Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **465,** 3291–3308 (2017)
Advance Access publication 2016 November 22

doi:10.1093/mnras/stw2944

Simulating galaxy formation with black hole driven thermal and kinetic feedback

Rainer Weinberger,[1]★ Volker Springel,[1,2] Lars Hernquist,[3] Annalisa Pillepich,[3,4] Federico Marinacci,[5] Rüdiger Pakmor,[1] Dylan Nelson,[6] Shy Genel,[7]† Mark Vogelsberger,[5] Jill Naiman[3] and Paul Torrey[5,8]‡

[1]*Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany*
[2]*Zentrum für Astronomie der Universität Heidelberg, ARI, Mönchhofstr. 12-14, D-69120 Heidelberg, Germany*
[3]*Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA*
[4]*Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*
[5]*Department of Physics, Kavli Institute for Astrophysics and Space Research, MIT, Cambridge, MA 02139, USA*
[6]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, D-85740 Garching bei München, Germany*
[7]*Department of Astronomy, Columbia University, 550 West 120th Street, New York, NY 10027, USA*
[8]*TAPIR, Mailcode 350-17, California Institute of Technology, Pasadena, CA 91125, USA*

Accepted 2016 November 11. Received 2016 October 16; in original form 2016 July 12

ABSTRACT

The inefficiency of star formation in massive elliptical galaxies is widely believed to be caused by the interactions of an active galactic nucleus (AGN) with the surrounding gas. Achieving a sufficiently rapid reddening of moderately massive galaxies without expelling too many baryons has however proven difficult for hydrodynamical simulations of galaxy formation, prompting us to explore a new model for the accretion and feedback effects of supermassive black holes. For high-accretion rates relative to the Eddington limit, we assume that a fraction of the accreted rest mass energy heats the surrounding gas thermally, similar to the 'quasar mode' in previous work. For low-accretion rates, we invoke a new, pure kinetic feedback model that imparts momentum to the surrounding gas in a stochastic manner. These two modes of feedback are motivated both by theoretical conjectures for the existence of different types of accretion flows as well as recent observational evidence for the importance of kinetic AGN winds in quenching galaxies. We find that a large fraction of the injected kinetic energy in this mode thermalizes via shocks in the surrounding gas, thereby providing a distributed heating channel. In cosmological simulations, the resulting model produces red, non-star-forming massive elliptical galaxies, and achieves realistic gas fractions, black hole growth histories and thermodynamic profiles in large haloes.

Key words: black hole physics – methods: numerical – galaxies: clusters: general – galaxies: evolution – galaxies: formation – cosmology: theory.

1 INTRODUCTION

In simulations of galaxy formation, feedback from active galactic nuclei (AGNs) is the most commonly invoked physical mechanism to explain the suppression of star formation in massive galaxies and the observed correlations between black hole masses and properties of their host galaxies. In particular, feedback from luminous quasars has been suggested to limit black hole growth and star formation during mergers at high redshift (Di Matteo, Springel & Hernquist 2005; Springel, Di Matteo & Hernquist 2005; Hopkins et al. 2006; Debuhr et al. 2010; Choi et al. 2014). Interacting galaxies trigger a redistribution of angular momentum and thus gas inflows into the nuclear region of galaxies (Hernquist 1989; Barnes & Hernquist 1996; Mihos & Hernquist 1996). These gas inflows then generate a cascade of gravitational instabilities (Hopkins & Quataert 2010; Emsellem et al. 2015), through which the supermassive black hole (SMBH) is fuelled and a fraction of the gravitational binding energy is released. This energy is sufficient to lower the star formation rate by several orders of magnitude (Di Matteo et al. 2005). However, it is not yet clear whether the released energy has a lasting effect on the whole galaxy and its star formation rate,

★ E-mail: rainer.weinberger@h-its.org
† Hubble Fellow.
‡ Hubble Fellow.

or just affects the innermost regions (Debuhr, Quataert & Ma 2011; Roos et al. 2015).

By applying semi-analytic modelling, Croton et al. (2006) pointed out that 'radio-mode' feedback, which provides an efficient source of energy in systems with hot, hydrostatic atmospheres, can simultaneously explain the low-mass drop-out rate in cooling flows, the exponential cutoff at the bright end of the galaxy luminosity function and the increased mean stellar age in massive elliptical galaxies. Bower et al. (2006) used a similar approach in their semi-analytic model. Sijacki et al. (2007) presented a unified sub-resolution model with energy input from both quasars and radio-mode feedback in hydrodynamical simulations and applied it to galaxy cluster formation. In this model, the second mode of feedback is active once the black hole accretion rate relative to the Eddington limit, $\dot{M}_{\rm BH}/\dot{M}_{\rm Edd}$, drops below a given value. The feedback energy injection is modelled by heating up spherical bubbles of gas in galaxy haloes, mimicking the observed radio lobes in galaxy clusters.

There are various implementations of 'quasar-mode' feedback in the literature. Debuhr et al. (2011) and Debuhr, Quataert & Ma (2012) use feedback from radiation pressure from luminous AGN, modelled by depositing momentum in surrounding simulation particles in idealized mergers. Choi et al. (2012, 2014, 2015) included mechanical and thermal energy and pressure from X-rays in their AGN feedback prescription and studied the effect on idealized mergers of disc galaxies and in cosmological 'zoom' simulations of elliptical galaxies while Wurster & Thacker (2013) performed a comparative study of these AGN models in merger simulations.

Likewise, many different approaches for 'radio-mode' activity have been taken, often using bipolar outflows in idealized simulations of hydrostatic haloes (Reynolds, Heinz & Begelman 2002; Basson & Alexander 2003; Omma et al. 2004; Ruszkowski, Brüggen & Begelman 2004; Zanni et al. 2005; Brighenti & Mathews 2006; Brüggen et al. 2007; Cattaneo & Teyssier 2007; Sternberg, Pizzolato & Soker 2007; Sternberg & Soker 2009; Gaspari et al. 2011a,b; Gaspari, Brighenti & Temi 2012; Li & Bryan 2014a,b; Li et al. 2015; Hillel & Soker 2016; Yang & Reynolds 2016a,b), or in cosmological simulations (Dubois et al. 2010, 2012, 2016). These methods assume that quenching is caused by the energy that is released from collimated jets and their associated radio lobes, which can be found in massive systems (Dunn & Fabian 2006). However, Meece, Voit & O'Shea (2016) show that these kinetic feedback implementations have a different impact in an idealized galaxy cluster setup compared to pure thermal injection.

The extensive body of literature on coupled AGN–galaxy evolution (including Granato et al. 2004; Kawata & Gibson 2005; Sijacki et al. 2007; Di Matteo et al. 2008; Hopkins et al. 2008a,b; Okamoto, Nemmen & Bower 2008; Somerville et al. 2008; Booth & Schaye 2009; Ciotti, Ostriker & Proga 2010; Debuhr et al. 2010; Teyssier et al. 2011; Dubois et al. 2012; Rosas-Guevara et al. 2015; Choi et al. 2014; Hirschmann et al. 2014; Khandai et al. 2015; Somerville & Davé 2015; Steinborn et al. 2015; Trayford et al. 2016, among others) has recently been complemented by a new generation of high-resolution cosmological simulations of galaxy formation in large volumes, such as Eagle (Schaye et al. 2015) and Illustris (Vogelsberger et al. 2014c). The corresponding implementations for black hole feedback in massive galaxies (in Illustris the radio-mode, while Eagle does not distinguish between modes) gather energy up to a predetermined threshold value, which parametrizes its burstiness and inject it instantaneously as thermal energy (see Sijacki et al. 2007; Booth & Schaye 2009, for Illustris and Eagle, respectively).

While the Illustris simulation – which forms the starting point of our work – has been remarkably successful in matching a wide range of galaxy properties, its results are in tension with a number of properties of observed haloes and galaxies. An important discrepancy arising from the AGN feedback model is the gas fraction of groups of galaxies and poor clusters, which is substantially too low in Illustris (Genel et al. 2014). At the same time, the stellar masses of the central galaxies in the simulated systems are too high. Employing a yet higher feedback efficiency of the BH radio mode to suppress star formation further would expel even more gas, and hence does not represent a viable solution. Alternatively, as part of our study, we made numerous attempts to improve the impact of the bubble model by adopting different choices for the parameters or by adding non-thermal pressure support in the form of magnetic fields, but without success. We therefore conclude that the particular AGN feedback model in Illustris is disfavoured, and a more radical change is in order.

This suggestion is supported by recent observational findings about the possible importance of kinetic winds driven during black hole accretion. For example, Cheung et al. (2016) find bisymmetric emission features in the centres of quiescent galaxies of stellar mass around $2 \times 10^{10}\,{\rm M}_{\odot}$, from which they infer the presence of centrally driven winds in typical quiescent galaxies that host low-luminosity active nuclei. They show that such 'red geyser' galaxies are very common at this mass scale, and that the energy input from the low activity of the SMBHs of these galaxies is capable of driving the observed winds, which contain sufficient mechanical energy to suppress star formation. This appears to be a BH feedback channel that is distinct from the radio galaxies at the centres of clusters, but as it affects many more galaxies at lower mass scales, it could well be more important for global galaxy evolution. Recently, Pontzen et al. (2016) found hot, AGN-driven outflows in post-merger galaxies, using the single-mode thermal AGN feedback model of Tremmel et al. (2016). Interestingly however, Genzel et al. (2014) and Förster Schreiber et al. (2014) have discovered wide-spread, powerful AGN-driven outflows in the majority ($\sim$70 per cent) of massive $z \sim 1$–2 star-forming galaxies. Because this phenomenon is so common, it likely arises from low-luminosity AGN with low Eddington ratios and thus appears consistent with a kinetic wind mode. Also, theoretically, there is good motivation for hot coronal winds from BH accretion flows. For example, Yuan & Narayan (2014) discuss such a scenario, which can be viewed as a small-scale version of the jet model of Blandford & Znajek (1977).

The motivation of our work is therefore to develop a revised model for black hole growth and feedback that takes these considerations into account. It is important to realize that the relevant time and length-scales of the detailed black hole physics are by far not resolved in cosmological simulations. Hence, the corresponding feedback models can only be implemented as so-called sub-resolution treatments that mimic the net effect of feedback on resolved scales. Besides the theoretical uncertainties involved, this approach comes with the drawback that the behaviour of the models can vary between different numerical methods because the scales at which the gas state is affected by the subgrid treatment are only marginally resolved. This is demonstrated for example in Sijacki et al. (2015) for the bubble heating model of Sijacki et al. (2007). We thus also aim to take recent improvements in the accuracy of the hydrodynamical modelling into account (Bauer & Springel 2012; Kereš et al. 2012; Sijacki et al. 2012; Vogelsberger et al. 2012; Pakmor et al. 2016).

The model presented here conjectures two modes of feedback from AGN in thermal and kinetic form, and in this sense is similar to

Dubois et al. (2012). While the kinetic part of their model is inspired by the sub-relativistic jet simulations of Omma et al. (2004), our approach does not directly aim to represent jets from AGNs that act on marginally resolved scales. Rather, we assume that the physical mechanisms that provide energy and momentum transport from black holes to their surroundings are reasonably efficient, and that their impact on large scales can be captured by depositing energy and momentum in small regions around halo centres. This approach does not address the microphysics of the origin of AGN feedback but aims to arrive at a robust parametrization of the effects of black holes on galaxy and galaxy cluster formation even at coarse resolution.

In what follows, we present a new model for SMBH growth and AGN feedback in cosmological simulations of structure formation implemented in the moving-mesh magnetohydrodynamics code AREPO (Springel 2010; Pakmor, Bauer & Springel 2011; Pakmor et al. 2016). In Section 2, we describe the model and its free parameters. Because the main modification to previous works lies in feedback injection at low-accretion rates, in Section 3, we discuss idealized tests of how the energy couples in this mode to the gas. We then continue in Section 4 with an investigation of its impact on cosmological simulations of galaxy formation. Section 5 is dedicated to a systematic exploration of the influence of the different model parameters on the results. Finally, we describe our findings and present our conclusions in Section 6. Appendix A specifies, for definiteness, details of our supernova feedback model, and Appendix B discusses numerical resolution dependences.

2 BLACK HOLE MODEL

Modelling AGNs in cosmological simulations poses several fundamental challenges. First, the detailed physical mechanisms of both accretion on to SMBHs (Hopkins & Quataert 2010, 2011; Anglés-Alcázar, Özel & Davé 2013; Gaspari, Ruszkowski & Oh 2013; Anglés-Alcázar et al. 2015, 2017; Curtis & Sijacki 2015, 2016; Emsellem et al. 2015; Rosas-Guevara et al. 2015) and the AGN–gas interaction (Huarte-Espinosa, Krause & Alexander 2011; Gaibler et al. 2012; Cielo et al. 2014; Costa, Sijacki & Haehnelt 2014; Roos et al. 2015; Bieri et al. 2017; Hopkins et al. 2016) are poorly understood, which makes it at present impossible to formulate a 'correct' treatment for simulations, independent of their resolution. Secondly, the extreme dynamic range posed by the problem, where a comparatively tiny accretion region around the black hole influences an entire galaxy or even a galaxy cluster and the surrounding intergalactic medium, vastly exceeds the capabilities of current numerical techniques so that much of the physics on the smallest scales needs to be coarsely approximated with sub-resolution models. Thirdly, the non-linear nature of galaxy formation intimately couples black hole accretion with other aspects of feedback, chiefly the regulation of ordinary star formation (Puchwein & Springel 2013). This makes it difficult to disentangle the impact of different astrophysical processes. While we first examine the behaviour of our model in well-defined idealized tests, we will primarily assess its performance through studies of its consequences in the full cosmological context.

Similar to Sijacki et al. (2007), we distinguish between states of high- and low-accretion rates. This follows the theoretical notion that there exist (at least) two physically distinct types of accretion flows on to massive black holes (e.g. Begelman 2014, and references therein): one at comparatively high rates in a classic disc mode (Shakura & Sunyaev 1973), the other at lower rates in a more spherical and hotter accretion flow (Shapiro, Lightman & Eardley 1976; Ichimaru 1977). These regimes have loosely been identified with 'quasar' and 'radio' modes in previous simulation work. The observed phenomenology of radio jets in galaxy clusters has often been interpreted as providing the dominant source of feedback, at least in the low-accretion radio-mode regime (McNamara & Nulsen 2007). This has also motivated, e.g. the bubble heating model in Sijacki et al. (2007) that was applied in the Illustris simulation and in other works. However, there are also theoretical indications pointing to the existence of kinetic winds in the low-accretion state (Igumenshchev & Abramowicz 1999; Stone, Pringle & Begelman 1999; Yuan & Narayan 2014; Yuan et al. 2015; Bu et al. 2016; Sądowski et al. 2016). These would be difficult to observe but could constitute an even more important feedback mechanism than the radio jets themselves. A central motivation of our work is to test this idea by replacing radio bubble feedback with a kinetic wind.

2.1 Accretion mode

We follow previous work and use the Eddington ratio as the criterion for deciding the accretion state of the black hole. Specifically, we assume SMBHs to be in the high-accretion state as long as their Bondi–Hoyle–Lyttleton accretion rate $\dot{M}_{\rm Bondi}$ (Hoyle & Lyttleton 1939; Bondi & Hoyle 1944; Bondi 1952) exceeds a fraction χ of the Eddington accretion rate $\dot{M}_{\rm Edd}$:

$$\frac{\dot{M}_{\rm Bondi}}{\dot{M}_{\rm Edd}} \geq \chi, \tag{1}$$

where

$$\dot{M}_{\rm Bondi} = \frac{4\pi {\rm G}^2 {\rm M}_{\rm BH}^2 \rho}{c_s^3}, \tag{2}$$

$$\dot{M}_{\rm Edd} = \frac{4\pi {\rm G M_{BH} m_p}}{\epsilon_r \sigma_{\rm T} c}. \tag{3}$$

Here, G denotes the gravitational constant, c the vacuum speed of light, m_p the proton mass and $\sigma_{\rm T}$ the Thompson cross-section. The factor ϵ_r is the radiative accretion efficiency. $M_{\rm BH}$ is the black hole mass, and ρ and c_s are the density and sound speed[1] of the gas near the black hole, respectively. They are obtained by averaging over a sphere with radius h in a kernel-weighted fashion around the black hole such that the enclosed number of cells in this sphere is approximately equal to a prescribed number:

$$n_{\rm ngb} \approx \sum_i \frac{4\,\pi h^3\, m_i}{3\, m_{\rm baryon}} w(r_i). \tag{4}$$

Here, $m_{\rm baryon}$ is the target mass of a gas cell, i.e. the gas mass resolution enforced by the refinement and derefinement operations of the hydrodynamic code, $n_{\rm ngb}$ is the prescribed number of neighbouring cells in this averaging and $w(r)$ is an SPH weighting kernel.

A sensible value for χ is expected to lie in the range $\sim$0.001–0.1, by analogy with X-ray binaries (e.g. Dunn et al. 2010). Previous works (Sijacki et al. 2007, 2015) have employed a fixed value of χ. Black holes at low redshift located in massive systems show clear signatures of being in a 'radio feedback state, (Dunn & Fabian 2006) which indicates low Eddington ratios. However, as we will show in Section 4, this does not occur in our simulations *unless* the black holes transition to the kinetic mode in the first place, which is not guaranteed. To favour this transition for the most massive black

[1] We use an effective sound speed, taking into account both thermal and magnetic signal propagation $c_{\rm s}^2 = c_{\rm s,therm}^2 + c_{\rm A}^2$, where $c_{\rm A} = \left(\boldsymbol{B}^2/4\pi\rho\right)^{1/2}$ is the Alfvén speed.

holes at late times (which tend to be found in the most massive haloes), we scale the threshold with black hole mass,

$$\chi = \min\left[\chi_0 \left(\frac{M_{\rm BH}}{10^8\,{\rm M}_\odot}\right)^{\beta}, 0.1\right] \quad (5)$$

with χ_0 and β as parameters. The pivot mass $10^8\,{\rm M}_\odot$ is degenerate with χ_0 and is therefore not set independently. We limit the threshold χ to a maximum of 0.1 to always allow any black hole (including the most massive ones) to reach the high-accretion state provided there is a large enough gas supply to fuel them. This would be expected for high-redshift quasars that have very massive black holes.[2] For $\beta > 0$, our scaling makes it more difficult for low-mass black holes to be in the kinetic mode, and vice versa. We expect this to support the occurrence of a rapid quenching transition in massive galaxies, and make it unlikely that low-mass galaxies will be strongly affected by kinetic feedback. The physics of the accretion mode transitions of SMBHs is poorly understood, making it difficult to parametrize it adequately in a coarse cosmological model. We have here opted for a heuristic model that is based on the only intrinsic black hole property we keep track of, the black hole mass, and which is selected pragmatically based on how well it reproduces observational trends. We note that it appears physically plausible that there are systematic trends with black mass scale in the accretion mode transition, given that radiative cooling physics breaks the scale invariance.

2.2 Accretion estimate and seeding of black holes

Note that in the above calculation of the Bondi accretion rate, we omit a boost-factor α that was used in older models to account for the unresolved ISM structure. When the latter is treated with a sub-resolution model that prescribes a high mean thermal support and an effective pressure, the Bondi rate is artificially biased low, slowing down especially the early growth of black holes. The boost factor was primarily introduced in the older models to compensate for this problem by ensuring that the Bondi growth time-scale for small mass seed black holes does not exceed the Hubble time. Since the actual accretion rate was however anyway limited to the Eddington rate, the latter is ultimately the governing rate for most of the growth. Furthermore, applying a boost factor for massive black holes in the low-accretion state, when their feedback generates a low-density, hot gas phase around them (that can be resolved, unlike the ISM), appears questionable. We therefore simplify our treatment by assuming that the black holes are always accreting at the pure Bondi rate, limited by the Eddington rate:

$$\dot{M}_{\rm BH} = \min\left(\dot{M}_{\rm Bondi}, \dot{M}_{\rm Edd}\right). \quad (6)$$

We note that for massive black holes at late times, the accretion rate is self-regulated, thus an additional factor in the accretion rate estimate has no overall effect in this regime apart from systematically shifting the black hole masses, i.e. here the boost factor is largely degenerate with the black hole masses reached. Only the early growth phase is strongly affected by the boost factor, but this phase depends sensitively on the black hole seed mass as well (see discussion below), and we use this dependence to make up for the omission of a boost factor.

Note that Vogelsberger et al. (2013) lowered the accretion rate estimate by a factor of $(P_{\rm ext}/P_{\rm ref})^2$ whenever $P_{\rm ext} < P_{\rm ref}$. Here, $P_{\rm ext}$ is the kernel-weighted gas pressure surrounding the black hole and $P_{\rm ref}$ is a reference pressure (Vogelsberger et al. 2013, their equation 23). While this was used in the Illustris simulation, we omit such a factor in this work. We ran simulations both with and without this factor and found no significant difference in the properties presented in this work. However, as this serves as a protection against rare cases of overly heated, underdense regions in galaxy centres, we plan to use it in future simulations that contain a larger sample of galaxies.

In our cosmological simulations, a black hole with mass $M_{\rm seed}$ is placed at the centre of a halo whenever the on-the-fly friend-of-friends halo finder identifies a halo more massive than a threshold mass $M_{\rm FOF}$ that does not yet contain a black hole. We note that to offset a potentially sluggish growth of black holes at high redshift, one can resort to a slightly larger seed mass, which then produces a similar result as using a boost-factor α. In order to remain close to our previous models, we use this here and adopt a black hole seed mass of $8 \times 10^5\,h^{-1}\,{\rm M}_\odot$ in our default model, which leads to a similarly fast growth at early times as our older models with $\alpha = 100$. Given the significant theoretical uncertainties in the early growth of SMBHs (e.g. Volonteri 2010), we consider the seed mass as a poorly constrained free parameter. We note that there are other models for black hole seed formation in cosmological simulations (e.g. Bellovary et al. 2011; Tremmel et al. 2016) that use thresholds of local gas properties such as metallicity, density and temperature. However, we decided for a seeding prescription depending solely on halo mass because of its simplicity and numerical robustness.

At the limited numerical resolution available in cosmological simulations, two-body discreteness effects and numerical N-body noise can displace black hole particles from halo centres. At the same time, the dynamical friction forces that should allow massive black holes to sink to the centres of dark matter haloes are not captured accurately by the simulation. To prevent black holes from artificially leaving the centres of haloes for long periods of time due to these effects, we resort to an ad hoc centring prescription designed to keep black holes very close to the potential minimum of their host dark matter haloes. To this end, at every global integration time-step (i.e. when the longest time-steps occurring in the whole simulation are synchronized in the nested time integration scheme), we determine the minimum gravitational potential in a region around the BH containing the equivalent of 1000 mass resolution elements. The BH particle is then shifted to this potential minimum (if not at the location of the BH already, which frequently happens), and its velocity is set to the mean mass-weighted velocity of the region. The latter minimizes any motion of the BH with respect to the central region of the halo. This method robustly prevents haloes from losing their central black hole, and it further adopts a scenario in which BH binaries are assumed to merge promptly. We use this approach here because of its numerical robustness and independence of resolution. However, there are more sophisticated treatments in the recent literature that use sub-resolution models for dynamical friction (e.g. Wurster & Thacker 2013; Hirschmann et al. 2014; Tremmel et al. 2016). We aim to use such a scheme in future high-resolution extensions of the present model.

2.3 Feedback

For the high-accretion state, we calculate the liberated feedback energy as

$$\Delta\dot{E}_{\rm high} = \epsilon_{\rm f,high}\epsilon_{\rm r}\dot{M}_{\rm BH}\,c^2, \quad (7)$$

where $\dot{M}_{\rm BH}$ is the estimated black hole mass accretion rate of the black hole with mass $M_{\rm BH}$, $\epsilon_{\rm r}$ is the radiative efficiency (i.e.

[2] Note that the volume of the simulations presented in this work (Section 4) is too small to host these kind of objects.

the canonical 0.1–0.2 of the accreted rest-mass energy that is released in the accretion process and not vanishing in the black hole), while $\epsilon_{\rm f,\,high}$ is the fraction of this energy that couples to the surrounding gas. For the low-accretion state, the feedback energy is parametrized as

$$\Delta \dot{E}_{\rm low} = \epsilon_{\rm f,kin} \dot{M}_{\rm BH} c^2. \tag{8}$$

Note that we use different coupling efficiencies, $\epsilon_{\rm f,\,kin}$ and $\epsilon_{\rm f,\,high}$, for the two modes, motivated by the different physical nature of the accretion modes; namely that the low-accretion state is thought to be radiatively inefficient. We keep the coupling efficiency in the high-accretion state at a constant value of $\epsilon_{\rm f,\,high} = 0.1$, resulting in an overall efficiency $\epsilon_{\rm r}\epsilon_{\rm f,\,high} = 0.02$, while we set a maximum value of $\epsilon_{\rm f,\,kin} = 0.2$ in the low-accretion mode, which assumes that the released rest-mass energy appears primarily in kinetic outflows. We note that our choice of $\epsilon_{\rm f,\,kin} = 0.2$ is within the physically plausible range, depending on the underlying physical mechanism. If, for example, the energy is delivered by small-scale jets produced by the Blandford–Znajek mechanism, the jet energy can be a factor of 10 or more larger than our adopted value for $\epsilon_{\rm f,\,kin}$ because of black hole spin (Yuan & Narayan 2014). In the opposite case, for non-spinning black holes, small-scale simulations of accretion (Yuan et al. 2015) provide a theoretical lower limit to $\epsilon_{\rm f,\,kin}$ of about 10^{-3}.

To protect against a potential runaway of the kinetic feedback mode that may drive the density to ever lower values (see also section 2.6.2 of Vogelsberger et al. 2013, for further discussion), we conjecture that at very low densities, the coupling efficiency $\epsilon_{\rm f,\,kin}$ eventually becomes weak. For simplicity, we assume that such a weakening occurs below a density $f_{\rm thresh}\rho_{\rm SFthresh}$, where $f_{\rm thresh}$ is a free parameter and $\rho_{\rm SFthresh}$ is the density threshold for star formation. If the surrounding density ρ drops below this value we reduce the coupling proportional to density. This then formally corresponds to a variable coupling efficiency in the low-accretion state,

$$\epsilon_{\rm f,kin} = \min\left(\frac{\rho}{f_{\rm thresh}\rho_{\rm SFthresh}}, 0.2\right). \tag{9}$$

Our standard value for this prescription is $f_{\rm thresh} = 0.05$, and we will show in Section 5 that the exact value of $\epsilon_{\rm f,\,kin}$ has hardly any impact on galaxy properties.

In the high-accretion state, we inject the feedback as pure thermal energy in a small local environment around the black hole, as in Springel et al. (2005) and our subsequent work (including Vogelsberger et al. 2013, as well as Illustris), while using our new model of kinetic feedback in the low-accretion state. In the latter case, we inject the energy as pure kinetic energy. Unlike in the high-accretion state, we hence input momentum but no immediate thermal energy to the gas. Technically, we inject both forms of feedback in a kernel-weighted manner into a prescribed number of gas neighbouring the BH, as determined by equation (4). This region is identical for imparting feedback and the calculation of the gas properties used in the accretion estimate.

Because we cannot spatially resolve small-scale jets and the accretions flows in our cosmological simulations, we add the momentum in a random direction. We have found that this approach is most robust for avoiding possible numerical artefacts that can be produced at poor resolution by more elaborate approaches for adding the momentum. For example, one may impart the momentum in a spherically symmetric fashion, radially away from the black hole, with zero total momentum (as vector sum) added per injection event. However, this can produce an artificial suppression of the gas density at the position of the black hole at the resolution we achieve here. Similarly, a biconical injection in opposite directions at the position of the black hole can create artificially depressed gas densities unless the 'jets' are well enough resolved. We therefore prefer random injection directions that change for every injection event, which we found to be least resolution dependent. In this case, detailed energy and momentum conservation is only obtained as a time average over the injection events.

Specifically, for an available kinetic feedback energy ΔE, we kick each gas cell j in the feedback region by

$$\Delta \boldsymbol{p}_j = m_j \sqrt{\frac{2\,\Delta E\, w(\boldsymbol{r}_j)}{\rho}}\,\boldsymbol{n}, \tag{10}$$

where $\Delta \boldsymbol{p}_j$ is the change in momentum of gas cell j, m_j denotes its mass and $\boldsymbol{r}_j$ is the distance vector from the black hole to the respective cell. The factor $\boldsymbol{n}$ is the unit vector in a randomly chosen injection direction, $w(\boldsymbol{r}_j)$ the value of the smoothing kernel, and ρ is the density estimate of the surrounding gas, as described in Section 2.1. The total momentum injection per feedback event is thus

$$\boldsymbol{p}_{\rm inj} = \sum_j m_j \sqrt{\frac{2\,\Delta E\, w(\boldsymbol{r}_j)}{\rho}}\,\boldsymbol{n}, \tag{11}$$

and the corresponding change in total energy of the gas (relative to the lab frame) is

$$E_{\rm inj} = \Delta E + \sum_j \left(\boldsymbol{p}_j \cdot \boldsymbol{n}\right) \sqrt{\frac{2\,\Delta E\, w(\boldsymbol{r}_j)}{\rho}}, \tag{12}$$

where $\boldsymbol{p}_j$ is the momentum of cell j before the injection event.

For a single injection event, this violates strict momentum conservation and will generally not increase the total energy by precisely ΔE. However, the average over many injection events leads to the desired energy injection and assures momentum conservation (i.e. $\langle \boldsymbol{p}_{\rm inj} \rangle = 0$), as the injection direction $\boldsymbol{n}$ is randomly chosen for each injection event and does not correlate with the flow direction of the surrounding gas. To make the occurrence of these injection events independent of the time-stepping, and also to make them powerful enough individually, we discretize the kinetic feedback mode by imposing a minimum energy that needs to accumulate in the kinetic accretion mode before the feedback is released. This is similar to the approach adopted in Illustris and Eagle for the BH feedback in large haloes.

In this work, we choose to parametrize the adopted energy threshold for the kinetic feedback in terms of a fiducial energy computed from the mass of the feedback region and the surrounding dark matter velocity dispersion. This identifies an energy per unit mass that is tied to the virial temperature of the halo. In fact, we could also construct the energy scale from the temperature of the surrounding gas. But the latter can be affected strongly by local cooling or previous feedback events, hence we prefer to use the dark matter velocity dispersion for increased robustness. We note that the velocity dispersion is also used for our supernova-driven wind feedback from star formation, which we adopt from Illustris in only a slightly modified form (see Appendix A). We parametrize the kinetic feedback threshold by

$$E_{\rm inj,min} = f_{\rm re}\,\frac{1}{2}\,\sigma_{\rm DM}^2\, m_{\rm enc}, \tag{13}$$

where $\sigma_{\rm DM}$ is the 1D dark matter velocity dispersion, $m_{\rm enc}$ is the gas mass in the feedback region and $f_{\rm re}$ is a free parameter that specifies the burstiness and thus the frequency of the reorientation of the kinetic feedback. If a larger value is chosen for $f_{\rm re}$, fewer feedback events occur, but they are individually stronger. Choosing

this scaling is partly numerically motivated, as it ensures that the resulting shocks are strong enough to be accurately captured by our finite-volume scheme. Without this threshold, low-luminosity black holes would drive very weak flows that would thermalize mainly via numerical dissipation effects, which is clearly undesirable. Part of the motivation is also physical because this scaling ensures that the specific energy of the wind does not significantly exceed the specific binding energy of the halo, and thus should not unbind a large amount of gas or overly disturb the thermodynamic state of the intrahalo gas.

3 KINETIC WIND DISSIPATION TESTS

To examine the dissipation mechanisms of the kinetic feedback model, we use idealized test simulations in a cubic box with constant density, temperature and pressure and a side length of 25 kpc. The fiducial values for density and temperature are $n = 10^{-1}\,\mathrm{cm}^{-3}$ and $T = 10^7$ K. We place a black hole at the centre and inject energy at a fixed rate of $10^{45}\,\mathrm{erg\,s}^{-1}$ in the kinetic mode. We run the simulations for 5 Myr, only solving the equations of hydrodynamics, switching off self-gravity, gas cooling and all galaxy formation sub-grid prescriptions such as star formation and feedback, metal enrichment, black hole seeding, etc. We run the simulations at two different resolutions: 32^3 initial cells, which roughly corresponds to the resolution of cosmological simulations (for $\rho = 10^{-1}\,\mathrm{cm}^{-3}$, the average mass of a gas cell is $7 \times 10^5\,\mathrm{M}_\odot$), and 256^3, to show the convergence properties.

Unlike in cosmological simulations, where we keep the number of neighbours in the feedback injection region roughly constant, we here fix the radius of the sphere in which the feedback is injected. We have tested both approaches and found that there is no substantial difference, except that tying the injection region to the number of neighbouring gas cells (equation 4) leads in this particular setup – in which self-regulation is disabled – to a slowly growing injection region, as the gas around the black hole is heated up by previous feedback events. To promote a clean study of the impact of the kinetic pulses on the gas, we prefer to keep the feedback injection region fixed to a sphere of 3.5 kpc radius around the black hole. We ensure that there is always a sufficient number of cells in this region by setting a maximum volume per cell, above which they are refined[3]. The physical size of the feedback injection region is kept the same for the 256^3 simulations, allowing a study of discretization effects in the gas. We note, however, that in cosmological simulations there are additional resolution dependences such as the scaling of the feedback injection region, which is discussed in Appendix B. As there is no dark matter in the present test simulations, we replace equation (13) with

$$E_{\mathrm{inj,min}} = f_{\mathrm{re}}\, u_{\mathrm{init}}\, m_{\mathrm{enc}}, \tag{14}$$

where u_{init} is the initial specific thermal energy. This means that we are assuming that the temperature of the gas in the initial state is equal to the virial temperature of the dark matter halo. Using this threshold, the total energy injected in 5 Myr suffices for seven injection events within the simulated time-span.

Fig. 1 shows a volume-weighted projection of the gas density, temperature, absolute velocity and the energy dissipation weighted Mach number of the shocks present after 5 Myr. For the temperature and density, we average over the logarithm of the corresponding quantity, while we average over the absolute value of the velocity to highlight the maximum velocities involved in the projection. The projected maps show that the model is very efficient in diluting the central regions near the black hole. This means that the accretion rate estimated in this regime would decrease immediately by orders of magnitude, resulting in a very tight self-regulation. As the projections were made shortly after the seventh injection occurred, we reach gas flows with very high velocities in the injection region, which slow down after leaving this immediate vicinity of the black hole.

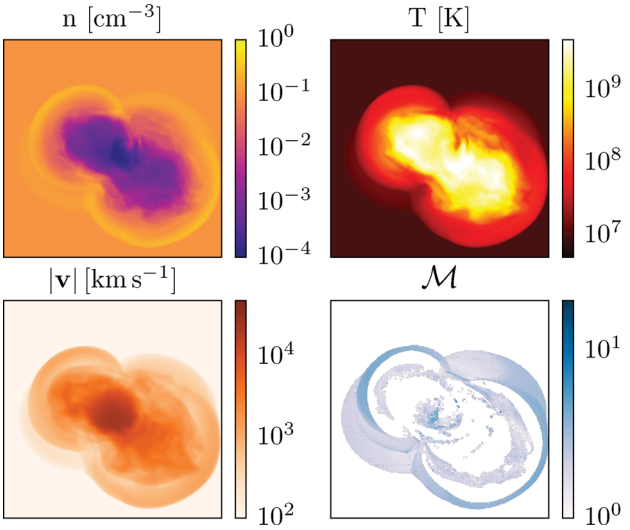

Figure 1. Thin projection (5 kpc in depth, 25 kpc on a side) of the 256^3, $n = 10^{-1}\,\mathrm{cm}^{-3}$, $T = 10^7$ K simulation after 5 Myr of evolution. The panels show volume-weighted density (top left), volume-weighted temperature (top right), absolute velocity (bottom left) and energy dissipation weighted Mach number (bottom right).

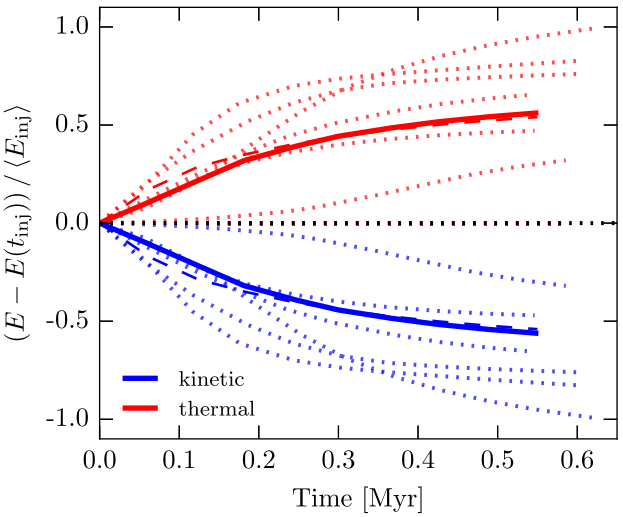

Figure 2. Evolution of the different energy components after kinetic energy injection. The dotted lines show individual injection events, the solid line their average, both in the simulation initially with 32^3 cells. The dashed line shows the average of the high-resolution test with 256^3 initial cells. On average, half of the feedback energy that was initially in kinetic form is thermalized after 0.5 Myr. This behaviour is converged at the resolution of cosmological simulations.

Fig. 2 shows the time evolution of the thermal and kinetic energy after individual injection events, as well as the average evolution

[3] Note that in cosmological simulations, we do not impose such a volume limit.

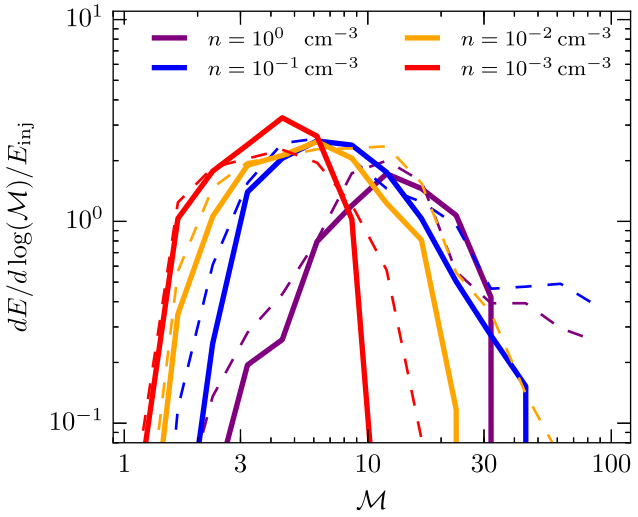

Figure 3. Energy dissipation as a function of shock Mach number $\mathcal{M}$ summed up over a simulation time of 5 Myr. The different colours denote different isobaric variations of the gas from relatively cool, dense (1 cm^{-3}, $T = 10^6$ K) to hot, dilute ($\rho = 10^{-3}$ cm^{-3}, $T = 10^9$ K). The solid lines show the simulation with an initial grid of 32^3 cells, comparable to the resolution of cosmological simulations, while the dashed lines indicate simulations with 256^3 cells to show the convergence of the analysis.

with time. We can detect an injection event simply by a jump in total energy of the system owing to very frequent simulation outputs. The first output after this jump defines the zero-point in Fig. 2, which means that the thermal energy increases and the kinetic energy decreases subsequently. The initial energy injection is purely kinetic; however within ∼0.5 Myr about half is dissipated into thermal energy, mostly via shock dissipation. As the direction of the momentum kicks change after every injection event, the AGN model does not build up a coherent gas flow that could reach several tens to a hundred kpc.

As a further analysis tool, we use the shock finder described in Schaal & Springel (2015) to detect shocks and calculate their Mach number $\mathcal{M}$ and energy dissipation rate for each snapshot. The bottom-right panel in Fig. 1 shows the energy dissipation weighted Mach number projection, which excludes all cells that do not belong to a shock, and Fig. 3 shows the corresponding energy dissipation as a function of Mach number. For a hot, dilute gas, the shock Mach number mostly remains below 10, while the shock strength increases with higher densities and correspondingly lower temperatures and sound speeds. Summing up all the energy that is dissipated in shocks, we are, up to a factor of order unity, able to reconstruct the feedback energy purely through post-processing analysis of the surrounding gas. This is possible for even moderate resolution, which opens up the possibility of studying the effects of shocks from AGNs on their surroundings even in cosmological simulations of galaxy formation, so that their behaviour can be compared to observations (e.g. Dopita et al. 2015; Medling et al. 2015). However, there are some technical challenges to this (see Schaal et al. 2016), in particular concerning the treatment of the unresolved ISM structure in these simulations. Therefore, for now, we restrict our analysis to the idealized setup and leave the study of shocks from AGN winds in cosmological simulations to future work.

The test simulations demonstrate that our kinetic feedback model can accelerate the gas in the injection region to several tens of thousands of km s^{-1}. This gas flow hits the surrounding medium and heats it via shock dissipation within time-scales of an Myr. A fraction of the energy will remain kinetic and ultimately decay via turbulent dissipation. This behaviour of the feedback injection is well converged, showing that our deposition of energy is not subject to significant numerical limitations on marginally resolved scales.

4 COSMOLOGICAL SIMULATIONS

Idealized test simulations such as those above cannot address the dynamics of self-regulated black hole growth. As this can only be meaningfully studied in calculations that follow cosmic structure formation and that also account for star formation, we now move on and examine the impact of our new model in hydrodynamical simulations of galaxy formation. This requires the full black hole model as described in Section 2, including the seeding of SMBHs and the estimate of their accretion rates. Also, because gas cooling and heating, star formation, stellar evolution and feedback, as well as the chemical enrichment of the interstellar medium, are all crucial ingredients of galaxy formation, we account for these processes using the respective models described in Vogelsberger et al. (2013). These are modified and extended as follows.

(i) We use isotropic winds from star formation with 10 per cent of the energy injected thermally (Marinacci, Pakmor & Springel 2014), instead of purely kinetically with a bipolar orientation as in Illustris.

(ii) We slightly adjust the scaling of the stellar wind model with redshift, metallicity and halo mass.

(iii) Updated chemical yields and an improved metal advection algorithm are used, which has however negligible influence on the results discussed here.

(iv) Ideal magnetohydrodynamics is included based on the Powell cleaning scheme (Pakmor et al. 2011; Pakmor & Springel 2013).

(v) An improved gradient estimator and time integration scheme for the hydrodynamics is used (Pakmor et al. 2016), which improves the accuracy of the AREPO code.

We briefly summarize the changes due to the modifications in the stellar wind model (i.e. the first two items) in Appendix A, as the interplay between stellar and AGN feedback affects the overall galaxy population (Puchwein & Springel 2013) as well as black hole growth rates (Dubois et al. 2015). The other modifications have a minor effect on the quantities examined in this work. We therefore focus this study on the black hole model and its parameters, and illustrate the relevance of the feedback efficiency, the accretion rate estimate and the black hole seeding model for the formation and evolution of galaxies as a function of their mass. A more detailed analysis of the other changes will be subject of a forthcoming paper (Pillepich et al., in preparation).

4.1 The simulations

We run a number of cosmological simulations of a periodic box with a side length of $30\,h^{-1}$ Mpc. As the large-scale modes of the matter power spectrum cannot be sampled in this comparatively small volume, our simulation does not contain structures as massive as the largest galaxy clusters observed in our Universe. However, we still follow the formation of 13 objects more massive than $10^{13}\,M_\odot$ and more than 100 haloes in the mass range between $10^{12}\,M_\odot$ and $10^{13}\,M_\odot$. This makes the simulations well suited for testing the AGN model and for studying its impact on the formation and evolution of massive galaxies at the resolution of the Illustris simulation.

Table 1. Overview of our primary simulations. All simulations, except for high res and low res are started from the same initial conditions. The only differences between the simulations and the fiducial case are the parameter values marked in bold. All simulations have a comoving volume of $(30\,h^{-1}\,\mathrm{Mpc})^3$.

Name	$n_{\mathrm{particles}}$ IN INITIAL CONDITIONS	m_{DM}, m_{baryon} $[10^5\,\mathrm{M}_\odot\,h^{-1}]$	$\epsilon_{\mathrm{f,kin}}$	ϵ_r	χ_0	β	M_{seed} $[\mathrm{M}_\odot\,h^{-1}]$	M_{FOF} $[\mathrm{M}_\odot\,h^{-1}]$	n_{ngb}	f_{re}
Fid	2×384^3	340, 67	0.2	0.2	0.002	2.0	8×10^5	5×10^{10}	128	20
No kin.	2×384^3	340, 67	0.2	0.2	**0.000**	2.0	8×10^5	5×10^{10}	128	20
High res	$\mathbf{2 \times 768^3}$	**42.5, 8.4**	0.2	0.2	0.002	2.0	8×10^5	5×10^{10}	**256**	20
Low res	$\mathbf{2 \times 192^3}$	**2720, 536**	0.2	0.2	0.002	2.0	8×10^5	5×10^{10}	**64**	20
Low ϵ_f	2×384^3	340, 67	**0.05**	0.2	0.002	2.0	8×10^5	5×10^{10}	128	20
Low ϵ_r	2×384^3	340, 67	**0.05**	**0.05**	0.002	2.0	8×10^5	5×10^{10}	128	20
High χ	2×384^3	340, 67	0.2	0.2	**0.008**	2.0	8×10^5	5×10^{10}	128	20
Low β	2×384^3	340, 67	0.2	0.2	0.002	**0.5**	8×10^5	5×10^{10}	128	20
Low M_{seed}	2×384^3	340, 67	0.2	0.2	0.002	2.0	$\mathbf{2 \times 10^5}$	5×10^{10}	128	20
High M_{FOF}	2×384^3	340, 67	0.2	0.2	0.002	2.0	8×10^5	$\mathbf{2 \times 10^{11}}$	128	20
High n_{ngb}	2×384^3	340, 67	0.2	0.2	0.002	2.0	8×10^5	5×10^{10}	**512**	20
Low f_{re}	2×384^3	340, 67	0.2	0.2	0.002	2.0	8×10^5	5×10^{10}	128	**5**

We adopt the cosmological parameters from the *Planck* intermediate results (Planck Collaboration XIII 2016), $\Omega_{\mathrm{M}} = 0.3089$, $\Omega_\Lambda = 0.6911$, $\Omega_{\mathrm{b}} = 0.0486$, $h = 0.6774$ and $\sigma_8 = 0.8159$ and use an Eisenstein & Hu (1998) matter power spectrum to produce initial conditions at redshift $z = 127$. The initial conditions contain 384^3 dark matter particles and the same number of gas cells at our default resolution. This implies an average gas cell and dark matter particle mass of $6.7 \times 10^6\,h^{-1}\,\mathrm{M}_\odot$ and $3.4 \times 10^7\,h^{-1}\,\mathrm{M}_\odot$, respectively, which is similar to the intermediate resolution Illustris simulation (Illustris-2 in Vogelsberger et al. 2014b). The corresponding softening length is 2 comoving kpc with a maximum value of 1 proper kpc for dark matter and stars. The softening for the gas cells depends on their volume and has a minimum of 0.25 comoving kpc. The moderate number of simulation particles allows us to study the effect of each parameter of the black hole model individually, but it also comes with a severe drawback: At this resolution, the star formation rate predicted by the employed Springel & Hernquist (2003) model is not fully converged for haloes below $10^{12.5}\,\mathrm{M}_\odot$ (Pillepich et al. 2014; Sijacki et al. 2015), as also shown in Appendix B. This entails important limitations, especially with regard to the comparison to observations.

To get a better idea of the behaviour at the low-mass end of AGN host galaxies, we run an additional simulation with 2×768^3 particles and cells and the same box size. In this run, all softening lengths are reduced by a factor of 2 compared to the fiducial setup. The implied resolution corresponds to the Illustris-1 high-resolution run. Additionally, we run a low-resolution test with 2×192^3 particles and the same side length of the simulation box. The softening is increased by a factor of 2 compared to the fiducial run. For each of the different resolutions, we also computed a dark matter only version to quantify the role of baryonic physics on the halo mass function.

We have also carried out a suite of simulations with 2×384^3 particles in which the parameters of the black hole model were systematically varied by a factor of 4 each in the direction that seemed most interesting. Table 1 gives an overview of these simulations and their parameters. The set of simulations also includes one simulation in which the black holes are always in the quasar mode, independent of their Eddington rate (labeled '`no kin.`'). Using identical initial conditions in our simulations allows a halo-by-halo comparison of all galaxy properties, facilitating a clean comparison of globally averaged properties and an interpretation of small changes in a meaningful way. We do so by matching the friend-of-friends groups in the different simulations in position space, followed by a verification that they are indeed the same structures by ensuring that they have at least half of their dark matter particles in common.[4] We discard the few per cent of haloes that could not be matched by these criteria and ignore them for the analysis; they are for the most part borderline cases where the friend-of-friends algorithm links two haloes across a feeble particle bridge in one simulation but not in the other. In addition to the matching of haloes across simulations at identical redshifts, we also match haloes of a given simulation at different times. We define the progenitor as the halo in the previous snapshot that contributes the most dark matter particles to a given halo, which allows us to study the evolution of individual haloes.

We base a substantial part of our analysis on a comparison of the same haloes in different simulations, thereby avoiding uncertainties due to the absolute halo abundance, which is affected significantly by box size and resolution effects. However, a detailed comparison to observational data requires a larger simulated volume and higher resolution. Achieving both at the same time is a computational challenge and is clearly beyond the scope of this paper. Simulations that reach this statistical power will be presented in future work.

4.2 Galaxy properties

4.2.1 Halo masses

We start by looking at the overall baryonic effect on the masses of individual haloes. Fig. 4 shows the mass of the haloes in our fiducial baryonic simulation in units of the mass of the corresponding halo in the dark matter only simulation, as a function of $M_{200,\mathrm{c}}$. The average mass fraction does not exceed unity for any halo mass, unlike in Vogelsberger et al. (2014b) for haloes of $M_{200,c} \sim 10^{11}\,\mathrm{M}_\odot$. However, also in our simulations, the mass of some individual haloes can scatter above the mass of their dark matter only counterparts. The masses of haloes with $M_{200,c} < 10^{11}\,\mathrm{M}_\odot$ are suppressed more than those of $10^{12}\,\mathrm{M}_\odot$ haloes, independent of the black hole feedback implementation. This is hence presumably caused by stellar

[4] This is done by checking their particle IDs, which are unique identifiers set in the initial conditions.

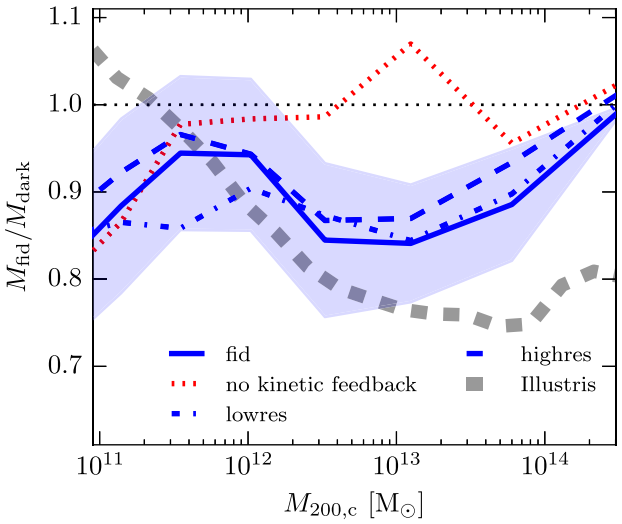

Figure 4. Average ratio of the mass of haloes in the full physics simulation to the mass of the corresponding halo in the dark matter only run, as a function of halo mass. The shaded region indicates the 1σ scatter of the results of our default simulation. The dashed grey line represents the result for the Illustris simulation (Vogelsberger et al. 2014b).

feedback. A mild decline in the halo masses relative to the dark matter only run occurs for haloes more massive than 10^{12} M$_\odot$. This drop can be clearly associated with the kinetic AGN feedback, as it is not present in the simulation without this mode. However, it is not as pronounced as in Vogelsberger et al. (2014b), which confirms that our feedback implementation is not as violent. The upturn at 10^{14} M$_\odot$ indicates a return to the universal baryon fraction for the most massive haloes but is based only on very few haloes. Better statistics will be needed to reliably establish the behaviour at these mass scales.

4.2.2 *Black holes*

As mentioned in Section 2.2, a black hole with mass $M_{\rm seed}$ is placed into a halo whenever the on-the-fly friend-of-friends halo finder identifies a structure that is more massive than a threshold mass $M_{\rm FOF}$ and does not yet contain a black hole. At seeding, the surrounding gas is usually dilute and the black hole accretes at low rates with a long Bondi growth time-scale. This means that the growth is slower than the growth in stellar mass and therefore, the corresponding galaxy evolves horizontally in the $M_{\rm BH}$–$M_{\rm bulge}$ diagram (Fig. 5, upper plot). After some time, enough gas piles up around the black hole and produces higher accretion rates, allowing the black hole to eventually grow more rapidly, aided also by the runaway character of Bondi growth due to its $\dot{M}_{\rm BH} \propto M_{\rm BH}^2$ scaling. Consequently, the slope in the $M_{\rm BH}$–$M_{\rm bulge}$ diagram steepens. This second phase continues until the feedback injection of the black hole into its surroundings becomes significant, at which point the black hole gas supply becomes self-regulated. In this final stage, the black holes grow less rapidly and are mostly in the low-accretion state again.

However, the slight change in slope in the $M_{\rm BH}$–$M_{\rm bulge}$ relation at $M_{\rm bulge} \approx 10^{10}$ M$_\odot$ is not due to the change of accretion mode, but rather due to the bulge-to-disc decomposition. We define the bulge mass as twice the stellar mass of the counter-rotating star particles within $0.1\,R_{200,\,c}$. Galaxies with $M_{\rm bulge} \approx 10^{10}$ M$_\odot$ have a large fraction of corotating stars (i.e. a disc) and correspondingly our estimate

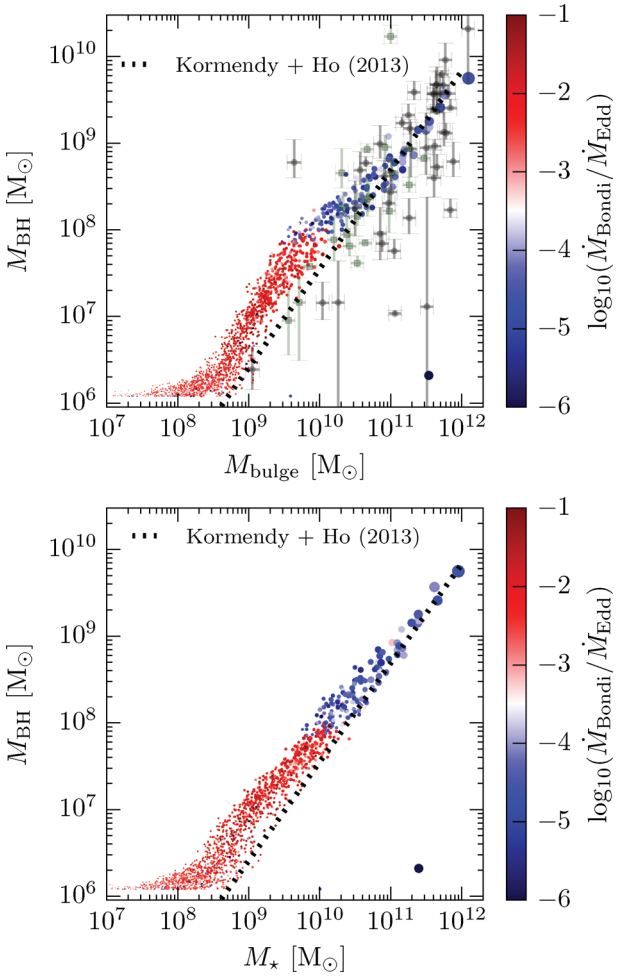

Figure 5. Black hole mass as a function of bulge mass (upper plot) and stellar mass within twice the half-mass radius (lower plot) for central galaxies in the high-resolution simulation. The size of the symbols is scaled with bulge mass for better visibility, and the assigned colour scale encodes the Eddington ratio. The dotted line is the fit to observational data. The symbols with error bars are observed ellipticals (black), and spirals or S0 galaxies with normal bulges (green), taken from Kormendy & Ho (2013). The bulge mass is estimated as twice the mass of the counter-rotating fraction of stars within $0.1\,R_{200,\,c}$. We note that this might slightly underestimate the bulge mass in the case of rotating bulges.

of the bulge mass is reduced, which shifts the corresponding points to the left in the upper plot of Fig. 5. In the $M_{\rm BH}-M_*$ plot (Fig. 5, lower plot), with M_* being the mass of all the stars within twice the stellar half-mass radius, such a change in slope does not show up.

Generally speaking, our systems with $M_* < 10^{10.5}$ M$_\odot$ tend to have slightly overly massive black holes compared to the observed relation, which indicates too early growth of the black holes, possibly caused by the increased seeding mass we use compared to earlier work (Sijacki et al. 2015). However, Volonteri et al. (2016) showed that the shape and scatter of the low-mass end changes significantly for different ways to measure $M_{\rm bulge}$. Considering this effect and the observational uncertainties, the discrepancy is not particularly worrisome. For high-mass systems, we follow the observed relation more closely, seemingly with little scatter. We leave a detailed analysis of the high-mass end to future work as it requires a larger sample of black holes.

The black hole population has a clear change in accretion rate at black hole masses of around 10^8 M$_\odot$ (colour coded in Fig. 5).

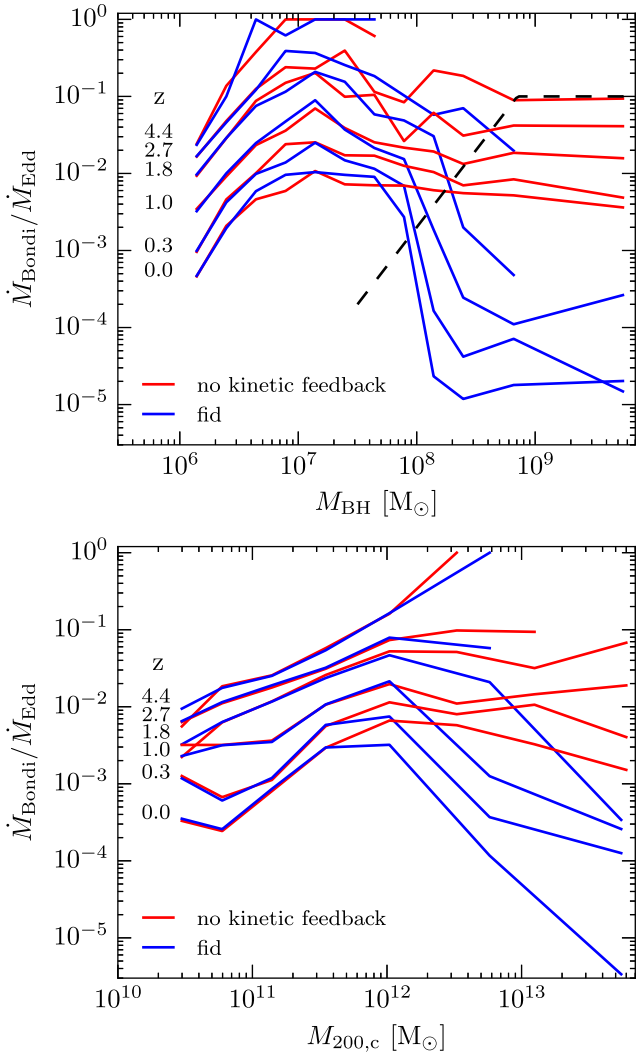

Figure 6. Median Eddington ratio as a function of black hole mass (top) and halo mass (bottom). The different lines show different redshifts. The dashed line in the top plot indicates our imposed transition point between low- and high-accretion states.

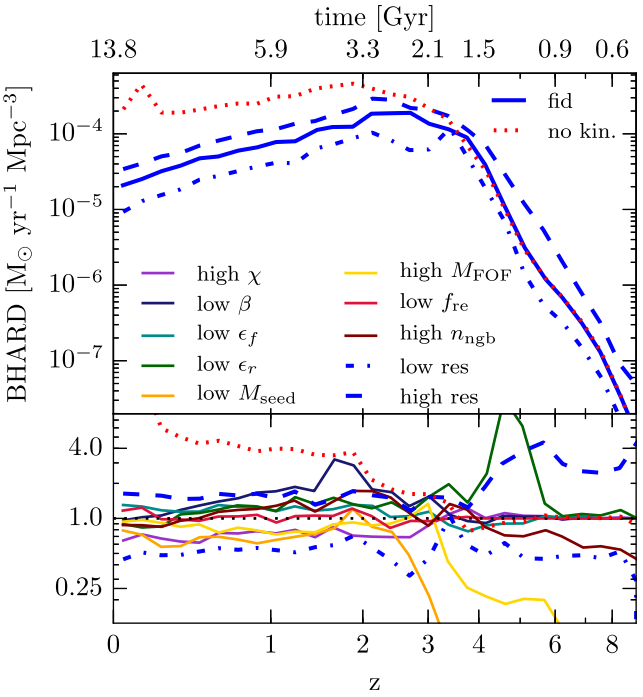

Figure 7. BHARD as a function of redshift for different runs. The lower panel shows the ratio relative to the fiducial run.

Fig. 6 shows the accretion rate in units of $\dot{M}_{\rm Edd}$ as a function of black hole mass for different redshifts. For comparison, we also show the Eddington factors in the run without kinetic feedback. One clear trend is the drop of the Eddington ratio with redshift, which is consistently present over the complete range of black hole masses. This is expected, as the black hole accretion rate density (BHARD) (Fig. 7) in M$_\odot$ yr^{-1} Mpc^{-3} decreases significantly towards low redshifts. The downturn towards the low black hole mass end shows the relatively slow initial growth of the black holes. This is partially due to a smaller amount of cold gas and partially due to the fact that in our implementation the Eddington ratio depends linearly on the black hole mass. A perhaps unexpected feature is that, without kinetic feedback, the Eddington factor does not significantly vary with black hole mass for black holes more massive than 10^7 M$_\odot$. This means that even the most massive black holes in galaxy group and cluster environments would have a good chance of accreting at high Eddington ratios, which is almost invariably also associated with high star formation rates.

This can be prevented by the kinetic feedback of the low-accretion mode, which is more efficient than the feedback in the high-accretion state. To ensure that a black hole and its surrounding gas transition to a self-regulated state with lower accretion rate, and to prevent newly seeded low-mass black holes from remaining in the low-accretion state, we employ a black hole mass dependent quasar threshold χ, shown as the dashed line in the top panel of Fig. 6. Once a black hole transitions to the kinetic feedback mode, its Eddington factor drops significantly as a consequence of the stronger feedback, making it likely to remain in this regime for an extended period of time. Note that the deviation of the median curve starts slightly to the left of the dashed line due to the scatter in the black hole properties, allowing some lower mass black holes to make the transition earlier than the mean.

The drop in Eddington ratio for high-mass systems also has an effect on the overall BHARD, shown in Fig. 7. In fact, for the simulations without kinetic feedback, there is no significant drop of the BHARD towards lower redshifts. There is an increased BHARD with higher resolution, which is related to the fact that the region in which the accretion rate is estimated is intentionally reduced with increasing resolution, which leads to systematically higher density estimates. Especially at early times, this leads to earlier and thus faster accretion, and consequently more massive black holes.

4.2.3 Stellar component

We now turn to the effect of black hole feedback on the stellar properties of galaxies. Fig. 8 shows the average star formation rate density (SFRD) as a function of redshift. At redshifts lower than $z = 3$, the fiducial simulation differs from the observed SFRD by about 0.3 dex. For our higher resolution simulation (dashed line), the SFRD is in better agreement with the observations. At low redshifts, the contribution from Milky-Way-sized galaxies dominates, which indicates a relatively poor convergence in their star formation rates. In Appendix B, we discuss this in more detail.

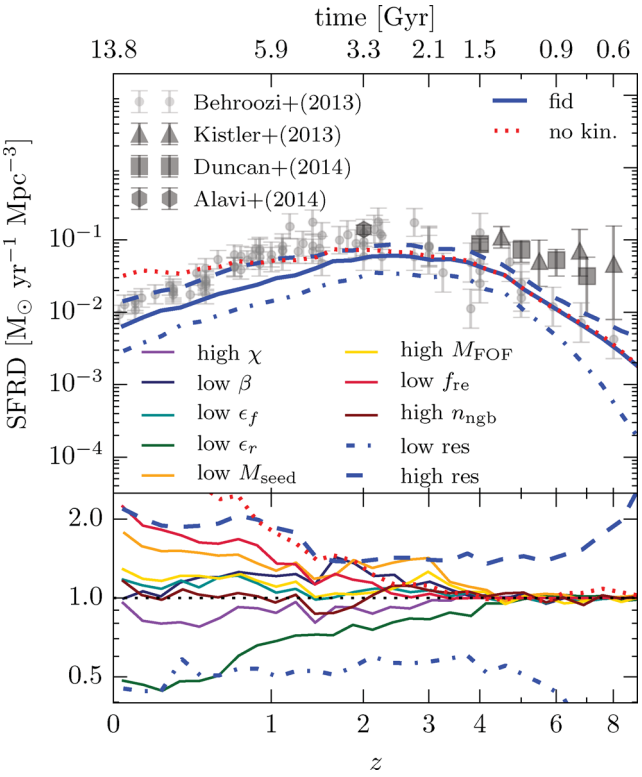

Figure 8. SFRD as a function of redshift. The grey dots, triangles, squares and hexagons are observational data from Behroozi, Wechsler & Conroy (2013a,b), Kistler, Yuksel & Hopkins (2013), Duncan et al. (2014) and Alavi et al. (2014), respectively. The lower panel shows the ratio of the SFRD relative to the fiducial run.

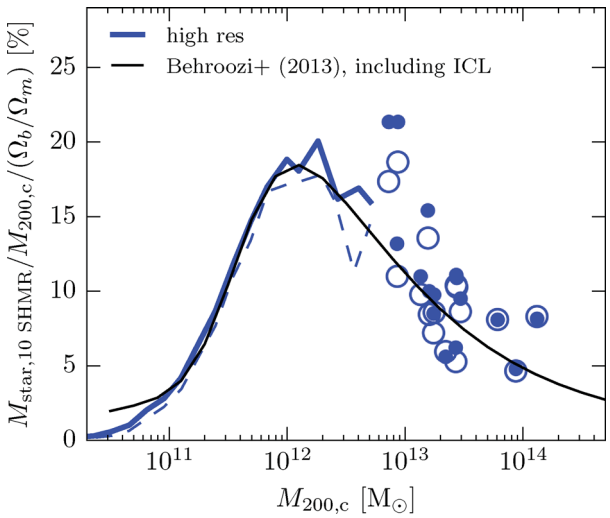

Figure 9. Stellar mass fraction as a function of halo mass for the high-resolution simulation. The stellar mass is calculated as the mass of all star particles within 10 stellar half-mass radii that do not belong to a subhalo. The black line is the corresponding fit to observations from Behroozi et al. (2013a,b) including the intracluster light. We use the simulation values of $\Omega_{\rm b}$ and $\Omega_{\rm m}$ for both simulation and literature data. The dashed line and open circles correspond to the same simulation data, but the halo mass $M_{200,\,c}$ is taken from the corresponding halo in the dark matter only simulation (see Munshi et al. 2013, for a discussion).

From an AGN-feedback point of view, galaxy formation can be divided into three epochs. At redshift $z > 5$, there are either no or only slowly accreting black holes with no significant impact on the host galaxy. During this stage, only stellar feedback regulates the star formation rate (e.g. Vogelsberger et al. 2013, their fig. 15). Correspondingly, changes in the black hole parameters have no effect on the SFRD. After this initial phase, the black holes enter the high-accretion regime and grow quickly, releasing a considerable amount of thermal feedback energy that suppresses star formation, in particular in galaxies with a final halo mass $>10^{12}\,{\rm M}_\odot$. At late times, from redshift $z = 2$ to the present day, the black holes switch to the low-accretion regime again, remaining in a self-regulated state in which both the stellar and AGN feedback balance cooling. The relative importance of AGN over stellar feedback depends on halo mass. While stellar feedback dominates in haloes up to the size of the Milky Way, more massive haloes are mainly regulated through AGN feedback.

The stellar mass fraction as a function of halo mass (Fig. 9) clearly shows the decrease in star formation efficiency with halo mass at the massive end, in good agreement with observations. We find it useful to compare the stellar mass including the diffuse intracluster light in the high-mass end in observations and theory. In this way, we are less sensitive to the choice of the aperture within which stellar masses are estimated. At the high halo-mass end, the data shown here are in reasonable agreement with Kravtsov, Vikhlinin & Meshscheryakov (2014) who pointed out the importance of outer stellar profiles in high-mass haloes in this type of analysis. Additionally, we plot both the halo mass from the full physics simulation as well as the halo mass of the corresponding halo in the dark matter only run, where the latter, i.e. the dashed line and open circles, should be compared to the results from abundance-matching. Fig. 4 shows the ratio of these two masses as a function of halo mass. In particular at around $M_{200,c} \approx 10^{13}\,{\rm M}_\odot$, it turns out to be crucial to take this effect into account.

One of the conjectured effects of AGN is that they can prevent the most massive galaxies from being blue and star forming, instead making them red and having an old stellar population. We use $B-V$ colour and mean stellar age (Fig. 10) as a measure for the efficiency of the feedback in the low-accretion state to accomplish this. To probe the relevance of the kinetic feedback mode for this, we compare our fiducial simulation with a simulation without the kinetic mode. The panels of Fig. 10 clearly show the need for this efficient mode to get 'red and dead galaxies with old stellar populations on the massive end of the galaxy population.

4.2.4 *Gas component*

The low gas fraction of haloes around $10^{13}\,{\rm M}_\odot$ has been identified as one of the main shortcomings of the Illustris simulation (Genel et al. 2014). The gathering of substantial amounts of feedback energy invoked in the bubble model of Illustris, and its explosive release once enough energy is available, does prevent the feedback energy from being quickly lost due to cooling, but it also expels a significant fraction of gas from the inner halo. This resulted in a gas fraction which is factor of a few too low in systems where the feedback is most efficient. In Fig. 11, we show the gas fractions within $R_{500,\,c}$ as a function of their mass $M_{500,\,c}$ is obtained with our new kinetic feedback model. Reassuringly, it does not expel too much gas from the inner halo, but rather heats it via shocks and drives turbulence in the halo core, leading to an overall good agreement with observations.

To further investigate the effect of AGN feedback on the gas properties it is instructive to look at the radial profiles of the gas

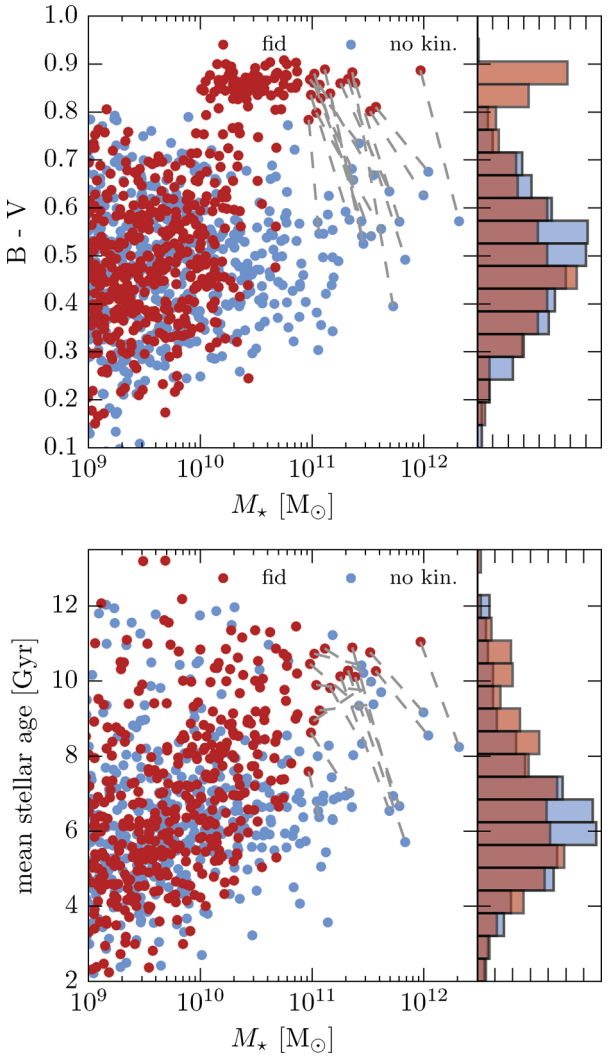

Figure 10. Top panel: $B-V$ colours as a function of stellar mass within twice the stellar half-mass radius. The red dots are from the fiducial simulations. The blue dots represent the haloes in the run without kinetic feedback. For the most massive systems, the dashed lines link the same haloes in the two different runs to emphasize the effect of the kinetic feedback mode on a halo-by-halo basis. The histogram on the side clearly shows the emergence of a red (larger $B-V$ values) population of galaxies due to the kinetic feedback model. Bottom panel: mass-weighted stellar age as a function of stellar mass within twice the stellar half-mass radius. Note that the choice of colours differs from the other figures.

distribution. To this end, we use the high-resolution simulation and plot the density, temperature and entropic function profiles in Fig. 12. For the most massive haloes with a mass around $10^{14}\,\mathrm{M}_\odot$, the temperature profiles are almost flat in the centre and the central entropic function $K = k_\mathrm{B}\,T n^{-2/3}$ has a value of around 50 keV cm^2. This confirms that the efficient quenching of star formation is not due to overly heating and diluting the central gas. For haloes less massive than $10^{13.5}\,\mathrm{M}_\odot$, the density profiles are more centrally peaked and the temperatures in the centres are lower, which indicates that these haloes might have some residual star formation. As the volume is relatively small, our simulations do not contain massive galaxy clusters for which we could compare the thermodynamic profiles with observations of local galaxy clusters. Such simulations of galaxy clusters and how they are impacted by different

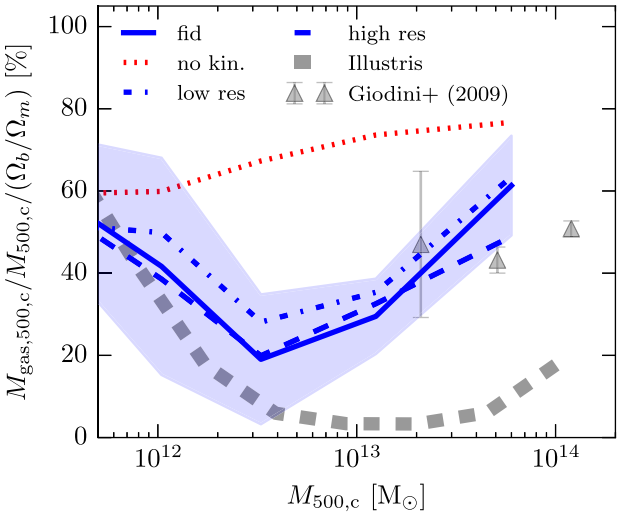

Figure 11. Gas mass fraction in $R_{500,\,c}$ as a function of halo mass $M_{500,\,c}$. The triangles show the binned data from Giodini et al. (2009). The dashed grey line represents the gas fractions in the Illustris simulation (Genel et al. 2014).

AGN models are analysed in detail in forthcoming work (Popa et al., in preparation).

5 DEPENDENCE ON MODEL PARAMETERS

We now investigate how robust the findings discussed in Section 4 are against changes in the parameters of our new black hole model. As we run identical initial conditions with several different parameter settings, we can compare their effects on a halo-by-halo basis. Fig. 13 shows the relative changes in the gas-, stellar- and black hole masses binned with respect to halo mass, as well as the star formation rate as a function of redshift for systems with different halo mass at $z = 0$. We shall first discuss the variations due to modifications of the efficiency parameters ϵ_f and ϵ_r, and then consider the other parameters in turn.

5.1 Global properties

For all the investigated changes of model parameters, there is generally only a weak change in the late time accretion rate density (Fig. 7). At higher redshift, the seeding parameters have however a significant impact on the accretion and growth history of the black holes. In our tests, we lower the seed mass M_seed or increase the halo mass M_FOF at which black holes are seeded, which both delay black hole growth. Lowering the radiative efficiency ϵ_r leads to a higher Eddington accretion limit and therefore an increased accretion rate at $z = 5$. As soon as the accretion rate is feedback regulated, it drops back to the fiducial rate. There is also an increase in the BHARD around $z = 2$ for the simulation with low quasar threshold slope β. This can be explained by a delayed transition to kinetic feedback for the black hole population.

We now focus on the change in the SFRD for different parameter settings, shown in Fig. 8. The most significant variations occur for the simulations with modified values of M_seed, ϵ_r or f_re. Changes in these parameters manifest themselves in the global SFRD after redshift $z = 4$, and the effect increases at later times. However, it is also evident that a factor of 2 change in spatial resolution has a stronger impact on the global SFRD than a factor of 4 change in any of the black hole parameters. The significant decline in SFRD

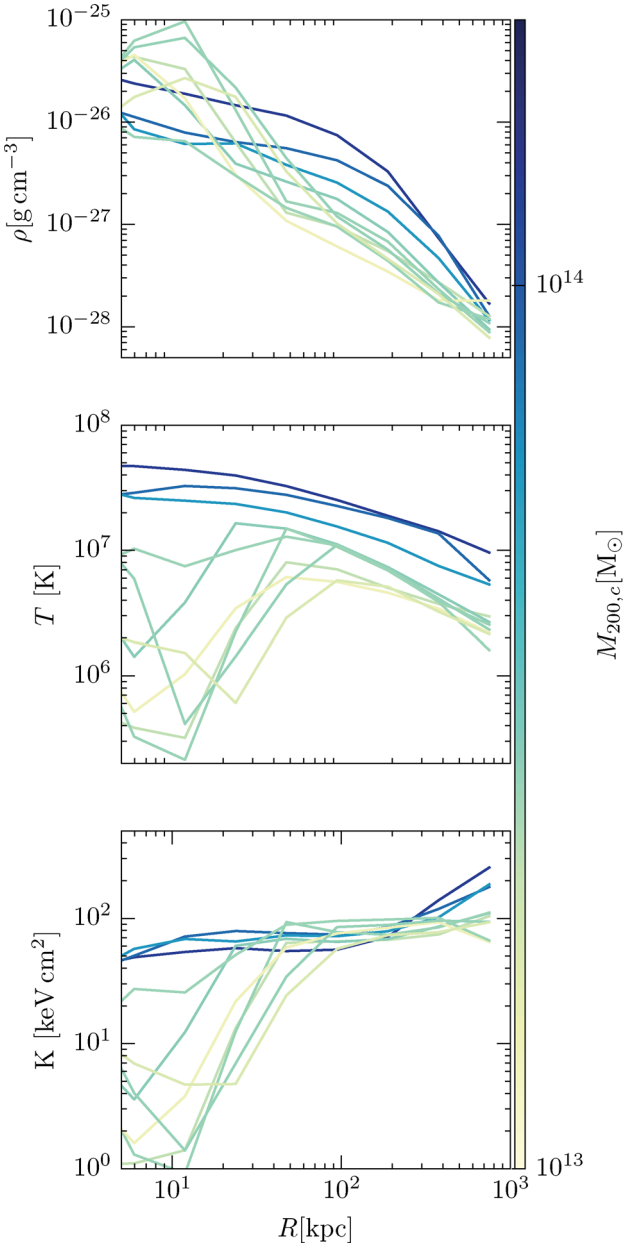

Figure 12. Mass-weighted density (top), temperature (middle) and entropic function (bottom) of the 10 most massive haloes in the high-resolution simulation. The line colour encodes $M_{200,c}$ of the halo.

for a lower ϵ_r can be explained by the increased accretion rate due to the lower feedback energy per accreted mass. The overmassive black holes in turn have a significantly larger impact on the star formation rate in the host galaxies. This will be further investigated in Section 5.2.1.

5.2 Halo-by-halo comparison

5.2.1 Efficiency parameters

The feedback efficiency $\epsilon_{\rm f,\,kin}$ is defined as the fraction of accreted rest mass energy that appears kinetically in the kinetic wind mode. Lowering this efficiency by a factor of 4 reduces the amount of feedback energy, which leads to an increase in the gas fraction and stellar mass, in particular in the high-mass systems. Having a higher gas fraction and star formation rate indicates that the central gas is denser and has lower temperatures than in the fiducial run. This leads to slightly increased black hole masses. Overall, the effect of this drastic change in kinetic feedback efficiency is rather small, which can be explained by the self-regulated nature of the feedback cycle.

The radiative efficiency ϵ_r determines how much of the accreted mass is converted to AGN luminosity in the quasar mode. To achieve a reduction of feedback energy for all black holes independent of accretion mode, we also reduce $\epsilon_{\rm f,\,kin}$ in the corresponding test. Lowering the overall feedback efficiency by a factor of 4 also increases the Eddington limit by the same factor, which leads to a significantly faster growth of the black holes (bump at $z = 5$ in Fig. 7), but leaves the injected feedback energy for a black hole accreting at that limit constant, given the same black hole mass. As the black holes accrete more, they become more massive and have therefore a more significant impact on their surroundings. This increases the quenching in all systems and expels gas from Milky-Way-sized galaxies. As the mass of the black hole increases, the threshold for them to be in the kinetic feedback mode also increases; i.e. more energy is injected in this mode. This might be an additional amplifying factor for the low gas fractions and the efficient quenching of these systems. The fact that a lowering of the quasar-threshold has similar effects (see below) indicates that this is indeed the case.

5.2.2 Accretion rate dependences

The quasar threshold χ determines whether a black hole is associated with the low- or high-accretion rate state. This means that in our test simulation (higher χ), the Bondi accretion estimate is four times higher when it transitions from the high-accretion state to the low-accretion state. This seems to have no effect on the initial growth of the black holes, which indicates that the black holes easily exceed the threshold at early times and accrete most of their mass in the high-accretion state. At lower redshift, however, the average Eddington rates decrease to a level where the increase of the threshold by a factor of 4 matters. As our model involves a mass dependence of χ, the higher Eddington rate threshold can also be interpreted as a lowering of the black hole mass for which, at a fixed Eddington factor, a black hole transitions between thermal and kinetic feedback modes. As the black hole mass correlates with halo mass, this means that the mass scale at which black holes are pre-dominantly in the kinetic mode is effectively shifted to lower masses. And because the kinetic feedback mode is comparatively more efficient at quenching a halo, this explains the dip in gas and stellar mass at around 3×10^{12} M$_\odot$. The lowered black hole masses above this mass scale can be explained by the fact that the black holes only grow significantly in the quasar mode. When the kinetic mode is switched on earlier, the black holes end up systematically less massive, provided they reach the transition threshold in the first place.

The slope β influences the adopted scaling of the Eddington ratio with black hole mass for setting the transition between quasar and kinetic mode. Reducing it by a factor of 4 as done in our test means that low-mass black holes (below 10^8 M$_\odot$) will be found more often in the kinetic mode, and higher mass black holes more often in the quasar mode, compared to our default run. This explains the relative increase in black hole mass towards high-mass haloes, keeping in mind that black holes predominantly grow in the quasar mode. Knowing that the quasar mode is less efficient at quenching, this also explains the larger stellar masses for high-mass systems as well as the fact that the additional stars form mainly at higher

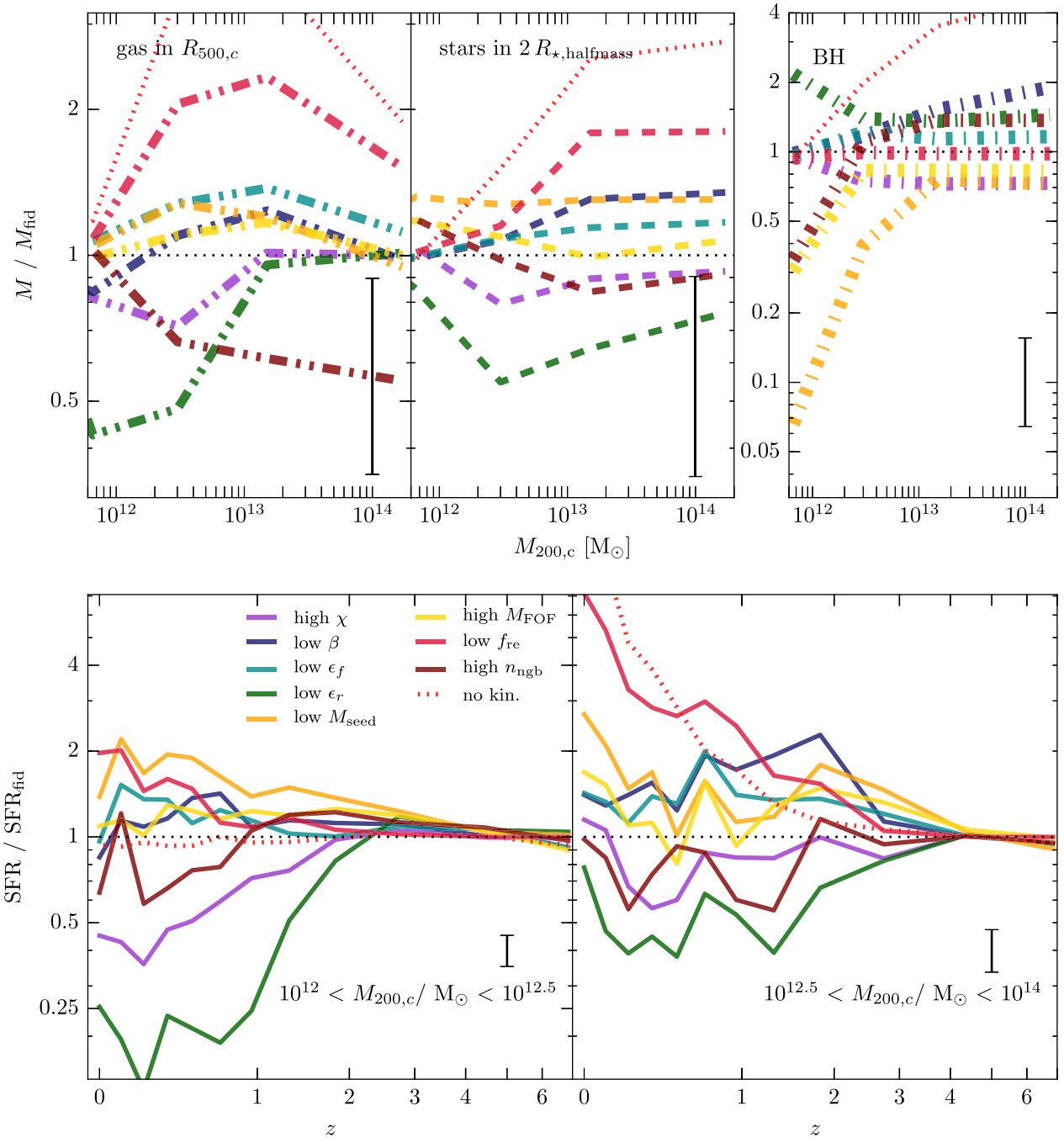

Figure 13. Upper panels (from left to right): gas mass in $R_{500,\,c}$, stellar mass in twice the stellar half-mass radius and black hole mass relative to their values in the fiducial run, compared on a halo-by-halo basis and binned as a function of the fiducial $M_{200,\,c}$. For clarity, we do not plot the individual scatter, but indicate with the black error bar the average scatter. Lower panels: average star formation rate of systems of different final mass (in the fiducial run) in units of their star formation rate in the fiducial simulation at a given redshift, on a halo-by-halo basis as a function of redshift. The error bar indicates the uncertainty of the mean.

redshift, at the time when the haloes are delayed in switching to the kinetic mode.

5.2.3 Black hole seeding parameters

The black hole seed mass M_{seed} is the initial mass given to the black holes when they are inserted in newly emerging haloes. If the black holes begin their evolution with smaller masses as in our test, their growth time-scale is considerably longer because the Bondi accretion rate scales as $\dot{M}_{\mathrm{Bondi}} \propto M_{\mathrm{BH}}^2$. This means that the black holes grow significantly later, as more gas needs to accumulate in the halo centres to start a rapid growth. This delayed growth implies that the black holes in $10^{12}\,\mathrm{M}_\odot$ systems have not yet ended their rapid accretion phase at $z = 0$. This explains why the mean black hole mass is an order of magnitude below the mass in the fiducial run. For the more massive haloes, the black holes formed earlier and already had enough time to catch up; however, they are still about 25 per cent less massive. The delayed growth also

implies that there is less feedback energy injected into the galaxies at all times, which increases the star formation rates and the stellar masses over the whole mass range of haloes. The fact that the gas fraction is comparatively higher, particularly at $3 \times 10^{12}\,\mathrm{M}_{\odot}$, is again due to the mass-dependent switch from quasar to the kinetic mode. As the black holes are less massive in the modified run, they mostly remain in the quasar mode, keeping a relatively high gas fraction while the corresponding black holes in the fiducial run have switched to kinetic feedback, which lowers the gas fractions by 20–30 per cent.

The halo mass $M_{\rm FOF}$ at which black holes are seeded has a similar effect: lowering $M_{\rm seed}$ and increasing $M_{\rm FOF}$ both lead to a delayed black hole growth. In our test simulation, we place the black holes only in haloes that have grown a factor of 4 more in mass compared to our fiducial simulation. This produces similar trends, but the effect is much weaker. The delay of the black hole growth is not as severe as in the previous case, which can be explained by the accretion rate dependence $\dot{M}_{\rm Bondi} \propto M_{\rm BH}^2$, which means that the black holes have a 16 times higher accretion rate for the same gas properties but are seeded in *four* times more massive haloes. This means that they do not need a similarly severe change of gas properties as the low-mass seeds to eventually grow into the Eddington-limited accretion phase.

All in all, the dependence on the seeding prescription reveals one of the most important theoretical uncertainties of the black hole modelling in cosmological simulations. The formation and the early growth of SMBHs are observationally as well as theoretically very poorly understood (see Volonteri 2010, for a review). However, as we just showed, they have a major impact on the evolution of galaxy properties. One way to reduce these uncertainties from a simulation point of view is to constrain the model with observations that also crucially depend on the seeding and early growth phase, such as the low-mass end of the $M_{\rm BH}$–σ and $M_{\rm BH}$–$M_{\rm bulge}$ relations, or the abundance of high-redshift quasars.

5.2.4 *Other parameter dependences*

The reorientation factor $f_{\rm re}$ determines the energy threshold at which a new kinetic feedback event along a new direction is injected. The specific parametrization we adopted sets the magnitude of the velocity kicks relative to the local dark matter velocity dispersion. The remarkable thing about lowering this parameter by a factor of 4 is that the black hole mass does not change at all, while the star formation rate and correspondingly the stellar masses as well as the gas mass increase significantly. This means that this factor substantially changes the efficiency of the kinetic feedback and therefore the properties of the high-mass haloes. The fact that the burstiness of the feedback has such a dramatic impact is in agreement with other works (e.g. Le Brun et al. 2014; Sijacki et al. 2015) and can in our case be explained by the fact that the velocity kicks directly determine the strength of the resulting shocks as well as the post-shock temperature. The faster the velocity, the higher the post-shock temperature and the lower the cooling losses during the process. With the adopted parameters, we reach velocity kicks up to several tens of thousand $\mathrm{km\,s^{-1}}$ in the largest haloes, which are realistic speeds for winds from optically thin accretion discs. This means that one could in principle try to constrain this parameter, both theoretically from small-scale GRMHD simulations of hot accretion flows (Yuan et al. 2015) as well as from observations (Tombesi et al. 2014).

The number of neighbours $n_{\rm ngb}$ sets the number of cells used for the density, sound speed and velocity estimates, as well as for the injection region of feedback energy. Increasing this number means that, at a fixed resolution, the radius out to which the gas properties are probed increases. As the gas properties change with radius, the accretion rate estimate tends to change as well. This has important consequences for the black holes in low-mass systems because here the black hole growth is delayed when we average over a four times larger number of cells, which can be seen in Fig. 7. This leads to a lower black hole mass for galaxies in $10^{12}\,\mathrm{M}_{\odot}$ haloes. For more massive haloes, the black hole mass increases by about 25 per cent, and correspondingly lower gas and stellar mass fractions are reached. However, Fig. 7 also reveals a higher BHARD between redshifts $z = 3$ and 0.5, which is responsible for the more massive black holes. This is because the quasar mode distributes the energy over more mass, leading to lower temperatures in the heated gas and higher radiative cooling losses. In the kinetic mode, the larger injection volume means that we implicitly increase the burstiness of the model, which increases its efficiency at quenching star formation. This explains the lower star formation rate for the most massive haloes and the lower star formation rate after $z = 1$ for haloes in the mass range $10^{12}\,\mathrm{M}_{\odot} < M_{200,\,c} < 10^{12.5}\,\mathrm{M}_{\odot}$.

6 CONCLUSIONS

In this study, we introduced a new model for SMBH growth and the associated feedback in cosmological simulations of galaxy formation. We distinguish between a state of high and a state of low accretion, which are associated with pure thermal or pure kinetic feedback, respectively. Unlike in previous work, we omit an artificial boost factor α in the accretion rate estimate to account for unresolved ISM structure, and instead adopt an accretion rate given by the Bondi formula throughout. The feedback energy in the high-accretion rate state is released with a continuous thermal feedback prescription. In the low-accretion state, we instead use pulsed kinetic feedback injection in random directions, which is the primary new element adopted in this study. We have shown in idealized simulations that this mode drives shocks in the surrounding gas, thermalizing a significant fraction of the AGN energy within an Myr.

In simulations of cosmological structure formation, our new model is able to significantly reduce star formation in the most massive haloes, leading to a stellar mass fraction in excellent agreement with observations, without overly heating and diluting the central gas. This resolves one of the central problems in the Illustris simulation. It also leads to massive galaxies with a red, old stellar population, living in haloes that have gas fractions in agreement with observations.

The star formation efficiency peaks in haloes with a few times $10^{12}\,\mathrm{M}_{\odot}$, in very good agreement with abundance matching expectations once we use the halo masses from dark matter only simulations for the comparison, as also used in the fits to observations on which the abundance models are based. The key to sustained quenching of massive haloes in our simulations is to ensure that the black holes in these systems transition to the low-accretion state and remain in it for most of their subsequent evolution. We encourage this behaviour by employing a BH mass-dependent Eddington ratio threshold for determining the accretion state, making it progressively easier for high-mass black holes to be in the kinetic mode. Once the black holes reach this mode, the more efficient coupling of the kinetic feedback and the self-regulated nature of gas accretion will typically keep the black holes accreting at low Eddington rates. Brief interruptions of this with episodes of quasar activity, triggered for example by significant inflows of cold gas during a galaxy merger, may nevertheless occur.

We analysed the impact of each of our black hole model parameters on the cosmic star formation rate history and the stellar, gas and black hole masses. To this end we varied each parameter by a factor of 4 and carried out otherwise identical simulations to our default model. We found that most of the parameters do not alter the global properties severely, but some of them can have a significant impact on a subset of haloes and galaxies over particular mass ranges. In these cases, the changes can be readily understood in terms of the tightly self-regulated nature of black hole growth that occurs in our models. We would like to emphasize that the assumption of the existence of a low-accretion rate state with efficient kinetic feedback is more important than the precise value of any of the model parameters.

The new AGN feedback model discussed here significantly improves the galaxy formation model explored previously in the Illustris simulation project, particularly at the high-mass end of the galaxy population. It therefore promises to be an excellent starting point for a new generation of hydrodynamical simulations of galaxy formation that allow much improved predictions for the bright end of the galaxy population, and for groups and clusters of galaxies, as well as their thermodynamic scaling relations. Future work with this model in high-resolution simulations of galaxy formation could potentially also shed light on the physical origin of observed centrally concentrated radio emission (Baldi, Capetti & Giovannini 2015, 2016), AGN-driven nuclear outflows (Förster Schreiber et al. 2014; Tombesi et al. 2014) and related phenomena.

ACKNOWLEDGEMENTS

The authors thank Peter Behroozi for providing his data and for useful advice, as well as Kevin Schaal for providing his shock finding algorithm. RW, VS and RP acknowledge support through the European Research Council under ERC-StG grant EXAGAL-308037. RW, VS and RP would like to thank the Klaus Tschira Foundation. RW acknowledges support by the IMPRS for Astronomy and Cosmic Physics at the University of Heidelberg. SG and PT acknowledge support provided by NASA through Hubble Fellowship grant HST-HF2-51341.001-A and HF2-51384.001-A, respectively, awarded by the STScI, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. LH acknowledges support from NASA grant NNX12AC67G and NSF grant AST-1312095. Simulations were run on the HazelHen supercomputer at the High-Performance Computing Center Stuttgart (HLRS) as part of project GCS-ILLU of the Gauss Centre for Supercomputing (GCS).

REFERENCES

Alavi A. et al., 2014, ApJ, 780, 143
Anglés-Alcázar D., Özel F., Davé R., 2013, ApJ, 770, 5
Anglés-Alcázar D., Özel F., Davé R., Katz N., Kollmeier J. A., Oppenheimer B. D., 2015, ApJ, 800, 127
Anglés-Alcázar D., Davé R., Faucher-Giguère C.-A., Özel F., Hopkins P. F., 2017, MNRAS, 464, 2840
Baldi R. D., Capetti A., Giovannini G., 2015, A&A, 576, A38
Baldi R. D., Capetti A., Giovannini G., 2016, Astron. Nachr., 337, 114
Barnes J. E., Hernquist L., 1996, ApJ, 471, 115
Basson J. F., Alexander P., 2003, MNRAS, 339, 353
Bauer A., Springel V., 2012, MNRAS, 423, 2558
Begelman M. C., 2014, preprint (arXiv:1410.8132)
Behroozi P. S., Wechsler R. H., Conroy C., 2013a, ApJ, 762, L31
Behroozi P. S., Wechsler R. H., Conroy C., 2013b, ApJ, 770, 57
Bellovary J., Volonteri M., Governato F., Shen S., Quinn T., Wadsley J., 2011, ApJ, 742, 13
Bieri R., Dubois Y., Rosdahl J., Wagner A., Silk J., Mamon G. A., 2017, MNRAS, 464, 1854
Blandford R. D., Znajek R. L., 1977, MNRAS, 179, 433
Bondi H., 1952, MNRAS, 112, 195
Bondi H., Hoyle F., 1944, MNRAS, 104, 273
Booth C. M., Schaye J., 2009, MNRAS, 398, 53
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
Brighenti F., Mathews W. G., 2006, ApJ, 643, 120
Brüggen M., Heinz S., Roediger E., Ruszkowski M., Simionescu A., 2007, MNRAS, 380, L67
Bu D.-F., Yuan F., Gan Z.-M., Yang X.-H., 2016, ApJ, 818, 83
Cattaneo A., Teyssier R., 2007, MNRAS, 376, 1547
Cheung E. et al., 2016, Nature, 533, 504
Choi E., Ostriker J. P., Naab T., Johansson P. H., 2012, ApJ, 754, 125
Choi E., Naab T., Ostriker J. P., Johansson P. H., Moster B. P., 2014, MNRAS, 442, 440
Choi E., Ostriker J. P., Naab T., Oser L., Moster B. P., 2015, MNRAS, 449, 4105
Cielo S., Antonuccio-Delogu V., Macciò A. V., Romeo A. D., Silk J., 2014, MNRAS, 439, 2903
Ciotti L., Ostriker J. P., Proga D., 2010, ApJ, 717, 708
Costa T., Sijacki D., Haehnelt M. G., 2014, MNRAS, 444, 2355
Croton D. J. et al., 2006, MNRAS, 365, 11
Curtis M., Sijacki D., 2015, MNRAS, 454, 3445
Curtis M., Sijacki D., 2016, MNRAS, 463, 63
Debuhr J., Quataert E., Ma C.-P., Hopkins P., 2010, MNRAS, 406, L55
Debuhr J., Quataert E., Ma C.-P., 2011, MNRAS, 412, 1341
Debuhr J., Quataert E., Ma C.-P., 2012, MNRAS, 420, 2221
Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604
Di Matteo T., Colberg J., Springel V., Hernquist L., Sijacki D., 2008, ApJ, 676, 33
Dopita M. A. et al., 2015, ApJ, 801, 42
Dubois Y., Devriendt J., Slyz A., Teyssier R., 2010, MNRAS, 409, 985
Dubois Y., Devriendt J., Slyz A., Teyssier R., 2012, MNRAS, 420, 2662
Dubois Y., Volonteri M., Silk J., Devriendt J., Slyz A., Teyssier R., 2015, MNRAS, 452, 1502
Dubois Y., Peirani S., Pichon C., Devriendt J., Gavazzi R., Welker C., Volonteri M., 2016, MNRAS, 463, 3948
Duncan K. et al., 2014, MNRAS, 444, 2960
Dunn R. J. H., Fabian A. C., 2006, MNRAS, 373, 959
Dunn R. J. H., Fender R. P., Körding E. G., Belloni T., Cabanac C., 2010, MNRAS, 403, 61
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Emsellem E., Renaud F., Bournaud F., Elmegreen B., Combes F., Gabor J. M., 2015, MNRAS, 446, 2468
Förster Schreiber N. M. et al., 2014, ApJ, 787, 38
Gaibler V., Khochfar S., Krause M., Silk J., 2012, MNRAS, 425, 438
Gaspari M., Melioli C., Brighenti F., D'Ercole A., 2011a, MNRAS, 411, 349
Gaspari M., Brighenti F., D'Ercole A., Melioli C., 2011b, MNRAS, 415, 1549
Gaspari M., Brighenti F., Temi P., 2012, MNRAS, 424, 190
Gaspari M., Ruszkowski M., Oh S. P., 2013, MNRAS, 432, 3401
Genel S. et al., 2014, MNRAS, 445, 175
Genzel R. et al., 2014, ApJ, 796, 7
Giodini S. et al., 2009, ApJ, 703, 982
Granato G. L., De Zotti G., Silva L., Bressan A., Danese L., 2004, ApJ, 600, 580
Hernquist L., 1989, Nature, 340, 687
Hillel S., Soker N., 2016, MNRAS, 455, 2139
Hirschmann M., Dolag K., Saro A., Bachmann L., Borgani S., Burkert A., 2014, MNRAS, 442, 2304
Hopkins P. F., Quataert E., 2010, MNRAS, 407, 1529
Hopkins P. F., Quataert E., 2011, MNRAS, 415, 1027
Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2006, ApJS, 163, 1
Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008a, ApJS, 175, 356

Hopkins P. F., Cox T. J., Kereš D., Hernquist L., 2008b, ApJS, 175, 390
Hopkins P. F., Torrey P., Faucher-Giguère C.-A., Quataert E., Murray N., 2016, MNRAS, 458, 816
Hoyle F., Lyttleton R. A., 1939, Proc. Camb. Phil. Soc., 35, 405
Huarte-Espinosa M., Krause M., Alexander P., 2011, MNRAS, 418, 1621
Ichimaru S., 1977, ApJ, 214, 840
Igumenshchev I. V., Abramowicz M. A., 1999, MNRAS, 303, 309
Kawata D., Gibson B. K., 2005, MNRAS, 358, L16
Kereš D., Vogelsberger M., Sijacki D., Springel V., Hernquist L., 2012, MNRAS, 425, 2027
Khandai N., Di Matteo T., Croft R., Wilkins S., Feng Y., Tucker E., DeGraf C., Liu M.-S., 2015, MNRAS, 450, 1349
Kistler M. D., Yuksel H., Hopkins A. M., 2013, preprint (arXiv:1305.1630)
Kormendy J., Ho L. C., 2013, ARA&A, 51, 511
Kravtsov A., Vikhlinin A., Meshscheryakov A., 2014, preprint (arXiv:1401.7329)
Le Brun A. M. C., McCarthy I. G., Schaye J., Ponman T. J., 2014, MNRAS, 441, 1270
Li Y., Bryan G. L., 2014a, ApJ, 789, 54
Li Y., Bryan G. L., 2014b, ApJ, 789, 153
Li Y., Bryan G. L., Ruszkowski M., Voit G. M., O'Shea B. W., Donahue M., 2015, ApJ, 811, 73
McNamara B. R., Nulsen P. E. J., 2007, ARA&A, 45, 117
Marinacci F., Pakmor R., Springel V., 2014, MNRAS, 437, 1750
Medling A. M. et al., 2015, MNRAS, 448, 2301
Meece G. R., Voit G. M., O'Shea B. W., 2016, preprint (arXiv:1603.03674)
Mihos J. C., Hernquist L., 1996, ApJ, 464, 641
Munshi F. et al., 2013, ApJ, 766, 56
Okamoto T., Nemmen R. S., Bower R. G., 2008, MNRAS, 385, 161
Omma H., Binney J., Bryan G., Slyz A., 2004, MNRAS, 348, 1105
Pakmor R., Springel V., 2013, MNRAS, 432, 176
Pakmor R., Bauer A., Springel V., 2011, MNRAS, 418, 1392
Pakmor R., Springel V., Bauer A., Mocz P., Munoz D. J., Ohlmann S. T., Schaal K., Zhu C., 2016, MNRAS, 455, 1134
Pillepich A. et al., 2014, MNRAS, 444, 237
Planck Collaboration XIII, 2016, A&A, 594, A13
Pontzen A., Tremmel M., Roth N., Peiris H. V., Saintonge A., Volonteri M., Quinn T., Governato F., 2016, preprint (arXiv:1607.02507)
Puchwein E., Springel V., 2013, MNRAS, 428, 2966
Reynolds C. S., Heinz S., Begelman M. C., 2002, MNRAS, 332, 271
Roos O., Juneau S., Bournaud F., Gabor J. M., 2015, ApJ, 800, 19
Rosas-Guevara Y. M. et al., 2015, MNRAS, 454, 1038
Ruszkowski M., Brüggen M., Begelman M. C., 2004, ApJ, 611, 158
Schaal K., Springel V., 2015, MNRAS, 446, 3992
Schaal K. et al., 2016, MNRAS, 461, 4441
Schaye J. et al., 2015, MNRAS, 446, 521
Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337
Shapiro S. L., Lightman A. P., Eardley D. M., 1976, ApJ, 204, 187
Sijacki D., Springel V., Di Matteo T., Hernquist L., 2007, MNRAS, 380, 877
Sijacki D., Vogelsberger M., Kereš D., Springel V., Hernquist L., 2012, MNRAS, 424, 2999
Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481
Springel V., 2010, MNRAS, 401, 791
Springel V., Hernquist L., 2003, MNRAS, 339, 289
Springel V., Di Matteo T., Hernquist L., 2005, MNRAS, 361, 776
Steinborn L. K., Dolag K., Hirschmann M., Prieto M. A., Remus R.-S., 2015, MNRAS, 448, 1504
Sternberg A., Soker N., 2009, MNRAS, 395, 228
Sternberg A., Pizzolato F., Soker N., 2007, ApJ, 656, L5
Stone J. M., Pringle J. E., Begelman M. C., 1999, MNRAS, 310, 1002
Sądowski A., Lasota J.-P., Abramowicz M. A., Narayan R., 2016, MNRAS, 456, 3915
Teyssier R., Moore B., Martizzi D., Dubois Y., Mayer L., 2011, MNRAS, 414, 195
Tombesi F., Tazaki F., Mushotzky R. F., Ueda Y., Cappi M., Gofford J., Reeves J. N., Guainazzi M., 2014, MNRAS, 443, 2154
Trayford J. W., Theuns T., Bower R. G., Crain R. A., Lagos C. d. P., Schaller M., Schaye J., 2016, MNRAS, 460, 3925
Tremmel M., Karcher M., Governato F., Volonteri M., Quinn T., Pontzen A., Anderson L., 2016, preprint (arXiv:1607.02151)
Vogelsberger M., Sijacki D., Kereš D., Springel V., Hernquist L., 2012, MNRAS, 425, 3024
Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2013, MNRAS, 436, 3031
Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2014a, MNRAS, 438, 3607
Vogelsberger M. et al., 2014b, MNRAS, 444, 1518
Vogelsberger M. et al., 2014c, Nature, 509, 177
Volonteri M., 2010, A&AR, 18, 279
Volonteri M., Dubois Y., Pichon C., Devriendt J., 2016, MNRAS, 460, 2979
Wurster J., Thacker R. J., 2013, MNRAS, 431, 2513
Yang H.-Y. K., Reynolds C. S., 2016a, ApJ, 818, 181
Yang H.-Y. K., Reynolds C. S., 2016b, ApJ, 829, 90
Yuan F., Narayan R., 2014, ARA&A, 52, 529
Yuan F., Gan Z., Narayan R., Sadowski A., Bu D., Bai X.-N., 2015, ApJ, 804, 101
Zanni C., Murante G., Bodo G., Massaglia S., Rossi P., Ferrari A., 2005, A&A, 429, 399

APPENDIX A: CHANGES IN THE WIND MODEL

In addition to the changes in the AGN model, we have implemented some alterations to the stellar wind feedback compared to the model described in Vogelsberger et al. (2013). They are introduced to address some of the shortcomings of the Illustris simulation in low-mass systems, such an excessive number of galaxies with blue star-forming rings, a too high stellar mass fraction in systems with halo masses $M_{200,\,c} < 10^{11.5}\,\mathrm{M}_{\odot}$ and a too mild decline in SFRD at low redshift (Vogelsberger et al. 2014b). We summarize these changes here for completeness and refer to Pillepich et al. (in preparation) for a detailed discussion.

First, we now use an isotropic wind injection with 10 per cent of the energy injected thermally and not the bipolar, purely kinetic approach employed in the Illustris project (Vogelsberger et al. 2014c). Furthermore, we slightly changed the wind velocity. We still use the scaling with local dark matter velocity dispersion as in equation 14 of Vogelsberger et al. (2013), but introduce an additional redshift dependent factor $[H_0/H(z)]^{1/3}$, which effectively yields a scaling of the wind velocity that depends purely on halo mass. Additionally, we set a minimum wind velocity of $v_{\mathrm{min}} = 350\,\mathrm{km\,s^{-1}}$ to prevent unrealistically high-mass loading factors in low-mass haloes. Taken together, the equation for the wind velocity is hence

$$v_w = \max\left[\kappa\,\sigma_{\mathrm{DM}}^{\mathrm{1D}}\,(H_0/H(z))^{1/3}\,,\,v_{\mathrm{min}}\right]. \tag{A1}$$

We choose the parameters such that the wind velocity of a given halo equals that of the previous Illustris model at a redshift $z \simeq 5$, implying that it then tends to increase slightly towards lower redshifts compared to Vogelsberger et al. (2014b). This is the main reason for the different scaling of the SFRD with redshift (Fig. A1). The minimum wind velocity is partially responsible for the sharp decline in star formation efficiency towards lower masses in Fig. 9. A summary of the adopted wind parameters and a comparison to those used in Illustris is given in Table A1.

Moreover, we use a higher baseline wind energy for gas of primordial abundance but now reduce the available energy with metallicity Z on the grounds that higher metallicity galaxies plausibly have larger radiative cooling losses of the supernova energy. A similar factor has also been used in the Eagle project (Schaye et al. 2015).

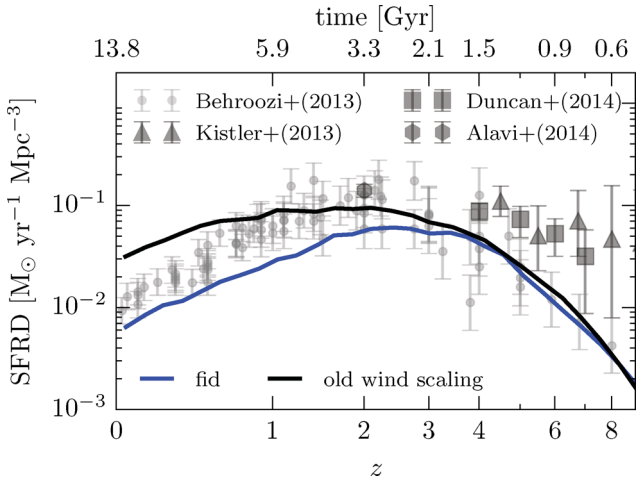

Figure A1. SFRD versus redshift for the fiducial run and a run with the old wind scaling.

Table A1. Comparison of wind parameters with Vogelsberger et al. (2013, 2014a) (V13).

Parameter	New	V13	Note
κ	7.4	3.7	Same velocity at $z \approx 5$
egy_w /egy_{0w}	3.6	1.09	Reduced through metallicity Dependence

The energy of the winds is reduced by a factor

$$f + (1-f)/\left[1 + (Z/Z_{\mathrm{red}})^{\gamma}\right], \quad \text{(A2)}$$

where $f = 0.25$, $\gamma = 2$ are free parameters and $Z_{\mathrm{red}} = 0.002$. This effectively lowers the efficiency of the supernova feedback in metal-enriched galaxies, and since most of the stars form there, the total injected wind energy is comparable to Illustris. However, the metal dependence leads to a higher relative efficiency of the wind feedback in low-mass systems, suppressing them more in comparison to Milky-Way-sized galaxies, which is an effect that seems required by the observational data and the low abundance of luminous dwarf galaxies.

All in all, our stellar feedback is somewhat stronger than in the Illustris simulation (Vogelsberger et al. 2014b), particularly at late times and in low-mass haloes. At high redshift, the strength of the stellar feedback is comparable, which also means that the $z = 0$ black hole masses are not significantly affected by the changes.

APPENDIX B: NUMERICAL CONVERGENCE

Achieving numerical convergence is a major challenge for full physics cosmological volume simulations due to the multiscale, multiphysics nature of the problem. Normally, convergence cannot be fully established, and hence the lack thereof represents an additional source of systematic uncertainty in the predictions. We attempt to quantify the magnitude of resolution effects here, using simulations with a box side length of $7.5\,h^{-1}$ Mpc with 2×384^3, 2×192^3 and 2×96^3 simulation particles and cells (dark matter + gas). In this small simulation volume, only low-mass haloes form, which makes an analysis of the global SFRD and BHARD meaningless, but it still allows us to estimate the uncertainties due to numerical convergence for the galaxies that happen to be present, especially because we can simulate them at resolutions higher than our standard high-resolution simulation with box side length of $30\,h^{-1}$Mpc.

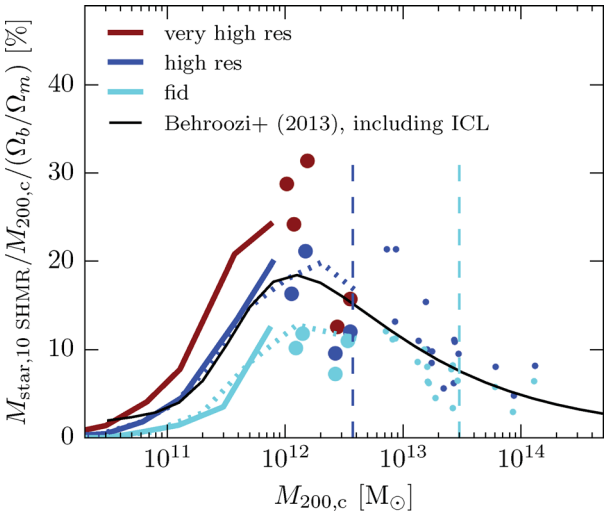

Figure B1. Stellar mass fraction as a function of halo mass for simulations of different resolution. The dotted lines and small dots indicate the simulations with $30\,h^{-1}$ Mpc side length, the solid lines and large dots test simulations with $7.5\,h^{-1}$ Mpc side length. Note that the colour coding is different compared with results presented in the main text. The dashed vertical lines are at $5 \times 10^5\,(m_{\mathrm{gas}} + m_{\mathrm{dm}})$, i.e. these haloes would have 5×10^5 simulation particles within $R_{200,\mathrm{c}}$ if their baryon content was equal to the cosmic baryon fraction.

We focus on the bias due to resolution effects at the low-mass end, as this regime has been the most severely affected when increasing the resolution in the Illustris simulation (Pillepich et al. 2014, fig. A1, upper panel). This is not surprising as the AREPO code ensures that the individual gas cells have approximately equal mass and consequently the number of gas cells within a halo decreases rapidly with decreasing halo mass. Fig. B1 shows the star formation efficiency as a function of halo mass for different resolutions. The vertical dashed lines correspond to $5 \times 10^5 (m_{\mathrm{gas}} + m_{\mathrm{dm}})$, which, depending on the gas fraction in the halo, translates to a few times 10^5 gas cells within $R_{200,c}$. Decreasing the number of gas cells in a halo, individual cells become so large that they average over significant regions of the ISM, producing lower average densities and hence longer gas consumption time-scales. This results in a numerically suppressed star formation rate and, over time, in a lower stellar mass fraction. As this convergence issue is present for the haloes at the peak of the star formation efficiency, it also manifests itself in the global SFRD.

A second resolution problem, related to black holes, is the radius h over which the gas properties are averaged to derive an accretion rate estimate and to inject the feedback energy. We adjust this radius such that it contains approximately a constant number of cells n_{ngb}. In Section 5, we presented the effect of increasing n_{ngb} by a factor of 4. Changing the particle number per dimension by a factor of 2, h also decreases, assuming constant n_{ngb}. This means that the volume over which the gas properties are averaged is a smaller volume at the centre of the galaxy and therefore usually denser, which leads to higher accretion rate densities, as seen in Fig. 7. If we increased n_{ngb} to keep approximately the same volume to average over, this effect would be smaller, but we would not benefit from the increased spatial resolution in the centre. For our simulation sequence, we aimed for a compromise by increasing n_{ngb} by a factor of 2 whenever the particle number per dimension is increased by a factor of 2.

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