow each other closely. The most notable difference can be observed for the mean μ, where a much higher μ is obtained on the satellites test set than on the other test sets. The higher error on the satellites test set can also be observed in Figure 21b and 22b.

Consequently, results seem to indicate that the relation between photometry and stellar mass can be slightly better retrieved from centrals than from satellites. However, the differences between the two categories are not large enough to draw a decisive conclusion from these experiments.

Finally, for SDSS no clear preference for either category training set can be observed as

the differences among the three categories are very small. Please note that the results on the SDSS set for training on both centrals and satellites are different than the results in previous analyses due to the fact that a smaller mass range is used here. From these results it can be concluded that the prediction on SDSS does not improve if the training set is restricted to either centrals or satellites.

	CV	**test *c+s***	**test *centrals***	**test *satellites***	**SDSS**
MAE	$5.06{\cdot}10^{-2}$	$4.44{\cdot}10^{-2}$	$4.29{\cdot}10^{-2}$	$5.00{\cdot}10^{-2}$	$1.84{\cdot}10^{-1}$
R^2	0.987	0.989	0.991	0.988	0.855
$\bar{R}^2$	0.985	0.988	0.989	0.986	0.835
μ	$5.25{\cdot}10^{-3}$	$-1.17{\cdot}10^{-3}$	$3.34{\cdot}10^{-3}$	$1.36{\cdot}10^{-2}$	$1.39{\cdot}10^{-1}$
σ^2	$3.85{\cdot}10^{-3}$	$3.45{\cdot}10^{-3}$	$2.92{\cdot}10^{-3}$	$3.92{\cdot}10^{-3}$	$2.93{\cdot}10^{-2}$

(a) Train on *centrals & satellites*

	CV	**test *c+s***	**test *centrals***	**test *satellites***	**SDSS**
MAE	$4.75{\cdot}10^{-2}$	$4.75{\cdot}10^{-2}$	$4.42{\cdot}10^{-2}$	$5.80{\cdot}10^{-2}$	$1.79{\cdot}10^{-1}$
R^2	0.988	0.988	0.989	0.985	0.860
$\bar{R}^2$	0.986	0.987	0.988	0.983	0.840
μ	$3.27{\cdot}10^{-3}$	$5.93{\cdot}10^{-3}$	$1.26{\cdot}10^{-2}$	$2.02{\cdot}10^{-2}$	$1.33{\cdot}10^{-1}$
σ^2	$3.49{\cdot}10^{-3}$	$3.73{\cdot}10^{-3}$	$3.13{\cdot}10^{-3}$	$4.61{\cdot}10^{-3}$	$2.92{\cdot}10^{-2}$

(b) Train on *centrals*

	CV	**test *c+s***	**test *centrals***	**test *satellites***	**SDSS**
MAE	$5.31{\cdot}10^{-2}$	$4.88{\cdot}10^{-2}$	$4.24{\cdot}10^{-2}$	$4.70{\cdot}10^{-2}$	$1.74{\cdot}10^{-1}$
R^2	0.986	0.989	0.991	0.990	0.857
$\bar{R}^2$	0.983	0.987	0.990	0.988	0.838
μ	$-4.51{\cdot}10^{-3}$	$-1.56{\cdot}10^{-2}$	$-1.08{\cdot}10^{-2}$	$2.69{\cdot}10^{-3}$	$1.22{\cdot}10^{-1}$
σ^2	$4.50{\cdot}10^{-3}$	$3.34{\cdot}10^{-3}$	$2.69{\cdot}10^{-3}$	$3.50{\cdot}10^{-3}$	$3.29{\cdot}10^{-2}$

(c) Train on *satellites*

Table 10: Evaluation of the `allcolours` neural network trained on 500 EAGLE centrals (10b), satellites (10c) and both centrals and satellites (10a) following a *uniform* mass distribution. The cross-validation (CV) results are shown in the first column for the galaxy category that the network is trained upon. The network is evaluated with 125 EAGLE centrals, satellites, both centrals and satellites ($c+s$) and the SDSS data set.

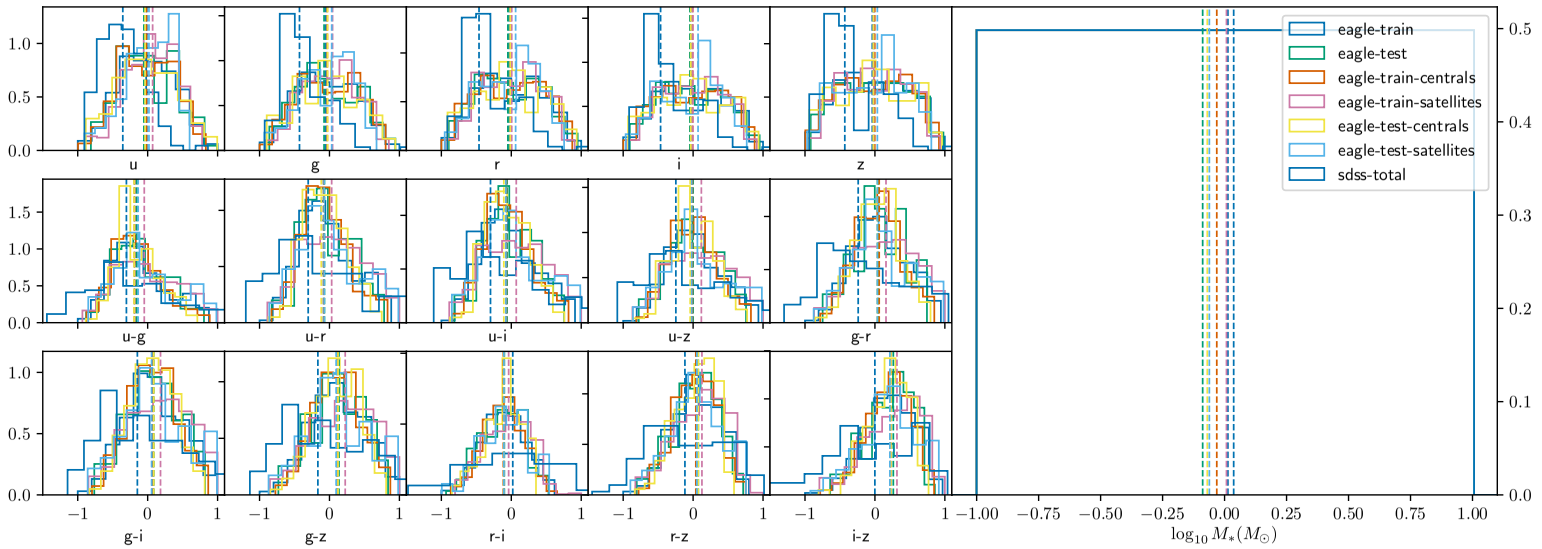
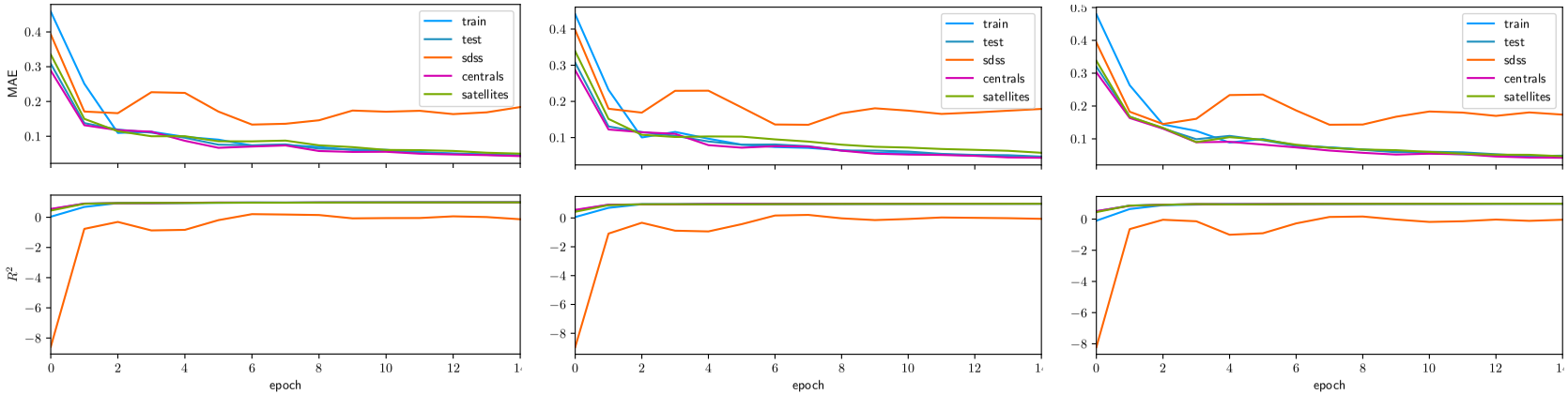
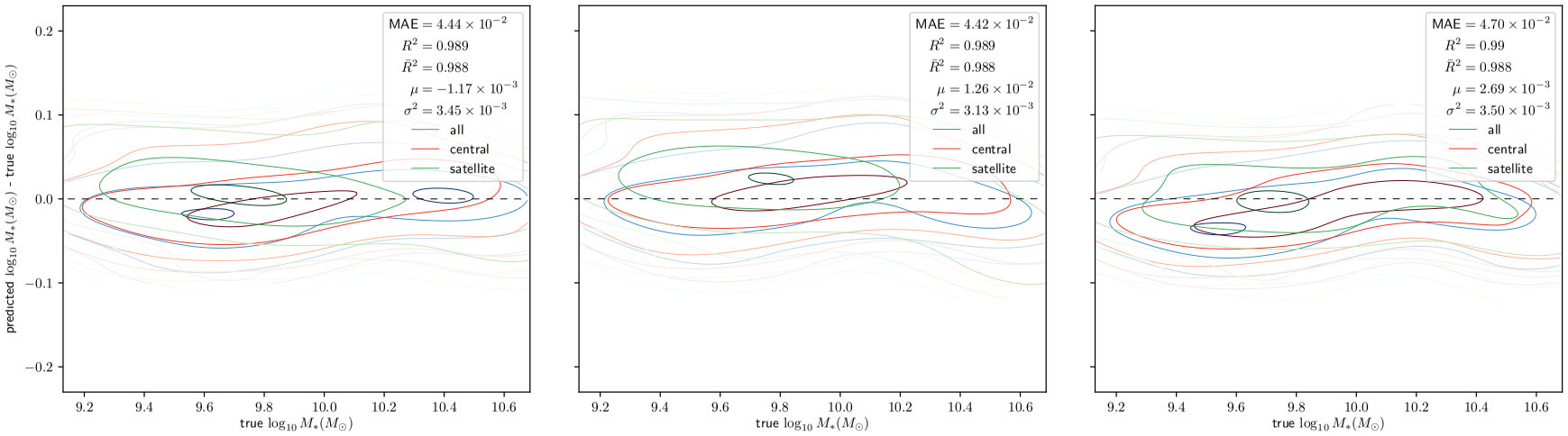


Figure 20: Distribution of features of the EAGLE central and satellite and SDSS galaxies after preprocessing. The left three rows show the distribution of photometry features and the large figure on the RHS shows the distribution of stellar mass. If no suffix is added to the label of the data set, this data set consists of both centrals and satellites. The vertical dashed lines show the median of distributions. All EAGLE datasets follow roughly the same photometry distributions.



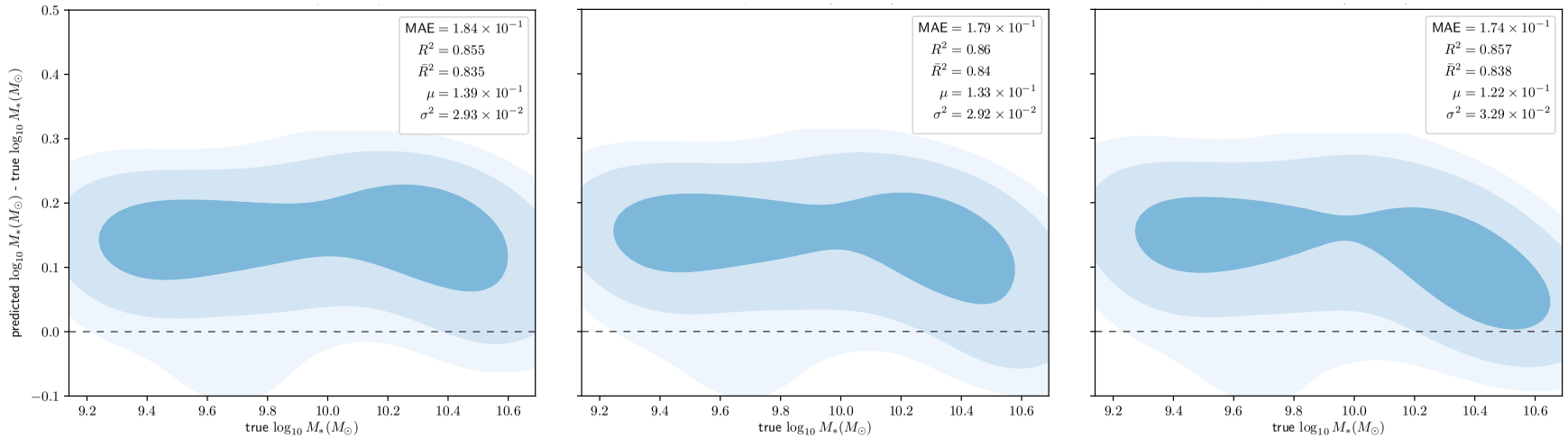
(a) Train on *c+s* **(b)** Train on *centrals* **(c)** Train on *satellites*

Figure 21: Evolution of the loss-function MAE and the R^2 statistic during training with the `allcolours` neural network on EAGLE centrals (21b), satellites (21c) and both centrals and satellites (21a). The measures on the three test sets, *centrals*, *satellites* and both centrals and satellites (label: *test*), are monitored throughout the training process. Additionally the measures on the SDSS data set is monitored as well.



(a) Train on *c+s* **(b)** Train on *centrals* **(c)** Train on *satellites*

Figure 22: Evaluation of the `allcolours` neural network trained on 500 EAGLE centrals (22b), satellites (22c) and both centrals and satellites (22a) following a *uniform* stellar mass distribution. The network is evaluated with 125 EAGLE centrals, satellites and both centrals and satellites. The figures show contours of the KDE distribution of the prediction error as a function of the true stellar mass. The error measures are shown in the top right corner for the test set containing the galaxy category that the network was trained upon.



(a) Train on *c+s* **(b)** Train on *centrals* **(c)** Train on *satellites*

Figure 23: Evaluation with SDSS galaxies of the `allcolours` neural network trained on 500 EAGLE centrals (23b), satellites (23c) and a set of both centrals and satellites (23a) following a *uniform* stellar mass distribution. The figures show contours the KDE distribution of the prediction error of the stellar mass as a function of the true stellar mass. The error measures on the SDSS data set are shown in the top right corner.

7 Feature importance

As analysed in Chapter 4, the choice of features is an important determinant in the prediction of the stellar mass. We have seen that the addition of colours to the input of the neural networks has increased the performance of the neural networks. We are interested to know which individual features contribute most to the prediction of stellar mass, to find out more about the process behind it.

Not all features (e.g. magnitudes or colours here) that are included in the neural network contribute equally to the predicted value of the network. Noisy or correlated features might not contribute to the prediction, and because an increase in the number of features requires an exponential increase in the number of samples that should be propagated through the network, uninformative features may reduce the performance of a neural network – a phenomenon known as the *curse of dimensionality* Bellman (1957); Leray and Gallinari (1998).

The importance of features in a neural network is determined by the weights of the network, but as all nodes are fully interconnected and hidden layers are present, it is impossible to map one weight to a specific input feature. The machine learning method of neural networks therefore resembles a 'black box': a method that is able to model complex relations well, but is not easily interpretable. Often within machine learning, a tradeoff has to be made between interpretability and accuracy.

To determine which features are most informative for the prediction of the stellar mass, it would make sense to try out all feature subsets as input features for the neural network, and choose the feature subset with the best performance. However, for j features, $2^j - 1$ feature subsets exist which implies that in our case ($j = 15$) we would have to cross-validate our network 32767 times, which is computationally unfeasible. A bottom-up sequential search is a solution to this problem, which is further explored in Section 7.3.

Several model-agnostic feature importance methods have been developed as well such as partial dependence, permutation dependence and the use of Shapley values. However, these methods rely on the manipulation (shuffling or permutation) of the existing data in such a way that for strongly correlated features, this artificially created data could not possibly exist. Therefore the results of these methods might be meaningless for data with strongly correlated features. Additionally, if features are statistically dependent, the contribution of one feature towards the prediction depends on the contribution of other features and these contributions cannot be separated. Therefore, in the following section (Section 7.1), the correlation of the data is evaluated. Consecutively, the method of Shapley values is implemented (Section 7.2), while acknowledging that the reliability of this method decreases for highly correlated data.

7.1 Correlation of the data

To assess the reliability of the methods described above that depend on uncorrelated features, the Pearson correlation coefficient ρ between all feature combinations X and Y,

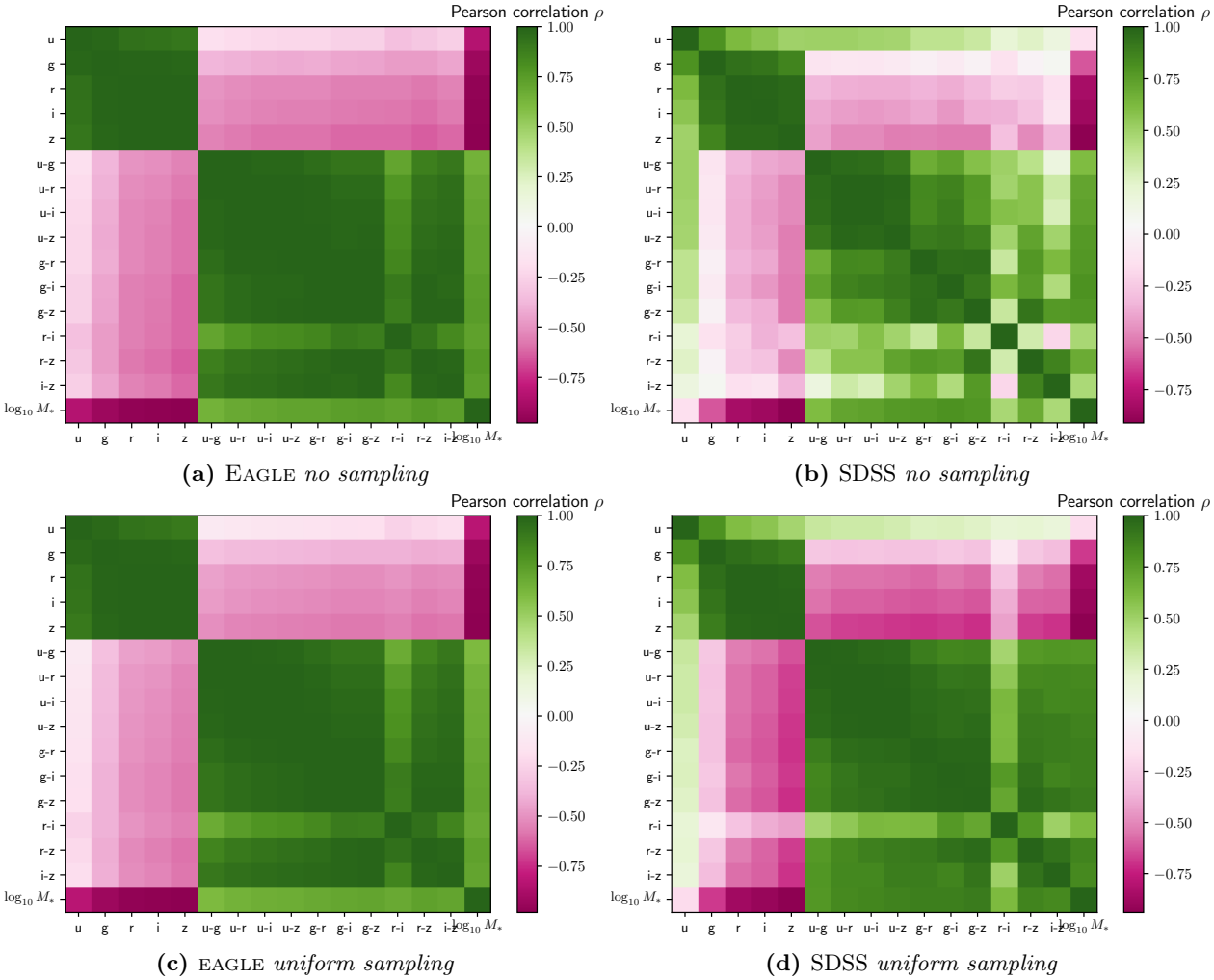
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\,, \tag{35}$$

is determined where σ_{XY} is the covariance between two features and σ is the standard deviation. The correlations for EAGLE and SDSS galaxies with no noise added to the data are illustrated in Figure 24. The top panels of the figure show the correlations for EAGLE and SDSS galaxies that are not sampled (Figures 24a and 24b, resp.), and the bottom panels of the figure show the correlations for EAGLE and SDSS galaxies that are sampled to follow a uniform

stellar mass distribution (Figures 24c and 24d, resp.). The figures show a two-dimensional histogram of all features plotted against each other, indicating the Pearson correlation coefficient ρ between all combinations of features. Positive and negative correlations are expressed with green and pink colours respectively, and a stronger correlation is coloured darker. The figures are symmetric along the diagonal axis and the diagonal illustrates the correlation of a feature with itself. The bottom row and rightmost column of each panel show the correlation of features with stellar mass.

From the figure it is apparent that magnitudes are highly positively correlated with other magnitudes, which is also accurate for the correlation between colours and colours. Colours and magnitudes demonstrate slightly lower negative correlation. For both EAGLE and SDSS, the $r - i$ colour indicates less correlation with features than other colours. Furthermore, we can observe that the correlation with stellar mass is stronger for magnitudes than for colours, with the z magnitude demonstrating the strongest correlation.

Overall, relatively high correlations between individual features are shown. Still we would like to research whether the method of Shapley values that depends on independent features can give comparable results to the sequential search. These methods are implemented in the next sections.



(a) EAGLE *no sampling* **(b)** SDSS *no sampling*

(c) EAGLE *uniform sampling* **(d)** SDSS *uniform sampling*

Figure 24: Pearson correlation coefficient between EAGLE (24a and 24c) features and SDSS (24b and 24d) features that are not sampled in any way (24a and 24b) and that are *uniformly* sampled and thus enforced to follow a uniform stellar mass distribution as described in Section 2.3 (24c and 24d).

7.2 Shapley values

The method of Shapley values (Shapley, 1953) originates from cooperative game theory, where it is used to calculate the marginal contribution of each player towards the payout of a cooperative game. Suppose a number of players play a cooperative game and need to determine how much each player has contributed to the collective payout. If all players join the game sequentially then at first sight a solution would be to determine the increase in payout after a new player has joined the game. For example, player A starts the game and receives a payout of 5, subsequently player B enters the game and the payout increases towards 8, and finally player C raises the payout towards 10. In that case the contribution of player A would be 5 and the contribution of player B would be 3 and the contribution of player C would be 2. However, if player B and C have similar dexterity, player B might get a higher prediction simply because he or she joined earlier in the game.

The method of Shapley values approaches this issue by determining the average marginal contribution of players to all possible subsets of players. The marginal contribution of a player i is defined as the difference between the payout of a subset of players S with this player $f_{S\cup\{i\}}(x_{S\cup\{i\}})$, subtracted by the payout of a subset of players without this player $f_S(x_S)$. Suppose there are in total $|F|$ players in this game and we would like to calculate the contribution of player i towards the total payout, then this contribution is given by the Shapley value ϕ_i,

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] , \tag{36}$$

where $|S|!$ is the number of permutations of a given subset $S \subseteq F\setminus\{i\}$ before player i joins the game, and $(|F|-|S|-1)!$ is the number of permutations of the remaining subset of players after i has joined the game.

This method can be translated to neural networks to determine the importance of individual features towards a prediction. The features that the neural network is trained upon (here the photometry) are the players in the above example, and the prediction of the neural network (e.g. the stellar mass) is the payout. The Shapley values are then an indicator of the contribution of a feature i towards the prediction (stellar mass) with respect to the average contribution.

Shapley values are implemented with the SHapley Additive exPlanations framework (SHAP Lundberg and Lee, 2017)[3]. The methods in this framework use an *explanation model* g: a model that is serves as an explanation for the original model $f(x)$. This explanation model is easier to understand than the original model and approximates the original model $g(x') \approx f(h_x(x'))$ with $x = h_x(x')$ a mapping function of the simplified input features x' to the original input features x. For the *additive feature attribution methods* that the SHAP framework consists of, the explanation model is a linear combination of input features,

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \,, \tag{37}$$

where ϕ_i is the relative importance of input x'_i towards the prediction and M is the total number of features in the model. For SHAP, this relative importance is calculated with the

[3]https://github.com/slundberg/shap

Shapley value for which the equation becomes,

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] , \tag{38}$$

where all features z' that are not evaluated (in the previous example: players that have not entered the game yet) are set to $z' = 0$ and $|z'|$ is then the number of features that are not set to zero (Lundberg and Lee, 2017). The missing features can not be simply ignored but must be set to a specific value since the architecture of the neural network is pre-defined. The value 0 of the missing features can be mapped to a reference value z_s with the mapping function $z_s = h_x(z')$. The prediction with this reference value,

$$f(z_s) = f(h_x(z')) = E\left[f(z) \mid z_s \right] , \tag{39}$$

is often approximated by assuming features are independent. The method of the SHAP framework used in this work is DeepSHAP, a method that combines both DeepLIFT (Shrikumar et al., 2017) and Shapley values. This method assumes independent features which is not the case for our data. Therefore we acknowledge that this method applied to our data is not fully reliable. For further details about this method, please refer to Lundberg and Lee (2017).

The data used for the DeepSHAP algorithm are EAGLE and SDSS galaxies that are sampled to follow a uniform stellar mass distribution. The EAGLE training set, consisting of 500 galaxies, is used as a *background* sample to approximate the conditional expectation described in Eq. (39). Consecutively, the SHAP values are determined on 125 galaxies of the EAGLE test set and 125 galaxies of the SDSS test set, given the trained neural network on the EAGLE training set.

Please note the following for the SHAP values of SDSS. The conditional expectation is determined using EAGLE galaxies, and the conditional expectation for the SDSS data might be slightly different from the conditional expectation determined with the EAGLE data. Additionally the neural network is trained with the EAGLE data set and the evaluation of the network on EAGLE and SDSS data differs, as analysed in Chapter 5. The SHAP values are determined using this trained neural network on EAGLE and the conditional expectation, therefore it is plausible that the SHAP values calculated with the SDSS data are different from those determined on EAGLE . Nonetheless, the difference between the SHAP values of EAGLE and SDSS could be indicative of the features that cause the discrepancy between EAGLE and SDSS.

The results of the DeepSHAP algorithm can be observed in Figure 25. The left-hand side of the figure shows the absolute SHAP value per feature averaged over the 125 evaluations for EAGLE and SDSS, sorted in descending order for the EAGLE galaxies. As can be observed from this figure, for the `nocolours` neural network, the z magnitude has the highest impact on the prediction of the stellar mass, according to the DeepSHAP algorithm. The right-hand side of the figure shows the SHAP values per feature for each of the evaluations on the EAGLE test set. Per feature, the values are distributed further in a vertical direction from the horizontal line when more galaxies have the same SHAP value, thus showing the distribution of SHAP values for each feature. If a SHAP value is negative, it means that it contributes negatively to the predicted value. The SHAP values are coloured, based on their corresponding feature value x_i, see Eq. (37).

The SHAP figures for the `nocolours` neural network can be interpreted as follows: the z magnitude contributes most to the prediction of the stellar mass. Additionally the z magnitude contributes negatively to the prediction of the stellar mass if the magnitude is high and positively if the magnitude is low. The fact that for the `nocolours` neural network the z

magnitude contributes most to the prediction of the stellar mass is unsurprising, as the z magnitude shows the strongest correlation with the stellar mass, as can be observed in Figure 24. Furthermore, the correlation between z magnitude and stellar mass is negative, supporting the observation of a negative SHAP value with a high z magnitude and vice versa. The importance of the remaining magnitudes is lower and relatively comparable. What is interesting is that Figure 25b suggests that not all magnitudes contribute negatively to the prediction of the stellar mass, even though all magnitudes are negatively correlated with stellar mass.

Once colours are added to the neural network, this images changes. For the `subsetcolours` neural network, the z magnitude remains the most important predictor for the value of the stellar mass, but the relative importance of the r magnitude has significantly increased. The contribution of the r and u magnitude is reversed, now corresponding to a negative correlation with the stellar mass. All colours contribute positively to the prediction of the stellar mass, implying that the importance of a colour increases for higher colour values. This corresponds to the positive correlation with stellar mass, as observed in Figure 24. Furthermore, for EAGLE , colours seem to be less important features in the prediction of stellar mass than magnitudes. For SDSS however, colours, specifically $g-r$ and $r-i$ colours, are features of significant importance, whereas for the `nocolours` neural network, the feature importance of EAGLE and SDSS is comparable. Note that all SHAP values for the `subsetcolours` neural network have decreased compared to the `nocolours` neural network. This makes sense as the sum of the SHAP values of all features in a prediction should be equal to one.

The `allcolours` neural network includes all colour combinations and demonstrates different results compared to the previous two networks. The z colour has become relatively unimportant, as well as all colour combinations with z, except for the $u-z$ colour that has a relatively high absolute SHAP value. For SDSS, this feature is by far the most important. Additionally, the g and i magnitude have become important features, while they are relatively unimportant features in previous analyses. The five most important features according to the DeepSHAP algorithm are g, i, $u-z$, r and $g-i$. It is interesting to note that the $r-i$ colour, a feature that has the lowest correlation with other features (24), indicates the lowest SHAP value for this network.

It is sensible that the relative importance of an individual magnitude decreases as more colours, related to that magnitude, are added to the network. Therefore a comparison of the results of the DeepSHAP algorithm across different neural networks can not be drawn.

Comparing the SHAP values of EAGLE and SDSS it is apparent that there are significant differences between the two, mainly in terms of the colours. The largest differences for the `subsetcolours` neural network occur between the $g-r$ and $r-i$ colour. For the `allcolours` neural network there are large differences for a lot of features. What is interesting to note is that the importance of the z magnitude, the feature that is strongest correlated with stellar mass, is exactly the same for both EAGLE and SDSS. This result suggests that the discrepancy between the stellar mass prediction of EAGLE and SDSS is caused by the relative differences between individual features for galaxies at a redshift of $z \sim 0.1$. The features with the largest difference in SHAP values between EAGLE and SDSS might indicate what causes the difference in the discrepancy between the prediction of EAGLE and SDSS.

The five features of the `allcolours` neural network, that most contribute to the prediction of the stellar mass, are,

$$\boldsymbol{x}^{(\mathrm{SHAP})} = \big(g, \quad i, \quad u-z, \quad r, \quad g-i\big) \ . \tag{40}$$

Using the `nocolours` neural network we test if using these five features instead of the five *ugriz* magnitudes achieves lower error measures. The fact that a different network is used

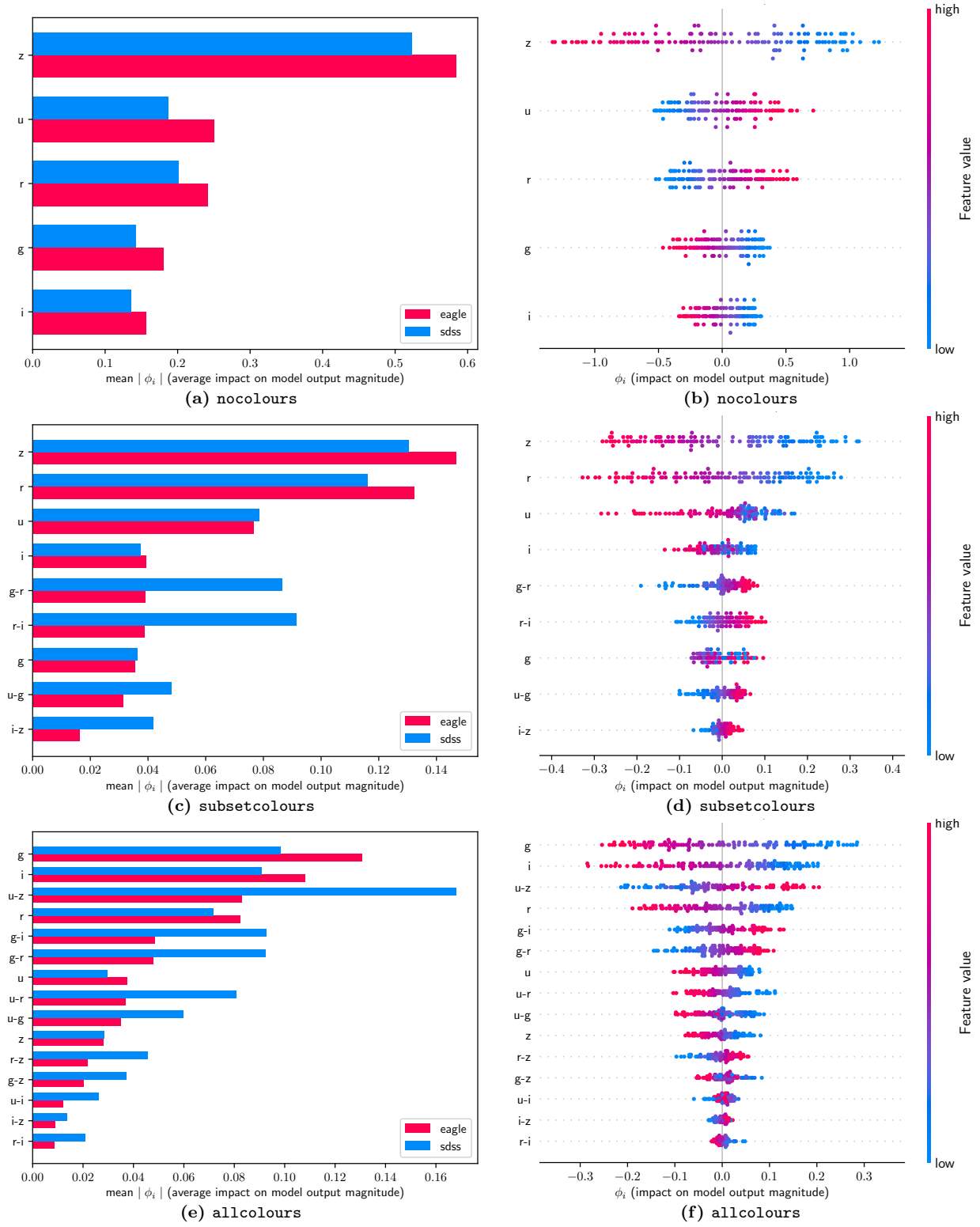
	CV	EAGLE test	SDSS test
MAE	$4.12{\cdot}10^{-2}$	$4.00{\cdot}10^{-2}$	$1.22{\cdot}10^{-1}$
R^2	0.992	0.990	0.920
$\bar{R}^2$	0.991	0.990	0.917
μ	$1.40{\cdot}10^{-3}$	$-1.24{\cdot}10^{-2}$	$1.09{\cdot}10^{-1}$
σ^2	$2.62{\cdot}10^{-3}$	$2.54{\cdot}10^{-3}$	$1.25{\cdot}10^{-2}$

Table 11: Evaluation of the `nocolours` neural network with the five most contributing input features found by SHAP, evaluated with five error measures as described in Section 3.3. The neural network is cross-validated (CV) on the EAGLE training set, and evaluated on the EAGLE test set and SDSS data set, all of which follow a uniform stellar mass distribution.

for training (`nocolours`) than the network that the contribution of features is determined upon with SHAP (`allcolours`) is justified by the experiment of Section 4.6.1 in which we show that cross-validation with the three different networks `nocolours`, `subsetcolours` and `allcolours` for the same input features obtains similar results. In this experiment the input features themselves cause differences in performance, rather than the neural networks.

The error measures obtained through evaluation on galaxies following a uniform stellar mass distribution, can be observed in Table 11. These error measures must be compared to the experiment of Section 5.2 in which we train the `nocolours` neural network with *ugriz* magnitudes on galaxies following a uniform stellar mass distribution, and not to the results of the benchmark experiment in Section 5.2.1 since a large mass range is used. Comparing these experiments it is apparent that the features found by SHAP obtain lower cross-validated errors on the EAGLE data set ($MAE = 4.12{\cdot}10^{-2}$) than the *ugriz* magnitudes ($MAE = 5.22{\cdot}10^{-2}$). For the SDSS galaxies, the performance with the features that contribute most according to SHAP ($R^2 = 0.920$), compared to the performance of *ugriz* magnitudes ($R^2 = 0.888$), increases as well. Unfortunately the discrepancy between EAGLE and SDSS galaxies is not resolved, as can be observed by the mean of the error distribution μ. Using the most contributing features of SHAP therefore gives more accurate predictions of the stellar mass.

From this experiment, it seems that SHAP is a valid method for determining the individual contribution of features towards a prediction and using a selection of the most contributing features found in a neural network. In the next section we will perform a sequential search for the best combination of features in the neural network. The features found with the sequential search can be used as a comparison for the importance of the features found with SHAP.



(a) `nocolours` (b) `nocolours`

(c) `subsetcolours` (d) `subsetcolours`

(e) `allcolours` (f) `allcolours`

Figure 25: Results of the DeepSHAP algorithm to determine the importance of individual features towards the prediction of stellar mass of the three neural networks `nocolours`, `subsetcolours` and `allcolours`. The absolute SHAP values for all features of the networks, as described in Section 7.2 are shown in the figures on the left-hand side, determined on 125 EAGLE and 125 SDSS galaxies with uniform stellar mass distribution. The distribution of SHAP values for each feature determined on 125 EAGLE galaxies with uniform stellar mass are shown in the figures on the right-hand side, where each SHAP value is coloured according to the value of the feature, thereby indicating the importance of this feature value if the feature value is high or low.

7.3 Bottom-up sequential search

As a comparison to the feature importance found with the SHAP framework, a *bottom-up sequential search* is performed. The bottom-up sequential search is an improved version of a search in which all subsets of the feature space F,

$$F = \big\{u,\quad g,\quad r,\quad i,\quad z,\quad u\text{-}g,\quad u\text{-}r,\quad u\text{-}i,\quad u\text{-}z,\quad g\text{-}r,\quad g\text{-}i,\quad g\text{-}z,\quad r\text{-}i,\quad r\text{-}z,\quad i\text{-}z\big\}\,, \tag{41}$$

are explored. Instead of exhaustively training a neural network on all subsets of the feature space and choosing the feature subset that obtains the lowest error, this method reduces the search space that is explored. In this method, for each iteration, only the three combinations of features that obtain the best results are further explored in the next iteration. The sequential search is executed in the following way.

In the first iteration ($n = 1$) of the sequential search the neural network is trained using only one input feature (thereby modifying the number of input nodes in the architecture of the network). The neural network is trained individually on the items of itemset I_1,

$$I_1 = \big\{\{u\},\{g\},\{r\},\{i\},\{z\},\{u\text{-}g\},\{u\text{-}r\},\{u\text{-}i\},\{u\text{-}z\},\{g\text{-}r\},\{g\text{-}i\},\{g\text{-}z\},\{r\text{-}i\},\{r\text{-}z\},\{i\text{-}z\}\big\}\,, \tag{42}$$

that correspond to all possible elements of F. Thus, in the first iteration of the sequential search, the neural network is cross-validated $|I_1|$ times. The training process consists of 5-fold cross-validation on 500 EAGLE galaxies of the EAGLE training set as described in Section 4.1, that are uniformly sampled as described in Section 2.3. The three feature sets of I_1 that retrieve the lowest cross-validated MAE on the EAGLE galaxies are placed into set B_1. For example, if the features $\{z\}$, $\{i\}$ and $\{r\}$ obtain the lowest, second-lowest and third-lowest cross-validated MAE, then $B_1 = \big\{\{z\},\{i\},\{r\}\big\}$.

In the second iteration of the sequential search, the neural network is trained using two input features. The feature subsets that the neural network is trained with are a combination of the three best features of the previous iteration $b_{1,1}, b_{1,2}, b_{1,3} \in B_1$ and all items f of the feature space F forming distinct set I_2,

$$\begin{aligned} I_2 = \{&b_{1,1}\cup\{u\}, b_{1,1}\cup\{g\}, b_{1,1}\cup\{r\},\ldots, b_{1,1}\cup\{r\text{-}i\}, b_{1,1}\cup\{r\text{-}z\}, b_{1,1}\cup\{i\text{-}z\},\\ &b_{1,2}\cup\{u\}, b_{1,2}\cup\{g\}, b_{1,2}\cup\{r\},\ldots, b_{1,2}\cup\{r\text{-}i\}, b_{1,2}\cup\{r\text{-}z\}, b_{1,2}\cup\{i\text{-}z\},\\ &b_{1,3}\cup\{u\}, b_{1,3}\cup\{g\}, b_{1,3}\cup\{r\},\ldots, b_{1,3}\cup\{r\text{-}i\}, b_{1,3}\cup\{r\text{-}z\}, b_{1,3}\cup\{i\text{-}z\}\}\,. \end{aligned} \tag{43}$$

In case $b_1 \cup f$ leads to an item of size one, it is removed from the set I_2. The three items of I_2 that lead to the lowest cross-validated MAE are added to set B_2. For example, if the feature sets $\{z, r\text{-}i\}$, $\{z, g\text{-}i\}$ and $\{i, g\text{-}r\}$ obtain the lowest cross-validated MAE, this set becomes $B_2 = \big\{\{z, r\text{-}i\},\{z, g\text{-}i\},\{i, g\text{-}r\}\big\}$. The set B_2 is then used to determine consecutive itemset I_3.

These steps can be generalised as follows. For each n^{th} iteration of the exploration, the neural network is trained using n input features, items of itemset I_n,

$$I_n = \big\{(b_{n-1}\cup f) : \forall b_{n-1}\in B_{n-1} \text{ and } \forall f\in F \text{ and } |b_{n-1}\cup f| = n\big\}\,. \tag{44}$$

The items that obtain the lowest cross-validated MAE are placed into B_n, which is used to determine the itemset I_{n+1} of the following iteration.

After all $|F|$ iterations of the sequential search, the three feature sets with the lowest cross-validated MAE on the EAGLE galaxies of all itemsets $B_1 \cup B_2 \cup B_3 \cup \ldots \cup B_{|F|}$ are returned. The pseudo code for this algorithm can be observed in Alg. (2).

Algorithm 2 Bottom-up sequential search

Define feature-space F
Define $B_0 = \{\}$
for $n = 1, 2, ..., |F|$ **do**
  $I_n = \left\{(b_{n-1} \cup f) : \forall b_{n-1} \in B_{n-1} \text{ and } \forall f \in F \text{ and } |b_{n-1} \cup f| = n\right\}$.
  for $i \in I_n$ **do**
    Cross-validate neural network with input features i
  $B_n = \{b_{n,1}, b_{n,2}, b_{n,3}\}$ with $b_{n,1}, b_{n,2}, b_{n,3} \in I_n$ for which the cross-validated MAE is lowest
Return the three feature sets $b \in (B_1 \cup B_2 \cup \ldots \cup B_{|F|})$ with the lowest cross-validated MAE

If we would perform an exhaustive search in which all feature subsets are explored, the cardinality of the union of all sets I_n explored would be $2^{|F|}$, and the time complexity of this algorithm would thus be of order $\mathcal{O}(2^{|F|})$. In the sequential search, roughly $\sum_{n=1}^{|F|} 3|F| = 3|F|^2$ feature subsets are explored, indicating that the time complexity of our algorithm is $\mathcal{O}(3|F|^2)$, which is significantly lower for $|F| = 15$. Therefore, our bottom-up sequential search reduces the search space significantly.

The bottom-up sequential search is performed with the three neural networks `nocolours`, `subsetcolours` and `allcolours`. The three best feature subsets for each network can be observed in Table 12. These feature subsets obtain the lowest cross-validated MAE on the EAGLE uniformly sampled training set of all feature subsets explored. For each neural network it is apparent that including all possible features of F is not preferred. Rather, a much lower number of features leads to better results, which coincides with the previously mentioned *curse of dimensionality*. Adding X extra dimensions to D-dimensional data increases the sparsity of the data over the entire $D + X$-dimensional volume. However, we must note that we have made the assumption that the most important parts of the feature search space have been explored with this sequential search.

From Table 12, it can be noticed that the features that are often preferred are z, $r - z$ and $r - i$. The z magnitude correlates strongest with stellar mass, therefore this observation is reasonable. This is not the case for $r - i$ and $r - z$ however, therefore the preference for these features cannot be explained by their correlations.

Comparing the features that occur often in the best features sets, found by the sequential search, with the SHAP values for the individual features (as observed in Figures 25e and 25f), we find that there is not much correspondence between the two methods. The z, $r - z$ and $r - i$ features have relatively low SHAP values while they are important features in the current analysis. This result might be explained by several factors. Either the SHAP values are less reliable due to correlations between features, or we have not explored the part of feature search space that obtains the lowest error with the bottom-up sequential search, or the best features found with the bottom-up sequential search could contain information of features with high SHAP values as all features that we are investigating are highly correlated. Another plausible explanation is the fact that we have only shown the three best feature sets, which obtain fairly similar results. The comparison between the two methods could be further extended by analysing how often features occur in the top-50 for each network, or by taking a weighted sum for each feature based on the error it obtains. Furthermore, comparing the results among the three networks, the best feature sets obtained by the sequential search seem to be different. However, as mentioned, only the top three feature sets are shown. Hence we do not know how high ranked each feature set is for other networks and no conclusion about the difference between the architectures can be drawn from this analysis.

The cross-validated error measures obtained by the best feature set of each network (#1 in Table 12) can be observed in Table 13. Comparing these results to the standard features of each network, as described in Section 2.3 (e.g. magnitudes for the `nocolours` neural network and all colour combinations plus magnitudes for the `allcolours` neural network) of which the cross-validation results are analysed in Section 5.2 and Table 7, it is apparent that for all three networks lower error measures are obtained with the input features of the sequential search. For the `nocolours` neural network, this could be expected since the six features are considered best by the sequential search, while only five (*ugriz* magnitudes) are default. For the `allcolours` neural network, the best option according to the sequential search consists of three features, while the default is fifteen features. Even for three features, better results are obtained than for fifteen features, clearly a result of the curse of dimensionality, whereby adding features increase the sparsity of the data over the feature volume.

				`nocolours` neural network				
1)	z	i-z	r-i	u-g	u-z	r-z	g-z	
2)	z	i-z	r-i	u-g	u-z	r-z	g-i	
3)	z	i-z	r-i	u-g	u-z	r-z	g-r	u-r
				`subsetcolours` neural network				
1)	i	r-z	z	u	r-i			
2)	i	i-z	u	g-z	r-i			
3)	i	r-z	z	u	i-z	r-i		
				`allcolours` neural network				
1)	z	r-z	u-r					
2)	z	r-z	u-z					
3)	z	r-z	u-i					

Table 12: Results of the bottom-up sequential search for the three neural networks `nocolours`, `subsetcolours` and `allcolours`. For each neural network, the three feature subsets with the lowest cross-validated MAE on the EAGLE training set are shown. The features of each subset are shown in the order that they were added to the subset.

	`nocolours`	`subsetcolours`	`allcolours`
MAE	$3.52{\cdot}10^{-2}$	$3.50{\cdot}10^{-2}$	$3.50{\cdot}10^{-2}$
R^2	0.994	0.994	0.994
$\bar{R}^2$	0.993	0.993	0.994
μ	$-2.95{\cdot}10^{-3}$	$-1.57{\cdot}10^{-3}$	$7.28{\cdot}10^{-3}$
σ^2	$1.88{\cdot}10^{-3}$	$1.96{\cdot}10^{-3}$	$1.94{\cdot}10^{-3}$

Table 13: Cross-validated error measures for the three `nocolours`, `subsetcolours` and `allcolours` neural networks implemented with the best feature sets obtained from the bottom-up sequential search.

8 Conclusion

The aim of this thesis is to find a relation between the photometry and star formation history, specifically stellar mass, of galaxies. We use the supervised machine learning method *neural networks* to find this relation on galaxies of the EAGLE `RefL0100N1504` simulation at $z \sim 0.1$ and their modelled spectra integrated over the SDSS *ugriz* bands. The SDSS Chang et al. (2015) catalogue is used to investigate whether this relation can be applied to observed galaxies.

Three neural networks are used in this thesis, the `nocolours` neural network that uses only *ugriz* magnitudes as input, the `subsetcolours` neural network that uses *ugriz* magnitudes plus $u-g$, $g-r$, $r-i$ and $i-z$ colours as input and the `allcolours` neural network that uses *ugriz* magnitudes plus all colour combinations of these filters as input. The architectures of the neural networks are optimised with 5-fold cross-validation on the Tree-structured Parzen Estimators (TPE) algorithm and the three neural networks `nocolours`, `subsetcolours` and `allcolours` obtain a mean absolute error (MAE) of $2.70{\cdot}10^{-2}$, $1.74{\cdot}10^{-2}$ and $1.75{\cdot}10^{-2}$ and an R^2 of 0.990, 0.996 and 0.996 respectively on the EAGLE test set.

To compare the performance of the neural network on the SDSS data, the EAGLE and SDSS galaxies are enforced to follow the same uniform stellar mass distribution after which the network is trained and tested on these galaxies. The performance of the three networks is evaluated with an R^2 of 0.987, 0.992 and 0.991 on the EAGLE galaxies and an R^2 of 0.888, 0.933 and 0.942 on the SDSS galaxies, respectively for the three neural networks, where an R^2 of one is ideal. It is apparent that the SDSS galaxies are continuously predicted to have a stellar mass that is too high. The subsequent experiments are designed to research this discrepancy such as adding noise to the EAGLE data and using a different SDSS stellar mass catalogue (that of Brinchmann et al., 2004). Nonetheless, all measures taken to reduce this discrepancy lead to higher errors of the network evaluated on EAGLE and thus also higher errors on SDSS.

Subsequently, the performance of the network is evaluated for centrals and satellites separately to research if lower errors on both the EAGLE and SDSS data set can be obtained if the network is exclusively tested on centrals or satellites. For the EAGLE data set a small preference towards centrals is observed and the results suggest that the photometry-stellar mass relation can be slightly better retrieved, but this difference is too minimal to be decisive. For the SDSS galaxies, no clear preference towards training on EAGLE centrals or satellites is observed.

Finally, the importance of individual features is determined using the SHAP framework. The features that contribute most to the prediction of the stellar mass according to SHAP are g, i, $u-z$, r and $g-i$. Training the `nocolours` neural network with these features on EAGLE galaxies following a uniform stellar mass distribution and comparing its performance to the `nocolours` neural network with *ugriz* magnitudes, lower errors for both EAGLE and SDSS are obtained. Furthermore, differences in the SHAP values of EAGLE and SDSS features are observed, possibly indicating which features cause the discrepancy in the prediction of stellar mass observed.

As a comparison, a bottom-up sequential search is performed to obtain the features for each network that give the lowest cross-validated mean absolute error on the EAGLE trained set. The features returned are not necessarily the features that contribute most according to SHAP. However, the cross-validated errors obtained with the features returned by the sequential search are lower than those obtained with the default features of each neural network, plausibly a result of the curse of dimensionality.

9 Discussion

An issue that arises in this thesis is the existing discrepancy in the evaluation of the three neural networks between EAGLE and SDSS galaxies. Several measures have been taken into account to ensure that the same assumptions are used for both data sets. Both data sets assume the same Chabrier (2003) initial mass function and in the calculation of the luminosity distance, the different cosmologies of EAGLE and SDSS have been taken into account. Additionally, on both data sets the same uniform stellar mass distribution has been enforced to ensure that no bias in the prediction error towards a certain stellar mass exists. Because SDSS consists of measurements that could be considered as values randomly sampled from a Gaussian distribution around their true value with standard deviation equal to the measurement error, the same method has been applied to the EAGLE galaxies to obtain "measured" EAGLE observables. The measurement errors are obtained from the SDSS catalogue by taking averages per observable bin. For the stellar mass noise, a bin size of 0.01 dex is used to ensure that no additional bias is incorporated in the data. Furthermore, the network is evaluated with a different SDSS catalogue (Brinchmann et al. (2004) instead of Chang et al. (2015)) to research if there might be any bias associated to the stellar mass modelling method. Moreover, the network is trained instead of tested on SDSS galaxies to analyse if the discrepancy might be caused by the incapability of the neural network itself to reproduce the photometry - stellar mass relation of these galaxies. This is not the case as the network is able to reproduce the photometry - stellar mass relation with approximately the same accuracy for the SDSS galaxies as for EAGLE .

Using all of the above methods, the discrepancy in the photometry-stellar mass relation of EAGLE and SDSS galaxies is not resolved. Even though noise is added to the EAGLE data and a uniform stellar mass distribution is enforced on both data sets, the photometry distributions of both data sets are still intrinsically different. Additionally, the neural network is able to accurately reproduce the photometry-stellar mass relation for SDSS, suggesting that a possible higher complexity of observational data either does not exist or is not a limiting factor. It is therefore plausible that the difference in performance of the network is caused by differences in the data; a possible indication that the neural network is *overfitting* on the same data set type that it is training upon. Please note the usage of the word *type*, as it is clear from the figures in Appendix D, showing the evolution of the error measures, that the network is not overfitting on the training set with respect to the test set, as the error measures of the EAGLE training and test set are similar. Rather it is overfitting – learning the intrinsic details of the data – on the complete EAGLE data set (or the complete SDSS data set when the network is trained on SDSS). Because the network conceivably learns these intrinsic properties of the data, the performance on a data set that is gathered through different means, and thus may have different intrinsic properties, will be lower. Several improvements that might reduce overfitting are for example adding regularisation or dropout to the neural network, and determining the best hyperparameters based on the performance on the SDSS data set rather than EAGLE , or using different modelling methods that have lower ability of learning complex details of the data. For these methods, a probable consequence is however that the performance on the EAGLE data will decrease. Therefore, we advise to investigate the origin of the differences observed between the two data sets themselves, rather than to apply a different modelling method.

The method chosen in this work - *neural networks* - is a method that is very useful in finding complex relations in data and it can approximate almost any kind of function. One caveat of this method is however that it behaves as a 'black box'. It does not provide much insight

into the relation that is found, as there are $m \times n$ free parameters (weights) for each pair of m and n dimensional subsequent layers present that define it and these weights do not translate directly to a specific input. In this work we have implemented various workarounds for this uninterpretability, but in future works it would be worth exploring other, more interpretable machine learning methods, such as random forests. Random forests is a method that uses an ensemble of decision trees and because of this structure can easily return the importance of individual input features. It would be interesting to research whether these methods can achieve similar performance as neural networks and at the same time provide more insight into the relation found.

The method that is used can be improved in a few ways. The evaluation of the network can be improved by using model quality estimators such as the log likelihood, Akaike's Information Criterion and the Minimum Description Length. These estimators are valuable in comparing the performance of across different statistical models that represent the original data. A few improvements in the optimisation of the network are optimising the learning rate of the network as it is an important parameter in the learning process and optimising the hyperparameters of the neural network on a uniform stellar mass distribution instead of the unsampled stellar mass distribution.

The research on this topic could be further extended by using other parts of the spectrum as well, as is done by for example Lovell et al. (2019). Using more information of the spectrum should theoretically give us more information on the star formation history of a galaxy.

Furthermore, it would be interesting to research how the photometry-mass relation found with neural networks translates over different redshifts; in this way a type of star formation history can be obtained. A simple option is training the network the same way as done in this work, but for different discrete redshift samples. One more complex option is to use redshift as an additional input and train the network on a galaxy sample with a range of redshifts.

Compared to other methods that are used to obtain the star formation history of galaxies, such as observational tracers and SED fitting, the relation between photometry and star formation history or stellar mass is learned from the data itself and relies on relatively few assumptions. Furthermore, this method uses as an advantage that once the relation between photometry and stellar mass is found, it is computationally inexpensive to apply it to newly observed galaxies.

It is difficult to quantitatively compare this work with other works that use machine learning to learn the relationship between photometry and star formation history, such as Stensbo-smidt et al. (2017), Lovell et al. (2019), Bonjean et al. (2019) and Delli Veneri et al. (2019), because these works use different catalogues, different redshift ranges, different input features and different error measures. Our work does not explore the prediction of stellar masses and star formation rates across different redshifts, but besides the work of Lovell et al. (2019), however, no other aforementioned work uses simulations to research the photometry-SFR relation. This work provides a path into using machine learning to find the relation between the spectrum of galaxies and star formation history from simulations and might possibly be used for the relation between other galaxy properties in simulations as well.

10 Acknowledgements

I would like to thank my supervisors, Camila Correa, Joop Schaye and James Trayford for their patient guidance, encouragement and advice throughout this research project. Even though the methods used in this project were novel, they showed great interest and understanding and we were able to have interesting discussions, which encouraged me to pursue this project. Specifically I would like to thank Camila for her guidance in this project. She took great effort to understand machine learning, was always available for help when I needed it and her patient explanations were incredibly helpful in understanding the project. Her positive outlook and encouragement inspired me in this project and I could not wish for a better supervisor.

I would like to thank my friend Esmee Stoop, for her tips and tricks about neural networks. As a master student graduating a year before me, she shared her experience on neural networks within astronomy, which helped me understand the methods that I used in this work.

I would like to thank the EAGLE group at Leiden Observatory for their tips and feedback during group meetings and for their interest in my research project.

This work has made use of the EAGLE simulations. I thank the EAGLE team for their outstanding work on the EAGLE simulations, without which this research would not have been possible. Additionally I would like to thank Chang et al. (2015) and Brinchmann et al. (2004) for making their catalogues publicly available. Furthermore, this work uses many open-source software packages that can be found in Appendix B.

Bibliography

Abazajian, K. N., J. K. Adelman-McCarthy, M. A. Agüeros, and others
2009. The Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 182(2):543–558.

Abramson, L. E., M. D. Gladders, A. Dressler, A. Oemler, B. Poggianti, and B. Vulcani
2016. Return to [Log-]Normalcy: Rethinking Quenching, The Star Formation Main Sequence, and Perhaps Much More. *The Astrophysical Journal*, 832(1):26.

Agarwal, S., R. Davé, and B. A. Bassett
2018. Painting galaxies into dark matter halos using machine learning. *Monthly Notices of the Royal Astronomical Society*, 478:3410–3422.

Aghanim, N., G. Hurier, J. Diego, M. Douspis, and E. Pointecouteau
2015. The Good, the Bad, and the Ugly: Statistical quality assessment of SZ detections. *Astronomy & Astrophysics*, 580:A138.

Baes, M., J. I. Davies, H. Dejonghe, S. Sabatini, S. Roberts, R. Evans, S. M. Linder, R. M. Smith, and W. J. De Blok
2003. Radiative transfer in disc galaxies - III. The observed kinematics of dusty disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 343(4):1081–1094.

Baes, M., J. Verstappen, I. De Looze, J. Fritz, W. Saftly, E. Vidal Pérez, M. Stalevski, and S. Valcke
2011. Efficient three-dimensional NLTE dust radiative transfer with skirt. *Astrophysical Journal, Supplement Series*, 196(2).

Bellman, R.
1957. *Dynamic Programming*, 1 edition. Princeton, NJ: Princeton University Press.

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl
2011. Algorithms for Hyper-Parameter Optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, Pp. 2546–2554.

Bergstra, J. and Y. Bengio
2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305.

Bilicki, M., T. H. Jarrett, J. A. Peacock, M. E. Cluver, and L. Steward
2014. Two Micron All Sky Survey Photometric Redshift Catalog: A Comprehensive Three-dimensional Census of the Whole Sky. *The Astrophysical Journal Supplement Series*, 210:9.

Bilicki, M., J. A. Peacock, T. H. Jarrett, M. E. Cluver, N. Maddox, M. J. I. Brown, E. N. Taylor, N. C. Hambly, A. Solarz, and B. W. Holwerda
2016. WISE ÃŮ SuperCOSMOS Photometric Redshift Catalog: 20 Million Galaxies over 3\pi Steradians. *The Astrophysical Journal Supplement Series*, 225:5.

Birnboim, Y. and A. Dekel
2011. Gravitational quenching by clumpy accretion in cool-core clusters : convective dynamical response to overheating. *Monthly Notices of the Royal Astronomical Society*, 415:2566–2579.

Bonjean, V., N. Aghanim, P. Salomé, A. Beelen, M. Douspis, and E. Soubrié
2019. Star Formation Rates and Stellar Masses from Machine Learning. *Astronomy & Astrophysics*, 622:A137.

Brinchmann, J., S. Charlot, S. D. M. White, C. Tremonti, G. Kauffmann, T. Heckman, and J. Brinkmann
2004. The physical properties of star-forming galaxies in the low-redshift Universe. *Monthly Notices of the Royal Astronomical Society*, 351:1151–1179.

Bruzual, G. and S. Charlot
2003. Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344:1000–1028.

Bryson, A. E.
1961. A Gradient Method for Optimizing Multi-Stage Allocation Processes. In *Proceedings of the Harvard University Symposium on Digital Computers and their Applications*.

Calzetti, D. and A. L. Kinney
1994. Dust Extinction of the Stellar Continua in Starburst Galaxies: The Ultraviolet and Optical Extinction Law. *The Astrophysical Journal*, 429:582–601.

Campbell, M., A. J. Hoane, and F.-h. Hsu
2002. Deep Blue. *Artificial Intelligence*, 134:57–83.

Camps, P. and M. Baes
2015. SKIRT: An advanced dust radiative transfer code with a user-friendly architecture. *Astronomy and Computing*, 9:20–33.

Carnall, A. C., J. Leja, B. D. Johnson, R. J. Mclure, J. S. Dunlop, and C. Conroy
2019. How to Measure Galaxy Star Formation Histories. I. Parametric Models. *The Astrophysical Journal*, 873(1):44.

Carrasco Kind, M. and R. J. Brunner
2013. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, 432:1483–1501.

Chabrier, G.
2003. Galactic Stellar and Substellar Initial Mass Function. *Publications of the Astronomical Society of the Pacific*, 115:763–795.

Chang, Y. Y., A. V. D. Wel, E. D. Cunha, and H. W. Rix
2015. Stellar Masses and Star Formation Rates for 1M Galaxies from SDSS+WISE. *Astrophysical Journal, Supplement Series*, 219(1):8.

Charlot, S., G. Kauffmann, M. Longhetti, L. Tresse, S. D. M. White, A. D. Paris, and B. Arago
2002. Star formation , metallicity and dust properties derived from the SAPM galaxy survey spectra. *Monthly Notices of the Royal Astronomical Society*, 330(4):876–888.

Cimatti, A., E. Daddi, and A. Renzini
2006. Mass downsizing and âĂIJ top-down âĂİ assembly of early-type galaxies. *Astronomy & Astrophysics*, 453:29–33.

Collister, A. A. and O. Lahav
2004. ANNz : Estimating Photometric Redshifts Using Artificial Neural Networks. *Publications of the Astronomical Society of the Pacific*, 116:345–351.

Conroy, C.
2013. Modeling the Panchromatic Spectral Energy Distributions of Galaxies. *Annual Review of Astronomy and Astrophysics*, 51:393–455.

Cowie, L. L., A. Songaila, and E. M. Hu
1996. New Insight on Galaxy Formation and Evolution from Keck Spectroscopy. *The Astronomical Journal*, 112(3):839–864.

Da Cunha, E., S. Charlot, and D. Elbaz
2008. A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. *Monthly Notices of the Royal Astronomical Society*, 388(4):1595–1617.

Dahl, G. E., T. N. Sainath, and G. E. Hinton
2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Proceedings of the 30th International Conference on Machine Learning*.

Davé, R., R. J. Thompson, and P. F. Hopkins
2016. MUFASA: Galaxy Formation Simulations With Meshless Hydrodynamics. *Monthly Notices of the Royal Astronomical Society*, 462(3):3265–3284.

Davies, L. J. M., S. P. Driver, A. S. G. Robotham, M. W. Grootes, C. C. Popescu, R. J. Tuffs, A. Hopkins, M. Alpaslan, S. K. Andrews, M. N. Bremer, S. Brough, M. J. I. Brown, M. E. Cluver, S. Croom, E. Cunha, L. Dunne, J. Loveday, A. J. Moffett, M. Owers, S. Phillipps, A. E. Sansom, E. N. Taylor, M. J. Michalowski, E. Ibar, M. Smith, and N. Bourne
2016. GAMA / H-ATLAS : A meta-analysis of SFR indicators - comprehensive measures of the SFR-M* relation and Cosmic Star Formation History at z $<$ 0 . 4. *Monthly Notices of the Royal Astronomical Society*, 461(1):458–485.

Davis, M., G. Efstathiou, C. S. Frenk, and S. D. M. White
1985. The Evolution of Large-Scale Structure in a Universe Dominated by Cold Dark Matter. *The Astrophysical Journal*, 292:371–394.

Dekel, A. and J. Silk
1986. The origin of dwarf galaxies, cold dark matter, and biased galaxy formation. *The Astrophysical Journal*, 303:39–55.

Delli Veneri, M., S. Cavuoti, M. Brescia, G. Longo, and G. Riccio
2019. Star formation rates for photometric samples of galaxies using machine learning methods. *Monthly Notices of the Royal Astronomical Society*, 386:1377–1391.

Deng, J., W. Dong, R. Socher, L.-j. Li, K. Li, and L. Fei-fei
2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition*, Pp. 248–255.

Dolag, K., S. Borgani, G. Murante, and V. Springel
2009. Substructures in hydrodynamical cluster simulations. *Monthly Notices of the Royal Astronomical Society*, 399:497–514.

Domíngues Sánchez, H., M. Huertas-Company, M. Bernardi, D. Tuccillo, and J. L. Fisher
2018. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676.

Dozat, T.
2016. Incorporating Nesterov Momentum into Adam. In *Workshop of the Fourth International Conference on Learning Representations*.

Duchi, J., E. Hazan, and Y. Singer
2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Fabian, A. C.
2012. Observational Evidence of AGN Feedback. *Annual Review of Astronomy and Astrophysics*, 50:455–489.

Fontanot, F., G. D. Lucia, R. S. Somerville, and P. Santini
2009. The Many Manifestations of Downsizing : Hierarchical Galaxy Formation Models confront Observations. *Monthly Notices of the Royal Astronomical Society*, 397(4):1776–1790.

Frazier, P. I.
2018. A Tutorial on Bayesian Optimization. *Recent Advances in Optimization and Modeling of Contemporary Problems*, Pp. 255–278.

Furlong, M., R. G. Bower, T. Theuns, J. Schaye, R. A. Crain, M. Schaller, C. D. Vecchia, C. S. Frenk, I. G. Mccarthy, J. Helly, and A. Jenkins
2015. Evolution of galaxy stellar masses and star formation rates in the Eagle simulations. *Monthly Notices of the Royal Astronomical Society*, 450(4):4486–4504.

Geach, J. E.
2012. Unsupervised self-organized mapping : a versatile empirical tool for object selection , classification and redshift estimation in large surveys. *Monthly Notices of the Royal Astronomical Society*, 419:2633–2645.

Gladders, M. D., A. Oemler, A. Dressler, B. Poggianti, B. Vulcani, and L. Abramson
2013. The imacs cluster building survey. IV. the log-normal star formation history of galaxies. *The Astrophysical Journal*, 770(1):64.

Glorot, X. and Y. Bengio
2010. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference On Artificial Intelligence and Statistics*, 9:249–256.

Glorot, X., A. Bordes, and Y. Bengio
2011. Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 15:315–323.

Groves, B., M. A. Dopita, R. S. Sutherland, L. J. Kewley, J. Fischera, C. Leitherer, B. Brandl, and W. van Breugel
2008. Modeling the PanâĂŘSpectral Energy Distribution of Starburst Galaxies. IV. The Controlling Parameters of the Starburst SED. *The Astrophysical Journal Supplement Series*, 176(2):438–456.

Hahnloser, R. H. R., R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung
2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.

Heavens, A., B. Panter, R. Jimenez, J. Dunlop, B. Hill, and E. Eh-hj
2004. The Complete Star Formation History of the Universe from stellar populations of nearby galaxies. *Nature*, 428(6983):625–627.

Henri, T.
1961. *Economic forecast and policy*. North-Holland Pub. Co.

Hernquist, L. and V. Springel
2003. An analytical model for the history of cosmic star formation. *Monthly Notices of the Royal Astronomical Society*, 341(4):1253–1267.

Hinton, G., N. Srivastava, and K. Swersky
2012. Neural Networks for Machine Learning - Lecture 6a: Overview of mini-batch gradient descent. Technical report, COURSERA.

Hopkins, A. M.
2018. On the evolution of star forming galaxies. *The Astrophysical Journal*, 615(1):209–221.

Huertas-Company, M., R. Gravet, J. S. Kartaltepe, G. Barro, M. Bernardi, S. Mei, F. Shankar, P. Dimauro, E. F. Bell, D. Kocevski, D. C. Koo, S. M. Faber, and D. H. Mcintosh
2015. A Catalog of Visual-Like Morphologies in the 5 CANDELS Fields using Deep Learning. *The Astrophysical Journal Supplement Series*, 221(1):8.

James, G., D. Witten, T. Hastie, and R. Tibshirani
2013. *An Introduction to Statistical Learning*, 7 edition. Springer.

Johansson, P. H., T. Naab, and J. P. Ostriker
2009. Gravitational Heating Helps Make Massive Galaxies Red and Dead. *The Astrophysical Journal*, 697:38–43.

Kamdar, M., M. J. Turk, and R. J. Brunner
2016. Machine Learning and Cosmological Simulations II : Hydrodynamical Simulations. *Monthly Notices of the Royal Astronomical Society*, 457(2):1162–1179.

Kauffmann, G., T. M. Heckman, S. D. M. White, S. Charlot, C. Tremonti, J. Brinchmann, G. Bruzual, E. W. Peng, M. Seibert, M. Bernardi, M. Blanton, J. Brinkmann, F. Castander, I. Csábai, M. Fukugita, Z. Ivezic, J. A. Munn, R. C. Nichol, N. Padmanabhan, A. R. Thakar, D. H. Weinberg, and D. York
2003. Stellar masses and star formation histories for 10ˆ5 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 341:33–53.

Kelley, H. J.
1960. Gradient Theory of Optimal Flight Paths. *ARS Journal*, 30(10):947–954.

Kennicutt, R. C.
1998. Star Formation in Galaxies along the Hubble Sequence. *Annual Review of Astronomy and Astrophysics*, 36:189–231.

Kennicutt, R. C. and N. J. Evans
2012. Star Formation in the Milky Way and Nearby Galaxies. *Annual Review of Astronomy and Astrophysics*, 50:531–608.

Khochfar, S. and J. P. Ostriker
2008. Adding Environmental Gas Physics to the Semianalytic Method for Galaxy Formation: Gravitational Heating. *The Astrophysical Journal*, 680:54–69.

Kingma, D. P. and J. Lei Ba
2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, Pp. 1–15.

Krakowski, T., K. Małek, M. Bilicki, A. Pollo, A. Kurcz, and M. Krupa
2016. Astrophysics Machine-learning identification of galaxies in the WISE ÃŮ SuperCOSMOS all-sky catalogue. *Astronomy & Astrophysics*, 596:A39.

Kroupa, P.
2001. On the Origin of the Initial Mass Function. *Monthly Notices of the Royal Astronomical Society*, 322:231–246.

Larson, R. B. and B. M. Tinsley
1978. Star Formation Rates in Normal and Peculiar Galaxies. *The Astrophysical Journal*, 219:46–59.

LeCun, Y.
1985. Une proc\'edure d'apprentissage pour r\'eseau a seuil asymmetrique (A Learning Scheme for Asymmetric Threshold Networks). *Proceedings of Cognitiva*, 85:599–604.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner
1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y., L. Bottou, G. B. Orr, and K.-R. M\"uller
1998. Efficient BackProp. In *Neural Networks Tricks of the Trade*. Springer.

Lee, S.-k., R. Idzi, H. C. Ferguson, R. S. Somerville, T. Wiklind, and M. Giavalisco
2009. Biases and Uncertainties in Physical Parameter Estimates of Lyman Break Galaxies from Broadband Photometry. *The Astrophysical Journal*, 184:100–132.

Leja, J., A. C. Carnall, B. D. Johnson, C. Conroy, and J. S. Speagle
2019. How to Measure Galaxy Star Formation Histories. II. Nonparametric Models. *The Astrophysical Journal*, 876(1):3.

Leray, P. and P. Gallinari
1998. Feature Selection with Neural Networks. *Behaviormetrika*, 26.

Li, C. and S. D. M. White
2009. The distribution of stellar mass in the low-redshift Universe. *Monthly Notices of the Royal Astronomical Society*, 398:2177–2187.

Lilly, S. J., O. Le Fèvre, F. Hammer, and D. Crampton
1996. The Canada-France Redshift Survey: The Luminosity Density and Star Formation History of the Univerze to z˜1. *The Astrophysical Journal*, 460:L1–L4.

Linnainmaa, S.
1970. *The Representation of the Cumulative Rounding Error of an Algorithm as a Tayler Expansion of the Local Rounding Errors*. PhD thesis, University of Helsinki.

Lovell, C. C., V. Acquaviva, P. A. Thomas, K. G. Iyer, E. Gawiser, and S. M. Wilkins
2019. Learning the Relationship between Galaxies Spectra and their Star Formation Histories using Convolutional Neural Networks and Cosmological Simulations. *arXiv e-prints*.

Lucie-Smith, L., H. V. Peiris, A. Pontzen, and M. Lochner
2018. Machine learning cosmological structure formation. *Monthly Notices of the Royal Astronomical Society*, 479(3):3405–3414.

Lundberg, S. M. and S.-I. Lee
2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Pp. 4765–4774. Curran Associates, Inc.

Madau, P. and M. Dickinson
2014. Cosmic Star Formation History. *Annual Review of Astronomy and Astrophysics*, 52:415–486.

Madau, P., L. Pozzetti, and M. Dickinson
1998. The star formation history of field galaxies. *The Astrophysical Journal*, 498:106–116.

Marchetti, T., E. M. Rossi, G. Kordopatis, A. G. A. Brown, A. Rimoldi, K. Youakim, and R. Ashley
2017. An artificial neural network to discover Hypervelocity stars : Candidates in Gaia DR1/TGAS. *Monthly Notices of the Royal Astronomical Society*, 470(2):1388–1403.

McCarthy, I. G., J. Schaye, S. Bird, and A. M. C. Le Brun
2016. The BAHAMAS project: calibrated hydrodynamical simulations for large-scale structure cosmology. *Monthly Notices of the Royal Astronomical Society*, 465(3):2936–2965.

McCulloch, W. S. and W. H. Pitts
1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McNamara, B. R. and P. E. J. Nulsen
2007. Heating Hot Atmospheres with Active Galactic Nuclei. *Annual Review of Astronomy and Astrophysics*, 45:117–176.

Minsky, M. and S. Papert
1969. *Perceptrons.* Oxford, England: M.I.T. Press.

Moore, B., N. Katz, G. Lake, A. Dressler, and A. Oemler
1995. Galaxy Harassment and the Evolution of Clusters of Galaxies. *Nature*, 379(6566):613–616.

Mosteller, F. and J. W. Tukey
1968. Data analysis, including statistics. In *Handbook of Social Psychology, Vol. 2*, G. Lindzey and E. Aronson, eds. Addison-Wesley.

Muzzin, A., D. Marchesini, M. Stefanon, M. Franx, H. J. McCracken, B. Milvang-Jensen, J. S. Dunlop, J. P. U. Fynbo, G. Brammer, I. Labbé, and P. G. van Dokkum
2013. The Evolution of the Stellar Mass Functions of Star-Forming and Quiescent Galaxies to z = 4 from the COSMOS/UltraVISTA Survey. *The Astrophysical Journal*, 777(1):18.

Noeske, K. G., B. J. Weiner, S. M. Faber, C. Papovich, D. C. Koo, R. S. Somerville, K. Bundy, C. J. Conselice, J. A. Newman, D. Schiminovich, E. L. Floc, A. L. Coil, G. H. Rieke, J. M. Lotz, J. R. Primack, P. Barmby, M. C. Cooper, M. Davis, R. S. Ellis, G. G. Fazio, P. Guhathakurta, and J. Huang
2007. Star Formation in AEGIS Field Galaxies since z = 1.1: The Dominance of Gradually Declining Star Formation, and the Main Sequence of Star-Forming Galaxies. *The Astrophysical Journal*, 660:43–46.

Papovich, C., M. Dickinson, and H. C. Ferguson
2001. The Stellar Populations and Evolution of Lyman Break Galaxies. *The Astrophysical Journal*, 559:620–653.

Parker, D. B.
1985. Learning-logic. Technical report, Sloan School of Management, MIT, Cambridge, MA.

Pashchenko, I. N., K. V. Sokolovsky, and P. Gavras
2018. Machine learning search for variable stars. *Monthly Notices of the Royal Astronomical Society*, 475(2):2326–2343.

Pearson, W. J., L. Wang, P. D. Hurley, K. Małek, V. Buat, D. Burgarella, D. Farrah, S. J. Oliver, D. J. B. Smith, and F. F. S. V. D. Tak
2018. Main sequence of star forming galaxies beyond the Herschel confusion limit. *Astronomy & Astrophysics*, 615:A146.

Peng, Y., R. Maiolino, and R. Cochrane
2015. Strangulation as the primary mechanism for shutting down star formation in galaxies. *Nature*, 521(7551):192–195.

Planck Collaboration
2014. Planck 2013 results. XVI. Cosmological parameters. *Astronomy & Astrophysics*, 571:A16.

Rafelski, M., J. P. Gardner, M. Fumagalli, M. Neeleman, H. I. Teplitz, N. Grogin, A. M. Koekemoer, and C. Scarlata
2016. The Star Formation Rate Efficiency of Neutral Atomic-Dominated Hydrogen Gas in the Outskirts of Star-Forming Galaxies from z ˜ 1 to z ˜ 3. *The Astrophysical Journal*, 825:87–108.

Reddi, S. J., S. Kale, and S. Kumar
2018. On the convergence of Adam and Beyond. In *International Conference on Learning Representations*, Pp. 1–23.

Rosenblatt, F.
1957. The Perceptron: A Perceiving and Recognizing Automaton. Technical report, Cornell Aeronautical Laboratory, Inc.

Rosenblatt, F.
1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386–408.

Rumelhart, D. E. and J. L. McClelland
1986. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, chapter 8, Pp. 318–362. MIT Press.

Schaye, J., R. A. Crain, R. G. Bower, M. Furlong, M. Schaller, T. Theuns, C. Dalla Vecchia, C. S. Frenk, I. G. Mccarthy, J. C. Helly, A. Jenkins, Y. M. Rosas-Guevara, S. D. White, M. Baes, C. M. Booth, P. Camps, J. F. Navarro, Y. Qu, A. Rahmati, T. Sawala, P. A. Thomas, and J. Trayford
2015. The EAGLE project: Simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554.

Schaye, J., C. D. Vecchia, C. M. Booth, R. P. C. Wiersma, T. Theuns, M. R. Haas, S. Bertone, A. R. Duffy, I. G. Mccarthy, F. V. D. Voort, J. Bank, A. T. Building, and M. Manchester
2010. The physics driving the cosmic star formation history. *Monthly Notices of the Royal Astronomical Society*, 402:1536–1560.

Schiminovich, D., T. K. Wyder, D. C. Martin, B. D. Johnson, S. Salim, M. Seibert, C. Hoopes, M. Zamojski, T. A. Barlow, K. G. Forster, M. A. Treyer, P. G. Friedman, P. Morrissey, S. G. Neff, T. A. Small, L. Bianchi, T. M. Heckman, Y.-w. Lee, B. F. Madore, B. Milliard, R. M. Rich, A. S. Szalay, B. Y. Welsh, and S. Yi
2007. The UV-Optical Color Magnitude Diagram. II. Physical Properties and Morphological Evolution on and off of a Star-Forming Sequence. *The Astrophysical Journal Supplement Series*, 173:315–341.

Searle, L., W. L. W. Sargent, and W. G. Bagnuolo
1973. The History of Star Formation and the Colors of Late-Type Galaxies. *The Astrophysical Journal*, 179:427–438.

Shapley, L.
1953. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39:1095–1100.

Shrikumar, A., P. Greenside, and A. Kundaje
2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 70th International Conference on Machine Learning*, Pp. 3145–3153.

Silk, J. and G. A. Mamon
2012. The current status of galaxy formation. *Research in Astronomy and Astrophysics*, 12(8):917–946.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, and K. Kavukcuoglu
2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7585):484–489.

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis
2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv e-prints*, Pp. 1–19.

Siudek, M., A. Pollo, T. Krakowski, A. Iovino, M. Scodeggio, T. Moutard, G. Zamorani, L. Guzzo, B. Garilli, B. R. Granett, M. Bolzonella, U. Abbas, C. Adami, D. Bottini, A. Cappi, O. Cucciati, I. Davidzon, P. Franzetti, A. Fritz, J. Krywult, V. L. Brun, D. Maccagni, F. Marulli, M. Polletta, R. Tojeiro, D. Vergani, A. Zanichelli, S. Arnouts, J. Bel, E. Branchini, J. Coupon, G. D. Lucia, O. Ilbert, C. P. Haines, L. Moscardini, and T. T. Takeuchi
2018. The VIMOS Public Extragalactic Redshift Survey (VIPERS). *Astronomy & Astrophysics*, 617:A70.

Smith, D. J. B. and C. C. Hayward
2015. Deriving star formation histories from photometry using energy balance spectral energy distribution modelling. *Monthly Notices of the Royal Astronomical Society*, 11(July):1–11.

Somerville, R. S. and R. Davé
2015. Physical Models of Galaxy Formation in a Cosmological Framework. *Annual Review of Astronomy and Astrophysics*, 53:51–113.

Springel, V.
2005. The cosmological simulation code GADGET-2. *Monthly Notices of the Royal Astronomical Society*, 364:1105–1134.

Springel, V., P. S. D. M. White, G. Tormen, and G. Kauffmann
2001. Populating a cluster of galaxies. I. Results at z = 0. *Monthly Notices of the Royal Astronomical Society*, 328:726–750.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov
2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Stensbo-smidt, K., F. Gieseke, C. Igel, A. Zirm, and K. S. Pedersen
2017. Sacrificing Information for the Greater Good: How to Select Photometric Bands for Optimal Accuracy. *Monthly Notices of the Royal Astronomical Society*, 464(3):2577–2596.

Stone, M.
1974. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Taigman, Y., M. A. Ranzato, T. Aviv, and M. Park
2014. DeepFace : Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Pp. 1701–1708.

Tinsley, B. M.
1972. Galactic Evolution: Program and Initial Results. *Astronomy & Astrophysics*, 20:383–396.

Trayford, J. W., P. Camps, T. Theuns, M. Baes, R. G. Bower, R. A. Crain, M. L. P. Gunawardhana, M. Schaller, J. Schaye, and C. S. Frenk
2017. Optical colours and spectral indices of z = 0.1 EAGLE galaxies with the 3D dust radiative transfer code SKIRT. *Monthly Notices of the Royal Astronomical Society*, 470(1):771–799.

Trayford, J. W., T. Theuns, R. G. Bower, J. Schaye, M. Furlong, M. Schaller, C. S. Frenk, R. A. Crain, C. D. Vecchia, and I. G. Mccarthy
2015. Colours and luminosities of z = 0.1 galaxies in the EAGLE simulation. *Monthly Notices of the Royal Astronomical Society*, 452:2879–2896.

Van Rossum, G. and F. Drake Jr
1995. *Python tutorial*. Amsterdam, The Netherlands: Centrum voor Wiskunde en Informatica.

Viquar, M., S. Basak, A. Dasgupta, and S. Agrawal
2018. Machine Learning in Astronomy: A Case Study in Quasar-Star Classification. *arXiv e-prints*.

Vogelsberger, M., S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, S. Bird, D. Nelson, and L. Hernquist
2014. Properties of galaxies reproduced by a hydrodynamic simulation. *Nature*, 509:177.

Walcher, J., B. Groves, and T. Budavári
2011. Fitting the integrated spectral energy distributions of galaxies. *Astrophysics and Space Science*, 331:1–51.

Werbos, P.
1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University.

Wetzel, A. R., J. L. Tinker, and C. Conroy
2012. Galaxy evolution in groups and clusters : star formation rates, red sequence fractions, and the persistent bimodality. *Monthly Notices of the Royal Astronomical Society*, 424(1):232–243.

White, S. D. M. and C. S. Frenk
1991. Galaxy Formation Through Hierarchical Clustering. *The Astrophysical Journal*, 379:52–79.

White, S. D. M. and M. J. Rees
1978. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183:341–358.

Wiegerinck, W., A. Komoda, and T. Heskes
1994. Stochastic dynamics of learning with momentum in neural networks. *Journal of Physics A: Mathematical and General*, 27(13):4425–4437.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Å. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young,

J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean
2016. Google's Neural Machine Translation System : Bridging the Gap between Human and Machine Translation. *CoRR*, Pp. 1–23.

York, D. G., J. Adelman, J. E. Anderson, Jr., S. F. Anderson, and others
2000. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1587.

Zeiler, M. D.
2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv e-prints*.

A List of abbreviations

ANN	Artificial Neural Network
AGN	Active Galactic Nuclei
CDM	Cold Dark Matter
CV	Cross-validation
DR	Data Release
EAGLE	Evolution and Assembly of GaLaxies and their Environment
EI	Expected Improvement
FoF	Friends-of-Friends
GSMF	Galaxy Stellar Mass Function
IMF	Initial mass function
IR	Infrared
ISM	Interstellar Medium
KDE	Kernel Density Estimation
MAE	Mean absolute error
ML	Machine Learning
MLP	Multi-layer Perceptron
MSE	Mean squared error
NN	Neural Network
RF	Random Forests
RSS	Residual sum of squares
SDSS	Sloan Digital Sky Survey
SED	Spectral Energy Distribution
SFH	Star Formation History
SFR	Star Formation Rate
sgd	Stochastic gradient descent
SHAP	SHapley Additive exPlanations
SPH	Smoothed Particle Hydrodynamics
SSFR	Specific Star Formation Rate
SN	Supernova(e)
TPE	Tree-structured Parzen Estimators
TSS	Total sum of squares
WISE	Wide-Field Infrared Survey

B Benchmark

This section contains a list of the soft- and hardware used in this thesis to allow for an adequate replication of the results. The code used in this thesis can be found at https://github.com/evavanweenen/eagle-SFH.

B.1 Hardware

Throughout this thesis the following hardware is used:

CPU	2.40 GHz Intel ®Xeon ®CPU E5620
Cache	32K L1d, 32K L1i, 256K L2, 12288K L3
Memory	5.8 GB RAM, 10 GB swap
Disk	mounted 1.5TB HDD
OS	Fedora 29
Kernel	Linux 5.1.6
Architecture	x86-64

B.2 Software

The methodology in this thesis is completed with `Python`[4] `3.6.8` (Van Rossum and Drake Jr, 1995). The following modules are used:

`numpy`[5] `1.16.1`	reading and manipulating data (arrays)
`scipy`[6] `1.2.0`	solving integrals
`keras`[7] `2.2.4`	programming of the neural network wrapper around `tensorflow`[8] `1.5.0`
`hyperas`[9] `0.4`	hyperparameter optimisation of neural network in `keras` wrapper around `hyperopt`[10] `0.1.1`
`scikit-learn`[11] `0.20.2`	normalisation, evaluation measures, gridsearch and cross-validation
`shap`[12] `0.28.3`	calculating shapley values in feature importance
`astropy`[13] `3.1.1`	implementing cosmology
`h5py`[14] `2.9.0`	reading hdf5 files
`matplotlib`[15] `3.0.2`	figures

[4] https://www.python.org/
[5] https://www.numpy.org/
[6] https://www.scipy.org/
[7] https://keras.io/
[8] https://www.tensorflow.org/
[9] https://github.com/maxpumperla/hyperas
[10] https://github.com/hyperopt/hyperopt
[11] https://scikit-learn.org/
[12] https://github.com/slundberg/shap
[13] https://www.astropy.org/
[14] https://www.h5py.org/
[15] https://matplotlib.org/

C Database queries

This section describes the EAGLE database query used.

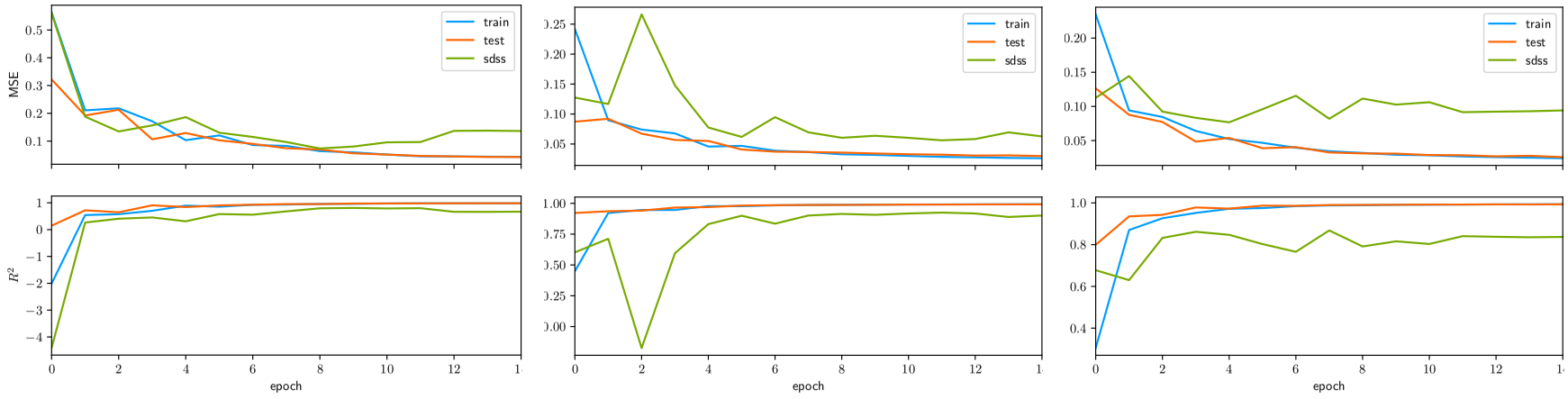
```
SELECT
    SH.TopLeafID as topleafid,
    SH.GalaxyID as galid,
    SH.Redshift as z,
    SH.SubGroupNumber as subgroup,
    AP.Mass_Star as m_star,
    DF.SDSS_u as dustyflux_sdss_u,
    DF.SDSS_g as dustyflux_sdss_g,
    DF.SDSS_r as dustyflux_sdss_r,
    DF.SDSS_i as dustyflux_sdss_i,
    DF.SDSS_z as dustyflux_sdss_z
FROM
    RefL0100N1504_SubHalo as SH,
    RefL0100N1504_Aperture as AP,
    RefL0100N1504_DustyFluxes as DF
WHERE
    -- Select aperture size to be 30 pkpc
    AP.ApertureSize = 30
    and SH.SnapNum = 27
    -- Join the objects in 3 tables
    and SH.GalaxyID = AP.GalaxyID
    and SH.GalaxyID = DF.GalaxyID
ORDER BY
    SH.TopLeafID,
    SH.GalaxyID asc,
    SH.Redshift asc
```

Table 14: SQL query to select the redshift, subgroupnumber, stellar mass at an aperture of 30 pkpc, and modelled SDSS dusty fluxes of galaxies at redshift $z \sim 0.1$ of the `RefL0100N1504` simulation of the EAGLE database

D Additional figures and tables

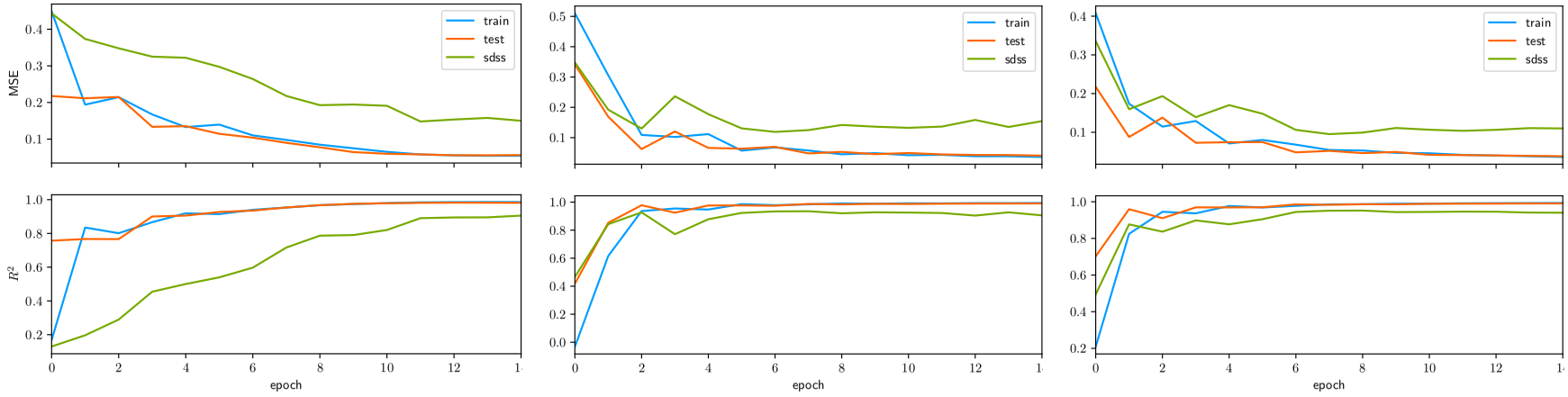
Optimiser	η	Additional parameters
sgd	0.01	$\mu = 0$
RMSprop	0.001	$\rho = 0.9$
Adagrad	0.01	
Adadelta	1.	$\rho = 0.95$
Adam	0.001	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Adamax	0.002	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Nadam	0.002	$\beta_1 = 0.9$, $\beta_2 = 0.999$

Table 15: Values of the optimiser parameters used in this thesis. The parameter η corresponds to the learning rate, μ to the the momentum, β_1 and β_2 to the decay and ρ to the fraction of gradient to keep. For all optimisers the fuzz factor preventing division by zero is $\epsilon = 10^{-7}$.



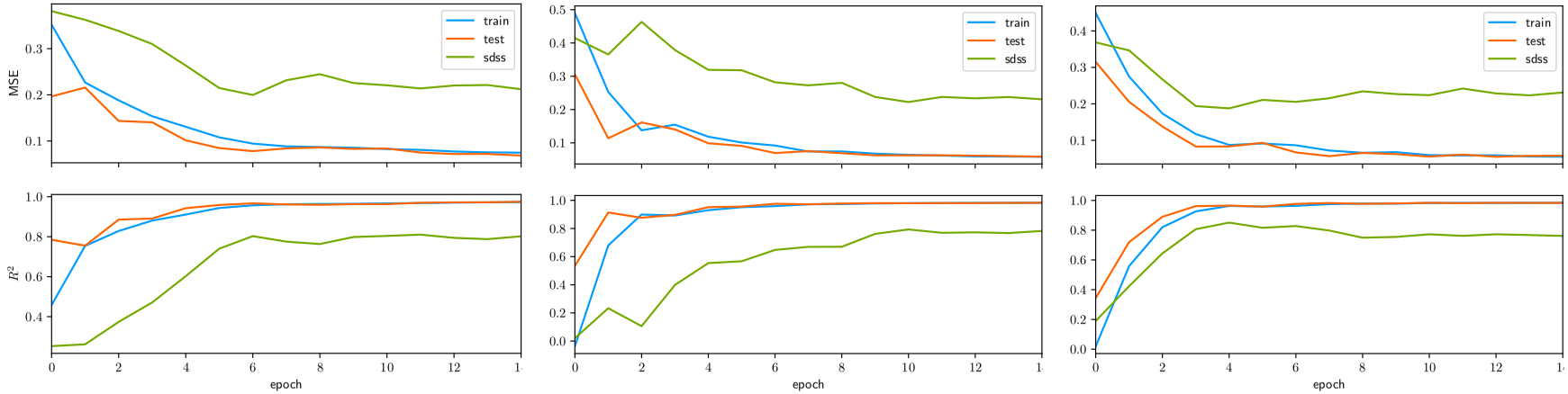
(a) `nocolours` (b) `subsetcolours` (c) `allcolours`

Figure 26: Evolution of the loss-function MAE and the monitored R^2 statistic during training on 500 **randomly sampled** EAGLE galaxies. The measures of the EAGLE training set, EAGLE test set and SDSS data set are indicated by the blue, orange and green line respectively. The EAGLE test set and SDSS data set are evaluated during each epoch of training with the current weights. The errors of the training and test set follow each other closely during the training process. The error of the SDSS data set however does not.

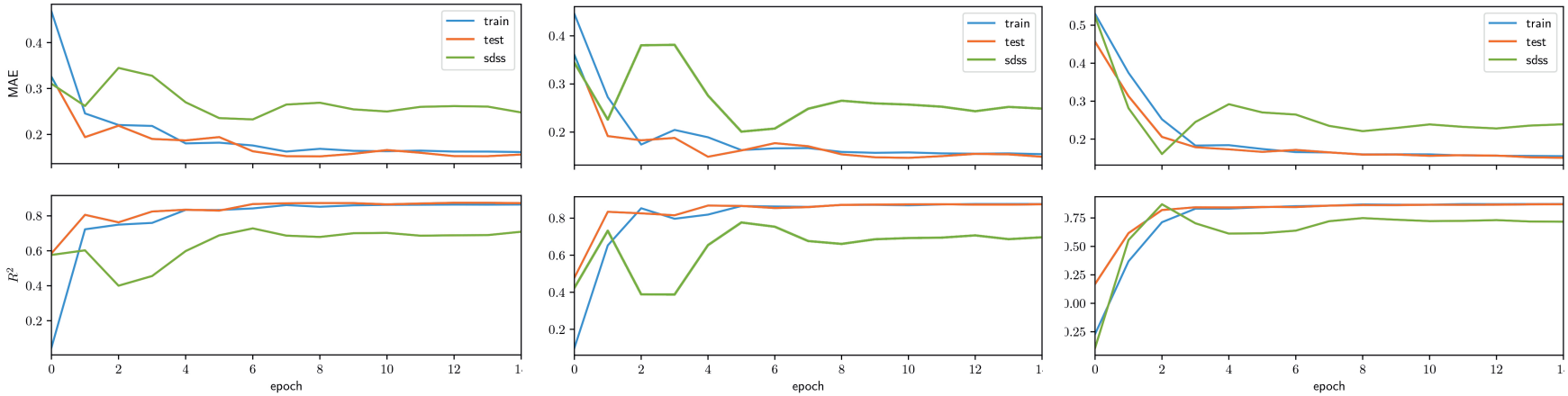


(a) `nocolours` (b) `subsetcolours` (c) `allcolours`

Figure 27: Evolution of the loss-function MAE and the monitored R^2 statistic during training on 500 **uniformly sampled** galaxies. The measures of the EAGLE training set, EAGLE test set and SDSS data set are indicated by the blue, orange and green line respectively. The EAGLE test set and SDSS data set are evaluated during each epoch of training with the current weights. The errors of the training and test set follow each other closely during the training process. The error of the SDSS data set however does not.



(a) `nocolours` - noisy photometry **(b)** `subsetcolours` - noisy photometry **(c)** `allcolours` - noisyphotometry



(d) `nocolours` - noisy stellar mass **(e)** `subsetcolours` - noisy stellar mass **(f)** `allcolours` - noisy stellar mass

Figure 28: Evolution of the loss-function MAE and the monitored R^2 statistic during training on EAGLE galaxies with **noisy photometry** (Figures 28a - 28c) and on EAGLE galaxies with **noisy stellar mass** (Figures 28d - 28f). Both error measures are monitored on the EAGLE test set and the SDSS data set. All data sets follow a uniform stellar mass distribution. The measures of the EAGLE training set, test set and SDSS data set are indicated by the blue, orange and green line respectively. The errors of the EAGLE training and test set follow each other closely during the training process. The error of the SDSS data set however does not.