Mon. Not. R. Astron. Soc. **370,** 1651–1665 (2006) doi:10.1111/j.1365-2966.2006.10608.x

Modelling the galaxy bimodality: shutdown above a critical halo mass

A. Cattaneo,[1,2,3★] A. Dekel,[1,2★] J. Devriendt,[4] B. Guiderdoni[4] and J. Blaizot[5]

[1] *Racah Institute of Physics, Hebrew University of Jerusalem, 91904 Jerusalem, Israel*
[2] *Institut d'Astrophysique, 98 Boulevard Arago, Paris 75014, France*
[3] *Astrophysikalisches Institut Potsdam, an der Sternwarte 16, 14482 Potsdam, Germany*
[4] *Centre de Recherche Astronomique de Lyon, 9 Avenue Charles André, 69561, St-Genis-Laval Cedex, France*
[5] *Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str.1, 85740 Garching, Germany*

Accepted 2006 May 23. Received 2006 May 21; in original form 2005 December 22

ABSTRACT

We reproduce the blue and red sequences in the observed joint distribution of colour and magnitude for galaxies at low and high redshifts using hybrid N-body/semi-analytic simulations of galaxy formation. The match of model and data is achieved by mimicking the effects of cold flows versus shock heating coupled to feedback from active galactic nuclei (AGNs), as predicted by Dekel and Birnboim. After a critical epoch $z \sim 3$, only haloes below a critical shock-heating mass $M_{\rm shock} \sim 10^{12}\,{\rm M}_{\odot}$ enjoy gas supply by cold flows and form stars, while cooling and star formation are shut down abruptly above this mass. The shock-heated gas is kept hot because being dilute it is vulnerable to feedback from energetic sources such as AGNs in their self-regulated mode. The shutdown explains in detail the bright-end truncation of the blue sequence at $\sim L_*$, the appearance of luminous red-and-dead galaxies on the red sequence starting already at $z \sim 2$, the colour bimodality, its strong dependence on environment density and its correlations with morphology and other galaxy properties. Before $z \sim 2$–3, even haloes above the shock-heating mass form stars by cold streams penetrating through the hot gas. This explains the bright star forming galaxies at $z \sim 3$–4, the early appearance of massive galaxies on the red sequence, the high cosmological star formation rate at high redshifts and the subsequent low rate at low redshifts.

Key words: shock waves – cooling flows – galaxies: evolution – galaxies: formation – galaxies: haloes – galaxies: ISM.

1 INTRODUCTION

Following the early indications for correlations between global galaxy properties along the Hubble sequence (e.g. Hubble 1926; Humason 1936), large statistical surveys, such as the Sloan Digital Sky Survey (SDSS) and the two-degree Field (2dF), have revealed the details of a robust bimodality, which divides the galaxy population into a 'red sequence' and a 'blue sequence'. The bimodality is seen in (i) the luminosity function (Bell et al. 2003; Baldry et al. 2004), (ii) the colour distribution and the joint distribution of colour and magnitude or stellar mass (Strateva et al. 2001; Baldry et al. 2004; Balogh et al. 2004; Hogg et al. 2004), (iii) the stellar age and the star formation rate (Kauffmann et al. 2003), (iv) the gas-to-stellar mass ratio (Kannappan 2004), (v) the bulge-to-disc ratio (Kauffmann et al. 2003), (vi) the environment density (Blanton et al. 2006) and (vii) the presence of an X-ray halo (Helsdon & Ponman 2003; Mathews & Brighenti 2003; Osmond & Ponman 2004).

In the colour–magnitude diagram, the red population and the blue population form two stripes. In both sequences, redder galaxies tend to be brighter, but more so for the blue sequence, which is also broader. The blue sequence is truncated at $M_{\rm r} \sim -22$, while the red one extends to brighter magnitudes. The division between the two classes of galaxies is associated with a critical stellar mass $M_{*{\rm crit}} \sim 3 \times 10^{10}\,{\rm M}_{\odot}$. Galaxies below $M_{*{\rm crit}}$ are typically blue, star forming spirals and reside in the field. Galaxies above $M_{*{\rm crit}}$ are dominated by red spheroids of old stars and live in dense environments. The faint red galaxies in dense environments define a third population (Blanton et al. 2006). This indicates that the mass of the host halo, rather than the stellar mass of the galaxy, determines whether a galaxy is blue or red.

Already at $z \gtrsim 1$, a separate population of bright red galaxies can be distinguished from fainter blue ones (Im et al. 2002; Bell et al. 2004; Moustakas et al. 2004; Weiner et al. 2005; Willmer et al. 2006; Faber et al. 2006), though their numbers at $z \sim 1$ were still small. A plausible scenario is that the growth in red galaxies was triggered by the quenching of star formation in blue galaxies, which caused them to migrate to the red sequence (Bell et al. 2004; Faber et al. 2006). Because star formation was still quite active in massive galaxies at $z \sim 2-4$ (Lyman break galaxies and SCUBA sources; Shapley et al. 2004; Smail et al. 2002; Chapman et al. 2003, 2004), this evidence suggests that star formation began to shut down in at

★E-mail: acattaneo@aip.de; dekel@phys.huji.ac.il

least some massive galaxies around $z \sim 2$ and has fully shut down in such objects by today.

Current models of galaxy formation have difficulties in explaining the bimodality features seen at the bright end. There is no natural explanation for the low rate of gas supply indicated by the low star formation rates (Kauffmann et al. 2003) and the low H I fractions (Kannappan 2004) of massive red galaxies compared to the high accretion rates and gas fractions inferred for blue spiral galaxies.

In the traditional scenario of galaxy formation (Rees & Ostriker 1977; Silk 1977; White & Rees 1978; Blumenthal et al. 1984; White & Frenk 1991), the infalling gas is shock-heated to the virial temperature of the dark halo near the virial radius. The inner gas that can cool radiatively on a free-fall time-scale falls in and forms stars, while the outer gas is held in quasi-static equilibrium until it eventually cools. The 'cooling radius' equals the virial radius today for haloes of mass $M_{\rm cool} \simeq 3 \times 10^{13}\ (Z/{\rm Z}_{\odot}) f_{\rm b}\,{\rm M}_{\odot}$, where Z is the hot gas metallicity and $f_{\rm b}$ is the cosmic baryon fraction. In more massive haloes the supply of cold gas gradually stops.

This argument has been successful in explaining the ballpark of the mass scale separating galaxies from groups, but when it is applied to compute galaxy colours, it predicts that massive E and S0 galaxies should lie on a continuous extension of the blue sequence rather than forming a separate sequence that is systematically redder by $u - r \simeq 0.2$ at $M_{\rm r} \lesssim -22$ as seen in the SDSS. Part of this failure arises from the gradual dependence of cooling on mass.

Analyses in spherical symmetry (Birnboim & Dekel 2003, hereafter BD03) and cosmological simulations (Kereš et al. 2005, hereafter K05) have shown that in haloes below a critical shock-heating mass $M_{\rm shock} \sim 10^{12}\,{\rm M}_{\odot}$ rapid cooling does not allow gas compression to support a stable virial shock. So the gas flows cold and unperturbed into the inner halo. The actual critical $M_{\rm shock}$ is smaller than the crude cooling mass by a factor of a few, and is comparable to the halo mass indicated by the observed bimodality. The new picture near the critical scale is of a multiphase intergalactic medium (IGM) with a cold phase embedded in a hot medium.

These studies indicate that the transition from efficient build-up by cold flows to complete dominance of the hot mode occurs relatively sharply over a narrow mass range about $M_{\rm shock}$. Binney (2004) and (Dekel & Birnboim 2006, hereafter DB06) noticed that this transition can be sharpened further once shock-heating triggers another heating mechanism that compensates for subsequent radiative losses. Feedback from active galactic nuclei (AGNs) is the leading candidate for the source of such heating (Tabor & Binney 1993; Ciotti & Ostriker 1997; Tucker & David 1997; Silk & Rees 1998; Granato et al. 2004; Ruszkowski, Brüggen & Begelman 2004; Brüggen, Ruszkowski & Hallman 2005; Di Matteo, Springel & Hernquist 2005, and references therein). The role of AGN feedback is supported by (i) the ubiquity of supermassive black holes in early-type galaxies (Magorrian et al. 1998; van der Marel 1999; Tremaine et al. 2002), (ii) the evidence from X-ray data that outflows can create deep cavities in the hot IGM of galaxy clusters (Fabian et al. 2003; McNamara et al. 2005), (iii) the estimate that a small fraction of the total energy radiated by an AGN over a cosmological time is sufficient and (iv) the possibility of a radiatively inefficient AGN mode at low accretion rates that allows pumping energy into the IGM without violating the strong constraints on the amount of blue light that can be released (e.g. Di Matteo et al. 2003, for the particular case of M87).

The critical halo mass for effective feedback does not seem to have its natural roots in the physics of the AGNs themselves. On the other hand, this scale arises naturally as the critical scale for shock heating. Near the shock-heating scale, the IGM is made of cold dense clumps in a hot dilute medium. This dilute medium is vulnerable to AGN feedback. The abrupt shutdown in the accretion of gas by galaxies above $M_{\rm shock}$ is triggered by shock heating and maintained over time by AGN feedback.

DB06 have studied the potential implications of the interplay between the cold/hot flow transition and different feedback sources and have predicted how this can reproduce the observed bimodality features. In this paper, we test this hypothesis in a more quantitative way, by incorporating the shutdown above a critical mass into simulations of galaxy formation that combine N-body gravity with semi-analytic modelling (SAM) of the baryonic processes. In Section 2 we present the semi-analytic/N-body code. In Section 3 we show that without the abrupt shutdown of gas accretion above a critical mass the SAM overpredicts the number of bright blue galaxies and fails to reproduce the red colours of massive early-type galaxies. In Section 4 we summarize the theory of cold flows versus shock heating, and describe how we implement these two features in the SAM. In Section 5 we explain our picture of the interaction between the massive black hole and the IGM. In Section 6, we find that the addition of such an abrupt shutdown brings the model into excellent agreement with the data. In Section 7, we demonstrate that the success of this model predominantly depends on the requirement that the shock-heated gas never cools and is robust to the detailed implementation of other physical ingredients in the model. In Section 8, we interpret the results of our simulations and propose a unified scenario for the origin of the blue sequence and the two parts of the red sequence. We compare our results with those of Croton et al. (2006) and Bower et al. (2006), who discuss similar models to the one in this paper. In Section 9 we summarize our main conclusions.

2 THE HYBRID SEMI-ANALYTIC/*N*-BODY CODE

GalICS (Galaxies In Cosmological Simulations; Hatton et al. 2003) is a method that combines high-resolution simulations of gravitational clustering of the dark matter with a semi-analytic approach to the physics of the baryons (gas accretion, galaxy mergers, star formation and feedback) to simulate galaxy formation in a Lambda cold dark matter (ΛCDM) Universe. The version of GalICS used for this article differs from the original version in the star formation law (equation 1), the outflow rate due to supernova feedback (equation 2) and the cut-off that prevents cooling on massive galaxies (equation 3).

2.1 The dark matter simulation

The cosmological N-body simulation has been carried out with the parallel tree code developed by Ninin (1999). The cosmological model is flat ΛCDM with a cosmological constant of $\Omega_{\Lambda} = 0.667$, a Hubble constant of $H_0 = 66.7\,{\rm km\,s^{-1}}$, and a power spectrum normalized to $\sigma_8 = 0.88$.

The simulated volume is a cube of $(150\,{\rm Mpc})^3$ with 256^3 particles of $8.3 \times 10^9\,{\rm M}_{\odot}$ each and a smoothing length of 29.3 kpc. The simulation produced 100 snapshots spaced logarithmically in the expansion factor $(1 + z)^{-1}$ from $z = 35.59$ to $z = 0$.

In each snapshot, we applied a friend-of-friend algorithm (Davis et al. 1985) to identify virialized haloes of more than 20 particles. The minimum halo mass is thus $1.65 \times 10^{11}\,{\rm M}_{\odot}$. The information about individual haloes extracted from the N-body simulation and passed to the semi-analytic model is in three parameters: the virial mass, the virial density and the spin parameter. Merger trees are computed by linking haloes identified in each snapshot with their progenitors in the one before, including all predecessors from which a halo has inherited one or more particles. We do not use the substructure provided by the *N*-body simulation. Once a halo becomes a subhalo of another halo, we switch from following its gravitational dynamics with the *N*-body integrator to an approximate treatment based on a semi-analytic prescription.

2.2 Galaxy formation through the accretion of gas

Newly identified haloes are attributed a gas mass by assuming a universal baryon fraction of $\Omega_b/\Omega_0 = 0.135$. All baryons start in a hot phase shock-heated to the virial temperature. Their density profile is that of a singular isothermal sphere truncated at the virial radius.

The cooling time is calculated from the density distribution of the hot gas using the cooling function of Sutherland & Dopita (1993). The gas for which both the cooling time and the free-fall time are shorter than the time-step Δt is allowed to cool during that time-step. Cooling is accompanied by inflow to maintain the shape of the hot gas density profile and is inhibited in haloes with positive total energy or with angular momentum parameter $\lambda > 0.5$.

Comparisons of semi-analytic and smoothed particle hydrodynamics (SPH) simulations (Benson et al. 2001; Yoshida et al. 2002; Helly et al. 2003) demonstrate that the two methods give similar results when the physical assumptions are the same. In particular, Cattaneo et al. (2006) find that the mass function and the accretion histories of galaxies computed with GalICS are very similar to those derived with an SPH method. This demonstrates that the cooling algorithm is not an important source of error.

2.3 Morphologies

GalICS models a galaxy with three components: a disc, a bulge and a starburst. Each galaxy component is made of stars and cold gas, while the hot is considered as a component of the halo. Only the disc accretes gas from the surrounding halo. The bulge grows by mergers and disc instabilities, while the starburst contains gas in transition from the disc to the bulge. The disc is assumed to be exponential, with the radius determined by conservation of angular momentum. The bulge is assumed to have a Hernquist (1990) profile and its radius is determined based on energy conservation.

Disc instabilities transfer matter from the disc to the bulge until the bulge mass is large enough to stabilize the disc (van den Bosch 1998). Dynamical friction drives galaxies to the centres of dark matter haloes and is the most important cause of mergers (while we also include satellite–satellite encounters). The fraction of the disc that transfers to the bulge in a merger grows with the mass ratio of the merging galaxies from zero, for a very minor, merger to unity, for an equal-mass merger.

This is of course a simplified picture of the dynamics of morphological transformations, but it provides results consistent with essential properties such as the Faber–Jackson relation and the fundamental plane of spheroids.

2.4 Star formation and feedback

The star formation law is the same for all components:

$$\dot{M}_* = \frac{M_{\rm cold}}{\beta_* t_{\rm dyn}}(1+z)^{\alpha_*}. \qquad (1)$$

The cold gas mass $M_{\rm cold}$ refers to the component in question, and $t_{\rm dyn}$ is the dynamical time (corresponding to half a rotation for a disc and the free-fall time for a spheroid). In the standard GalICS model, star formation is activated when the gas surface density is $\Sigma_{\rm gas} > 20\, m_{\rm p}\,{\rm cm}^{-2}$ ($m_{\rm p}$ is the proton mass). The star formation efficiency parameter has a fiducial value of $\beta_* = 50$ (Guiderdoni et al. 1998), which is assumed to be the same at all redshifts, $\alpha_* = 0$. Gas in transit from the disc to the spheroid passes through a starburst phase in which the star formation rate grows by a factor of 10 before the gas is converted into bulge stars. This high star formation rate is obtained by assuming that the starburst radius is 10 times smaller than the final bulge radius (see the hydrodynamic simulation in Cattaneo et al. 2005). The properties of the galaxy population at low redshifts are insensitive to the starburst star formation time-scale as long as it is much shorter than the Hubble time.

The mass of newly formed stars is distributed according to the Kennicutt (1983) initial mass function. Stars are evolved between snapshots using substeps of at most 1 Myr. During each substep, stars release mass and energy into the interstellar medium. Most of the mass comes from the red giant and the asymptotic giant branches of stellar evolution, while most of the energy comes from shocks due to supernova explosions. The enriched material released in the late stages of stellar evolution is mixed with the cold phase, while the energy released from supernovae is used to reheat the cold gas and return it to the hot phase in the halo. Reheated gas is ejected from the halo if the potential is shallow enough. The rate of mass-loss through supernova-driven winds $\dot{M}_{\rm w}$ is determined by the equation

$$\frac{1}{2}\dot{M}_{\rm w}v_{\rm esc}^2 = \epsilon_{\rm SN}\eta_{\rm SN}E_{\rm SN}\dot{M}_*, \qquad (2)$$

where $E_{\rm SN} = 10^{51}$ erg is the energy of a supernova, $\eta_{\rm SN} = 0.0093$ is the number of supernovae every 1 $M_\odot$ of stars and $v_{\rm esc}$ is the escape velocity (Dekel & Silk 1986). In GalICS $v_{\rm esc} \simeq 1.84 v_{\rm c}$ for discs and $v_{\rm esc} = 2\sigma$ for bulges/starbursts. The supernova efficiency $\epsilon_{\rm SN}$ is a free parameter, which determines the fraction of the supernova energy used to reheat the cold gas and drive an outflow at the escape speed. We determine its value by requiring that the model matches the luminosity (Fig. 1) and the gas fraction of a galaxy that lives in a Milky Way type halo. Our best-fitting value $\epsilon_{\rm SN} \simeq 0.2$ agrees with those adopted in other SAMs (Somerville & Primack 1999; Cole et al. 2000).

2.5 STARDUST

The STARDUST model (Devriendt, Guiderdoni & Sadat 1999) gives the spectrum of a stellar population as a function of age and metallicity and includes a phenomenological treatment of the reprocessing of light by dust. GalICS uses STARDUST to compute the spectrum of each component and output galaxy magnitudes. Dust absorption is computed with a phenomenological extinction law that depends on the column density of neutral hydrogen, the line of sight and the metallicity of the obscuring material. The reemitted spectrum is the sum of four templates (big and small carbon grains, silicates and polycyclic aromatic hydrocarbons). Their relative weights are chosen to reproduce the relation between bolometric luminosity and infrared colours observed locally in *IRAS* galaxies. In GalICS we consider only the self-absorption of each component. The mean column density is determined assuming spherical symmetry for the starburst and the bulge, while in the disc component there is a dependence on a randomly selected inclination angle.

3 FAILURE OF THE STANDARD MODEL

Semi-analytic models based on the standard scenario (Section 2) have difficulties in explaining the early formation of bright elliptical galaxies followed by passive evolution. At $z = 0$,

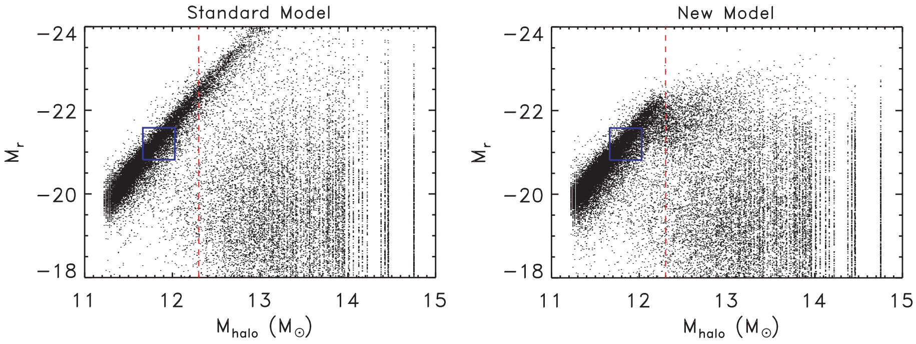


Figure 1. Galaxy luminosity versus halo mass for the 'standard' model (left-hand panel) and the 'new' model (right-hand panel). The blue square refers to the Milky Way (Binney & Merrifield 1998; Dehnen & Binney 1998). The vertical dashed line marks the shock-heating scale. The main difference between the two models is in the break in the relation at $M_{\rm halo} > M_{\rm shock}$ due to the shutdown above $M_{\rm crit}$ (equation 3). Some differences also show up below $M_{\rm shock}$ because of the shutdown of cooling in galaxies where the bulge is the dominant component.

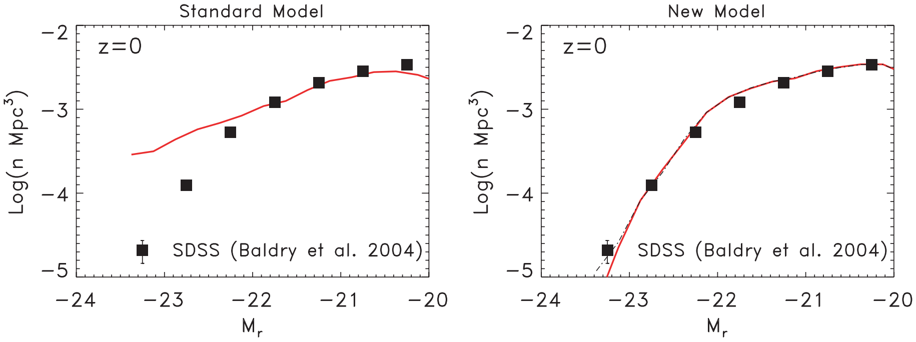


Figure 2. Luminosity function in the *r* band. The model predictions (solid line, red) for the 'standard' model (left-hand panel) and for the 'new' model with shutdown in massive haloes (right-hand panel) compared to the observed luminosity function (symbols) observed by SDSS (Baldry et al. 2004). The dotted–dashed line (right-hand panel) illustrates the effect of lowering the critical redshift from $z_{\rm c} = 3.2$ to $z_{\rm c} = 3$.

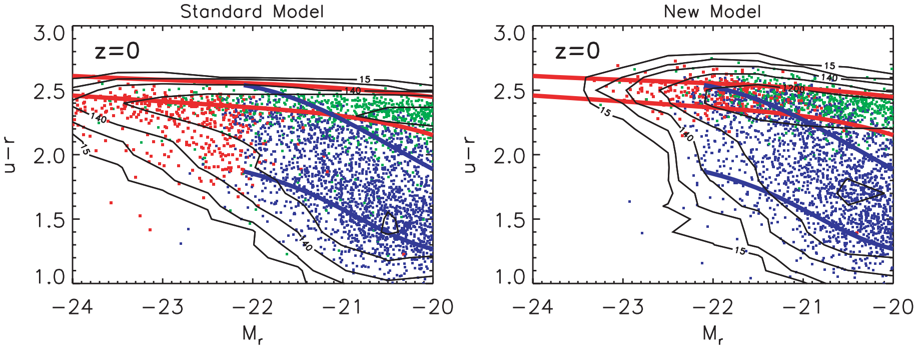


Figure 3. Colour–magnitude diagram at $z = 0$ using $u - r$ versus $M_{\rm r}$ (SDSS magnitudes). Model predictions (dots), 'standard' (left-hand panel) and 'new' (right-hand panel), compared to the observed loci of the red/blue sequence in the SDSS data (Baldry et al. 2004) marked by the pairs of red and blue curves, which show the $\pm 1\sigma$ dispersion around the mean. Galaxies in haloes below $M_{\rm shock}$ are marked blue; they are typically the only galaxy in their halo. Galaxies in haloes above $M_{\rm shock}$ are marked red if they are the central galaxy of their halo and green if they are satellites. The contours, directly comparable to fig. 2 of Baldry et al. (2004), refer to the density of points in units of mag^{-2} Mpc^{-3}) and highlight the bimodality.

the luminosity function shown in Fig. 2 demonstrates that the standard model predicts too many bright galaxies compared to SDSS.

Figs 3 and 4 illustrate in two different ways that the model bright ellipticals are also not red enough. In Fig. 3 we see that the brightest galaxies (which are also the central galaxies of massive haloes) fall below the colour range of red galaxies in the SDSS. Fig. 4 displays the galaxy colour distribution in different magnitude bin. A comparison of the predictions of the standard model with the SDSS data in the $-23 < M_{\rm r} < -22.5$ bin shows that (i) there are many more galaxies in the model than in the data, (ii) the predicted peak of the colour distribution is bluer by $u - r \simeq 0.2$ and (iii) the model distribution is skewed towards blue colours.

At $z \sim 1$, Fig. 5 shows that the model fails to predict the existence of bright red galaxies. Instead, the most luminous galaxies are predicted to lie on an extension of the blue sequence, in conflict with

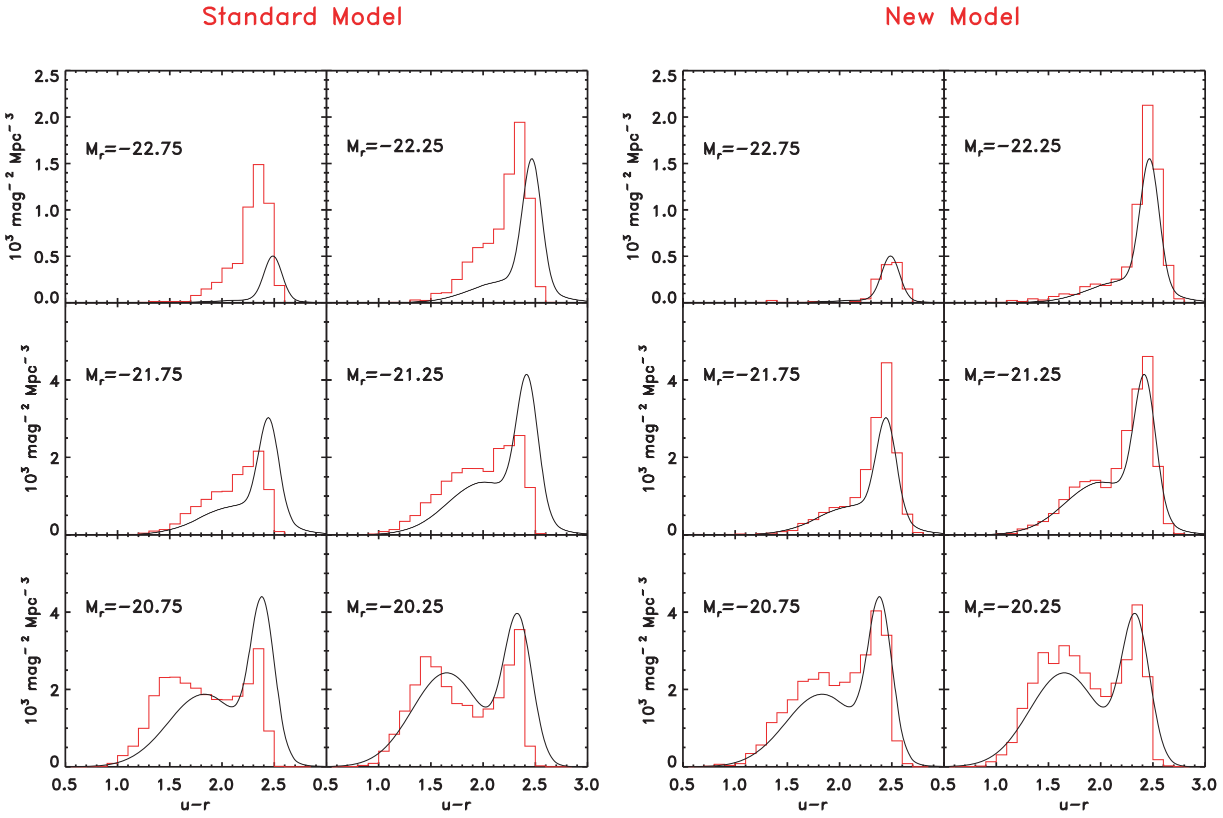


Figure 4. $u - r$ colour distribution in r-band magnitude bins. The model predictions (red histogram), 'standard' (left-hand panel) and 'new' (right-hand panel), compared to the SDSS data (black smoothed histogram) from Baldry et al. (2004). The 'new' model parameters are $M_{\rm shock} = 2 \times 10^{12}\,{\rm M}_{\odot}$ and $z_{\rm c} = 3.2$ (equation 3). The star formation efficiency increases at high redshift with a power of $\alpha_* = 0.6$ (equation 1) and the energetic efficiency of supernova feedback is $\epsilon_{\rm SN} = 0.2$ (equation 2).

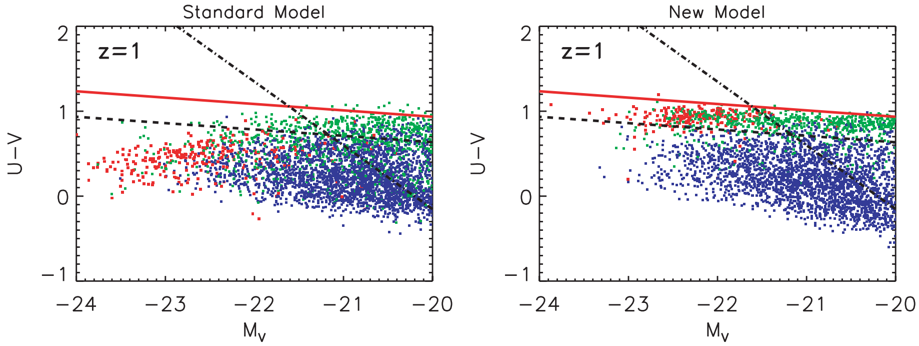


Figure 5. Colour–magnitude diagram at $z = 1$. $U - V$ versus M_V. The 'standard' and 'new' models are the same as in Fig. 3. The red solid line and the black dashed line refer to the centre and the lower bound of the red sequence for the $1.0 < z < 1.1$ bin in the COMBO-17 data (Bell et al. 2004). The dot–dashed line marks the estimated completeness limit of that survey.

the COMBO-17 data (Bell et al. 2004), which reveal the bimodality already at $z \sim 1$.

Fig. 6 shows the failure of the standard model in matching the luminosity function of LBGs (Steidel et al. 1999) at $z \sim 3$. With its star formation rate, the model cannot explain the presence of massive star forming galaxies at $z \sim 3$. In the whole computational volume, there is only a handful of galaxies with $M_V < -23$ (Fig. 7). The star formation history shown in Fig. 8 demonstrates again that the model star formation is not sufficient at high redshifts. These discrepancies are generic to the standard scenario and are not specific to the way this scenario is implemented in GalICS. Blaizot et al. (2004) were able to use GalICS without the modifications proposed here and they did find a reasonable agreement with the Steidel et al. (1999) luminosity function at $z \sim 3$, but they used a non-standard feedback model, which spoils the fit to joint distribution of colour and magnitude at $z \sim 0$.

4 COLD FLOWS VERSUS SHOCK HEATING

A stable extended shock can exist in a spherical system when the pressure that develops in the post-shock gas is sufficient to balance the gravitational attraction towards the halo centre. For adiabatic

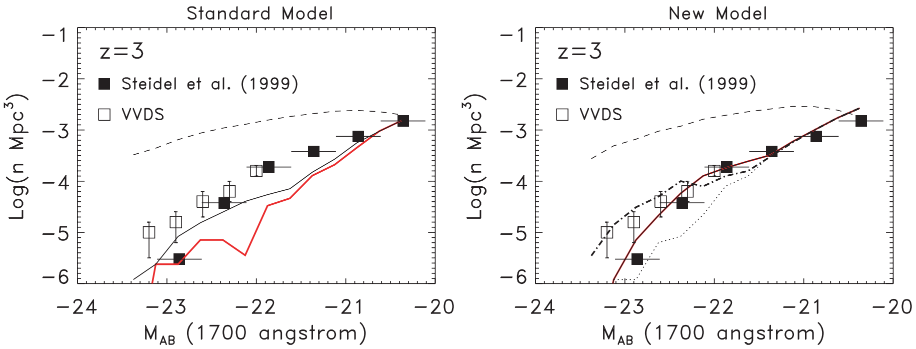


Figure 6. Luminosity function at $z \simeq 3.1$. The magnitude is rest-frame 0.17 μm. The model predictions (solid, red line), 'standard' (left-hand panel) and 'new' (right-hand panel), compared to the data (symbols) inferred from R-band observations of LBGs (Steidel et al. 1999). Superimposed are data from the VVDS. The solid black line in the left-hand panel refers to an extreme variant of the 'standard' model where all the gas cools instantly and feedback is practically shut off. Even with this extreme cooling the model fails to form enough bright star forming galaxies at $z \sim 3$, demonstrating that a higher star formation efficiency is necessary at high z. The dot–dashed line in the right-hand panel refers to the 'new' model with $z_{\rm c} = 3.0$ rather than the fiducial $z_{\rm c} = 3.2$, showing that the high-z predictions of the 'new' model are sensitive to the critical redshift after which $M_{\rm crit} = M_{\rm shock}$. The thin dotted line in the right-hand panel corresponds to $z_{\rm c} \to \infty$, namely $M_{\rm crit} = M_{\rm shock}$ for all z. The thin black dashes are the red lines without dust extinction.

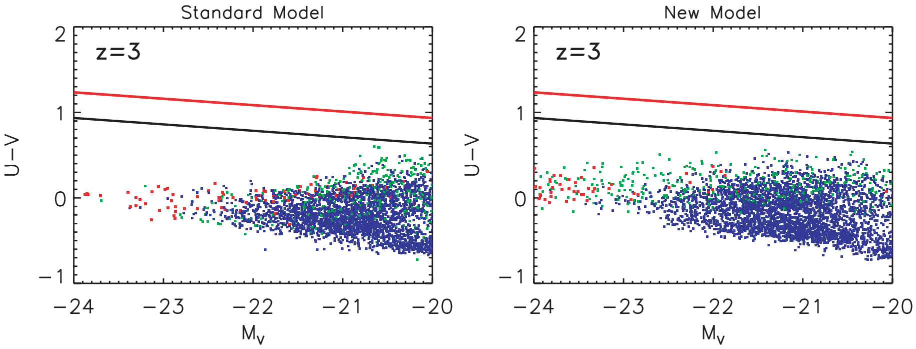


Figure 7. Colour–magnitude diagram at $z = 3$. The 'standard' and 'new' models are the same as in Fig. 3. The lines, shown for reference only, are the same as in Fig. 5 describing the COMBO-17 red sequence at $z \simeq 1$.

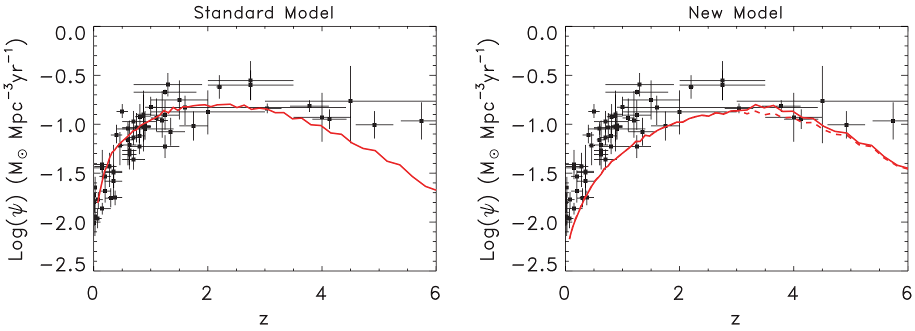


Figure 8. Global star formation history. The model predictions (solid, red line), 'standard' (left-hand panel) and 'new' (right-hand panel) are compared to the estimates derived from different sources, using a common obscuration correction when necessary (symbols; Hopkins 2004). The dashed line shows the difference that assuming a constant critical mass at all redshifts makes to the 'new' model.

contraction, this requirement translates into a condition on the adiabatic index $\gamma \simeq 5/3$, defined as the ratio between the specific heats at constant pressure and constant volume.

BD03 have extended this result to the case where energy is lost at a rate q by replacing the adiabatic index γ with the effective polytropic index $\gamma_{\rm eff} \equiv {\rm d}\ln p/{\rm d}\ln\rho = \gamma - (\rho/\dot{\rho})(q/e)$, where e is the internal energy. This can be rewritten as $\gamma_{\rm eff} = \gamma - t_{\rm comp}/t_{\rm cool}$, where $t_{\rm cool} \equiv e/q$ is the cooling time of the post-shock gas and $t_{\rm comp} \equiv \rho/\dot{\rho} = -(\nabla v)^{-1} = -r_{\rm s}/(3u_1)$ is its compression time, with $r_{\rm s}$ the shock radius and u_1 the post-shock radial velocity. The first equality is the continuity equation and the second assumes that the post-shock radial flow pattern is homologous. Using perturbation analysis, BD03 showed that the criterion for a stable shock is $\gamma_{\rm eff} > 10/7$ (for a monatomic gas in spherical symmetry), which translates to $t_{\rm cool}^{-1} < t_{\rm comp}^{-1}$. This defines a critical halo mass, above which the gas is shock-heated near the virial radius, and below which the gas flows cold and unperturbed into the inner halo, where it may eventually shock. This result has been confirmed by BD03 using spherical hydrodynamical simulations. A similar conclusion was obtained earlier in one-dimensional simulations of pancake formation (Binney 1977).

DB06 have applied this theory in a cosmological context. They have assumed that the pre-shock gas traces the dark matter distribution in its free fall, and the strong shock condition determines the post-shock infalling speed, density and temperature. They infer a shock-heating scale of $M_{\rm shock} \sim 6 \times 10^{11}\,{\rm M}_{\odot}$ for a shock in the inner halo ($r \sim 0.1 r_{\rm vir}$), and $M_{\rm shock} \sim 3 \times 10^{12}\,{\rm M}_{\odot}$ for a shock at the virial radius.

Cosmological simulations with smoothed particle hydrodynamics (Fardal et al. 2001; K05) and with an adaptive Eulerian mesh (DB06) have shown that the phenomenon is not an artefact of the spherical or planar symmetry. Fig. 6 of K05, based on a simulation with zero metallicity, shows that the infall is almost all cold at $M_{\rm halo} \leqslant 10^{11}\,{\rm M}_{\odot}$, drops to 50 per cent at $M_{\rm shock} \simeq 3 \times 10^{11}\,{\rm M}_{\odot}$, and for $z \leqslant 2$ is almost all hot at $M_{\rm halo} \geqslant 10^{12}\,{\rm M}_{\odot}$. This is an underestimate of the critical mass by a factor of the order of 2 because of the assumed zero metallicity. At higher redshifts the cold fraction does not vanish even above the critical mass. At $z = 3$ it is $\sim$20 per cent even in the most massive halo contained in the computational box, $\gtrsim 10^{13}\,{\rm M}_{\odot}$.

We model the complex multiphase behaviour in an idealized way: a sharp transition from 100 per cent cold flow to a complete shutdown of the accretion at a critical halo mass $M_{\rm crit}(z)$. We Parametrize the critical mass by

$$M_{\rm crit} = M_{\rm shock} \times \max\left\{1,\ 10^{1.3(z-z_{\rm c})}\right\}. \qquad (3)$$

At redshifts below $z_{\rm c}$ the critical mass is the shock-heating mass, which is expected to be of the order of $10^{12}\,{\rm M}_{\odot}$. We treat its exact value as a free parameter determined by the best fit to different constraints. The growth of $M_{\rm crit}$ at high z is not meant to reflect a change in the threshold scale for spherical, virial shock heating, which is predicted to be quite insensitive to redshift (K05; DB06, fig. 7). Instead, it mimics the upper bound for haloes that permit substantial narrow cold streams above the threshold mass for a spherical shock. The functional form for the z dependence is motivated by the predictions of a simplified model by DB06 (fig. 7 and equation 40), although the critical redshift $z_{\rm c}$, predicted to be $\sim$1–3, is adjusted for best fit of the high-redshift data. While the increase of $M_{\rm crit}$ at high z is important for a good match of the $z \sim 3$ data, it has a negligible effect on the success of our model at low redshifts (Section 6).

5 NEW ELEMENTS IN THE GALAXY FORMATION SCENARIO

5.1 Shutdown by virial shock heating and AGN feedback

The critical stellar mass of $3 \times 10^{10}\,{\rm M}_{\odot}$ associated with the observed bimodality corresponds to a critical halo mass of $\lesssim 10^{12}\,{\rm M}_{\odot}$. Star forming galaxies are mainly confined to haloes below this mass and most galaxies in haloes above this threshold are evolving passively (Kauffmann et al. 2004). The main reason for the failure of the standard model at low redshift is the cooling of hot gas on to the central galaxies of haloes that are more massive than the critical scale. The traditional cooling scale (Rees & Ostriker 1977) is somewhat too big and is not associated with a transition that is sharp enough to explain the sharp colour bimodality and the steep drop of the luminosity function beyond L_*.

The gas that cools and accumulates in the central galaxies of massive haloes will inevitably form stars unless there is a mechanism that reheats it or removes it. AGNs provide the most plausible source of energy for the job (Binney & Tabor 1995; Ciotti & Ostriker 1997; Tucker & David 1997; Silk & Rees 1998; Nulsen & Fabian 2000; Springel, Di Matteo & Hernquist 2005). The supporting evidence comes from several lines of argument such as (i) the abundance of supermassive black holes in massive spheroids together with the correlation between black hole mass and bulge mass (Marconi & Hunt 2003; Häring & Rix 2004) or the velocity dispersion of the bulge (Merritt & Ferrarese 2001; Tremaine et al. 2002), (ii) the concordance between the quasar epoch and the stellar ages of early-type galaxies (Granato et al. 2001; Cattaneo & Bernardi 2003), (iii) the connection of black hole growth and elliptical galaxies with mergers (Toomre & Toomre 1972; Stockton 1999) and (iv) the observation of powerful outflows from active galaxies (e.g. McNamara et al. 2005).

Supernova feedback is not powerful enough to affect the gas in massive haloes significantly unless the rate of formation of massive stars is especially high, but then these galaxies would no longer be red. A similar argument can apply to feedback from quasars since early-type galaxies are 10^4 times more common than quasars at low redshifts (Wisotzki, Kuhlbrodt & Jahnke 2001). This is a serious problem for models in which AGNs are supposed to interact with the halo gas through radiative processes such as radiation pressure or Compton scattering. However, there is evidence showing that AGN outflows are possible not only during the main phase of black hole growth, which was very brief and happened at high redshift, but also when the black hole accretion rate is so low that the AGN is not optically luminous (see e.g. the adiabatic inflow–outflow solution by Blandford & Begelman 1999 and the study of the jet in M87 by Di Matteo et al. 2003). In the inflow–outflow model, the density of the accretion flow is too low for the gas to radiate efficiently. The accretion power is released mechanically, through particle jets, rather than radiatively, through the emission of optical/ultraviolet light. In this scenario, one possible way for the AGN to heat the gas is through shocks produced by the propagation of jets (e.g. Reynolds, Heinz & Begelman 2001; Fabian et al. 2003; Omma et al. 2004), although the problem of how the heat generated by this process is distributed in the hot gas remains open (see also Begelman & Nath 2005; Fabian et al. 2005, and references therein).

One difficulty in connecting AGN feedback to the galaxy bimodality is that the critical scale does not seem to affect black hole growth. For example, the correlation between black holes and bulges extends down to small bulges as in the Milky Way (Marconi & Hunt 2003; Häring & Rix 2004), while cooling seems to shut down only in massive galaxies.

Our proposal here (and in DB06) is that AGN feedback is switched on by the change in the large-scale black hole environment occurring once the gas is shock heated above the shock-heating scale. While the cold phase is fragmented in dense clouds, the hot phase is distributed much more uniformly in a low-density medium. Outflows from the AGN will preferentially propagate through the less resistive hot gas that fills the space between the clouds and leave the cold clouds behind instead of blowing them away (e.g. McKee & Ostriker 1977, in the case of supernova blast waves). The dilute hot gas is thus more vulnerable to the propagation of shock fronts. This serves as the basis for the new scenario simulated in the current paper.

We thus adopt the assumption that all massive galaxies contain a supermassive black hole, which can accrete gas and deposit energy in the surrounding gas, but the capacity of the gas to react to this injection depends on the presence of a shock-heated component. As soon as this condition is fulfilled, the gas begins to expand and the black hole accretion rate drops until it stabilizes at a value that over time compensates the radiative losses of the hot gas. The onset of this self-regulated accretion cycle prevents the shock-heated gas from ever cooling. We model this physical scenario by interpreting the critical mass that separates the regimes of cold and hot accretion equation (3) as a sharp cooling cut-off.

When most of the mass is in the hot phase, the cold clouds also become vulnerable to feedback. We therefore assume that when $M_{\rm halo} > M_{\rm crit}$ the cold gas in the central galaxy is reheated to the virial temperature of the halo. Black hole and supernova winds from gas-rich galaxy mergers will also sharpen the transition to a hot IGM because the cold streams in free fall will go through a shock when they hit the winds blown from the galaxy. We account for this scenario by preventing cold flows on galaxies in which the bulge is the main component as a massive bulge correlates with a massive black hole and, therefore, with a history of high AGN activity. This bulge-related condition has been introduced because it improves the details of the quantitative agreement with the SDSS colour–magnitude distribution, but we shall see below (Section 7) that it only plays a minor role in curing the discrepancies and achieving a good match with the observed features. We therefore include it as a provisional feature. However, differently from the halo mass criterion, we do not consider it an essential aspect of our model.

5.2 Cold streams and star formation at high redshifts

The standard scenario underpredicts the number of bright galaxies in the rest-frame 0.17 μm luminosity function at $z \sim 3$ (Fig. 6). This is due to a slow conversion of gas into stars, as it is not cured even when all the gas is allowed to cool in free fall from rest and supernova feedback is practically shut off. The shutdown of cooling in the most massive haloes will only accentuate the problem. The problem can be cured if the star formation recipes derived from local observations are revised to take into account the different way by which galaxies accrete their gas at high z.

In our 'new' model, the formation of stars at any redshift is restricted to galaxies undergoing cold-mode accretion (Section 4). However, there is a fundamental difference between the high and the low-redshift case. At high redshift, there is a lot of cold gas that streams in at the virial velocity along the dark matter filaments. At low redshift, the accretion rate decreases and the formation of stars continues until the gas reservoirs accumulated at earlier epochs have dried out.

At high redshift, the collision of the cold streams among themselves and with the galactic disc produces bursts of star formation analogous to the bursts resulting from the collision of two disc galaxies or cold gas clouds. Under the conditions that allow a cold flow, these collisions are expected to produce isothermal shocks, and the rapid cooling behind the shocks generates dense cold slabs in which the Jeans mass becomes small and stars can form efficiently. The detailed physics of star formation under these conditions is yet to be worked out, but this argument strongly suggests that the conversion of gas into stars must have been more efficient at high z than now that most of the cold gas is in dynamically relaxed discs. We mimic this effect in the 'new' model by introducing a $\sim(1+z)^{\alpha_*}$ increase in star formation efficiency equation (1) with α_* a free parameter determined by fitting the luminosity function of LBGs at $z \sim 3$.

6 RESULTS

We have revised the GalICS SAM to accommodate the new features discussed above, in particular (i) the high star formation rate due to cold streams at high redshift, (ii) the abrupt shutdown of galaxy accretion and star formation due to shock heating in haloes above a critical mass and (iii) the maintenance of this shutdown by AGN feedback that couples to the hot gas. This involves three new free SAM parameters, which are fine-tuned to reproduce the observational constraints. They are: (i) the shock-heating mass scale, $M_{\rm shock}$, (ii) the critical redshift $z_{\rm c}$ prior to which $M_{\rm crit}$ is growing with z equation (3) and (iii) the power index α_* in the redshift dependence of the star formation rate equation (1). The shock-heating scale is the single most important parameter determining the fit at low and intermediate redshifts. As seen in the $M_{\rm r}$–$M_{\rm halo}$ diagram of Fig. 1, the critical halo mass translates into a maximum luminosity for a star forming galaxy, where the main inclined stripe marking the brightest galaxies intersects the vertical red line. This parameter is tuned to match the number density of blue ($u - r \lesssim 2$) galaxies with $M_{\rm r} \sim -22.25$ as observed in the SDSS. The other two parameters are determined by fitting the luminosity function of LBGs at $z \sim 3$ (Steidel et al. 1999). The obtained best-fitting values are $M_{\rm shock} = 2 \times 10^{12}\,{\rm M}_\odot$, $z_{\rm c} = 3.2$, $\alpha_* = 0.6$.

With this choice of the parameter values, we recover an excellent match to the $z = 0$ luminosity function (Fig. 2) and to the bimodal colour–magnitude distribution in the SDSS (Figs 3 and 4). Fig. 3 shows that the blue sequence is properly truncated at $M_{\rm r} \sim -22.25$. The central galaxies of massive haloes are also positioned correctly in the SDSS red sequence colour range (Fig. 3). This figure shows the model predictions at $M_{\rm r} < -20$. The incompleteness at fainter magnitudes comes from the fact that the N-body simulation used to construct the dark matter merger trees does not resolve haloes with $M_{\rm halo} \lesssim 2 \times 10^{11}\,{\rm M}_\odot$ (Fig. 1).

The colour histograms (Fig. 4) show that the simulated joint colour–magnitude distribution agrees with the observed distribution both in shape and in normalization. We now reproduce the red population observed at $z \sim 1$ in the COMBO-17 survey and the fact that the brightest galaxies are indeed on the red sequence (Fig. 5).

The model provides a good fit to the cosmological star formation history as observed, e.g. by Giavalisco et al. (2004), but only a fair fit to some of the other observational estimates, especially at $z \lesssim 1$ (Hopkins 2004; Fig. 8). The difference from assuming constant $M_{\rm crit}$ at all z is small because only a small fraction of the cosmic star formation is in haloes above $M_{\rm shock}$. At $z < z_{\rm c}$ the model with constant $M_{\rm crit}$ and the 'new' model are identical. At $z \gg z_{\rm c}$ the simulation contains no haloes with $M_{\rm halo} > 2 \times 10^{12}\,{\rm M}_\odot$.

It also provides an excellent fit to the comoving number density of bright star forming galaxies at $z \sim 3$ (Fig. 6). In this comparison, one should keep in mind that LBGs at $z \sim 3$ suffer from substantial dust extinctions and the quality of the fit is extinction-model dependent (Section 2). A more extensive comparison of model and data at high z is desirable but is beyond the scope of the current paper, which concentrates on the low-z bimodality. Our current comparison to high-z data is only a preliminary effort to identify the qualitative trends.

7 SENSITIVITY TO MODEL INGREDIENTS

Here, we explore the contribution of different ingredients to the success of the model in fitting the observed colour–magnitude distribution.

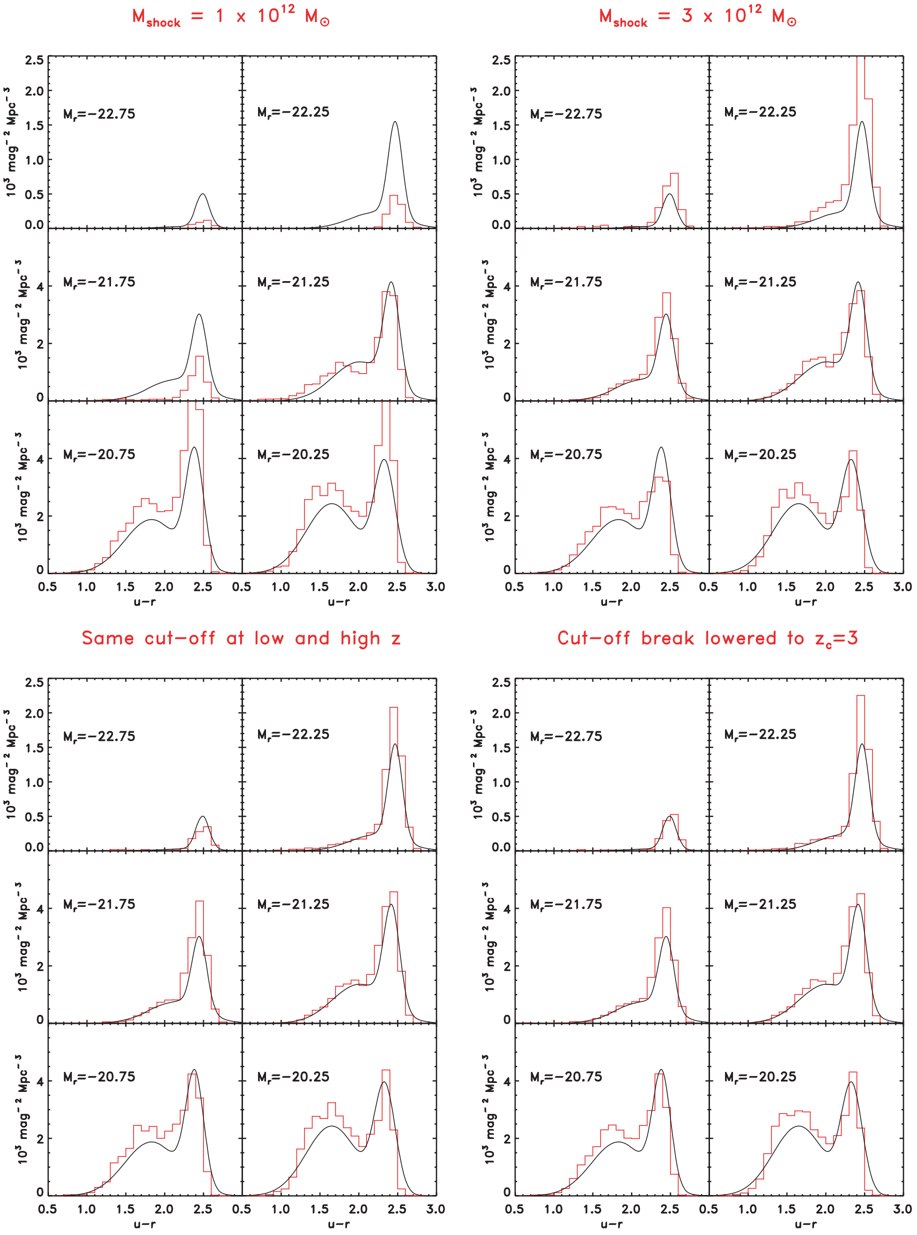
7.1 The critical mass at low z

The shutdown above the critical halo mass is the main new feature of our model, and thus $M_{\rm shock}$ is the key parameter. Fig. 1 compares the distribution of galaxy luminosity versus halo mass in the 'standard' model and our 'new' model. Preventing cooling and star formation in haloes above $M_{\rm shock}$ keeps the magnitudes of these galaxies fainter than $M_{\rm r} \sim -23$. The abrupt change in the mass-to-light ratio at the critical scale separates the red sequence from the blue sequence and sets an upper limit to the luminosity of galaxies in the blue sequence. The best fit to the data, with $M_{\rm shock} = 2 \times 10^{12}\,{\rm M}_{\odot}$, is shown in Figs 4 and 9 shows how by lowering (or raising) $M_{\rm shock}$ to $10^{12}\,{\rm M}_{\odot}$ (or $3 \times 10^{12}\,{\rm M}_{\odot}$) the model reproduces too few (or too many) galaxies at $M_{\rm r} = -22.25$.

7.2 The critical mass at high *z*

At high *z*, our fiducial 'new' model with $z_{\rm c} \sim 3$ allows star formation even in haloes above $M_{\rm shock}$ equation (3), mimicking cold streams in hot haloes. This change is responsible for the appearance in Fig. 7 of a population of bright galaxies with a high star formation rate at $z \sim 3$ (compare to Sawicki & Thompson 2005), while the red sequence of passively evolving galaxies has not formed yet.

The model predictions for LBGs are sensitive to the value of $z_{\rm c}$. When the increase of $M_{\rm crit}$ at high redshifts is not taken into


$M_{\rm shock}$ = 1 x 10^{12} $M_{\odot}$

$M_{\rm shock}$ = 3 x 10^{12} $M_{\odot}$

Same cut-off at low and high z

Cut-off break lowered to z_c=3

Figure 9. The effect of varying different model ingredients on the joint distribution of colour and magnitude at $z = 0$. Shown is the $u - r$ colour distribution in *r*-band magnitude bins for the model predictions (red histogram) compared to the SDSS data (black smoothed histogram) from Baldry et al. (2004). The models are the fiducial 'new' model of Fig. 4 with one ingredient varied as follows: (a) $M_{\rm shock} = 10^{12\,{\rm M}\odot}$ (compared to $2 \times 10^{12\,{\rm M}\odot}$ in the fiducial model). (b) $M_{\rm shock} = 3 \times 10^{12}\,{\rm M}_{\odot}$. (c) $z_{\rm c} \to \infty$, i.e. same $M_{\rm crit}$ at all *z*. (d) $z_{\rm c} = 3$, somewhat lower than the fiducial value. (e) No dominant-bulge requirement for shutdown of cooling and star formation. (f) No ejection of the remaining cold gas when cooling is shut off. (g) No increase in star formation efficiency at high redshift ($\alpha_* = 0$). (h) An increased efficiency of supernova feedback ($\epsilon_{\rm SN} = 0.25$).

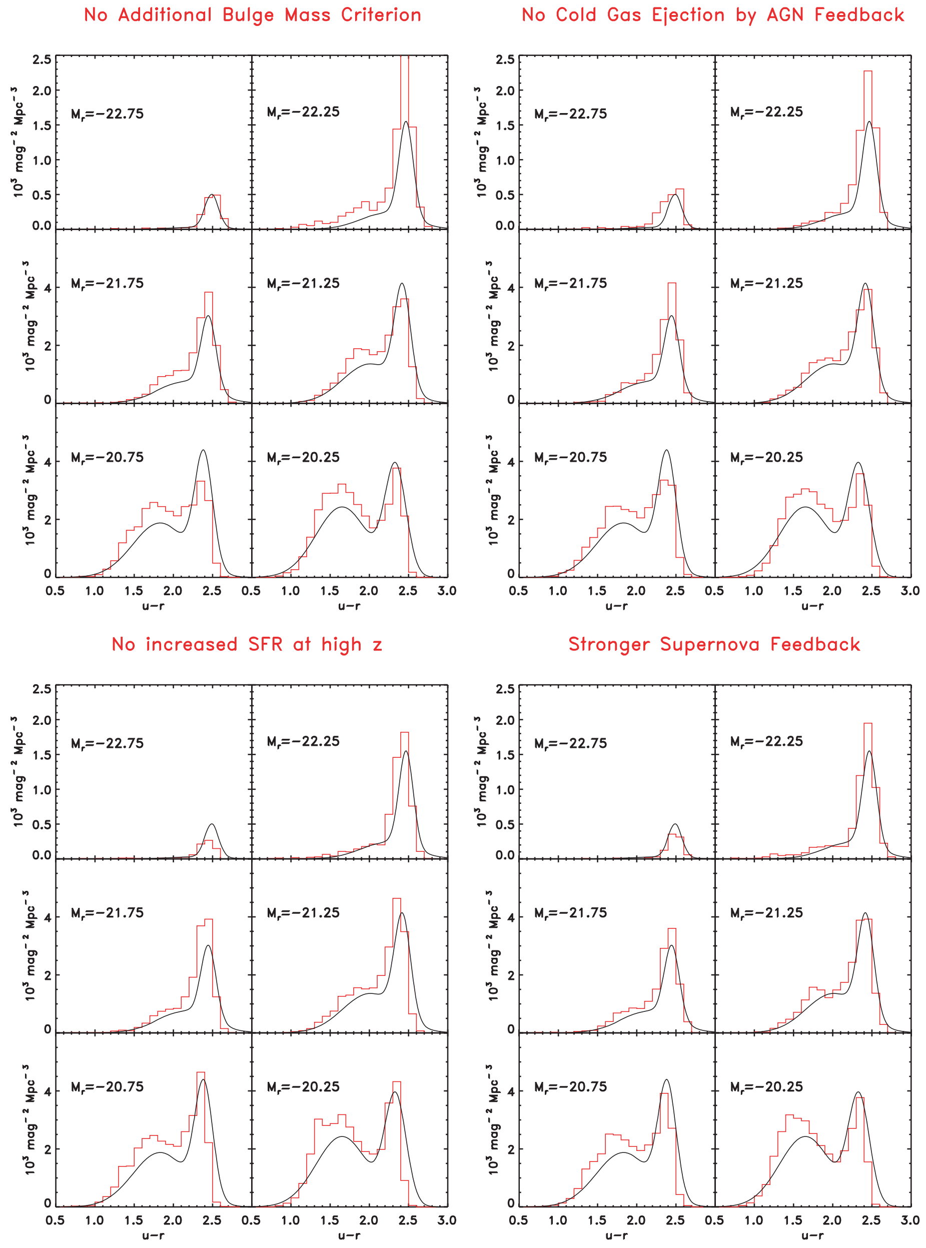


Figure 9 – *continued*

account, the predicted luminosity function at $z \sim 3$ falls short at the bright end compared to that of bright LBGs ($z_c \rightarrow \infty$ in Fig. 6). On the other hand, when z_c is taken to be smaller than the fiducial value of 3.2, the predicted bright end of the luminosity function overshoots that of LBGs as observed by Steidel et al. (1999). If, however, the actual comoving density of bright LBGs is somewhat higher, as indicated by the VIMOS/VLT Deep Survey (VVDS) team (private communication), then $z_c = 3.0$ becomes our best-fitting value (Fig. 6). With this value, the match to the r-band luminosity function at $z = 0$ becomes even better (Fig. 2), but in general the effect of z_c on the results at low redshifts is negligible (Figs 9c and d).

7.3 A bulge criterion for shutdown

Our fiducial model allows no cooling or star formation in galaxies where the bulge is the dominant component. In practice, given the resolution limit of the N-body simulation, the critical shutdown scale is reduced, in this case, to the resolution scale of $\sim 2 \times 10^{11}$ M$_\odot$. This scale is significantly lower than the best-fitting value of $M_{\rm shock}$, but is only slightly lower than the critical halo mass as predicted by theory (BD03; K05; DB06) and as observed (Kauffmann et al. 2003). As explained in Section 4, our $M_{\rm shock}$ is closer to an upper limit than to a mean value for the physical shock-heating scale.

When the effect of the bulge fraction is eliminated from the criterion for shutdown (Fig. 9e), the model shows a small excess of bright galaxies ($M_{\rm r} \sim -22.25$), both blue and red. This is because the more efficient cooling allows more massive star forming galaxies, which in turn speeds up the merger rate due to more efficient dynamical friction. We learn that the bulge criterion has some effect, but it basically serves as a secondary criterion that helps fine-tuning the match to the data. The fit is fairly adequate without it.

7.4 Cold gas ejection in shock-heated haloes

In the fiducial model, once the halo mass grows above the critical mass, all the gas remaining in the central galaxy is heated to the virial temperature and all modes of star formation are shut off. The idea is that when most of the gas is hot, AGN outflows heat, destroy or blow away the cold clouds as well. When the cold gas is not ejected from massive galaxies but rather allowed to form stars later, the model predicts a small excess of blue galaxies in the brightest bin (Fig. 9f). This is similar to the effect of eliminating the bulge criterion, but even less noticeable.

7.5 Efficient star formation at high redshift

We have shown that increasing the star formation efficiency at high z is necessary to reproduce the luminosity function of LBGs. When the star formation is kept constant with redshift, instead, star formation is postponed to later epochs. The model predicts a shortage of galaxies in the brightest bin and the colours become slightly too blue, but these are weak effects (Fig. 9g).

7.6 Supernova feedback

Our fiducial model for supernova feedback is similar to the model used in most semi-analytic models based on Dekel & Silk (1986). It is more efficient in less massive galaxies with shallower potential wells. The naive expectation is that a stronger supernova feedback would make the galaxies redder by removing gas and stopping star formation. Instead, we find that supernova feedback makes galaxies slightly bluer by ejecting metals from the galaxies into the IGM (Fig. 9h). The metal-enriched gas is later accreted on to the hierarchically assembled more massive galaxies, and it is kept there as supernova feedback becomes less efficient. Thus, supernova feedback ends up transferring metals from low-mass galaxies into massive galaxies.

We conclude that the main element responsible for the improved match of the model to the colour–magnitude data at low redshifts is the abrupt shutdown above a critical mass of $\sim 10^{12}\,{\rm M}_{\odot}$. All the other changes in the model recipes, including the bulge criterion, are of secondary importance. They serve for fine-tuning the model in order to achieve a nearly perfect match.

8 CORRELATIONS WITH OTHER PROPERTIES

The observed bimodality of the galaxy population, which is very apparent in the colour–magnitude diagram, also involves all other global properties of galaxies. We already saw in Fig. 3 that there is a strong correlation between being blue or red and being below or above the critical halo mass. In Fig. 10, we explore a variety of such correlations as predicted by the 'new' model at $z = 0$. In the same colour–magnitude diagram shown in each panel, colour refers to a different galaxy property. The properties considered here include environment density via halo mass, bulge fraction, stellar age, stellar metallicity, total stellar mass and star formation rate.

A useful way to interpret the various correlations displayed in Fig. 10 is to consider the evolutionary track of a typical galaxy in the colour–magnitude diagram. Panels (a) and (e) show that the blue sequence is a sequence of growing halo mass and stellar mass from bottom right-hand side to top left-hand side. As a halo is growing in mass, starting from the smallest mass resolved by the N-body simulation, gas is accumulating in the disc of the central galaxy and stars are forming. The galaxy becomes brighter because its stellar mass increases, but it also gets redder because a growing fraction of the stellar mass is in old stars. In this evolutionary phase, the galaxy moves along a stripe which roughly coincides with the blue sequence of the SDSS marked by the blue lines in Fig. 10.

A galaxy reaches the top of the blue sequence when its host halo mass becomes comparable to $M_{\rm crit}$. Soon after, it ceases to accrete gas, star formation shuts off, and the galaxy evolves passively to become red and dead. Most galaxies never reach the top of the blue sequence because they merge into a halo much more massive than their own before their original halo exceeds the critical mass. These galaxies are typically of low mass compared to the halo into which they have merged, and therefore their dynamical friction time for sinking into the new halo centre is long. They become the 'satellite' galaxies populating groups and clusters of galaxies. Since satellite galaxies are assumed not to accrete gas, they fairly quickly exhaust their remaining gas and become passive. Their brightness fades and their colour becomes redder, as the most massive, bluest stars are the first to die. These evolutionary tracks can be deduced, for example, from the stellar mass panel of Fig. 10, where galaxies evolve passively along the equal colour stripes, which stretch roughly perpendicular to the blue sequence, from bottom left-hand side to top right-hand side. By mentally superimposing these diagonal equal-stellar-mass stripes on the panel of Fig. 10 referring to halo mass, one can verify that the evolution up along such stripes involves an increase in halo mass, consistent with merging into a more massive halo. A similar inspection of Fig. 3 shows that this is indeed a transition from a central galaxy to a satellite.

Fading into the red sequence need not be the end point. Red galaxies, in hot massive haloes, do not accrete gas, but they can still merge with other galaxies. The highest merger rates are for massive galaxies in the bright part of the red sequence, as higher mass implies stronger dynamical friction. Thus the more massive a galaxy is when it leaves the blue sequence, the more it will continue growing by merging after it has joined the red sequence. Repeated mergers with other massive galaxies allow the formation of giant ellipticals with stellar masses of $\sim 10^{12}\,{\rm M}_{\odot}$, which would be unattainable through simple accretion of gas in the presence of a limit on cooling. Most of this growth is due to dissipationless merging within the red sequence. This interpretation of the role of merging and passive evolution is supported by the dominance of central galaxies in the bright part of the red sequence and of satellite galaxies in the faint side, separated roughly at L_*, the characteristic luminosity of a galaxy at the top of the blue sequence (Fig. 3). It is also supported by the variation of morphology with position in the colour–magnitude diagram (Fig. 10b). Blue galaxies are mainly spirals, with bulge-to-disc ratio that increases with luminosity. The morphological composition along the red sequence also varies from mostly spiral galaxies at $M_{\rm r} \sim -20$ to mostly elliptical galaxies at $M_{\rm r} \sim -22.5$, evidence for an increasing merger rate from faint to bright galaxies.

The correlation of colour and magnitude with morphology is a generic outcome of the merger scenario and is not an artefact of the bulge criterion for shutdown. This is demonstrated in Fig. 11,

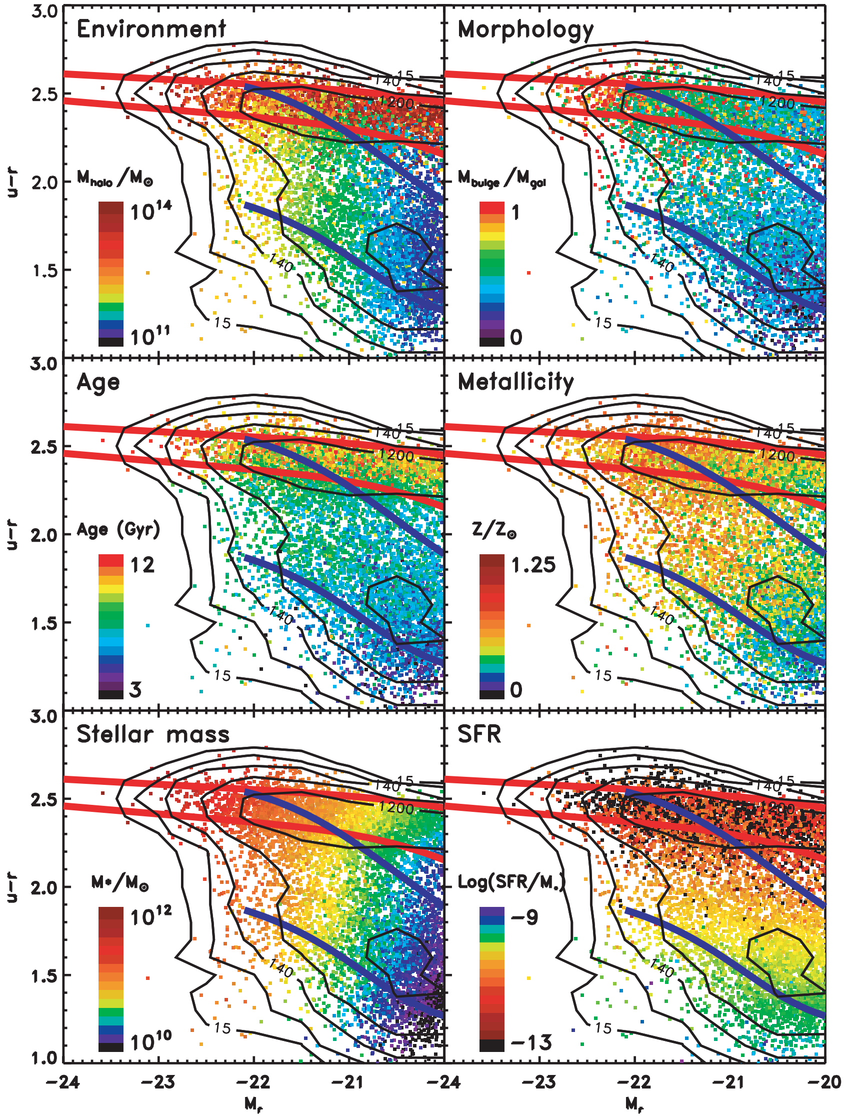


Figure 10. Correlation of colour–magnitude with other properties in the 'new' model at $z = 0$. As in Fig. 3, the points mark individual galaxies in the simulation and the contours refer to their number density in the colour–magnitude plane (mag^{-2} Mpc^{-3}), while the blue and red lines mark the blue and red sequences in the SDSS. All the above are the same in all six panels. The colour coding in each panel represents a different galaxy property as follows: (a) Mass of the host halo (log scale), which is correlated with the environment density. (b) Morphology, quantified by the bulge-to-total mass ratio. (c) Mass-weighted stellar age. (d) Mass-weighted stellar metallicity. (e) Total stellar mass (log scale). (f) Star formation per year per unit stellar mass (log scale). Black corresponds to galaxies with no star formation at all.

in which the bulge criterion was eliminated. The central galaxies of group and cluster size haloes, shown as red dots at the bright-end of the red sequence in Fig. 3 are those with the most intense merger history and therefore with the highest bulge-to-disc ratio. A qualitatively similar correlation with morphology is valid already in the 'standard' model without cooling shutdown.

The presence of two populations within the red sequence manifests itself also in the environment panel of Fig. 10. The galaxies that live in the most massive haloes (10^{13-14} M$_\odot$) dominate both the bright end and the faint end of the red sequence, but the middle range between $M_{\rm r} = -21$ and -22.5 shows many middle-of-the-range haloes of 10^{12-13} M$_\odot$ and even smaller. This means that as the halo mass grows in time the galaxy can evolve according to one of the following two tracks: it may either keep on growing until it becomes a bright E/S0 galaxy or fade to become an early-type satellite with a significant disc. This prediction of our model is consistent with the SDSS results (Blanton et al. 2006), where the number density in the 1 h^{-1} Mpc environment of galaxies is measured to peak both at the bright end and at the faint end of the red sequence while the morphology shows a gradient along the red sequence.

We learn from the star formation rate panel of Fig. 10 that model galaxies are red because they are not forming stars and thus lack a young stellar population. Indeed, the distinction between the red sequence and the blue sequence is predominantly due to age rather than metallicity, as can be deduced by comparing the gradients at a given luminosity in the age panel and the metallicity panel of Fig. 10. However, the colour gradients along the red sequence and the blue sequence themselves are driven by the metallicity gradient more than by the age gradient.

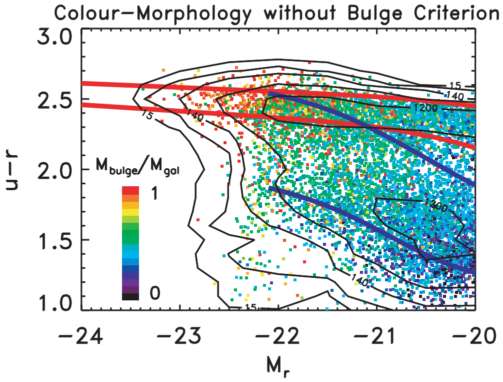


Figure 11. Morphology variations in a model with no bulge criterion for cooling shutdown. The similarity to the bulge panel in Fig. 10 indicates that the bulge criterion is of secondary importance to the correlations between colour, luminosity and morphology.

9 CONCLUSION

We have addressed the origin of the robust division of the galaxy population into two major types: the blue sequence of gas-rich galaxies with a young stellar population and the red sequence of gas-poor galaxies dominated by old stars. While the blue sequence is truncated beyond a characteristic stellar mass of $\sim 3 \times 10^{10}\,\mathrm{M}_{\odot}$, the red sequence extends to higher stellar masses. This bimodality is strongly correlated with the morphological type and with the galaxy density in the environment. Large galaxy surveys have revealed that the bimodality is reflected in almost every other property of galaxies and that the separation between the two types is fairly sharp. High-redshift surveys show that a fraction of the red sequence is in place already at $z \sim 1$–2, and that massive star forming galaxies exist at $z \sim 2$–4.

We find that models of galaxy formation can successfully reproduce the bimodality once a complete shutdown of cooling and star formation is imposed in haloes above a critical mass of $\sim 10^{12}\,\mathrm{M}_{\odot}$ starting at $z \lesssim 3$. The time evolution of the bimodality is reproduced when allowing partial cooling and efficient star formation above the critical mass at higher redshifts. By incorporating these simple new features in hybrid semi-analytic/*N*-body simulations, we have achieved an unprecedented simultaneous match to the joint distributions of galaxy properties at low and high redshifts, practically confirming the qualitative predictions of DB06.

The motivation for the sharp transition at a critical halo mass comes from theoretical analysis (BD03; Binney 2004; DB06) and cosmological simulations (K05; DB06) showing a sharp transition from galaxy build-up by cold filamentary flows in small haloes to virial shock heating of the halo gas in massive haloes. The abrupt shutdown in the supply of cold gas for star formation, due to the abrupt appearance of stable virial shocks as opposed to the smooth variation of cooling time with mass, is responsible for the very red colours of the red sequence, the colour gap between the red and the blue sequence, and the sharp truncation of the blue sequence above L_*.

An important input for the success of the model is that, once the gas is heated to the virial temperature, it never cools. In this model shock heating serves as the trigger for efficient AGN feedback, which maintains the gas hot. The long-term effect may be associated with a self-regulated accretion cycle maintained by the coupling between the central black hole and the hot gas. The luminous quasar mode is associated with the rapid build-up of massive black holes probably due to the high gas inflow rate triggered by galaxy mergers. The resulting high output power may be important in terminating star formation and ending the quasar phase itself (e.g. Di Matteo et al. 2005), but this phase is short-lived and cannot serve to maintain the gas hot for long periods. The self-regulation is achieved when the accretion rate drops to an equilibrium value at which the energy ejected from the black hole and absorbed by the gas compensates for the radiative losses (Cattaneo et al., in preparation). The mechanical efficiency of this mode is close to 100 per cent, while in the quasar mode most of the output power goes into radiation. In the quiet mode, the dense clumps of cold gas are not swift in responding to the external injection of mechanical energy – they can be destroyed or blown away only when the AGN is sufficiently powerful as in a quasar. On the other hand, the hot gas, which feeds the black hole in a gradual fashion, is more dilute, and can therefore respond effectively to the injected energy. Therefore, the self-regulated mode of hot accretion may be responsible for a significant fraction of the mechanical energy affecting the gas, even if the hot gas contributes only a small fraction of the final black hole mass.

In parallel to this work, Croton et al. (2006) and Bower et al. (2006) have implemented related semi-analytic models, in which they also address the origin of the galaxy bimodality and the effects of AGN feedback in galaxy formation. Croton et al. (2006) focus on the growth of black holes and AGNs as feedback sources. In particular, they assume a transition from rapid to slow cooling and associate the accretion of cold and hot gas with an ‘optical’ and a ‘radio’ mode of black hole growth, respectively. The latter is proposed to provide the relevant AGN feedback process, which determines a gradual shutdown of cooling in massive haloes. Bower et al. (2006) and we focus instead on the thermal properties of the gas as the trigger for an abrupt shutdown above a threshold halo mass, which translates into the bimodality scale in the galaxy properties. In our study, the transition from cold filamentary flows to a hot spherical halo provides a physical model for the sharp shutdown at the critical scale. The key to our successful modelling of the galaxy properties is this robust shutdown, regardless of the details of AGN feedback. This simple addition to the model results in an unprecedented quantitative agreement with the observed joint distribution of magnitude and colour for galaxies. The fact that the three independent modelling efforts manage to reproduce the key observational constraints after incorporating a shutdown based on related but not identical physical processes confirms the robustness of the shutdown scenario simulated here.

Our ‘new’ picture introduces a strong correlation between the spectral galaxy type and the host halo mass, while the morphological type is driven by the galaxy merger history as in the ‘standard’ model. Thus, while the new model gas fraction and colour of massive galaxies are different, the morphological distribution and the correlation between morphology and environment density are reproduced as in the standard model. The natural correlation between the mass of the host halo and the density of the environment, according to the ‘halo model’ of galaxy distribution (Yan, Madgwick & White 2003; Kravtsov et al. 2004; Abazajian et al. 2005), is reflected in the correlation between spectral and morphological type.

A few comments on the evolution in the colour–magnitude diagram, as summarized in Fig. 12 (see also Faber et al. 2006 and DB06), are the following ones. Spiral galaxies grow discs by cold filamentary streams at the centres of haloes below the critical mass. As the accreted gas is converted into stars, the galaxies become brighter and redder along the blue sequence. A galaxy leaves the blue sequence either when its halo grows above the critical mass or when it merges into such a halo and becomes a satellite galaxy. In

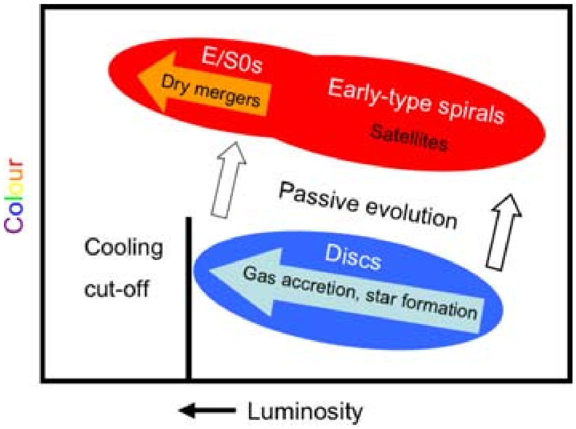


Figure 12. A schematic sketch of galaxy evolution. Galaxies grow along the blue sequence through cold filamentary flows until they stop accreting gas because their host halo has grown above the critical shock-heating mass. At that point they move into the red sequence by becoming red and fading in luminosity. The shutdown of cooling above the critical halo mass sets an upper limit to the luminosity of blue galaxies and explains the characteristic galaxy luminosity L_*. Galaxies get to the E/S0 bright end of the red sequence either by dissipationless ('dry') mergers along the red sequence, or by gas-rich ('wet') mergers from the top of the blue sequence. This simple cartoon ignores the complications associated with the temporary reddening by dust extinction in star forming galaxies.

both cases, the galaxy ceases to accrete gas from its hot environment and rapidly turns red and dead. Small satellite galaxies tend to fade as they redden and turn into the red discs observed at the faint end of the red sequence (Blanton et al. 2006). Big central spirals whose haloes have grown above the critical mass continue their growth along the red sequence through dissipationless mergers and evolve into the giant ellipticals at the bright end of the red sequence. Gas-rich mergers may also exhaust the cold gas reservoir of the merging galaxies and transform two blue spirals into a red elliptical, but this is a less frequent path in the formation of massive early-type galaxies. This picture implies that the colour difference between the blue and red sequences is primarily due to stellar age, while the colour–magnitude gradient within the red sequence is largely a metallicity effect, since deeper potential wells retain their metals more effectively. The colour–magnitude gradient along the blue sequence is a combination of the two effects.

A detailed study of the origin of the red sequence via evolutionary tracks in the colour–magnitude diagram is described by Cattaneo, Dekel & Faber (in preparation). At high redshifts, the enhanced star formation rate that we have incorporated in galaxies of all masses helps curing the difficulty of the standard scenario in forming enough stars in massive galaxies at $z \sim 2$–4, as indicated by the observations of LBGs and SCUBA sources. This new ingredient is motivated by the build-up of galaxies via nearly supersonic filamentary flows (BD03; K05; DB06). The collisions of these streams among themselves and with the central discs are likely to trigger efficient bursts of star formation, in analogy to collisions of discs or cold clouds. However, the fate of cold streams in hot haloes and the resulting star formation are yet to be studied in detail. We note that our scenario helps explaining the 'downsizing' of the formation of massive galaxies, where more massive galaxies tend to form stars earlier and over shorter periods. The most massive galaxies, those above the shock-heating mass, have formed their stars at $z >$ 2–3 by cold flows in hot media, and stopped forming stars at lower redshifts. Galaxies below the critical mass continue to form stars at lower redshifts.

While the simple prescriptions that we have used for star formation and its shutdown are clearly only crude approximations, their success in simultaneously matching all the observed bimodality features at different redshifts indicates that they fairly represent the effects of the complex physical processes involved. This motivates a detailed study of the processes of cold-flow-induced star formation and the subsequent shutdown above the critical mass at late times due to the coupling between virial shock heating and AGN feedback.

ACKNOWLEDGMENTS

We acknowledge stimulating discussions with J. Binney, Y. Birnboim, S. M. Faber, A. Kravtsov, C. Maulbetsch, E. Neistein, J. R. Primack, M. J. Rees and R. Somerville. This research has been supported by ISF 213/02 and NASA ATP NAG5-8218. AC has been supported by a Golda Meir Fellowship at HU. AD acknowledges support from a Miller Visiting Professorship at UC Berkeley, a Visiting Professorship at UC Santa Cruz and a Blaise Pascal International Chair by the École Normale Supérieure at the Institut d'Astrophysique de Paris.

REFERENCES

Abazajian K. et al., 2005, ApJ, 625, 613
Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Balogh M. L., Baldry I. K., Nichol R., Miller C., Bower R., Glazebrook K., 2004, ApJ, 615, L101
Begelman M. C., Nath B. B., 2005, MNRAS, 361, 1387
Bell E. F., McIntosh D. H., Katz N., Weinberg M. D., 2003, ApJS, 149, 289
Bell E. F. et al., 2004, ApJ, 608, 752
Benson A. J., Pearce F. R., Frenk C. S., Baugh C. M., Jenkins A., 2001, MNRAS, 320, 261
Binney J., 1977, ApJ, 215, 483
Binney J., 2004, MNRAS, 347, 1093
Binney J., Merrifield M., 1998, Galactic Astronomy. Princeton Univ. Press, Princeton, NJ
Binney J., Tabor G., 1995, MNRAS, 276, 663
Birnboim Y., Dekel A., 2003, MNRAS, 345, 349 (BD03)
Blaizot J., Guiderdoni B., Devriendt J. E. G., Bouchet F. R., Hatton S. J., Stoehr F., 2004, MNRAS, 352, 571
Blandford R. D., Begelman M. C., 1999, MNRAS, 303, L1
Blanton M. R., Eisenstein D. J., Hogg D. W., Zehavi I., 2006, preprint (astro-ph/0411037)
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nat, 311, 517
Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645
Brüggen M., Ruszkowski M., Hallman E., 2005, ApJ, 630, 740
Cattaneo A., Bernardi M., 2003, MNRAS, 344, 45
Cattaneo A., Combes F., Colombi S., Bertin E., Melchior A.-L., 2005, MNRAS, 359, 1237
Cattaneo A. et al., 2006, MNRAS, submitted
Chapman S. C., Blain A. W., Ivison R. J., Smail I. R., 2003, Nat, 422, 695
Chapman S. C., Smail I., Blain A. W., Ivison R. J., 2004, ApJ, 614, 671
Ciotti L., Ostriker J. P., 1997, ApJ, 487, L105+
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Croton D. J. et al., 2006, MNRAS, 365, 11
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dehnen W., Binney J., 1998, MNRAS, 294, 429
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Dekel A., Silk J., 1986, ApJ, 303, 39
Devriendt J. E. G., Guiderdoni B., Sadat R., 1999, A&A, 350, 381

Di Matteo T., Croft R. A. C., Springel V., Hernquist L., 2003, ApJ, 593, 56
Di Matteo T., Springel V., Hernquist L., 2005, Nat, 433, 604
Faber S. M. et al., 2006, ApJ, submitted (astro-ph/0506044)
Fabian A. C., Sanders J. S., Allen S. W., Crawford C. S., Iwasawa K., Johnstone R. M., Schmidt R. W., Taylor G. B., 2003, MNRAS, 344, L43
Fabian A. C., Reynolds C. S., Taylor G. B., Dunn R. J. H., 2005, MNRAS, 363, 891
Fardal M. A., Katz N., Gardner J. P., Hernquist L., Weinberg D. H., Davé R., 2001, ApJ, 562, 605
Giavalisco M. et al., 2004, ApJ, 600, L103
Granato G. L., Silva L., Monaco P., Panuzzo P., Salucci P., De Zotti G., Danese L., 2001, MNRAS, 324, 757
Granato G. L., De Zotti G., Silva L., Bressan A., Danese L., 2004, ApJ, 600, 580
Guiderdoni B., Hivon E., Bouchet F. R., Maffei B., 1998, MNRAS, 295, 877
Häring N., Rix H., 2004, ApJ, 604, L89
Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, MNRAS, 343, 75
Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., Pearce F. R., 2003, MNRAS, 338, 913
Helsdon S. F., Ponman T. J., 2003, MNRAS, 340, 485
Hernquist L., 1990, ApJ, 356, 359
Hogg D. W. et al., 2004, ApJ, 601, L29
Hopkins A. M., 2004, ApJ, 615, 209
Hubble E. P., 1926, ApJ, 64, 321
Humason M. L., 1936, ApJ, 83, 10
Im M. et al., 2002, ApJ, 571, 136
Kannappan S. J., 2004, ApJ, 611, L89
Kauffmann G. et al., 2003, MNRAS, 341, 54
Kauffmann G., White S. D. M., Heckman T. M., Ménard B., Brinchmann J., Charlot S., Tremonti C., Brinkmann J., 2004, MNRAS, 353, 713
Kennicutt R. C., 1983, ApJ, 272, 54
Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2 (K05)
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottloeber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Magorrian J. et al., 1998, AJ, 115, 2285
Marconi A., Hunt L. K., 2003, ApJ, 589, L21
Mathews W. G., Brighenti F., 2003, ARA&A, 41, 191
McKee C. F., Ostriker J. P., 1977, ApJ, 218, 148
McNamara B. R., Nulsen P. E. J., Wise M. W., Rafferty D. A., Carilli C., Sarazin C. L., Blanton E. L., 2005, Nat, 433, 45
Merritt D., Ferrarese L., 2001, ApJ, 547, 140
Moustakas L. A. et al., 2004, ApJ, 600, L131
Ninin S., 1999, PhD thesis, Univ. Paris 11
Nulsen P. E. J., Fabian A. C., 2000, MNRAS, 311, 346
Omma H., Binney J., Bryan G., Slyz A., 2004, MNRAS, 348, 1105
Osmond J. P. F., Ponman T. J., 2004, MNRAS, 350, 1511
Rees M. J., Ostriker J. P., 1977, MNRAS, 179, 541
Reynolds C. S., Heinz S., Begelman M. C., 2001, ApJ, 549, L179
Ruszkowski M., Brüggen M., Begelman M. C., 2004, ApJ, 615, 675
Sawicki M., Thompson D., 2005, ApJ, 635, 100
Shapley A. E., Erb D. K., Pettini M., Steidel C. C., Adelberger K. L., 2004, ApJ, 612, 108
Silk J., 1977, ApJ, 211, 638
Silk J., Rees M. J., 1998, A&A, 331, L1
Smail I., Ivison R. J., Blain A. W., Kneib J.-P., 2002, MNRAS, 331, 495
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Springel V., Di Matteo T., Hernquist L., 2005, ApJ, 620, L79
Steidel C. C., Adelberger K. L., Giavalisco M., Dickinson M., Pettini M., 1999, ApJ, 519, 1
Stockton A., 1999, in Barnes J. E., Sanders D. B., eds, Proc. IAU Symp. 186, Galaxy Interactions at Low and High Redshift. Kluwer, Dordrecht, p. 311
Strateva I. et al., 2001, AJ, 122, 1861
Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253
Tabor G., Binney J., 1993, MNRAS, 263, 323
Toomre A., Toomre J., 1972, ApJ, 178, 623
Tremaine S. et al., 2002, ApJ, 574, 740
Tucker W., David L. P., 1997, ApJ, 484, 602
van den Bosch F. C., 1998, ApJ, 507, 601
van der Marel R. P., 1999, in Barnes J. E., Sanders D. B., eds, IAU Symp. 186, Galaxy Interactions at Low and High Redshift. Kluwer, Dordrecht, p. 333
Weiner B. J. et al., 2005, ApJ, 620, 595
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
Willmer C. N. A. et al., 2006, ApJ, in press (astro-ph/0506041)
Wisotzki L., Kuhlbrodt B., Jahnke K., 2001, in Márquez I., Masegosa J., del Olmo A., Lara L., García E., Molina J., eds, QSO Hosts and Their Environments. Kluwer, Dordrecht, p. 83
Yan R., Madgwick D. S., White M., 2003, ApJ, 598, 848
Yoshida N., Stoehr F., Springel V., White S. D. M., 2002, MNRAS, 335, 762

This paper has been typeset from a TEX/LATEX file prepared by the author.