Empirical constraints on the formation of early-type galaxies

Benjamin P. Moster,1,2⋆ Thorsten Naab,2 Simon D. M. White2
1Universitäts-Sternwarte, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 München, Germany
2Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Straße 1, 85748 Garching, Germany

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present constraints on the formation and evolution of early-type galaxies (ETGs) with the empirical model emerge. The parameters of this model are adjusted so that it reproduces the evolution of stellar mass functions, specific star formation rates, and cosmic star formation rates since \( z \approx 10 \) as well as ‘quenched’ galaxy fractions and correlation functions. We find that at fixed halo mass present-day ETGs are more massive than late-type galaxies, whereas at fixed stellar mass ETGs populate more massive halos in agreement with lensing results. This effect naturally results from the shape and scatter of the stellar-to-halo mass relation and the galaxy formation histories. The ETG stellar mass assembly is dominated by ‘in-situ’ star formation below a stellar mass of \( 3 \times 10^{11} M_\odot \) and by merging and accretion of ‘ex-situ’ formed stars at higher mass. The mass dependence is in tension with current cosmological simulations. Lower mass ETGs show extended star formation towards low redshift in agreement with recent estimates from IFU surveys. All ETGs have main progenitors on the ‘main sequence of star formation’ with the ‘red sequence’ appearing at \( z \approx 2 \). Above this redshift, over 95 per cent of the ETG progenitors are star-forming. More than 90 per cent of \( z \approx 2 \) ‘main sequence’ galaxies with \( m_\ast > 10^{10} M_\odot \) evolve into present-day ETGs. Above redshift 6, more than 80 per cent of the observed stellar mass functions above \( 10^{9} M_\odot \) can be accounted for by ETG progenitors with \( m_\ast > 10^{10} M_\odot \). This implies that current and future high redshift observations mainly probe the birth of present-day ETGs. The source code and documentation of emerge are available at github.com/bmoster/emerge.

Key words: cosmology: dark matter, theory – galaxies: evolution, formation, statistics, stellar content

1 INTRODUCTION

In the \( \Lambda \)CDM framework, the Universe consists of dark energy that can be described by a cosmological constant \( \Lambda \), dynamically cold dark matter (CDM), and a small fraction of baryonic matter. In this scenario, structure formation proceeds hierarchically through gravitationally driven collapse and the merging of smaller structures. In the standard picture of galaxy formation, gas can cool within virialised dark matter haloes and form stars at their centres (White & Rees 1978; Fall & Efstathiou 1980; Blumenthal et al. 1984). Once a halo falls into a more massive halo and merges, its galaxy becomes a satellite and may eventually merge with the central galaxy of the main halo impacting its properties. In this way, galaxies can grow both by converting their gas into stars (in-situ), and by accreting stars from merging satellites (ex-situ). As the properties of the haloes determine how much gas is available for star formation and how efficiently this conversion can proceed, the formation histories of the haloes directly impact the formation histories of their galaxies. Consequently, the relation between the stellar mass of a galaxy and the mass of the halo in which it is embedded depends on the average formation time of the halo.

Observed galaxies can be arranged into a continuous sequence of types, as originally introduced by Hubble (1926). Early type galaxies (ETGs) on the one end typically have elliptical shapes with little internal structure, while late type galaxies (LTGs) on the other end show diverse substructure such as discs with spiral arms. ETGs are the non-star-forming (or ‘quenched’), dynamically hot and massive systems of the present-day galaxy population, whereas LTGs tend to be actively star-forming, dynamically cold and less...
massive. In the current picture, massive galaxies are likely to have violent merger histories leading to the elliptical morphologies of ETGs. Weak gravitational lensing studies show that ETGs tend to live in more massive haloes than their LTG counterparts of the same stellar mass (Mandelbaum et al. 2006, 2016). This is further evidence for different formation histories of ETGs and LTGs. All archaeological stellar population studies indicate that the most massive present-day ETGs with log(m*/M⊙) ≥ 11 formed the vast majority of their stars at high redshift (z ≥ 2) on short timescales of the order of 1 Gyr (Gavazzi et al. 2002; Heavens et al. 2004; Thomas et al. 2005, 2010; McDermid et al. 2015). For lower mass ETGs there is a trend towards lower formation redshifts and longer formation time scales. In contrast, it has been shown that low-mass dark matter haloes assemble early, and massive haloes assemble late (Lacey & Cole 1993; Sheth & Tormen 2004; Boylan-Kolchin et al. 2009). This apparent discrepancy has been termed ‘anti-hierarchical’ galaxy formation or ‘downsizing’ (Cowie et al. 1996). The dichotomy has been addressed by De Lucia et al. (2006), who show how this effect naturally arises in the hierarchical paradigm. While the assembly of massive galaxies is hierarchical, AGN feedback leads to low conversion efficiencies in massive haloes, such that all stars formed at early times, before the onset of AGN feedback. Further, in a hierarchical cosmological model the formation time and the assembly time of stars in galaxies can be decoupled, as galaxies also grow by merging (e.g. Oser et al. 2010). Since most of the mergers that bring in a significant amount of stellar mass happen at late cosmic times, the assembly times are typically much later than the formation times of the stars.

The cosmological formation and assembly histories of ETGs have been studied with various theoretical approaches. The most common methods employ a specified set of physical processes for the baryons to evolve an initial distribution of dark matter and gas. In hydrodynamic simulations the gas component is followed along with the dark matter, and the physical processes such as gas cooling, star formation, and several feedback mechanisms are modelled on a chosen physical resolution scale (Springel et al. 2005; Teyssier et al. 2013; Hopkins et al. 2014; Wang et al. 2015). Semi-analytic models (SAMs) separate the formation of galaxies from the formation of gravitationally bound structures by post-processing halo merger trees with physically-motivated recipes (White & Frenk 1991; Kauffmann et al. 1993; Cole et al. 1994; Somerville & Primack 1999). Due to the limited achievable resolution, both approaches need to adopt parameterised models, and adjust the free parameters using observational constraints. As these methods rely on empirical evidence, they are not fully ‘ab initio’. An alternative approach is taken by empirical models (Mo & White 1996; Wechsler et al. 1998; Peacock & Smith 2000; Yang et al. 2003; Zehavi et al. 2004; Vale & Ostriker 2004; Conroy et al. 2006; Moster et al. 2010; Behroozi et al. 2010; Guo et al. 2010; Moster et al. 2013, 2018; Behroozi et al. 2013c, 2019), which use parameterised relations between the properties of observed galaxies and those of simulated haloes. The parameters of these relations are constrained by requiring a number of statistical observations to be reproduced. This approach has the advantage of accurately matching observations by construction, so that the evolution of observed galaxy properties can be followed. We refer to the reviews by Naab & Ostriker (2016) for hydrodynamical simulations, Baugh (2006) and Somerville & Davé (2015) for SAMs, and Wechsler & Tinker (2018) for empirical models. These very different approaches qualitatively agree that massive ETGs build up rapidly by gas dissipation and in-situ star formation at high redshift (z > 2). At this phase, the galaxies are still star-forming and populate the ‘main sequence of star formation’. Towards lower redshifts the star formation ceases, the galaxies move to the ‘red sequence’, and grow their stellar mass further by merging with ex-situ formed systems. Here the importance of ex-situ growth increases with mass and dominates above a certain mass scale. In models that implement the baryonic physics, like hydrodynamical simulations, the transition from star-forming to non-star-forming and the mass scale for accretion domination typically vary with the physical processes included but also with numerical resolution and the assumed sub-resolution model recipes, or simulation codes. The quenching of star formation and the existence of such a merger driven growth phase for massive ETGs, however, is clearly indicated by the observed evolution of galaxy stellar mass functions (SMFs Lilly et al. 1995; Faber et al. 2007) and evidence from size growth (Trujillo et al. 2006; van Dokkum et al. 2008; van der Wel et al. 2014), outer stellar surface density profiles (D’Souza et al. 2014), present-day merger signatures (Duc et al. 2011), or changes in the galaxy scaling relations for high mass systems (Bernardi et al. 2003).

Here we present results on the formation and assembly histories of the ETG population of the empirical galaxy formation model EMERGE. This model follows the formation of individual dark matter haloes in cosmological simulations and links the growth rate of the halo to the star formation rate (SFR) of the galaxy at its centre. The parameterised conversion efficiency determines how efficiently baryons in the halo can be converted into stars. By construction the model successfully reproduces the evolution of the SMF, specific and cosmic star formation rates, the fractions of ‘quenched’ galaxies, and mass-dependent galaxy correlation functions. We show that EMERGE predicts the observed evolution of the blue and red sequence with reasonable present-day archeological star formation histories for ETGs. Moreover, EMERGE reproduces the observed trend for present-day quenched galaxies to live in higher mass halos. Above a certain mass scale the growth of ETGs is dominated by merging and we compare this trend to current hydrodynamical simulations and a SAM. Finally we follow the main progenitors of present-day ETGs back in redshift and demonstrate their growing importance for the star-forming galaxy population up to redshift (z ~ 8).

This paper is organised as follows. In Section 2 we describe the methodology of EMERGE including the employed dark matter simulation and observational data. In Section 3 we present our main results: in section 3.1 we present the stellar-to-halo mass (SHM) relation of active and passive galaxies and compare the model to the lensing constraints. In section 3.2 the predictions for the fraction of mass formed ex-situ are compared to other models. We determine the star formation and assembly histories of ETGs in section 3.3 and the build-up of the star formation bimodality in section 3.4.
In section 3.5 we study the progenitors of ETG up to high redshift. Finally, we present our conclusions in section 4.

Throughout this work we assume a Planck ΛCDM cosmology with \( \left( \Omega_m, \Omega_\Lambda, h, \alpha_s, \sigma_8 \right) = (0.308, 0.692, 0.0484, 0.6781, 0.9677, 0.8149) \). We employ a Chabrier (2003) initial mass function (IMF) and convert all stellar masses and SFRs into solar masses. We use the terms ETG, passive galaxies, and star-forming galaxies interchangeably. Similarly, we use the terms LTG, active galaxies, and star-forming galaxies interchangeably.

2 METHOD

We follow the growth of galaxies in dark matter haloes with the empirical model EMERGE as presented in Moster et al. (2018). This code populates simulated halo merger trees with galaxies reproducing observed statistical galaxy properties such as stellar mass functions, star formation histories, and clustering. Consequently, the two main pillars of the model are cosmological \( N \)-body simulations from which dark matter haloes and subhaloes in each snapshot. Halo merger trees were generated with the ConsistentTrees (Behroozi, Wechsler & Wu 2013a), we identified the minimally resolved halo mass is \( M_{\text{min}} = 10^{10} M_\odot \). With a minimal particle number of 100 for each halo mass were derived using the criterion by Bryan & Norman (1998). With a minimal particle number of 100 for each halo.

2.1 Simulation

The model EMERGE traces dark matter haloes through cosmic time and assigns a SFR to the galaxy at its centre depending on its mass and growth rate. We have extracted halo merger trees from a cosmological \( N \)-body simulation with 200 Mpc side length. The adopted cosmological parameters are consistent with the latest results by the Planck Collaboration et al. (2018): \( \Omega_m = 0.308, \Omega_\Lambda = 0.692, \Omega_b = 0.0484, H_0 = 67.81 \text{ km s}^{-1} \text{ Mpc}^{-1}, n_s = 0.9677 \), and \( \sigma_8 = 0.8149 \). The initial conditions for the simulation were generated with the MUSIC code (Hahn & Abel 2011) and a power spectrum computed with the CAMB code (Lewis, Challinor & Lasenby 2000). The simulation contains \( 10^{12} \) collisionless particles with a mass of \( 2.92 \times 10^8 M_\odot \) and was run with periodic boundary conditions from redshift \( z = 63 \) to 0 using the TreePM code GADGET3 (Springel 2005) with a gravitational softening of 3.3 kpc. We saved 94 snapshots which are equally spaced in scale factor (\( \Delta a = 0.01 \)). Using the halo finder ROCKSTAR (Behroozi, Wechsler & Wu 2013a), we identified the dark matter haloes and subhaloes in each snapshot. Halo masses were derived using the criterion by Bryan & Norman (1998). With a minimal particle number of 100 for each halo the minimally resolved halo mass is \( \log_{10}(M_{\text{min}}/M_\odot) = 10.5 \). Merger trees were generated with the ConsistentTrees code (Behroozi et al. 2013b). The term ‘main halo’ is used to refer to distinct haloes that are not located within a larger halo, while all other haloes are called ‘subhaloes’. We further assume that a ‘central galaxy’ is located at the centre of a main halo, and a ‘satellite galaxy’ within each subhalo.

2.2 Observations

Given the properties of the halo as extracted from the simulation, the model assigns properties to its galaxy such that a range of observed statistical galaxy properties are reproduced. The observational constraints used in this work are SFRs, cosmic SFR densities, specific SFRs, quenched galaxy fractions, and projected galaxy correlation functions. All units are converted into physical units with \( h = 0.6781 \) and a Chabrier (2003) IMF. Systematic errors are taken into account by calculating the variance between different observations and adding the result quadratically to each data point. For the SFRs this error is \( \sigma = 0.15 \text{ dex at } z \leq 0.5 \) and \( \sigma = 0.3 \text{ dex beyond} \). The error found for the cosmic SFR density is \( \sigma = 0.1 \text{ dex} \), while for the specific SFRs we find \( \sigma = 0.15 \text{ dex} \), independent of stellar mass and redshift. For the systematic error of the quenched fractions we use \( \sigma = 10 \text{ per cent} \), while for the clustering we use \( \sigma = 0.15 \text{ dex} \). We have used all the data that was used in (Moster et al. 2018), and refer to their tables 1 to 5 for more information. In addition we have have used the SMFs by Wright et al. (2018), Beare et al. (2019) and Bhatawdekar et al. (2019), the cosmic SFR densities by Katsianis et al. (2017) and Liu et al. (2018), and the specific SFRs by Davidzon et al. (2018). We account for intrinsic scatter in the mass estimates relative to the true mass arising from limited photometric information by drawing stellar masses from a lognormal distribution with a mean value given by the model mass and a scatter of \( \sigma(z) = 0.08 + 0.06z \), capped at a maximum value of 0.32 dex at \( z = 4 \), when comparing to observed stellar masses.

2.3 Galaxy formation model

The simulated dark matter haloes are populated with galaxies by the empirical model EMERGE, such that the observed data is reproduced. To this end, the SFR of the galaxy at the centre of a dark matter halo is set to the product of the halo’s baryonic growth rate and the instantaneous baryon conversion efficiency \( \epsilon \):

\[
\frac{dm_b}{dt}(M, z) = \frac{dm_h}{dt} \epsilon(M, z) = \frac{f_b}{\Omega_b/\Omega_m} \frac{dM}{dt} \epsilon(M, z),
\]

where \( f_b = \Omega_b/\Omega_m \) is the universal baryonic fraction. In this way, the baryonic growth rate signifies how much material becomes available for galaxy formation, while the conversion efficiency determines how effectively it is converted into stars. The halo growth rates are computed from the merger trees and averaged over one dynamical time. For negative halo growth rates we set the baryonic growth rate to zero, and we do not correct for pseudo-evolution.

The conversion efficiency \( \epsilon \) describes how efficiently the gas that falls into the halo is converted into stars, and thus combines the effects of cooling, star formation, and various feedback processes. Following an Occam’s razor approach to find the simplest model that is able to reproduce the observed data, we have tested several parametrizations for the efficiency, which we assessed with various model selection criteria, such as the Bayesian Evidence. Consequently, we parametrise the efficiency as a double power law with normalisation \( \epsilon_N \), characteristic halo mass \( M_1 \), and slopes \( \beta \) and \( \alpha \) :
While we do not impose any a-priori restrictions on the values of the slopes, we find the values of $\beta$ and $\gamma$ to be positive at all cosmic times, i.e. at low/high masses the efficiency increases/decreases with increasing halo mass.

Once the growth rate $\dot{M}$ and the conversion efficiency $\epsilon(M, z)$ have been computed for a halo at a specific time, Emerge can compute the SFR of the central galaxy $m_\star(M, \dot{M}, z)$ using equation (1). Integrating this rate over cosmic time, while taking into account the fraction of mass that is being lost as a consequence of dying stars, yields the stellar mass formed in-situ. In addition, the stellar mass of central galaxies grows by the accretion of satellite galaxies, i.e. by ex-situ mass added during mergers. In our model, galaxies in haloes with the same mass and growth rate at a given redshift have the same SFR, but the stellar mass is different depending on the full formation history of the haloes. The star formation history of each galaxy is then a consequence of the specific path a halo has taken through the halo mass-redshift plane. Typically, at early cosmic times the SFR of a galaxy will be low, as both the growth rate of its halo and the conversion efficiency are low. When the halo mass approaches the characteristic mass $M_1$, the conversion efficiency is maximal, resulting in the peak of the star formation history. As the halo becomes more massive, the conversion efficiency decreases again, resulting in quenched central galaxies.

For growing dark matter haloes, the evolution of the stellar content is determined by the conversion efficiency. Once a halo starts to lose mass, which roughly corresponds to the time its galaxy becomes a satellite, we consider three physical processes that regulate galactic growth: satellite quenching, stripping and merging. For each process the simplest model that is able to reproduce the observational data has been chosen, though the complexity of the models can be increased if new observational data require it.

Once a halo starts to be accreted by a larger halo, its mass begins to decline as a result of tidal stripping. Consequently the infall of gas onto its galaxy stops, such that the cold gas reservoir gets used up by star formation, quenching the galaxy. We take this process into account by employing the ‘delayed-then-rapid’ quenching model, i.e. once the halo stops growing, the SFR of its galaxy is kept constant for a time $\tau$ and then set to zero. This quenching time is parameterised with respect to the halo’s dynamical time $t_{\text{dyn}} = R_{\text{vir}}/v_{\text{vir}}$ and is longer for low mass satellites:

$$t_{\text{dyn}} = R_{\text{vir}}/v_{\text{vir}}$$

$$\tau = t_{\text{dyn}} \cdot \max \left( \frac{M_{\text{sat}}}{10^{10} M_\odot}, 1 \right).$$

The parameters $\tau_0$ and $\tau_s$ are mostly constrained by the observed fraction of quenched galaxies, as shorter quenching times lead to higher quenched fractions. If the halo mass surpasses its previous peak, the halo growth rate and the instantaneous conversion efficiency again determine the SFR (eqn. 1).

If a subhalo has lost enough of its mass through tidal stripping, its gravitational potential is no longer able to protect the stars in its centre from stripping, such the galaxy is destroyed. This process is implemented by moving its stars to the stellar halo (ICM) around the central galaxy once the mass of its subhalo has dropped below a fraction $f_s$ of the peak mass:

$$M < f_s \cdot M_{\text{peak}}$$

The stripping parameter $f_s$ is mostly constrained by small-scale clustering, as faster stripping leads to fewer satellites on small scales, and consequently to lower small-scale clustering. Once a halo becomes a subhalo, we move all the mass in its ICM to the ICM of the host halo.

Once a subhalo has lost all of its orbital energy due to dynamical friction, the satellite galaxy will merge with the central galaxy. At this point we let a fraction of satellite stars $f_{\text{esc}}$ escape to the halo as diffuse stellar material, resulting in a remnant mass of

$$m_{\text{rem}} = m_{\text{cen}} + m_{\text{sat}} \cdot (1 - f_{\text{esc}}),$$

and an ICM mass of $m_{\text{ICM, new}} = m_{\text{ICM, old}} + f_{\text{esc}} m_{\text{sat}}$. The escape fraction $f_{\text{esc}}$ is mostly constrained by the evolution of the massive end of the SFR at low redshift, as a high escape fraction will lead to a slow growth of massive galaxies.

When the mass of a subhalo has been stripped below the resolution limit of the simulation, it can no longer be identified and its galaxy becomes an ‘orphan’. We keep the orphans until their dynamical friction time $t_f$ has elapsed (employing Boylan-Kolchin et al. 2008) and then let them merge with their central galaxies. If a host halo of an orphan merges with a larger halo, we recompute the dynamical friction time with respect to the new central galaxy, and let the orphan merge with it once this new time has elapsed.

To constrain the model parameters we construct mock observations and calculate a model value for each individual observed data point, so that we can directly compare both. To calculate the quenched fractions, we define a galaxy to be

| Parameter | Best-fit | Upper 1$\sigma$ | Lower 1$\sigma$ |
|-----------|----------|----------------|----------------|
| $M_0$     | 11.339   | +0.005         | -0.008         |
| $M_0$     | 0.692    | +0.010         | -0.009         |
| $\epsilon_0$ | 0.005     | +0.001         | -0.001         |
| $\epsilon_z$ | 0.689     | +0.003         | -0.003         |
| $\beta_0$ | 3.344    | +0.084         | -0.101         |
| $\beta_z$ | -2.079   | +0.127         | -0.134         |
| $\gamma_0$ | 0.966     | +0.002         | -0.003         |
| $f_{\text{esc}}$ | 0.388     | +0.002         | -0.002         |
| $f_s$     | 0.122    | +0.001         | -0.001         |
| $\tau_0$ | 4.282    | +0.015         | -0.020         |
| $\tau_s$ | 0.363    | +0.014         | -0.014         |

Notes: All masses are in units of $M_\odot$.
quenched if its sSFR is below a redshift dependent threshold given by \( \Psi < 0.3 r_H^{-1} \), where \( r_H \) is the Hubble time at that redshift. We calculate the real space galaxy two-point correlation functions \( \xi(r) \) using kd-trees following Moore et al. (2001) and Dolence & Brunner (2008), and derive the projected correlation functions \( w_p(r_p) \) at the same projected radii as the observations. We compare the observed data points \( \tilde{\omega} \) to the modelled values \( \tilde{\mu}(\theta) \) for a set of parameters \( \theta \) to get the difference \( \Delta = \tilde{\omega} - \tilde{\mu}(\theta) \). With the covariance matrix of the observed data \( C \), we then compute \( \chi^2 = \Delta^T C^{-1} \Delta \) and assign a likelihood to the model \( L = \exp(-\chi^2/2) \). Parameter space is then explored with two different sampling techniques (Elson et al. 2007; Goodman & Weare 2010) to find the most likely parameters and their credibility intervals, which are presented in Table 1.

To efficiently explore parameter space even for large dark matter simulations, Emerge has been adapted to run on massively parallel computers with distributed memory. First the halo merger trees are loaded and distributed among a set of NC computing cores (processors) using the Message Passing Interface (MPI) protocol. The trees in this first ‘universe’ are copied to more sets of NC cores with MPI, to create a total of \( N_U \) parallel universes on total number of \( N_U \cdot NC \) cores. Parameter space is then explored with \( N_W \) walkers, such that at any time \( N_U \) walkers are processed in parallel, requiring \( N_W/N_U \) steps for all walkers. Each core first computes the galaxy properties for all haloes in its trees, which can be further parallelised with OpenMP if the cores support hyperthreading. All cores in one universe combine their information to calculate the model statistics and the resulting model likelihood. Combining all these different parallelisation techniques, Emerge is able to achieve a near-perfect scaling. For our 200 Mpc box and Xeon Gold 2GHz CPUs, around 30 universes can be computed per core hour. The source code and documentation of Emerge are available on GitHub\(^1\).

3 RESULTS

The empirical model Emerge was designed to follow the assembly of galaxies in dark matter haloes while being in good agreement with global statistics, i.e., SMFs, cosmic SFR densities, specific SFRs, quenched fractions, and galaxy clustering. We will use this model to investigate the different formation mechanisms of active and passive galaxies, and to study the progenitors of ETGs.

3.1 The SHM relation for active and passive galaxies

The SFR of a galaxy is given by the product of the growth rate of its halo and the conversion efficiency, which in turn depends on the halo’s mass and redshift. Therefore, the stellar mass of a galaxy is a direct consequence of the individual formation history of its dark matter halo. In M13 we have shown that independent of halo mass, passive galaxies embedded in haloes of a specific mass have formed more stellar mass than their active counterparts. At fixed halo mass, the conversion efficiency is the same in all haloes, so that the SFR of a galaxy only depends on the halo’s growth rate. Haloes with a low growth rate thus host quenched galaxies, but they must have had a high growth rate at earlier times to reach the same final halo mass, which means that

\(^1\) github.com/bmoster/emerge
they hosted galaxies with a high (specific) SFR, i.e. active galaxies. Conversely, haloes with high growth rates at late times host active galaxies, but must have had low growth rates (hosting passive galaxies) at earlier times to reach the same final halo masses. However, as the conversion efficiency evolves with redshift, this does not imply that also the stellar mass reaches the same value at late times. Since the conversion efficiency is higher at high redshift, haloes with high growth rates at early times host galaxies with very high specific SFRs, such that the resulting final stellar mass is higher compared to galaxies in haloes with low growth rates at high redshift. While those haloes have higher growth rates (and host active galaxies) at late times, the lower conversion efficiency results in SFRs that cannot compensate the lower SFRs at high redshift, such that the final stellar masses are lower. We show the resulting SHM relation for each galaxy and the mean stellar mass at fixed halo mass for active, passive, and all central galaxies in Figure 1.

Our model EMERGE thus predicts that at fixed halo mass, passive galaxies have a higher stellar mass than active galaxies. This result seems to be in disagreement with the findings of weak lensing studies (e.g. Mandelbaum et al. 2006, 2016), which show that at fixed stellar mass, passive galaxies tend to live in more massive haloes than active galaxies. However, simply inverting the average SHM relation at fixed halo mass does not result in the correct SHM relation at fixed halo mass because of the scatter. We demonstrate this in Figure 2, where we have colour-coded all active galaxies in blue and all passive galaxies in red. The vertical dashed lines indicate a halo mass bin between log M_h/\(M_\odot\) = 12.7 and 12.9, while the black, red, and blue crosses show the average stellar mass of all, active, and passive galaxies in this halo mass bin, respectively. As we noted above, passive galaxies have a higher stellar mass than active galaxies (log m_*/\(M_\odot\) = 10.9 vs. 10.7). In this rather high halo mass bin, most galaxies are quenched, so that the average stellar mass of all galaxies is close to the average mass of passive galaxies.

The horizontal dashed lines in Figure 1 indicate a stellar mass bin with the same width as the halo mass bin (0.2 dex). The bin edges log m_*/\(M_\odot\) = 10.9 and 11.1 have been chosen, such that the average halo mass for all galaxies is the same as the average halo mass of the halo mass bin (log M_h/\(M_\odot\) = 12.8). Here, we notice the opposite trend, i.e. active galaxies are located in lower mass haloes than passive galaxies (log M_h/\(M_\odot\) = 12.6 vs. 12.8). This is a consequence of two effects: first, the fraction of active galaxies is higher at lower halo (and stellar) masses, and second, the scatter in stellar mass at fixed halo mass is also higher at lower halo mass. Consequently, there are more active galaxies in low-mass haloes that are scattered to higher stellar masses compared to active galaxies in massive haloes that are scattered to lower masses. We further note, that for the same halo mass the average stellar mass is higher at fixed halo mass than at fixed stellar mass. This also results from a combination of higher scatter at lower halo masses and a larger number of low-mass haloes, such that more galaxies are scattered up from lower halo masses than scattered down from higher halo masses. The same stellar mass is thus reached at lower halo mass, which leads to a higher SHM ratio.

We further investigate difference in the SHM relation at fixed halo mass and fixed stellar mass in Figure 3. For all masses, the SHM ratio is higher when computed at fixed stellar mass compared to the ratio obtained at fixed halo mass. At the massive end, the SHM ratio at fixed halo mass is lower for active galaxies compared to passive galaxies, while the SHM ratio at fixed stellar mass is higher for active galaxies than for passive galaxies. At intermediate masses, the SHM relation is very similar for active and passive galax-
ties, both computed at fixed stellar and fixed halo mass. At the low-mass end we find that the SHM ratio at fixed halo mass is still higher for passive galaxies, but we also find that the SHM ratio at fixed stellar mass is higher for passive galaxies, in contrast to the SHM ratio at the massive end. This means that EMERGE predicts that at the massive end, passive galaxies of given stellar mass are located in more massive haloes than active galaxies, while at the low-mass end, passive galaxies are hosted by less massive haloes than active galaxies. The trend for fixed stellar mass reverses at log \( M_h / M_\odot \approx 11.5 \), and log \( m_*/M_\odot \approx 9.5 \).

Finally, we compare the predictions of EMERGE directly to the observationally measured SHM relation at fixed stellar mass. To this end, we compute the average halo mass for active and passive galaxies in the same stellar mass bins as Mandelbaum et al. (2016). For each bin, we use the model at the same redshift as the median redshift in the observations. We convert all halo masses to our definition. The resulting SHM relations for active and passive galaxies and the data by Mandelbaum et al. (2016) and Mandelbaum et al. (2006) are presented in Figure 4. The predictions of EMERGE are in very good agreement with the lensing constraints. As noted before, the SHM ratio for galaxies with log \( m_*/M_\odot > 9.5 \) is lower for passive galaxies, i.e. at the same stellar mass, they live in more massive haloes than their active counterparts. The model further predicts that this trend inverts for lower masses. Overall, we take this agreement with the data as a further confirmation of the model’s objective, i.e. to link galaxies to dark matter haloes such that observational data is reproduced.

### 3.2 The growth of massive galaxies through accretion

In a hierarchical structure formation scenario, dark matter haloes grow by continuously accreting smaller haloes. In this picture, galaxies do not only gain mass by converting their accreted gas into stars (in-situ growth), but also by accreting other galaxies (ex-situ growth). To better understand how massive galaxies assemble their mass, we compute the ex-situ mass fraction \( f_{ex} \) for each individual galaxy at \( z = 0 \), as the ratio between the remaining stellar mass that has been accreted and the remaining total stellar mass. We show the results for all bound stellar mass, i.e. mass in the central galaxy plus the mass in the ICM, in Figure 5, and the results for the stellar mass in the central galaxy in Figure 6. In both Figures, the lower abscissa indicates the peak halo mass for each galaxy, while the upper abscissa indicates the stellar mass (total bound for Figure 5 and central for Figure 6) with an average halo mass that corresponds to the lower abscissa. While this is not directly equivalent to the stellar mass of each individual galaxy because of scatter, the upper abscissa still yields a good proxy for how the ex-situ mass fraction scales with stellar mass. The dots represent individual systems as computed by EMERGE, while the black solid line gives the average \( f_{ex} \) for a fixed halo mass. The wide light grey area indicates the 1σ scatter in this relation, and the narrow dark grey area gives the uncertainty of the average resulting from observational uncertainties. The latter has been derived by computing the average ex-situ mass fraction for all final MCMC walkers, and then calculating the standard deviation.

The average ex-situ mass fraction for all bound mass is very low in low-mass haloes with log \( M_h / M_\odot < 12 \), although there are individual galaxies with \( f_{ex} \) as high as 90 per cent. Starting at log \( M_h / M_\odot = 12 \) where \( f_{ex} \) is 5 per cent, the ex-situ mass fraction increases quickly and reaches 57 per cent at log \( M_h / M_\odot = 13 \) and 89 per cent at log \( M_h / M_\odot = 14 \). Beyond this mass, \( f_{ex} \) increases more slowly and approaches 100 per cent. We contrast this with the results of hydrodynamical simulations including Illustris (magenta line: Vogelsberger et al. 2014), Horizon-AGN (red line Dubois et al. 2014), Magneticum (green line: Hirschmann et al. 2014), and Illustris TNG (solid blue line: TNG300, dashed blue line: TNG100; Pillepich et al. 2018), and we convert all halo masses to our definition. We also include the semi-analytic model by Henriques et al. (2019) in our comparison (yellow line). For low-mass haloes, we find a very good agreement with the results of Illustris and TNG, although they predict slightly higher ex-situ fraction than EMERGE. For more massive haloes, Illustris has slightly lower values for \( f_{ex} \). Similarly, TNG follows the empirical prediction very closely below log \( M_h / M_\odot \leq 13.5 \), but above this halo mass the simulated galaxies have somewhat lower ex-situ mass fractions for all bound stellar mass. Horizon-AGN and Magneticum show much higher ex-situ fraction for low halo masses compared to the other models, while in massive haloes their ex-situ fraction is lower than in EMERGE, closer to the results of Illustris and TNG. The semi-analytic model agrees well with our predictions, although the semi-analytic ex-situ fraction is slightly lower than the empirical results.

The ex-situ fraction is very sensitive to the amount of
feedback in these models, as shown by Choi et al. (2015) and Henriques et al. (2019). Because of the shape of the conversion efficiency, feedback in low-mass haloes is stronger in satellites than in centrals. Consequently, if (SN) feedback in low-mass haloes is less efficient, the stellar mass in satellites compared to the stellar mass in centrals is higher, which leads to higher ex-situ fractions. Conversely, in massive haloes feedback is typically stronger in the centrals than in the satellites, such that less efficient (AGN) feedback implies higher stellar mass in the centrals compared to the satellites leading to lower ex-situ fractions. Given this context, it is likely that the SN feedback in Horizon-AGN and Magneticum is comparatively low, such that the stellar mass in the satellites is higher than in the other models. At the massive end, the hydrodynamical simulations have somewhat higher conversion efficiencies than emerge, which is also apparent from their stellar mass functions that are higher for massive galaxies. Consequently, they form more stars in the central galaxies and their ex-situ fractions are slightly lower than our result suggests.

Within the central galaxy only, the ex-situ mass fraction is much lower for each halo mass. Similar as before, below a halo mass of $\log(M_{\text{halo}}/M_\odot) = 12$, the ex-situ mass fraction is smaller than 1 per cent with very few individual galaxy reaching much higher values. Compared to all bound stars, $f_{\text{ex}}$ increases less strong with halo mass in the central galaxy (almost linearly with log halo mass), and reaches $\sim 30$ per cent at $\log(M_{\text{halo}}/M_\odot) = 13$ and $\sim 60$ per cent at $\log(M_{\text{halo}}/M_\odot) = 14$. Above $\log(M_{\text{halo}}/M_\odot) = 15$, the ex-situ fraction in the central galaxy starts to saturate at larger than 80 per cent. We again contrast these results with hydrodynamical simulations including Illustris (magenta line), Magneticum (green line), and TNG (solid blue line: TNG300, dashed blue line: TNG100). However, it is not straight-forward to compare the ex-situ fraction for the central galaxy only. The stellar mass in emerge corresponds to observationally obtained masses (typically using model
emerge to our model directly. For the central galaxy (yellow line), which can be compared to the empirical radius of $30\, \text{kpc}$, comparison is impossible, but we show the results at a typical radius of $30\, \text{kpc}$. We also show the semi-analytic result for the central galaxy (yellow line), which can be compared to our model directly.

All simulations show significantly higher ex-situ fractions in the central galaxy compared to EMERGE and the semi-analytic model. The closest agreement is with Illustris, which has only slightly higher ex-situ fractions. For haloes between $\log_\odot(M_*/M_\odot) = 12$ and 14.5, the ex-situ mass fractions in TNG are considerably higher than the empirical constraint. Magneticum has an even higher ex-situ fraction, especially for low-mass haloes, where it deviates from all other models. The semi-analytic results again agree quite well with our model, although the ex-situ fraction in the SAM is somewhat lower. At low to intermediate halo masses, the disagreement with the hydrodynamical simulation is again likely caused by more massive accreted satellite galaxies due to less efficient feedback in low-mass galaxies. The second reason why the simulations show higher ex-situ fractions is related to the distribution of the stars in those galaxies, which is relevant for the comparison based on the central stellar mass. Unlike for all bound stellar mass, where it is straightforward to select the star particles, for the central stellar mass it is important up to which radius star particles are selected. In this comparison, the central stellar mass in the simulations has been defined to include all stars within 30 kpc of the centre, while for EMERGE and the semi-analytic model the masses are based on observed model magnitudes. The higher ex-situ fractions in the simulations thus likely imply that the simulation results are based on a larger radius than the empirical constraints.

So far, we have discussed the ex-situ fractions as a function of halo mass, and found that they are close to zero for low-mass haloes with $\log_\odot(M_*/M_\odot) < 12$, and then rise continuously to over 80 per cent over 3 order of magnitudes in halo mass. Now we investigate more closely, how the ex-situ fractions grow with stellar mass, and determine at which stellar mass accretion becomes the dominant growth channel for galaxies. For this we compute the average ex-situ fraction and the corresponding scatter in bins of stellar mass in the central galaxy at $z = 0$, and show the results in the left-hand panel of Figure 7. The dots again represent individual systems as computed by EMERGE, while the black solid line gives the average $f_{\text{ex}}$ for a fixed central stellar mass at $z = 0$. The wide light grey area indicates the $1\sigma$ scatter in this relation, and the narrow dark grey area gives the uncertainty of the average resulting from observational uncertainties as derived from the final MCMC walkers. The long-dashed line shows the average ex-situ fraction at fixed central stellar mass at $z = 1.0$, and the short-dashed line shows the same quantity at $z = 2$. The coloured dotted horizontal lines represent fixed ex-situ fractions of 20, 40, and 60 per cent. In low-mass galaxies, we find that the ex-situ fraction is negligible, while at $\log_\odot(m_*/M_\odot) = 11$ it reaches 20 per cent. This value strongly increases with stellar mass and reaches 40 per cent at $\log_\odot(m_*/M_\odot) = 11.3$, and 60 per cent at $\log_\odot(m_*/M_\odot) < 11.5$. These results are in very good agreement with the conclusions of Bernardi et al. (2019), who find that mergers start to matter for galaxies with $m_* = 3 \times 10^{11} M_\odot$. At higher redshift, the dependence of the ex-situ fraction on the central stellar mass is very similar, although given values for $f_{\text{ex}}$ are reached at a slightly higher stellar mass.

To investigate the dependence of the transition mass for which mergers contribute a significant fraction of the stellar mass, we compute the stellar mass at which a given ex-situ fraction is reached as a function of redshift. In the right-hand panel of Figure 7, we show the results for six specific ex-situ fractions. The colours of the lines correspond to the same
colours in the left-hand panel. We find that for a given ex-situ fraction, the corresponding central stellar mass increases weakly with redshift. A low ex-situ fraction of 10 per cent is reached by galaxies with \( \log(m_*/M_\odot) = 10.9 \) at \( z = 0 \), while at \( z = 1 \) galaxies need to have a mass of \( \log(m_*/M_\odot) = 11.1 \), and a mass of \( \log(m_*/M_\odot) = 11.2 \) at \( z = 2 \). Galaxies for which the stellar mass growth is dominated by accretion, i.e. an ex-situ fraction of more than 50 per cent, need to be more massive than \( \log(m_*/M_\odot) = 11.4 \) at \( z = 0 \), and more massive than \( \log(m_*/M_\odot) = 10.6 \) at \( z = 2 \). This shows that only the most massive galaxies in the Universe are mainly growing through accretion of satellites, especially at high redshift.

### 3.3 Formation of stars in ETGs

Having determined through which channels galaxies grow their stellar mass, we now turn to the question of when this mass is formed. In M18, we have shown that massive galaxies have star formation histories that peak at a higher redshift compared to low-mass galaxies. In this section we focus on two aspects: First, when do galaxies in haloes of a fixed mass assemble their stellar mass, i.e. at which redshift the galaxy reaches a certain fraction of its final mass, and second, at what time did the stellar populations in the galaxy form (including the stars that formed ex-situ)?

#### 3.3.1 Stellar mass assembly

To address the first issue, we compute the redshift at which the central galaxy in haloes of a fixed mass has assembled a given fraction of its final stellar mass for the first time. In this context, assembly means that the stars are part of the central galaxy, but can have formed either within the central galaxy or in an accreted satellite galaxy. We show the results in Figure 8 for 25 (dashed lines), 50 (solid lines), and 75 per cent (dotted lines). The mean assembly redshift for all galaxies are given by the black lines, while the mean 50 per cent assembly redshift for active and passive galaxies are given by the blue and red line, respectively. For comparison, we also show the mean redshift at which the dark matter haloes have assembled 25, 50, and 75 per cent of its final mass (green lines). The points in the background show the 50 per cent assembly redshift for individual galaxies, colour-coded by specific SFR (c.f. Figure 1).

While low-mass haloes tend to assemble their mass relatively early – 50 per cent of the final halo mass of \( 10^{13} M_\odot \) has assembled by \( z = 1.5 \) – the stellar mass in the central galaxy is in place only relatively late – 50 per cent of the final stellar mass has assembled by \( z = 1 \). Conversely, in more massive haloes the final halo mass assembles late (\( z = 1 \) for 50 per cent in haloes of \( 10^{13} M_\odot \), while the final stellar mass is in place much earlier (\( z = 2 \) for 50 per cent). This effect is sometimes called ‘anti-hierarchical’ galaxy formation, and results from the shape of the instantaneous conversion efficiency. While low-mass haloes are already in place at high redshift, the conversion of gas into stars is inefficient until the halo has gained enough mass. Only then will the galaxy be able to form stars efficiently, resulting in a late assembly redshift. In more massive haloes, this situation is reversed, i.e. while the haloes get most of their mass late, they have passed through the peak of the conversion efficiency at higher redshift, such that the central galaxy was able to form stars efficiently. Consequently, most of the stellar mass is forming early, while at late times, the haloes become massive enough to have a low conversion efficiency, such that the galaxies are quenched and only very little stellar mass assembles late, practically only through merging satellites.

In very massive haloes though, the situation changes again, as the assembly is increasingly dominated by merging satellites (see section 3.2). While star formation still mainly happens at high redshift, satellites are predominantly accreted late, so that the resulting assembly redshifts are shifted to lower values. This also explains why the mean 25 per cent assembly redshift is still very high (\( z > 3 \) in haloes of \( \sim 10^{14} M_\odot \)), but the mean 75 per cent assembly redshift is considerably lower (at \( z < 0.5 \)). Independent of halo halo mass, active galaxies assemble later than passive galaxies at fixed halo mass, which simply follows from the enhanced SFRs of active galaxies at late times compared to passive galaxies.

#### 3.3.2 Star formation histories of ETGs

Instead of looking at the time when stellar mass finally assembles in ETGs, we now turn towards the formation of the stars in the first place. For this we select all passive galaxies at \( z = 0 \) and divide this sample into five stellar mass bins. We then take all stellar mass that is still in place at \( z = 0 \) (and has not been lost by dying stars), and compute the fraction that has formed in a certain lookback time bin of width \( \Delta t = 1 \) Gyr. The results are shown in the left-hand panel of Figure 9 for all five mass bins (solid lines). The right-hand panel of Figure 9 shows the cumulative mass fractions as a function of lookback time for the same mass bins, i.e. which fraction of the stars present at \( z = 0 \) has formed before a
specific lookback time. This shows that in massive galaxies, most of the stars that are still present today have formed at early cosmic times. For the most massive galaxies with \( \log(m_*/M_\odot) \) between 11.5 and 12.0 about 30 per cent of the remaining stars form between a lookback time of 12.5 and 13.5 Gyr, and another 30 per cent between 11.5 and 12.5 Gyr. In the right-hand panel, we can identify that 50 per cent of the stars in this mass bin have formed before a lookback time of 11.8 Gyr, i.e. within the first 2 Gyr of the Universe.

The lower the stellar mass of the galaxies, however, the later its stars formed. Intermediate-mass galaxies with \( \log(m_*/M_\odot) \) between 10.5 and 11.0 form most of their stars at a lookback time of around 10 Gyr, which corresponds to the time when half of their present stellar mass had been formed. For low-mass galaxies with \( \log(m_*/M_\odot) \) between 9.5 and 10.0 the bulk of the star formation happens at a lookback of 8 Gyr and half of their \( z = 0 \) stellar mass has formed before a lookback time of 7.3 Gyr. As all galaxies in the samples are passive, the fraction of the stellar mass formed at late times \( (t_\text{gb} < 5 \text{ Gyr}) \) is generally very low.

Galactic archeology allows to derive the formation times of stellar populations by observational means directly. In Figure 9, we compare our results to formation times obtained observationally by Thomas et al. (2010) based on SDSS, and by McDermid et al. (2015) based on the ATLAS3D survey (rectangles and dashed lines). The mean stellar ages in the SDSS data have been obtained from Lick absorption line indices, while the formation time scales are based on the measured \( a/Fe \) element ratios. These parametrised results show the same trend for higher mass galaxies forming their stars at higher redshift but miss the extended histories for towards lower redshift for low mass galaxies. The star formation histories in the ATLAS3D data have been derived by fitting a linear combination of single stellar populations (SSPs) model spectra to the observed galaxy spectra integrated within apertures covering up to one effective radius. To deal with the inherent degeneracy of inferring the star formation history from integrated light, linear regularisation is used such that the resulting star formation history is as smooth as possible while still reproducing the observed spectrum. The observationally derived star formation histories agree fairly well with our model predictions. However, we find that the observed data is somewhat more extreme than our model. Massive galaxies form even earlier in the observations, while low-mass galaxies tend to have even later star formation compared to our model. One difficulty in this comparison is obviously the sample selection. While the ATLAS3D sample has been selected morphologically as ETGs, we have selected our sample as passive galaxies with a cut in the specific SFR. This means that while the observed sample still mainly consists of passive galaxies, it will contain a fraction of active galaxies, especially at the low-mass end. As active galaxies have much younger stellar populations than passive galaxies, these few active galaxies in the sample cause the fraction of stars formed at late cosmic times to increase, while at the same time the relative fraction of stars formed at early times decreases. This effect accounts for most of the difference between modelled and observationally inferred star formation histories for low and intermediate-mass galaxies.

In summary, our empirical model EMERGE implies that the most massive galaxies have the oldest stellar populations, i.e. they have star formation histories that peak at very early cosmic times. In contrast, galaxies of lower mass contain younger stellar populations as their star formation histories peak later. This ‘downsizing’ effect is a result of
Figure 10. Model prediction for the dependence of the star formation rate on the stellar mass at six redshifts from $z = 0$ to $z = 8$ (from top left to bottom right). The colour of the SFR $- m_*$ map indicates how many galaxies can be found in any location, i.e. there are many galaxies in the red areas and no galaxies in the blue areas. The white dashed lines corresponds to the specific SFR threshold that separates active and passive galaxies at each redshift. The symbols with error bars show the results found by surveys: diamonds for SDSS (Renzini & Peng 2015), crosses for 3D-HST (Whitaker et al. 2012, 2014), circles for MUSE-HUDF (Boogaard et al. 2018), and squares for CANDELS (Salmon et al. 2015).

the shape of the baryon conversion efficiency, as massive haloes and galaxies are most efficient in converting baryons to stars at early cosmic times when their halo reaches a mass of around $10^{11.5} M_\odot$, while low-mass haloes reach this most efficient mass only at late cosmic times. However, if we do not look at when the stars were formed in the first place, but when they assemble in the final galaxy, we find that the most massive galaxies have rather late assembly times. This can be explained by stellar growth in massive galaxies being dominated by ex-situ star formation. While the vast majority of the stars in massive galaxies have formed very early, a large fraction of those stars formed in other galaxies and has only been accreted onto the final galaxy at a later time. Therefore, the galaxies that have assembled most of their stars the earliest are not the most massive galaxies, but those that live in haloes of $\log_10(M_h/M_\odot) \approx 13$, and have typical stellar masses of $\log_10(m_*/M_\odot) \approx 11$. Independent of this, we find that at a fixed halo mass, passive galaxies form and assemble earlier than active galaxies.

3.4 The build-up of the star-formation bimodality

So far we have investigated how galaxies build up their mass and found that massive galaxies typically form their stars at relatively early cosmic times. Comparing our results to observations that have not been used in constructing the empirical model, such as the SHM relation for active and passive galaxies from lensing and the star formation histories in ATLAS$^3D$, we obtained a very good agreement. Now we check if the predictions by EMERGE are able to meet other independent observational constraints on the SFR of galaxies. For this we determine how the SFR depends on stellar mass at different redshifts and confront this with observations from major surveys. Once these constraints have been met, we investigate how the progenitors of passive galaxies follow these scaling relations and when they start to depart from it.

3.4.1 The relation between stellar mass and star formation rate

Before we can compare the relation between stellar mass and SFR to observations, we need to define an ‘observational SFR’ for model galaxies that have extremely low or zero SFRs. In practice, these galaxies would not have any signal in the bands that are typically used to derive observed SFRs. To still be able to show these galaxies in observational data, it is common to set their SFRs to the maximum possible value, i.e. the detection limit of the survey. We simulate this effect by adopting a minimal specific SFR before we
apply any observational uncertainty. We apply a minimal specific SFR of $10^{-12}\,\text{yr}^{-1}$ and a scatter of 0.3 dex to derive our mock SFR from the model SFR. In Figure 10, we show the relation between SFR and stellar mass derived with EMERGE. The colour scale indicates how many galaxies with a given stellar mass-to-SFR combination can be found, i.e. the red areas are populated by many galaxies, while there are no galaxies in the blue regions of the plot. The white dashed lines represent the specific SFR threshold used to divide active and passive galaxies at each redshift.

At high redshift of $z \sim 8$, we observe that the star formation main sequence is building up. At this cosmic time, the relation is still very wide with a large scatter, resulting from highly fluctuating SFRs of individual galaxies. At later times, these fluctuations become smaller and the main sequence becomes tighter. At $z \sim 2$, the main sequence is fully in place and comprises the vast majority of all galaxies. At this redshift, we observe two effects. The first effect is the bending of the main sequence at high stellar masses towards lower SFRs. This downturn can be explained by the onset of feedback in massive central galaxies. The main sequence follows a power-law up to the stellar mass where the conversion efficiency starts to decrease again, which corresponds to the characteristic mass (‘knee’) of the SMF. Here, the conversion of baryons into stars becomes less efficient, resulting in lower SFRs. At $z \sim 2$, many galaxies cross this stellar mass, such that the bend in the main sequence becomes apparent. We tested an alternative implementation of the conversion efficiency in EMERGE, where instead of gradually decreasing with halo mass, the conversion efficiency instantaneously drops to zero once a threshold halo mass has been crossed. In this model we do not find any bending in the main sequence, indicating that this effect is caused by a gradual decline of the conversion efficiency, i.e. feedback becomes stronger with halo mass instead of acting above a certain halo mass. The second effect we notice, is the emergence of a cloud of passive galaxies after $z \sim 2$. In EMERGE, these are galaxies living in haloes that have stopped growing, such that the galaxies have eventually quenched. Most of these galaxies are satellites.

Towards lower redshift, we observe that the bend at the massive end of the main sequence becomes more apparent. While the haloes of those galaxies still grow, their conversion efficiency becomes much lower. Consequently, their SFRs drop and their stellar mass does no longer grow, which results in lower SFRs at fixed stellar mass, and an even stronger bend in the main sequence. At the same time, more galaxies get instantly quenched in haloes that have stopped growing (or started to lose mass). As a result, the passive cloud becomes more prominent towards lower redshift. Finally, at $z = 0$, the conversion of gas into stars in massive galaxies is so inefficient that the SFRs in those galaxies are extremely low, so that they are also located in the passive cloud. This area has become much more prominent at $z = 0$, and includes centrals as well as satellites. The bend in the main sequence has moved to lower stellar masses, while at the same time there remain many more massive galaxies on the main sequence that no longer accrete new baryonic material, but convert the baryons that have already accreted.

We compare these results to the observational findings of SDSS (Renzini & Peng 2015, diamonds), 3D-HST (Whitaker et al. 2012, 2014, crosses), MUSE-HUDF (Boogaard et al. 2018, circles), and CANDELS (Salmon et al. 2015, squares) in Figure 10. At high redshift the model is in excellent agreement with the CANDELS observations. At intermediate redshift ($z = 1 - 2$), the model agrees well with the observational constraints, although the observations tend to find slightly higher values for the SFR on the main sequence. The model prediction for the bending of the main sequence around $\log(m_*/M_\odot) \geq 10$ is in very good agreement with the observations. This indicates that the conversion efficiency at the massive end has to decline gradually with halo mass instead of instantaneously dropping above a certain halo mass, and that feedback at the massive end acts more strongly with increasing halo mass. At low redshift, we find very good agreement with SDSS and 3D-HST. As EMERGE is able to reproduce these individual constraints up to high redshifts, we will use the predictions for the SFRs in combination with the information how different galaxy populations are linked through cosmic time to further investigate the progenitors of ETGs.

Before we study when and how passive galaxies were quenched and moved onto the red cloud, we focus on the relation between stellar mass and SFR for the present galaxy population, and compare our results to observations. Figure 11 presents a mock observation for local galaxies obtained with EMERGE by applying the previously stated minimum specific SFR and observational uncertainty. Further, to take into account the completeness limits of SDSS, we remove all galaxies for which the product of the stellar mass and SFR is lower than $10^6.5 M_\odot$ yr$^{-1}$, such that galaxies with a low stellar mass are only included in the plot if they have a relatively high SFR. We further show the SDSS results by Renzini & Peng (2015) as diamonds (c.f. their Figure 4), both for the star formation main sequence and the passive
Figure 12. Distribution of specific SFRs of the main progenitors of presently passive galaxies. The dashed lines with different colours show the specific SFR distributions for fixed progenitor redshifts. The thin vertical lines indicate the threshold that distinguishes active from passive galaxies for each redshift. The top panel gives the results for intermediate mass galaxies, and shows that almost all of their progenitors have high SFRs before $z = 1$. The bottom panel gives the results for massive galaxies, and shows that their progenitors have almost exclusively high SFRs only before $z = 2$.

could. The error bars correspond to the observed 1σ scatter in those relations. We find a very good agreement between data and model predictions, although we note that the slope in the observed main sequence is slightly flatter. At this stage it remains unclear, whether this discrepancy results from a bias in the data, or from a deficiency in the model. The passive cloud is centred around $\log(m_*/M_\odot) = 10.5$ and $\log(\psi/M_\odot yr^{-1}) = -1.3$, both in the data and in the model.

3.4.2 Star formation in progenitors of passive galaxies

Having verified that EMERGE predicts SFRs that are in agreement with observational constrains up to high redshift, we now focus on the build-up of the star-formation bimodality and investigate how massive galaxies end up on the passive cloud at $z = 0$. For this we divide all model galaxies including satellites in the present-day Universe into an active and a passive sample, and mainly concentrate on the latter. We then trace them through cosmic time and study the properties of their progenitors, and how they evolve with redshift.

In Figure 12, we show the distribution of specific SFRs of the main progenitors of presently passive galaxies. To derive these results, we select all passive galaxies in a specified stellar mass bin at $z = 0$ and follow their main branches (tracing the most massive progenitor) up to the indicated redshifts. We then compute the specific SFR for each main progenitor and derive the probability density function at any redshift. These specific SFR distributions are plotted for each redshift (coloured lines) for intermediate $z = 0$ stellar masses of $10.0 \leq \log(m_*/M_\odot) < 10.5$ in the upper panel and massive galaxies with $11.0 \leq \log(m_*/M_\odot) < 11.5$ in the lower panel. As above, we have applied a minimum specific SFR of $10^{-12} yr^{-1}$ and an observational uncertainty of $0.3 dex$. The dotted vertical lines indicate the threshold that divides active and passive at each redshift.

The specific SFRs of the selected galaxies at $z = 0$ are low and centred around the adopted minimum value, as we have only included passive galaxies. Consequently, all galaxies have specific SFRs below the division threshold. We note that intermediate mass sample shows a slightly wider distribution for the passive mode compared to the massive sample. The reason for this are galaxies that have recently departed the main sequence and have not yet fully settled in the passive cloud, i.e. their SFRs are still higher than the minimum value. In contrast, the massive sample includes much fewer of these recently quenched galaxies and therefore shows a narrower distribution. The specific SFR distributions of the main progenitors move towards higher values with higher redshift. The progenitors of the intermediate mass sample still have a small mode of quenched galaxies around the minimum specific SFR value at $z = 0.5$. The vast majority of the progenitors are actively forming stars with the mode centred around a specific SFR of $\log(sSFR/yr^{-1}) = -9.0$. At higher redshift this mode further moves to higher specific SFR values and reaches $\log(sSFR/yr^{-1}) = -9.0$ at $z = 2$. At the same time, there are no passive progenitors, so that we can conclude that all of the main progenitors of presently passive galaxies.
progenitors of intermediate mass passive galaxies lie on the star formation main sequence at $z \sim 1.5$.

In contrast, massive galaxies are passive and have low specific SFRs much earlier. The progenitors of the massive sample are mostly already passive at $z = 0.5$ with the mode of the distribution at $\log(sSFR/\text{yr}^{-1}) = -10.8$. Only at $z = 1$, more than half of the main progenitors are active and have specific SFRs centred around $\log(sSFR/\text{yr}^{-1}) = -10.1$. The mode of the specific SFR moves to higher values with increasing redshift, although the peak is at lower values for the progenitors of the massive sample compared to the progenitors of the intermediate mass sample, which is again a manifestation of the bending in the star formation main sequence. At $z = 2$ almost all progenitors of the massive sample are active, and at $z \sim 2.5$, all they all lie on the star formation main sequence.

Using these results, we can compute at which redshift the galaxies in the passive sample become quenched. As galaxies usually do not have monotonic star formation histories, they can be quenched multiple times with periods of active star formation in between. Therefore, we compute the redshift at which passive galaxies at $z = 0$ were quenched for the first time, and the redshift when they were quenched for the last time in their history. We show the resulting quenching redshifts in Figure 13 as a function of the stellar mass at $z = 0$. The blue and red lines correspond to the peak in the distribution of the first and last quenching redshifts at fixed $z = 0$ stellar mass, respectively. The blue and red shaded areas show the upper and lower 1σ values for both distributions, i.e. 68 per cent of all passive galaxies are quenched within this interval.

Low-mass galaxies are typically quenched very late. While galaxies with a stellar mass of $\log(m_*/M_\odot) = 9.5$ can be quenched for the first time as early as $z = 2$, most of them are quenched for the first time only by $z = 0.3$, and very few of them after $z = 0.1$. In contrast, the redshift when those galaxies are quenched for the final time is much lower. Some galaxies were finally quenched by $z = 0.6$, while the majority became passive by $z = 0.1$, and a some were only quenched very recently. Intermediate mass galaxies with $\log(m_*/M_\odot) \sim 10.75$ tend to have higher quenching redshifts than low-mass galaxies. While the first of them are quenched a little later around $z = 1.5$, the majority are quenched for the first time before $z = 0.75$, and almost all of them went through a passive episode by $z = 0.3$. The final quenching redshift for this sample is typically around $z = 0.4$, while most of them became fully passive by $z = 0.2$. The most massive galaxies with stellar masses of $\log(m_*/M_\odot) = 12$ were quenched very early. Quenching can start as early as $z = 2.5$ and most galaxies have undergone a passive period by $z = 1.7$. Most progenitors have been completely passive since $z = 1.5$, and only very few of them have formed stars after $z = 1$.

This analysis shows that before $z \sim 2$ almost all progenitors of presently passive galaxies have continuously formed stars and were located on the main sequence. The most massive galaxies have departed from the main sequence a long time ago (at $z \sim 1.5$) and were passive ever since, while galaxies with a lower stellar mass at $z = 0$ stayed on the main sequence much longer, and only left it rather recently (after $z \sim 0.5$). To quantify this behaviour in more detail, we compute the fraction of all main progenitors of presently passive galaxies that are still on the main sequence. For this we select all passive galaxies within a $z = 0$ stellar mass bin and record how many of their main progenitors are still actively forming stars at any given redshift, having specific SFRs that are higher than the threshold. We show the re-
sulting fractions as a function of redshift in Figure 14 for five \( z = 0 \) stellar mass bins (dashed coloured lines).

The fraction of main progenitors of low-mass passive galaxies with a \( z = 0 \) stellar mass of \( 9.5 \leq \log(m_*/\text{M}_\odot) < 10.0 \) that are on the main sequence rises quickly with redshift. At \( z = 0.1 \), already 50 per cent of the progenitors are located on the main sequence. At \( z = 0.5 \), 90 per cent of the progenitors of low-mass passive galaxies are on the main sequence, and at \( z = 1.3 \) virtually all (99 per cent) of the progenitors are forming stars. Intermediate-mass galaxies with a \( z = 0 \) stellar mass of \( 10.5 \leq \log(m_*/\text{M}_\odot) < 11.0 \) leave the main sequence earlier. Half of their progenitors have left the main sequence by \( z = 0.4 \), while 90 per cent of them are still part of the main sequence at \( z = 1 \), and almost all (99 per cent) are active before \( z = 1.5 \). The most massive passive galaxies with a \( z = 0 \) stellar mass of \( 11.5 \leq \log(m_*/\text{M}_\odot) < 12.0 \) have left the main sequence a long time ago. Around half of their progenitors have become passive by \( z = 1.3 \), while 90 per cent are still active before \( z = 2 \). Before \( z = 2.5 \) all progenitors of the most massive galaxies are still on the main sequence. This shows that all passive galaxies have main progenitors on the main sequence at some redshift.

Having analysed the progenitors of passive galaxies at \( z = 0 \) and determined when those galaxies become quenched, we now invert the question. Given the populations of galaxies that lie on the main sequence at a given redshift, we investigate how many of those active galaxies become quenched by \( z = 0 \). For this we first select active galaxies that are on the main branch at a specified redshift and have a specified stellar mass. Instead of using stellar mass bins, we here employ stellar mass thresholds, to distinguish this analysis clearly from the previous one, as now the stellar mass is selected at a given (progenitor) redshift and not at the (descendant) redshift of zero. For each of the galaxies in the samples we then compute the fraction that has left the main branch and become passive by \( z = 0 \). We show the resulting quenched fractions in Figure 15 for four stellar mass thresholds. For the redshift indicated in the abscissa, the ordinate gives the fraction of all galaxies that are more massive than the threshold mass at this redshift, which will leave the main sequence by \( z = 0 \).

More than 90 per cent of all galaxies on the main sequence at \( z = 2 \) that are more massive than \( \log(m_*/\text{M}_\odot) = 9.5 \) will leave the main sequence by the present day. Even at a lower redshift of \( z = 1 \), more than 85 per cent of all galaxies on the main sequence with \( \log(m_*/\text{M}_\odot) = 9.5 \) will be passive at \( z = 0 \). More massive galaxies tend to have an even higher probability of being quenched and leaving the main sequence. Virtually all massive galaxies with \( \log(m_*/\text{M}_\odot) \geq 11 \) on the main sequence at \( z = 1 \) will be passive today. Even at \( z = 0.5 \), 95 per cent of all active massive galaxies will be quenched by \( z = 0 \). At low redshift the values for the quenched descendant fractions drop steeply. Still, almost independent of stellar mass, about half of all galaxies that are on the main sequence at \( z = 0.1 \) will leave the main sequence by \( z = 0 \). This shows that most of the galaxies that are observed on the main sequence at intermediate to high redshift \( (z \geq 1) \) do not stay on the main sequence throughout their lives, but instead will eventually become passive and settle in the red cloud by \( z = 0 \). This again confirms that the massive galaxies that are observed to have very high SFRs at \( z = 1 – 2 \) are the progenitors of present-day passive ETGs.

### 3.5 The progenitor populations of ETGs

In the previous section we focused on the properties of the main progenitors of present ETGs, i.e. we traced the most massive progenitors in the merger trees through cosmic time. In this section we investigate the full distribution of the progenitors including the smaller galaxies that merged onto the main branch of the merger tree to study which galaxies at high redshift have accumulated to form the present population of passive galaxies. We further analyse how frequent the ETG progenitors are at high redshift, and how much of the total galaxy population they account for. First we select all passive galaxies at \( z \geq 0 \) and create two samples: one massive sample in which all galaxies are more massive than \( \log(m_*/\text{M}_\odot) = 11 \), and a sample that includes all galaxies that are more massive than \( \log(m_*/\text{M}_\odot) = 10 \). We then trace the merger trees of all galaxies in these samples to high redshift and record all progenitors including the main progenitors and the progenitors that have merged. Using the populations at six redshifts, we determine their SMF. We also divide these populations into active and passive progenitors (with our standard specific SFR threshold) and determine their respective SMFs. Finally, we only select the main progenitors (including active and passive) and determine their SMF.

The resulting SMFs are shown in Figure 16 for progenitors of passive galaxies with a \( z = 0 \) stellar mass of \( \log(m_*/\text{M}_\odot) > 11 \), and in Figure 17 for progenitors of passive galaxies with a \( z = 0 \) stellar mass of \( \log(m_*/\text{M}_\odot) > 10 \). In both Figures, the black solid lines show the total SMFs of all galaxies at each redshift. The brown solid lines show the abundance of all progenitors of the \( z = 0 \) samples, while the dashed blue and red lines divide this into components for active and passive progenitors, respectively. The green lines indicate the results obtained when only considering the main (most massive) progenitors of each \( z = 0 \) galaxy. The top panels show the SMFs at six redshifts, while the bottom panel shows the relative fraction of each population, i.e. their contribution to the total SMF at this redshift. The dotted horizontal lines in the lower panels indicate fractions of 80 and 50 per cent of the total abundance of galaxies.

For both samples, we find that the progenitors of presently passive galaxies account for the massive end of the SMF. At low redshift \( (z = 0.5) \) they contribute 50 per cent to the total SMF above the selected threshold mass, and over 80 per cent above slightly higher stellar masses of \( \log(m_*/\text{M}_\odot) = 10.1 \) and 11.1, respectively. At high redshift, the massive end consists almost exclusively of the progenitors of passive galaxies. At \( z = 4 \) the progenitors of the massive sample account for over 80 per cent of the SMF above a mass of \( \log(m_*/\text{M}_\odot) = 10.5 \), while at \( z = 8 \) the ETG progenitors account for over 80 per cent of the SMF above \( \log(m_*/\text{M}_\odot) = 8 \). For the sample with a lower \( z = 0 \) threshold mass, we find that at \( z = 4 \) more than 80 per cent of the total SMF above \( \log(m_*/\text{M}_\odot) = 9.5 \) consists of the progenitors of this sample, and at \( z = 8 \) more than 80 per cent of the SMF is made from the progenitors. This indicates that most of the galaxies that contribute to the SMFs found in present surveys can be accounted for by the progenitors of present-day ETGs with \( \log(m_*/\text{M}_\odot) > 10 \).

The low-mass end of SMFs at low to intermediate redshift \( (z \leq 4) \) contains very few progenitors of passive galax-
Empirical constraints on ETG formation

Figure 16. The SMFs of all progenitors of present quenched galaxies with log($m_*/M_\odot$) > 11 at different redshifts (upper panels). The blue and red lines show the SMFs of all active and passive progenitors, respectively, the brown lines show the total SMFs of all progenitors and the green lines show the SMFs of the main progenitors. In comparison, the black lines show the total SMFs of all galaxies. The bottom panels indicate the fraction that the progenitors of passive galaxies contribute to the total SMF at each redshift. The dotted horizontal lines give fractions of 80 and 50 per cent.

Figure 17. Same as Figure 16, but for progenitors of present quenched galaxies with log($m_*/M_\odot$) > 10.
ies. The progenitors of the massive sample contribute less than the total SMF below \( \log(m_*/M_\odot) = 10 \) at \( z = 0.5 \), and less than 10 per cent below \( \log(m_*/M_\odot) = 10 \) at \( z = 2 \). The progenitors of the sample that include intermediate-mass galaxies, the contribution to the low-mass end of the SMFs is also negligible, although slightly higher than for the massive sample. At \( z = 0.5 \) these progenitors contribute up to a few per cent below \( \log(m_*/M_\odot) = 10 \), and at \( z = 2 \) they contribute about 50 per cent at \( \log(m_*/M_\odot) = 9.5 \) and 10 per cent at \( \log(m_*/M_\odot) = 8 \). At high redshift the progenitors of presently passive galaxies significantly contribute to the total SMFs. The progenitors of the mass sample make up 50 per cent of the \( z = 4 \) SMF at \( \log(m_*/M_\odot) = 9.5 \), and 50 per cent of the \( z = 8 \) SMF at \( \log(m_*/M_\odot) = 7 \). Similarly, at \( z = 4 \), 50 per cent of the SMF at \( \log(m_*/M_\odot) = 8 \) consists of the progenitors of the sample with a lower threshold mass. This shows that at low redshift the progenitors of presently passive galaxies contribute little to the low-mass end of the observed SMFs, while a significant fraction of the observed low-mass galaxies at high redshift end up in present-day ETGs.

We further find that the progenitors of both samples at low redshift are mostly active at the low-mass end, and mostly passive at the massive end. At \( z = 0.5 \) the majority of progenitors that are more massive than \( \log(m_*/M_\odot) = 11 \) are passive, while below this mass most progenitors are actively forming stars. The contribution of passive galaxies diminishes at higher redshift and is insignificant above \( z \gtrsim 2 \), i.e. almost all progenitors are forming stars at high redshift. For the contribution of the main progenitors to the total SMF, we find that at low redshift the main progenitors only account for the massive end of the SMF. The small contribution of low-mass progenitors to the SMF exclusively consists of galaxies that will merge onto a larger galaxy later. However, at high redshift we can identify a number of low-mass galaxies that are the main progenitors of presently passive galaxies. While there are more secondary progenitors than main progenitors at the low-mass end, roughly half of the progenitors at the massive end at high redshift are main progenitors. Moreover, they can contribute up to 50 per cent of the galaxies in the SMF at the massive end. Our analysis highlights that a large fraction of the observed galaxies at high redshift will eventually be accreted by a larger galaxy that is passive by \( z = 0 \).

4 CONCLUSIONS

We analysed the formation and evolution of ETGs with the empirical model \textsc{emerge}, which follows the evolution of galaxies in individual dark matter haloes through cosmic time. For each halo in a cosmological N-body simulation, we computed the growth rate and calculate the SFR of the galaxy within the halo as the product of the halo growth rate, which specifies how much material becomes available, and the instantaneous baryon conversion efficiency, which specifies how efficiently this material can be converted into stars. Within \textsc{emerge}, the latter summarises all of the baryonic physics of galaxy formation and depends on halo mass and redshift. When a halo stops growing, the SFR is left constant for a specified amount of time and is then rapidly quenched. Galaxies in haloes that have lost a significant fraction of their mass are stripped. Once satellites lose their kinetic energy due to dynamic friction, they merge with the central galaxy and eject a fraction of their stars to the halo. We integrated the star formation history of each galaxy taking into account mass loss from dying stars to derive its stellar mass.

The model is constrained by a variety of observational data, such as SMFs, specific SFRs, the cosmic SFR density, fractions of quenched galaxies and clustering as a function of stellar mass. All model parameters were determined using an MCMC ensemble sampler by requiring that all observed data be reproduced simultaneously. This allowed us to connect the observations at different epochs and to follow the growth of stellar mass in galaxies as constrained by the observations. For each individual galaxy we determined if it was active or passive using a specific SFR threshold that corresponds to 30 per cent of the Hubble parameter at the observed redshift. The conclusions we can draw from \textsc{emerge} are therefore a direct consequence of the evolution of the observational data used to constrain the model over cosmic time in a ΛCDM Universe. As a result, we expect similar conclusions for any ΛCDM-based models, which fit the same observational data, e.g. empirical models such as the UniverseMachine (Behroozi et al. 2019), or state-of-the-art semi-analytical models such as the model by Henriques et al. (2019). We expect somewhat more discrepant results from current hydrodynamic simulations since these typically fit the data somewhat less well, although future simulations that are in better agreement with observations are likely to derive similar results. In our analysis, we focused on ETG properties and derived the following conclusions:

(i) \textit{The stellar-to-halo mass relation in ETGs:} We found that at fixed halo mass, passive galaxies have a higher stellar mass, as they preferentially formed at higher redshift when the instantaneous conversion efficiency was higher. However, we also found that this relation cannot be simply inverted, and that at fixed stellar mass, ETGs live in more massive haloes than their LTG counterparts. The reason for this effect is the scatter in the SHM relation, which causes many active galaxies in haloes of lower mass to have a high stellar mass. In massive haloes there are fewer active galaxies with higher stellar mass, so that the average halo mass at fixed stellar mass is biased towards lower masses for active galaxies. We compared our results to the observational weak lensing constraints by Mandelbaum et al. (2006, 2016) and found excellent agreement.

(ii) \textit{The growth of ETGs through accreted satellites:} We calculated the fraction of the stellar mass that has formed ex-situ for each galaxy and determined the average ex-situ fraction as a function of halo and stellar mass. Because of the shape of the SHM ratio low-mass haloes typically accrete satellites that are much less massive than the central galaxy, such that the ex-situ fraction is low. However, in massive haloes the conversion efficiency is suppressed, such that the stellar mass of accreted satellites is of the order of the central galaxy’s mass, resulting in a high ex-situ fraction. When taking into account the mass of the central galaxy and the mass in the ICM, we found that ex-situ star formation starts to become relevant for haloes with \( \log M_\text{halo}/M_\odot \gtrsim 12 \) and reaches 50 per cent around \( \log M_\text{halo}/M_\odot \sim 13 \). For the most massive haloes, the ex-situ fraction approaches unity. Taking only into account the stellar mass in the central galaxy (i.e.
excluding the ICM), we found lower ex-situ fractions that reach 50 per cent around \( \log M_s/M_\odot \sim 13.5 \) and 80 per cent for the most massive haloes. Comparing our results to other models, we found that ENSUEF predicts a lower ex-situ fraction for intermediate-mass haloes with \( \log M_s/M_\odot \sim 13-14 \). Finally, we found that the ex-situ fraction is a strong function of stellar mass. Below \( \log(m_*/M_\odot) = 11 \) the ex-situ fraction is negligible, but then rises quickly to 60 per cent at \( \log(m_*/M_\odot) = 11.5 \), in agreement with Bernardi et al. (2019).

The transition mass for which accreted stellar mass becomes relevant is only a weak function of redshift.

(iii) Star formation and mass assembly in ETGs: Tracing the stellar mass and SFRs of the main progenitors of present-day galaxies through cosmic time, we found that low-mass galaxies tend to assemble late although their dark matter haloes typically assemble at high redshift. In contrast, massive galaxies tend to assemble early, while their dark matter haloes assemble late. This effect that has been called ‘anti-hierarchical’ galaxy formation or ‘downsizing’ can be easily explained invoking the conversion efficiency. While low-mass haloes are already in place at high redshift, their galaxies can convert gas into stars only very inefficiently until the halo becomes massive enough, delaying star formation in low-mass galaxies. In massive haloes, the galaxies can only grow efficiently at high redshift, before the halo becomes too high and only allows for low SFRs at late times. Only for the most massive galaxies we found relatively late assembly times. While the stars in those haloes formed at high redshift a large fraction gets accreted by the massive galaxy only at late times. Further, we found star formation histories that peak at higher redshift for more massive systems. For the most massive systems \( \log(m_*/M_\odot) > 11.5 \), we found that most stars have formed before \( z = 3 \), and over 90 per cent of the present stellar mass had formed by \( z = 1 \). Our results are in agreement with estimates from IFU surveys derived through galactic archeology.

(iv) The build-up of the star formation bimodality: We derived the relation between stellar mass and SFR up to high redshift and found that the ‘main sequence of star formation’ starts to form before \( z = 8 \) and is in place by \( z = 4 \) comprising the vast majority of galaxies. Around \( z = 2 \), we found that the main sequence develops a bend at the massive end, as star formation in the most massive galaxies is becoming inefficient. We also found that this downturn can only be explained by a gradual decline of the conversion efficiency, i.e. feedback has to become stronger gradually, while a model where feedback quenches massive galaxies instantaneously does not result in a bending. The ‘red sequence’ is beginning to emerge at \( z \sim 2 \) and grows until \( z = 0 \), when star formation in massive galaxies has become very inefficient. Our results are in good agreement with major surveys. We further investigated when present-day ETGs are quenched, and found that low-mass galaxies become passive very late, while massive galaxies finally stop forming stars by \( z = 1 \). Above this redshift, most ETG main progenitors are located on the main sequence. Conversely, more than 90 per cent of the galaxies on the main sequence at \( z = 2 \) with \( \log(m_*/M_\odot) > 10 \) evolve into present-day ETGs.

(v) The progenitor populations of ETGs: Following the full distribution of ETG progenitors including galaxies that merge onto them, we derived the SMF of all ETG progenitors, as well as active and passive ETG progenitors. We found that the total ETG progenitor population accounts for the massive end of the SMF up to high redshift. At low redshift, the ETG progenitors contribute very little to the low-mass end of the SMF. However at high redshift, most of the galaxies in the SMF, even at low stellar mass, are the progenitors of ETGs. Above redshift 6, more than 80 per cent of the SMF above \( \log(m_*/M_\odot) = 9 \) can be accounted for by progenitors of present-day ETGs with \( \log(m_*/M_\odot) > 10 \). In contrast, we found that the main progenitors of those ETGs only account for less than 50 per cent of the \( z = 6 \) SMF above \( \log(m_*/M_\odot) = 9 \).

Our final conclusion is that current and future high redshift observations mainly probe the formation of present-day ETGs.

ACKNOWLEDGEMENTS

We thank Yohan Dubois, Bruno Henriques, Richard McDermid, Rhea-Silvia Remus, Annalisa Pillepich, and Daniel Thomas for sharing their data. We are also grateful to Peter Behroozi, Andreas Burkert, Michael Fall, Andrew Hearin, Andrei Kravtsov, Aura Obreja, Joseph O’Leary, Jerry Ostriker, Ulrich Steinwandel, and Frank van den Bosch for enlightening discussions. The cosmological simulations used in this work were carried out at the Freya Cluster at the Max Planck Computing and Data Facility in Garching, and the Leibniz Supercomputing Centre. BPM acknowledges an Emmy Noether grant funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – MO 2979/1-1.

REFERENCES

Baugh C. M., 2006, Reports on Progress in Physics, 69, 3101
Beare R., Brown M. J. I., Pimbblet K., Taylor E. N., 2019, ApJ, 873, 78
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, ApJ, 762, 109
Behroozi P. S., Wechsler R. H., Wu H.-Y., Busha M. T., Klypin A. A., Primack J. R., 2013b, ApJ, 763, 18
Behroozi P. S., Wechsler R. H., Conroy C., 2013c, ApJ, 770, 57
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MN- RAS, p. 1134
Bernardi M., et al., 2003, AJ, 125, 1849
Bernardi M., Domínguez Sánchez H., Brownstein J. R., Drory N., Sheth R. K., 2019, arXiv e-prints, p. arXiv:1904.11996
Bhathawdekar R., Conselice C. J., Margalef-Bentabol B., Duncan K., 2019, MNRAS, 486, 3805
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517
Boogaard L. A., et al., 2018, A&A, 619, A27
Boylan-Kolchin M., Ma C.-P., Quataert E., 2008, MNRAS, 383, 93
Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., London G., 2009, MNRAS, 398, 1150
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Chabrier G., 2003, PASP, 115, 763
Choi E., Ostriker J. P., Naab T., Oser L., Moster B. P., 2015, MNRAS, 449, 4105
Cole S., Aragon-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, MNRAS, 271, 781
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
