Core Formation in High-z Massive Haloes: Heating by Post Compaction Satellites and Response to AGN Outflows

Avishai Dekel[1,2]⋆, Jonathan Freundlich[1,3], Fangzhou Jiang[1,4,5], Sharon Lapiner[1], Andreas Burkert[6], Daniel Ceverino[7], Xiaolong Du[5], Reinhard Genzel[8], Joel Primack[9]

[1] *Racah Institute of Physics, The Hebrew University, Jerusalem 91904 Israel*
[2] *SCIPP, University of California, Santa Cruz, CA 95064, USA*
[3] *Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France*
[4] *TAPIR, California Institute of Technology, Pasadena, CA 91125, USA*
[5] *Carnegie Observatories, 813 Santa Barbara Street, Pasadena, CA 91101, USA*
[6] *Ludwig-Maximilians Universitat Munchen, Department fur Physik, Scheinerstr. 1, D-81679 Munchen, Germany*
[7] *Departamento de Fisica Teorica, Facultad de Ciencias, Universidad Astronomia de Madrid, Cantoblanco, 28049 Madrid, Spain*
[8] *Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse 1, 85738 Garching, Germany*
[9] *Physics Department, University of California, Santa Cruz, Santa Cruz, CA 95064, USA*

4 June 2021

ABSTRACT
Observed rotation curves in star-forming galaxies indicate a puzzling dearth of dark matter in extended flat cores within haloes of mass $\geqslant 10^{12} M_\odot$ at $z \sim 2$. This is not reproduced by current cosmological simulations, and supernova-driven outflows are not effective in such massive haloes. We address a hybrid scenario where post-compaction merging satellites heat up the dark-matter cusps by dynamical friction, allowing AGN-driven outflows to generate cores. Using analytic and semi-analytic models (SatGen), we estimate the dynamical-friction heating as a function of satellite compactness for a cosmological sequence of mergers. Cosmological simulations (VELA) demonstrate that satellites of initial virial masses $> 10^{11.3} M_\odot$, that undergo wet compactions, become sufficiently compact for significant heating. Constituting a major fraction of the accretion onto haloes $\geqslant 10^{12} M_\odot$, these satellites heat-up the cusps in half a virial time at $z \sim 2$. Using a model for outflow-driven core formation (CuspCore), we demonstrate that the heated dark-matter cusps develop extended cores in response to removal of half the gas mass, while the more compact stellar systems remain intact. The mergers keep the dark matter hot, while the gas supply, fresh and recycled, is sufficient for the AGN outflows. AGN indeed become effective in haloes $\geqslant 10^{12} M_\odot$, where the black-hole growth is no longer suppressed by supernovae and its compaction-driven rapid growth is maintained by a hot CGM. For simulations to reproduce the dynamical-friction effects, they should resolve the compaction of the massive satellites and avoid artificial tidal disruption. AGN feedback could be boosted by clumpy black-hole accretion and clumpy response to AGN.

Key words: black holes — dark matter — galaxies: discs — galaxies: formation — galaxies: haloes — galaxies: mergers

1 INTRODUCTION

As seen in cosmological gravitating N-body simulations with no baryons, dark-matter (DM) haloes robustly produce cuspy inner density radial profiles with a central negative log slope $\alpha \sim 1$ (Navarro, Frenk & White 1997, NFW). In contrast, dwarf galaxies are observed kinematically to have inner flat DM cores, $\alpha \sim 0$ (e.g. Flores & Primack 1994; Burkert 1995; de Blok et al. 2001, 2008; Oh et al. 2011b,a, 2015). This is commonly understood both analytically and in simulations in terms of bursty supernova (SN) feedback (e.g. Pontzen & Governato 2012; Dutton et al. 2016; Freundlich et al. 2020a), which is capable of effectively

⋆ E-mail: dekel@huji.ac.il

© 2002 RAS

ejecting the gas from the central regions in haloes below a critical halo virial velocity of $V_v \sim 100 \, \mathrm{km \, s^{-1}}$, where the supernovae energy deposited in the ISM is comparable to the binding energy of the gas in the central halo potential well (Dekel & Silk 1986). The inner DM halo responds to the shallowing of the gravitational potential well due to the central mass loss by expanding and producing a flat core.

Very surprisingly, pioneering kinematic *observations* by Genzel et al. (2020), of 41 massive star-forming disc galaxies at $z = 0.65 - 2.5$, indicate the occurance of low central DM fractions (f_{dm}) and DM cores that extend beyond the stellar effective radii to $\sim 10 \, \mathrm{kpc}$. The data include near-infrared observations with SINFONI and KMOS at the ESO-VLT, as well as LBT-LUCI, and sub-millimeter observations from IRAM-NOEMA. The kinematics is determined from ionized gas traced by $H\alpha$ and molecular gas traced by CO. The baryons in the selected galaxies are rotationally supported ($V_{\mathrm{rot}}/\sigma > 2.3$), lying in the Main Sequence of star-forming galaxies (SFGs), with stellar masses $M_s = 10^{9.8-11.4} M_\odot$. There is a correlation between low f_{dm} and a high bulge-to-disc ratio. By properly inverting the relations based on abundance matching of observed galaxies and simulated ΛCDM DM haloes (Behroozi, Wechsler & Conroy 2013; Moster, Naab & White 2018), the corresponding halo virial mass range is $M_v = 10^{11.6-12.9} M_\odot$. In the high part of the redshift range, $z \sim 1.2 - 2.5$, more than two thirds of the galaxies show low DM fractions within the effective radius ($f_{\mathrm{dm}} < 0.3$), of which about one half are indicated to have extended cores. The cores tend to appear in this sample for $M_v \geqslant 10^{12} M_\odot$, with the median at $M_v = 10^{12.4} M_\odot$, while less massive galaxies tend to have higher DM fractions. At the lower redshift band, $z \sim 0.65 - 1.2$, the portion of low DM fraction galaxies is reduced to about a third, but their vast majority are with extended cores. At these redshifts, the DM fractions are typically not as low as at higher redshifts, and the mass dependence of this phenomenon is clearer, with a few galaxies of high DM fractions below $10^{12} M_\odot$. This adds to earlier indications for low central DM fractions in massive star-forming disks (e.g. Wuyts et al. 2016). The analysis is being extended to 100 galaxies (Price et al. 2021, in preparation), strengthening the appearance of DM-deficient extended cores especially in massive SFGs at $z \sim 2$. The appearance of low DM fractions in extended cores is very surprising because in such massive haloes, and at such redshifts, the gravitational potential is likely to be too deep for SN feedback to effectively remove central gas for gravitationally pushing the DM out and generating cores (Dekel & Silk 1986).

Beyond the unavoidable observational uncertainties, the non-trivial dynamical analysis of the observed rotation curves, which reveals the apparent low DM fractions and extended cores, is based on various assumptions and is thus associated with further uncertainties. These include, for example, the assumption of an exponential disk and an uncertain correction term to the centripetal force by a pressure gradient ("asymmetric drift" Burkert et al. 2010). Furthermore, a cuspy NFW profile is assumed in the fit of a DM halo to the

data, and when the corresponding central DM fraction conflicts with the observations, the conclusion is that the assumed NFW cusp should be replaced by a core. These uncertainties are expected to be reduced in a subsequent analysis, e.g., by using a more flexible model profile that allows a core. However, the uncertainties that have been carefully considered so far do not seem to systematically weaken the general need for a dearth of central DM and extended cores. Thus, despite the uncertainties, the intriguing implications of the pioneering results call for a special theoretical effort to explore potential processes that may push the DM out and generate cores in massive galaxies, while the stellar systems remain intact.

One should mention that Sharma, Salucci & van de Ven (2021) report from the KROSS survey somewhat higher fdm values, but in fact their results are largely consistent with Genzel et al. (2020). They obtain higher DM fractions because because of multiple reasons: (a) they refer to larger radii of $(1.9 - 3)R_e$ as opposed to R_e, (b) their KROSS sample is at $z \sim 1$, while the low f_{dm} values are observed by Genzel et al. (2020) mostly at $z \sim 2$, with a typical reduction of $0.2 - 0.3$ between these redshifts, and (c) their galaxies are typically of lower masses than the galaxies of Genzel et al. (2020), while it is shown in Price et al. (2021, in preparation) that f_{dm} is strongly decreasing with mass.

Several different mechanisms could in principle work toward generating cores in massive galaxies. These include, for example, mergers of SN-driven cored building blocks (Dekel, Devor & Hetzroni 2003), dynamical friction (DF) by the DM on merging satellites (El-Zant, Shlosman & Hoffman 2001) and AGN-driven outflows (e.g. Martizzi et al. 2012; Peirani et al. 2017), which may work in massive galaxies in analogy to supernova-driven outflows that generate cores in low mass galaxies. The puzzling issue is that none of the currently available cosmological simulations seem to give rise to such low central DM fractions and extended cores in massive haloes, despite the fact that all the above mentioned physical elements, including mergers, dynamical friction, supernovae and AGNs, are supposed to be incorporated in the simulations. Examples showing cusps in massive galaxies are in the TNG simulation (Wang et al. 2020) and in the FIRE-2 simulations (Lazar et al. 2020).

The effect of *AGN feedback* on the DM profile in the central regions of clusters of galaxies has been studied using cosmological simulations (e.g. Martizzi et al. 2012) and isolated-halo simulations (Martizzi, Teyssier & Moore 2013). It has been found that the repeating episodes of gas ejection and recycling can indeed generate a core of radius $\sim 10 \, \mathrm{kpc}$, but with the simulated halo masses of $10^{14} M_\odot$ and $1.4 \times 10^{13} M_\odot$ in these studies, the virial radii are $\sim 1.2 \, \mathrm{Mpc}$ and $\sim 0.6 \, \mathrm{Mpc}$ respectively, so the core radii extend only to $1 - 2\%$ of the virial radii as opposed to the desired $\sim 10\%$. No cores, or cores of limited extent, have also been found in massive galaxies in the Horizon-AGN simulation (Peirani et al. 2017, 2019) and in the NIHAO-AGN simulations (Macciò et al. 2020), both including AGN feedback. A

preliminary inspection of the haloes of massive galaxies in the TNG50 cosmological simulation, where a relatively strong AGN feedback is implemented (Weinberger et al. 2018), indicates cores that extend to only $1-2\%$ of the virial radius (Sandro Tacchella, private communication). It seems that AGN feedback in cosmological simulations, as implemented so far, fails to reproduce cores of $\sim 0.1 R_v$ extent as deduced from observations by Genzel et al. (2020). Potential ways to remedy this could be to have the DM halo pre-heated and thus be closer to escape before the mass ejection, and/or to boost the strength of the AGN feedback itself and tighten its coupling with the gas in the inner halo.

The effect of *dynamical friction* on the cusps of massive haloes has been discussed by El-Zant, Shlosman & Hoffman (2001). They used semi-analytic simulations with hundreds of satellites that sum up to 10% of the host mass of $10^{12} M_\odot$ during $2-3$ Gyr, which is roughly a halo virial time at $z=0$. This led to almost-flat cores that extend to $(4-6)$ kpc, which at $z=0$, with $R_v \sim 300$ kpc, corresponds to only $(0.01-0.02) R_v$. In El-Zant et al. (2004), they used N-body simulations with multiple satellites that amount to $3-20\%$ of the host-halo mass. Semi-cores of log slope -0.35 that extend out to $R_c \sim 0.06 R_v$ were generated in ~ 6 Gyr. However, in both cases, as well as in El-Zant (2008), where a universal profile is maintained under satellite sinkage, the models assumed point-mass satellites and ignored their tidal stripping, thus overestimating the effect of DF on the host cusp. While this has been a stimulating proof of concept, the situation studied is not a representative cosmological population of satellites that could reproduce the realistic effect of DF on the host cusps. The impression is that DF by itself is not likely to generate flat cores as extended as observed.

Clearly, efficient satellite penetration that may result in DM heating requires that the incoming *satellites* would be massive, compact, and on relatively radial orbits, such that tidal stripping would be minimal prior to reaching the host central regions where the DF is to increase the energy and angular momentum of the DM cusp. The questions to be verified are whether the realistic satellites that merge with the host haloes of the relevant masses at the relevant redshifts are sufficiently massive, compact and on proper orbits for effective DF heating or flattening of the host DM cusps. It is also interesting to find out how the deposited energy is divided between heating the cusp and partially flattening the profile into a core.

We study here a promising *two-staged scenario* where DF heating or flattening and AGN feedback are both boosted, and they are put together in a hybrid scenario where the DF that acts on compact merging satellites heats up or flattens the inner DM halo, making it more susceptible to expansion under clumpy AGN-driven outflows.

We note that standard NFW cusps are kinematically cold, as Jeans equilibrium implies that the (isotropic) velocity dispersion profile is decreasing toward the center like $\sigma^2 \propto r$, while the gravitational potential gets deeper. This means that the dark matter is more tightly bound in the inner cusp, which implies that it would be hard to unbind it by central mass ejection. This is why kinematically heating the DM particles by DF prior to the mass ejection, bringing them closer to the escape velocity from the cusp, could be a key to DM expansion and core formation.

We demonstrate below that in such a scenario the observed preferred host-halo masses and preferred redshifts for DM cores are natural outcomes. We note that for host haloes of $M_v \geqslant 10^{12} M_\odot$ at $z \sim 1-3$, the likely massive merging satellites are with haloes of $m_v \geqslant 10^{11.3} M_\odot$, which are above the mass threshold for drastic compaction events into blue nuggets (Zolotov et al. 2015; Tomassetti et al. 2016; Tacchella et al. 2016a,b). This *golden mass* (Dekel, Lapiner & Dubois 2019) is determined by the combined effect of supernova feedback at lower masses (Dekel & Silk 1986) and virial shock heating of the circum-galactic medium (CGM) aided by AGN feedback at higher masses (Birnboim & Dekel 2003; Dekel & Birnboim 2006). The satellites above the golden mass are therefore likely to be more *compact* than lower-mass satellites that typically feed lower-mass hosts. These compact satellites could be optimal for deep penetration into the host cusp with less tidal stripping, allowing effective DF heating or flattening in the host cusp. The highly compact baryonic cusp of the satellite should allow the buildup of a central stellar system that would remain intact under the AGN-driven gas outflows that push the less-concentrated DM outward. AGN feedback is indeed expected to be most effective above the golden mass of $M_v \sim 10^{12} M_\odot$, where the black-hole (BH) growth is no longer suppressed by supernova feedback (Dubois et al. 2015), the BH growth is made possible by the hot-CGM (Bower et al. 2017; Anglés-Alcázar et al. 2017b), and the rapid BH growth is triggered by a wet-compaction event (Dekel, Lapiner & Dubois 2019; Lapiner, Dekel & Dubois 2020). This is confirmed by observations of luminous AGN fractions above the golden mass (Förster Schreiber et al. 2019) The preferred appearance of massive bulges in low f_{dm} galaxies is indeed consistent with the association of cores with the bulge-forming compaction events that tend to occur above the golden mass.

The preferred high redshift for DM cores could naturally emerge from the higher merger rate, higher gas fraction, and therefore more effective compaction events at higher redshifts. Furthermore, at higher redshifts, the dark-matter cusps are maintained "hot" under the dynamical friction by a cosmological sequence of satellites, and are thus more susceptible to core formation by AGN outflows, while at lower redshifts the cusps cool and expand and thus make the AGN outflows less effective.

A companion paper (Burkert et al. 2021) addresses the same problem of DM core formation emphasizing somewhat different angles and a similar hybrid scenario of DF heating followed by response to outflows. This version of the scenario appeals instead to DF heating by inward migrating clumps in violently unstable discs, and it bases the feasibility tests of the two parts of the scenario on simple N-body simulations tailored for the purpose.

The paper is organized as follows. In §2 we derive from the VELA simulations the fiducial mass profiles of pre-compaction and post-compaction satellites below and above the golden mass. In §3 we use a toy model to make analytic estimates of the DF heating as a function of satellite compactness. In §4 we utilize SatGen semi-analytic simulations to explore the DF heating as a function of satellite compactness and orbit, as well as for a cosmological sequence of satellites. In §5 we address the post-heating relaxation of the cusp, cooling and expansion, and use simplified N-body simulations to study the interplay between DF heating and density flattening, and evaluate the cooling timescale. In §6 we use the CuspCore analytic model to study the response of the pre-heated haloes to gas ejection by AGN feedback. In §7 we discuss our results and refer to the prospects of reproducing them in cosmological simulations. In §8, we summarize our conclusions. In appendices A–J (available as supplementary material online), we bring supportive descriptions of the analytic models and simulations used, and add figures relevant to cases that are complementary to our fiducial cases.

2   COMPACT SATELLITES ABOVE A THRESHOLD MASS IN COSMOLOGICAL SIMULATIONS

A key for the strength of the cusp heating by dynamical friction is the initial compactness of the merging satellite, which determines the mass with which it penetrates into the host-halo cusp. A mass threshold for compact satellites would be translated to a host-halo mass threshold for the formation of extended cores in the hosts. Indeed, cosmological simulations reveal that most galaxies undergo a major event of wet compaction to a "blue nugget" when they are near a golden halo mass of $m_{\rm v} \simeq 10^{11.3} M_\odot$ (Zolotov et al. 2015). Following Tacchella et al. (2016a), we use galaxies from the VELA zoom-in cosmological simulations (briefly described in Appendix §B, available as supplementary material online), in order to derive the mass profiles of haloes pre compaction versus post compaction, namely below and above the golden mass, respectively (Zolotov et al. 2015; Tomassetti et al. 2016; Huertas-Company et al. 2018). These will serve us in evaluating the compactness of the merging satellites as a function of their mass.

Figure 1 shows the evolution of mass profiles for four representative simulated VELA galaxies, at a sequence of output times before and after the major compaction events, which occur in different galaxies at different redshifts but typically when the halo mass is near $m_{\rm v} \sim 10^{11.3} M_\odot$. The figure shows the profiles of the total mass, consisting of dark matter, stars and gas, in comparison with the mass profiles of the DM alone. In both cases, we see rather universal profiles both before and after the compaction events, showing that the overall compactness is growing significantly during the wet-compaction process, both for the DM and for the total mass.

In order to quantify the universal mass profiles, relevant here in particular to the merging satellites, we use the Dekel-Zhao profile (Dekel et al. 2017; Freundlich et al. 2020b, DZ)[1] (summarized in Appendix §A, available as supplementary material online). This is a two-parameter functional form, with a flexible inner slope, and with analytic expressions for the profiles of density, mass and velocity as well as potential and kinetic energies (and lensing properties). It has been shown to fit DM haloes in the NIHAO cosmological simulations with baryons better than the other commonly used two-parameter profiles (such as the generalized-NFW and the Einasto profiles). This is being confirmed in the Auriga, Apostle and EAGLE simulations (Marius Cautin, private communication). In the DZ profile, the mean density within a sphere of radius r, $\bar{\rho}(r)$ (denoted later $\bar{\rho}_{\rm sat}(\ell)$ for the satellites), with a virial mass $M_{\rm v}$ ($m_{\rm v}$) inside a virial radius $R_{\rm v}$ ($\ell_{\rm v}$), is characterized by two shape parameters, an inner slope α and a concentration c,

$$\bar{\rho}(r) = \frac{\bar{\rho}_{\rm c}}{x^\alpha \, (1 + x^{1/2})^{2(3-\alpha)}} \,, \quad x = \frac{r}{R_{\rm v}} \, c \,, \quad (1)$$

in which the constant is related to the parameters (c, α) by

$$\bar{\rho}_{\rm c} = c^3 \mu(c, \alpha) \, \bar{\rho}_{\rm v} \,, \quad \bar{\rho}_{\rm v} = \frac{M_{\rm v}}{(4\pi/3) R_{\rm v}^3} \,, \quad (2)$$

$$\mu(c, \alpha) = c^{\alpha-3} \, (1 + c^{1/2})^{2(3-\alpha)} \,. \quad (3)$$

The mean density contrast of the haloes with respect to the cosmological background in the EdS cosmological regime (approximately valid at $z > 1$) is $\bar{\rho}_{\rm v}/\rho_{\rm u} \sim 200$, and by definition $\bar{\rho}_{\rm v}$ is the same for the host and for the satellites when they were still isolated. To be used below, the associated mass encompassed within a sphere of radius r is

$$\frac{M(r)}{M_{\rm v}} = \frac{1}{c^3 \bar{\rho}_{\rm v}} x^3 \bar{\rho}(x) = \frac{\mu}{\bar{\rho}_{\rm c}} \, x^3 \, \bar{\rho}(x) \,, \quad (4)$$

and the log slope of the mass profile is

$$\nu(r) = \frac{d \log M}{d \log r} = \frac{3 - \alpha}{1 + x^{1/2}} \,. \quad (5)$$

We fit a Dekel-Zhao profile with the free parameters (c, α) to each of the VELA simulated mass profiles, separately for the DM and for the total mass. The best-fit is obtained by minimizing residuals in equally spaced log bins in the range $(0.01-1)R_{\rm v}$ and $(0.03-1)R_{\rm v}$ for the DM and total mass respectively, assuming that the most relevant mass profile is near $(0.03-0.1)R_{\rm v}$. Figure 2 shows the median and standard deviation for the two profile parameters over all the 34 simulated galaxies as a function of time, via the expansion factor $a = (1+z)^{-1}$, with respect to the time of the blue-nugget peak (BN) at the end of the wet compaction process. Based on this figure, we adopt for the fiducial pre-compaction DM profile the DZ parameters $(c, \alpha) = (3, 0.5)$, and for the fiducial post-compaction profile $(c, \alpha) = (7, 1)$. Other

---

[1] Available for implementation in https://github.com/JonathanFreundlich/Dekel_profile .
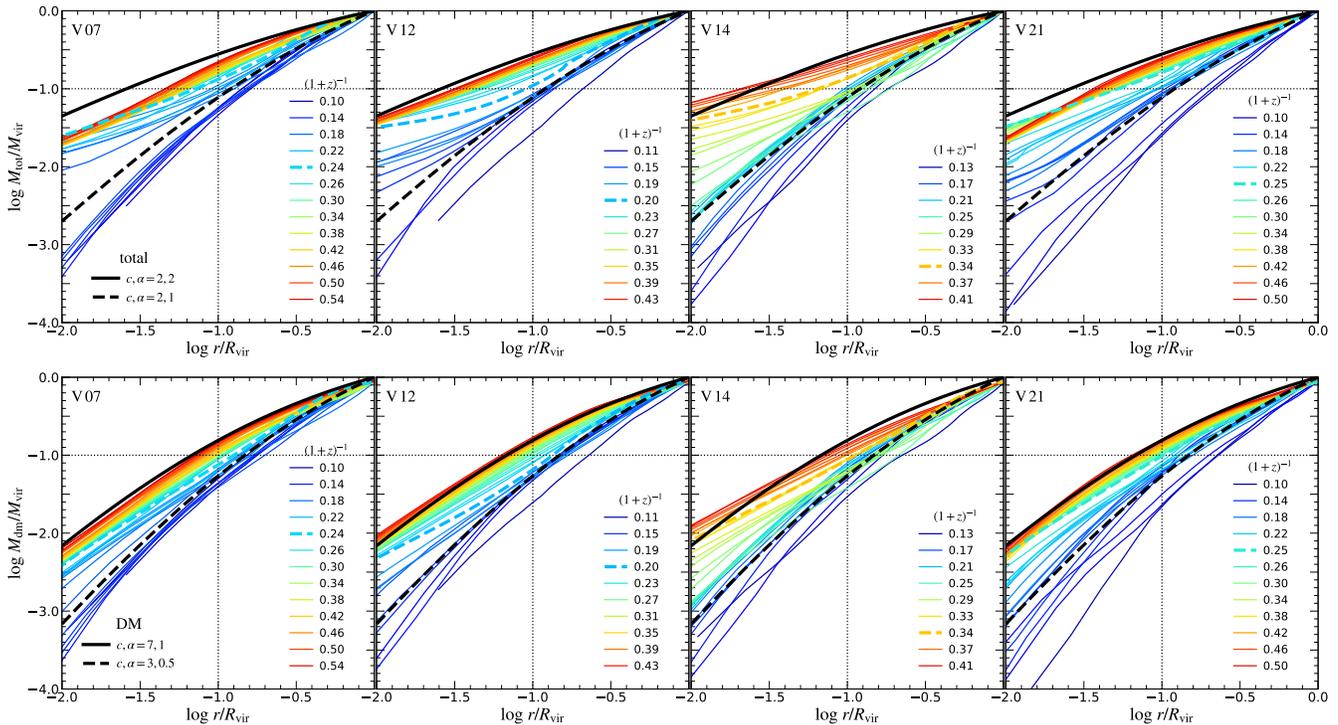
Figure 1. Total mass profiles for representative VELA galaxies at a sequence of times, before and after the major compaction events. **Top:** for the total mass of dark matter, stars and gas. **Bottom:** for the dark matter alone. Marked are the corresponding expansion factors $a = (1+z)^{-1}$, separated by $\Delta a = 0.03$ (which roughly corresponds to $\Delta t \sim 400\,\mathrm{Myr}$ at $a = 0.25$), with the blue-nugget highlighted by a thick dashed curve. We see universal shapes of the profiles pre-compaction and post-compaction, with a significant increase in global compactness during the compaction process, both for the total mass and for the DM alone. The black curves mark our fiducial DZ fits for the pre-compaction (dashed) and post-compaction (solid) profiles, as deduced from Fig. 2. We conservatively adopt the DM profiles for our merging satellites.
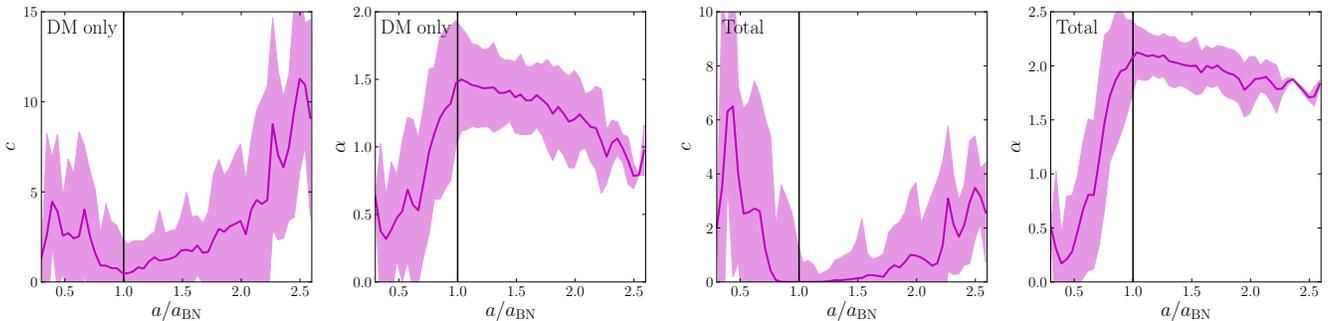

Figure 2. Dekel-Zhao profile fits to the DM mass profiles in all 34 VELA simulated galaxies. Shown are the median and standard deviation of the parameters c and α as a function of time (expansion factor $a = (1+z)^{-1}$) with respect to the time of the blue nugget at the end of the major compaction. This quantity is a proxy for mass, below and above a halo mass of $m_{\mathrm{v}} = 10^{11.3} M_\odot$. We see different typical profiles well before and after the compaction event, namely below and above the critical mass. Left: for the DM mass profiles, where we crudely adopt $(c, \alpha) = (3, 0.5)$ and $(7, 1)$ as our fiducial profiles pre-compaction and post-compaction, respectively. Right: for the total mass profiles, where we crudely adopt $(c, \alpha) = (5, 0.5)$ and $(3, 1.8)$, respectively. The corresponding profiles are shown in Fig. 1.

choices, relevant at different times before or after the compaction, make only little differences to the effective diffuseness and compactness of the mass distribution in these two distinct phases of evolution. For the total mass we adopt $(c, \alpha) = (5, 0.5)$ and $(3, 1.8)$ for the diffuse and compact satellites, respectively.

While the DZ best-fit parameters are quite different for the DM and for the total mass profiles, the actual mass profiles turn out not to be that different. The diffuse haloes roughly match an NFW profile with $c_{\mathrm{NFW}} \simeq 5$, with the total mass about 50% higher than the DM mass interior to $0.067 R_{\mathrm{v}}$. The compact haloes, interior to the same radius, have more mass than the diffuse haloes by a factor of a few, with the total mass twice as large as the DM mass. We find below that the effect of using the DM mass in the merging satellites

is not very different from the effect of using the total mass. In order for our analysis to be on the conservative side in terms of the DF heating, we adopt as our fiducial satellite profile the typical profile of the dark matter alone.

As described in eq. (A12) to eq. (A15) of Appendix §A (available as supplementary material online), based on equations 11-14 of Freundlich et al. (2020b), an alternative, sometimes more accessible pair of shape parameters is the concentration c_2, referring to the radius where the log local density slope is -2 (as in the NFW profile), and the negative log local density slope s_1 at some small radius r_1 (e.g. at 1% of the virial radius). While there is a valid DZ profile for any values of c (>0) and α (<3), a valid profile is not guaranteed for all arbitrary values of c_2 and s_1. For example, c_2 is not defined for an isothermal sphere or a steeper density profile.

The alternative DZ parameters corresponding to the universal profiles are $(c_2, s_1) = (3, 0.94)$ and $(15.6, 1.52)$ for the fiducial diffuse and compact satellite haloes respectively, using the DM mass alone. The inner slopes s_1 are respectively somewhat flatter and steeper than an NFW cusp (which has $s_1 = 1.06$ and 1.27 for $c_{\mathrm{NFW}} = 3$ and 15.6 respectively), with a smaller and larger concentration compared to the $c_{\mathrm{NFW}} \simeq 5$ that fits typical simulated haloes in the relevant mass range at $z \sim 2$. For the total mass, the alternative DZ parameters are $(c_2, s_1) = (4.5, 1.31)$ and $(>100, 2.2)$, the latter indicating that the slope is steeper than 2 in the whole relevant radius range.

The above implies that deeper satellite penetration, that are associated with more significant host-cusp heating by dynamical friction, are expected once the host halo is more massive than $M_{\mathrm{v}} \sim 10^{12} M_{\odot}$, such that a significant fraction of the accreting mass is in post-compaction nuggets above $m_{\mathrm{v}} \sim 10^{11.3} M_{\odot}$. On the other hand, much less penetration and heating is expected for hosts of $M_{\mathrm{v}} < 10^{11} M_{\odot}$, where most of the merging satellites tend to be pre-compaction. This will be studied next.

3   DYNAMICAL-FRICTION HEATING BY SATELLITES: A TOY MODEL

We start with a simple analytic toy model in order to obtain a crude estimate for the energy deposited in the host-halo cusp as a result of the dynamical friction exerted on penetrating satellites as a function of their compactness. These estimates will be improved in §4 using semi-analytic simulations.

3.1   Satellite penetration - tidal stripping

In order to evaluate the energy deposited in the host cusp by dynamical friction, we should first estimate the satellite mass as it penetrates the cusp after it had suffered tidal stripping along its orbit. We first evaluate the mass within the satellite stripping radius when the satellite is at the cusp half-mass radius, crudely assuming an onion-shell outside-in stripping outside the stripping radius. When doing so, we neglect changes in the density and kinematics inside the tidal radius (crudely referred to as "tidal heating") that may affect the stripping. These are incorporated in the treatment by SatGen in §4. We then apply a simple correction to the mass within the tidal-radius, to be calibrated by simulations, in order to take into account effects such as the deviations of the tidal force from spherical symmetry about the satellite and for the fact that the stripping is not instantaneous but rather takes a dynamical time.

3.1.1   The tidal radius

The tidal radius (King 1962), where self-gravity balances the tidal and centrifugal forces along the line connecting the centers of host and satellite, obeys the equation

$$\bar{\rho}_{\mathbf{sat}}(\ell_{\mathrm{t}}) = \bar{\rho}(r) \left[2 - \nu(r) + \epsilon^2(r)\right], \qquad (6)$$

where $\bar{\rho}_{\mathbf{sat}}(\ell)$ and $\bar{\rho}(r)$ are the mean density profiles of the satellite (before it entered the host virial radius R_{v}) and the host, respectively. Here $\nu(r)$ is the local log slope of the host mass profile $M(r)$, namely $\nu(r) = d\ln M/d\ln r$. It is $\nu(r) = 3 - \alpha(r)$, where $\alpha(r)$ is minus the local log slope of $\bar{\rho}(r)$. For example, $\nu = 3, 2$ and 1 for a flat core, an NFW cusp and an isothermal-sphere, for which $\alpha = 0, 1$ and 2, respectively.

The quantity $\epsilon(r)$, representing the centrifugal force that helps the stripping, is the local circularity of the satellite orbit at r, namely the ratio of tangential to circular velocity

$$\epsilon(r) = \frac{V_{\mathrm{tan}}(r)}{V_{\mathrm{circ}}(r)}, \qquad (7)$$

with $\epsilon = 0$ and 1 for local radial and circular orbits, respectively. This local circularity is typically not the same as its value at R_{v}, upon entry to the halo, which commonly serves as one of the two parameters characterizing the orbit. The value of the local ϵ may vary as the satellite is orbiting within the host halo, and we do not know a priori its effective range of values. We therefore use $\epsilon = 0.5$ as a fiducial value, recalling that it could range from zero to above unity.

3.1.2   The penetrating satellite mass

We use for the satellites the Dekel-Zhao profile, eq. (1). As shown in eq. (A11) of Appendix §A (available as supplementary material online), the fraction $f = m/m_{\mathrm{v}}$ of satellite mass within a sphere of radius ℓ about the satellite center with respect to the initial total satellite mass, obeys

$$(f/\mu) \left[(f/\mu)^{-[2(3-\alpha)]^{-1}} - 1\right]^6 = \bar{\rho}_{\mathbf{sat}}(\ell)/\bar{\rho}_{\mathbf{sat,c}}. \qquad (8)$$

Using the tidal condition, eq. (6), we obtain an equation for the mass fraction $f_{\mathrm{t}}(r)$ within the tidal radius,

$$\left(\frac{f_{\mathrm{t}}}{\mu}\right) \left[\left(\frac{f_{\mathrm{t}}}{\mu}\right)^{-[2(3-\alpha)]^{-1}} - 1\right]^6 = \frac{[2 - \nu(r) + \epsilon^2(r)]\,\bar{\rho}(r)}{\mu\,c^3\,\bar{\rho}_{\mathrm{v}}}. \qquad (9)$$

Recall that μ is a function of the satellite parameters c and α. For a given host mass profile $M(r)$, and its log slope $\nu(r)$, and for a given satellite orbit circularity $\epsilon(r)$, this equation can be solved numerically for $f_{\rm t}(c,\alpha)$. We note that the solution depends only on the profiles, and it does not depend explicitly on the total masses or the ratio $m_{\rm v}/M_{\rm v}$.

One can see that for a given α (< 3), in the limit $c \to \infty$, where $\mu \to 1$, the equation yields $f_{\rm t} \to 1$. Similarly, for a given c, in the limit $\alpha \to 3$ (from below), again $\mu \to 1$ and $f_{\rm t} \to 1$. Thus, the mass fraction within the tidal radius is increasing with either c or α, approaching no stripping as $c \to \infty$ or $\alpha \to 3$. Analytic solutions of eq. (9) can be obtained for certain values of α, e.g., there is a simple explicit solution for $\alpha = 0$, and solutions for $\alpha = 1$ and $\alpha = 2$ via solutions of cubic polynomial equations.

The actual bound satellite mass at r, for the purpose of dynamical friction, may deviate from the mass within the tidal radius. Among other effects, this is because the tidal stripping is not spheri-symmetric, because it is not instantaneous, occurring over a dynamical timescale at r, and because of the effects on the inner satellite. For example, Green, van den Bosch & Jiang (2021) assumed that the actual mass stripping rate is crudely modeled as $\dot{m} = A m(> \ell_{\rm t})/t_{\rm dyn}$, where $t_{\rm dyn}$ is approximated by the dynamical time of the halo at r, and found a best fit to N-body simulations with $A \simeq 0.55$.[2] We can therefore consider the mass within the tidal radius to be an underestimate of the actual satellite mass, possibly by a factor of ~ 2 or more at the relevant radii within the cusp if the modeling above is valid. We crudely model this correction in the cusp as

$$f = B f_{\rm t}. \tag{10}$$

The value of B can be calibrated by matching the results of the toy model to the SatGen simulations described in §4, where any evolution of the inner satellite is included. We adopt $B = 2$ as our fiducial value, but this choice has no qualitative effect on the results.

3.1.3 The satellite mass in a given host cusp

The host halo is also described by a Dekel-Zhao profile, eq. (1) and eq. (A1), with parameters $(c_{\rm h}, \alpha_{\rm h})$. The log slope of the mass profile, $\nu(r)$, to be used in eq. (9), is given for the DZ profile by eq. (5). We evaluate $\bar{\rho}(r)$ and $\nu(r)$ at the half-mass radius of the cusp of radius $R_{\rm c} = 10\,{\rm kpc}$, which is roughly $R_{\rm h} = 7\,{\rm kpc}$, and insert them in eq. (9) in order to solve for the satellite mass fraction there.

Our fiducial case is an initial host halo of $M_{\rm v} = 10^{12.5} M_\odot$ at $z = 2$ (with virial radius and velocity $R_{\rm v} = 150\,{\rm kpc}$ and $V_{\rm v} = 300\,{\rm km\,s^{-1}}$). Such a halo, if unperturbed by baryons, is well fit by an NFW profile, eq. (A20) (Navarro, Frenk & White 1997), with

a moderately steep cusp, and an NFW concentration $c_{\rm NFW} = 5$ (Bullock et al. 2001). This NFW profile is matched best across $r = (0.01 - 1) R_{\rm v}$ by a DZ profile of $(c_{\rm h}, s_{\rm 1h}) = (5.04, 0.908)$, or $(c_{\rm h}, \alpha_{\rm h}) = (7.13, 0.216)$.[3] In this case we neglect earlier baryonic effects on the DM halo, either steepening by adiabatic contraction or partial flattening by outflows.

We note that the assumed "cusp radius" of $R_{\rm c} = 10\,{\rm kpc}$ is rather crude, motivated by the desired "core radius" based on the observational estimates. Just for reference, in the fiducial NFW host, the local negative log slope of the density profile at this radius is $\simeq 1.56$.

As an alternative case with an initially steeper cusp we use $(c_{\rm h}, \alpha_{\rm h}) = (1.13, 1.29)$ or $(c_{\rm h}, s_{\rm 1h}) = (5.04, 0.91)$. This steep-cusp profile is typical to haloes of $M_{\rm v} \gtrsim 10^{12} M_\odot$ from the NIHAO cosmological simulations with baryons, as analyzed in Freundlich et al. (2020b). We will mention the results for the steep-cusp halo below, but the figures referring to this case are presented in Appendix §G (available as supplementary material online).

The left panel of Fig. 3 shows the bound mass fraction f as a function of the DZ parameters (c, α) for the satellite profile, as obtained by solving eq. (9) (with the correction of eq. (10) assuming $B = 2$). This is shown here for the fiducial NFW host and in Fig. G1 (available as supplementary material online) for the steep-cusp host.

As described in §2, we learn from the VELA simulations that the DM profiles of typical diffuse and compact haloes, below and above $m_{\rm v} \sim 10^{11.3} M_\odot$, are well fit by a DZ profile with $(c, \alpha) = (3, 0.5)$ and $(c, \alpha) = (7, 1)$, respectively. These parameters for the satellites are marked in the figure. Also marked are the parameters for the total mass profiles, $(c, \alpha) = (5, 0.5)$ and $(3, 1.8)$, for the diffuse and compact satellites respectively.

For the NFW host halo, with the fiducial DM satellite profiles, we obtain $f = m/m_{\rm v} \simeq 0.38$ and 0.05 for the compact and diffuse satellites, respectively. Thus, the fiducial massive, post-compaction satellite penetrates into the cusp with a significant fraction of its original mass, which promises significant cusp heating. On the other hand, the fiducial low-mass, pre-compaction satellite is more significantly stripped and it penetrates to the halo cusp with a small fraction of its mass, such that it is not expected to provide much cusp heating.

For the NFW halo but with the somewhat more compact total satellite profiles, we obtain $f = m/m_{\rm v} \simeq 0.74$ and 0.12. This is a deeper penetration, as expected from the larger satellite compactness due to the baryons, but the qualitative difference between the compact and diffuse satellites remains the same.

For the steep-cusp halo, we obtain $f \simeq 0.23$ and

---

[2] In these simulations, the satellite bound mass is defined using an iterative un-binding algorithm (van den Bosch & Ogiya 2018), which considers the energy in the satellite frame, with the gravitational potential energy involving satellite particles only.

[3] The very inner slope at $0.01 R_{\rm v}$, $s_{\rm 1h}$, is apparently slightly smaller than unity, but the fit is overall good across the cusp and in the most relevant region of $r \lesssim 0.1 R_{\rm v}$. Different DZ fits, that enforce a perfect fit at $0.01 R_{\rm v}$ and have an almost similar overall quality, are possible, but they do not significantly affect the results.

0.01 for the fiducial compact and diffuse satellites, respectively. Here, as expected from the steeper cusp, the satellites penetrate to the cusp with a somewhat smaller fraction of their mass. Nevertheless, the penetration of the compact satellite to the steep cusp is still with a noticeable fraction of its original mass, so it still has the potential to provide significant heating, to be estimated next.

3.2 Energy deposited in the host cusp

Given the penetrating mass fraction f, we wish to estimate the energy deposited in the host cusp by dynamical friction, with respect to the kinetic energy within this cusp. This will tell whether dynamical friction is capable of significantly heating the cusp, and possibly partly flatten the density profile, in preparation for core formation by AGN feedback.

3.2.1 A single satellite

We crudely approximate the work done by dynamical friction on a satellite within the cusp by

$$W_c = s\, R_c\, F\,, \qquad (11)$$

where F is the typical DF force exerted by the dark matter on the satellite in the cusp, R_c is the characteristic cusp radius, encompassing mass M_c, and $s \gtrsim 1$ is a factor that characterizes the effective length of the path of the satellite inside the cusp. The DF force is approximated by the Chandrasekhar formula (Chandrasekhar 1943) (as long as the cusp has not flattened to a core),

$$F = 4\pi \ln\Lambda\, G^2\, \frac{\rho\, m^2}{V^2}\, H(u)\,, \qquad (12)$$

where ρ is the typical DM density within the cusp, m is the bound satellite mass when in the cusp, V is the satellite orbital velocity, and $\ln\Lambda$ is the Coulomb logarithm. The latter is commonly assumed to be approximately $\ln(M_v/m) \sim 3$. The correction factor $H(u)$ that limits the effect to particles that move slower than the satellite, assuming a Maxwellian distribution of velocities, is

$$H(u) = \mathrm{erf}(u) - \frac{2u}{\sqrt{\pi}}\, e^{-u^2}\,, \quad u = \frac{V}{\sqrt{2}\sigma}\,, \qquad (13)$$

where σ is the velocity dispersion in the cusp. For example, $H(u=0.55)=0.1$ and $H(u=0.7)=0.2$.

We note that the use of the Chandrasekhar formula involves certain inaccuracies, beyond the uncertainties in the values adopted for the parameters, which could either strengthen or weaken the force. One is an uncertainly in the satellite mass that is relevant for the purpose. For example, when the mass is measured in a given way, Green, van den Bosch & Jiang (2021) indicate that a multiplicative factor $\beta \sim 0.75$ in eq. (12) yields best fit to results deduced from the Bolshoi cosmological N-body simulation. This should be better calibrated using carefully designed simulations where the dynamical friction is studied in more detail. On the other hand, self-friction within the satellite is ignored, and so is the binding effect of the baryons within the satellite. Dynamical

friction is known to be significantly reduced when the satellite is orbiting in a very flat host core, but here we limit ourselves to affecting a steep cusp in its development toward a somewhat flatter cusp. In summary, our toy-model treatment of dynamical friction should be considered as a crude approximation only, good for an order-of-magnitude estimate.

Adopting the mean density within R_c and the circular velocity at R_c, $\rho = (4\pi/3)^{-1} M_c/R_c^3$ and $V = V_c = (GM_c/R_c)^{1/2}$, we get

$$W_c = 1.8\, s_2\, (\ln\Lambda)_3\, H(u)_{0.1}\, \frac{Gm^2}{R_c}\,, \qquad (14)$$

where the subscripts denote the fiducial values assumed for the different quantities in this expression ($s_2 = s/2$ etc.).

The DF work estimated in eq. (14) is to be compared to the kinetic energy in the cusp, $K_c \simeq (1/2)\, M_c\, \sigma^2$. We assume for the host $M_v = 10^{12.5} M_\odot$, whose virial radius is $R_v = 150\,\mathrm{kpc}\, M_{12.5}^{1/3}\, (1+z)_3^{-1}$, in which $M_{12.5} = M_v/10^{12.5} M_\odot$ and $(1+z)_3 = (1+z)/3$. For an NFW profile with $c_{\mathrm{NFW}} = 5$ at $z=2$, a cusp of $R_c = 10\,\mathrm{kpc}$ corresponds to $r = 0.067 R_v$ and $x_c = 0.33$, so from eq. (A20) the cusp mass is $M_c = 0.04 M_v$. This implies that $V_c = 0.77 V_v$. We read from Figure 1 of Freundlich et al. (2020b) that at $r = 0.067 R_v$ we have $\sigma \simeq V_v$ (and that indeed $V_c \simeq 0.78 V_v$). These imply for the NFW host that[4]

$$K_c \simeq 0.5\, M_c V_v^2 \simeq 0.83\, M_c V_c^2 \simeq 6 \times 10^{15} M_\odot\, \mathrm{kpc}^2\, \mathrm{Gyr}^{-1}\,. \qquad (15)$$

We also learn that $u \simeq 0.54$, such that $H(u) \simeq 0.1$ is a reasonable estimate. For the steep-core host, we find that K_c is 1.25 times larger.

Using eq. (14) and eq. (15), we obtain

$$\frac{W_c}{K_c} = 2.2\, s_2\, (\ln\Lambda)_3\, H(u)_{0.1}\, \frac{m^2}{M_c^2}\,. \qquad (16)$$

Thus, for a single satellite of initial mass m_v in the NFW host of mass M_v, we obtain

$$\frac{W_c}{K_c} = 0.55\, s_2\, f_{0.2}^2\, \left(\frac{m_v}{M_v}\right)_{0.1}^2\,. \qquad (17)$$

The value of s here absorbs the uncertainties in the three parameters s, $\ln\Lambda$ and $H(u)$. Here, f ($\equiv 0.2 f_{0.2}$) is to be obtained, e.g., by solving eq. (9) given the profile shapes of the halo and the satellite, with the possible correction of eq. (10). Note that f in this model is independent of m_v/Mv. We learn that a single satellite of $m_v \geqslant 0.13 M_v$ may be capable of heating the cusp, with $W_c \geqslant K_c$, once more than 0.2 of its mass survives tidal stripping before entering the cusp. For the steep-core host, the numerical factor in eq. (17) is 1.25 times lower because of the higher K_c.

The middle column of Fig. 3 shows the toy-model estimates of W_c/K_c, based on eq. (17), for a single satellite of $m_v = 0.1 M_v$ as a function of the satellite's DZ profile parameters (c, α), and for the NFW host, assuming

---

[4] Note that $1\,\mathrm{kpc\, Gyr}^{-1} \simeq 0.955\,\mathrm{km\, s}^{-1}$.
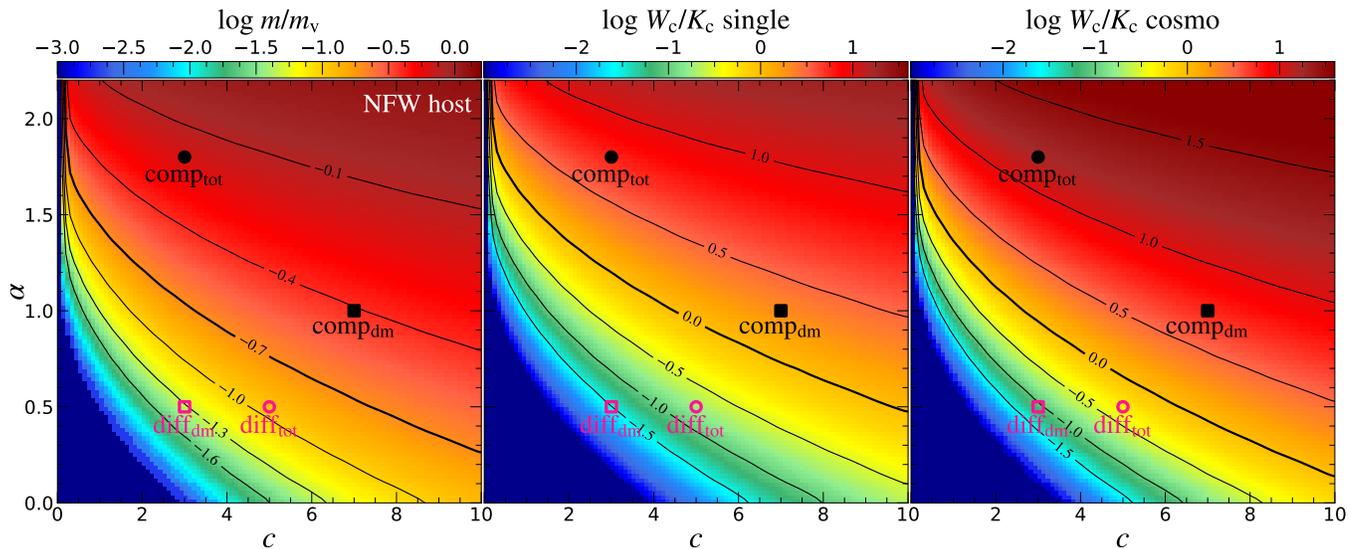
Figure 3. Toy-model estimates for satellite penetration and energy deposited in the host cusp by dynamical friction, as a function of the satellite compactness via the Dekel-Zhao profile parameters of concentration and inner slope (c, α). The host-halo profile is NFW with a moderate inner slope $s_1 \simeq 1$ and $c_{\mathrm{NFW}} = 5$. Figure G1 (available as supplementary material online) shows the same for or a steeper cusp with DZ parameters $s_1 = 1.5$ and $c_2 = 5$. We consider the cusp radius to be $R_c = 10\,\mathrm{kpc} \simeq 0.067 R_v$. The effective circularity of the satellite orbit in the cusp is assumed to be $\epsilon = 0.5$. **Left:** The fraction of bound satellite mass $f = m/m_v$ when at the half-mass radius of the host cusp, $r = 0.7 R_c = 7\,\mathrm{kpc}$, as obtained by solving eq. (9), corrected by eq. (10) with $B = 2$. **Middle:** The energy deposited in the host cusp of $R_c = 10\,\mathrm{kpc}$ by dynamical friction acting on a single satellite with $m_v/M_v = 0.1$, assuming $s = 2$ in eq. (17). The energy is normalized by the cusp initial kinetic energy, W_c/K_c. **Right:** The energy deposited in the cusp by a cosmological sequence of satellites during one virial crossing time, as obtained from eq. (21) with $\tau = 1$. The fiducial DZ profiles of pre-compaction and post-compaction satellites, as determined from the VELA simulations in §2, below and above the golden mass of $m_v \sim 10^{11.3} M_\odot$, are marked by open-magenta and filled-black symbols respectively. The squares refer to the fiducial profiles of DM only, with $(c, \alpha) = (3, 0.5)$ and $(7, 1)$ respectively. For comparison, the circles refer to the slightly more compact total mass profiles, with $(c, \alpha) = (5, 0.5)$ and $(3, 1.8)$ respectively. For an NFW host we read for the fiducial diffuse and compact satellites respectively $m/m_v \sim 0.05, 0.39$, $W_c/K_c \sim 0.03, 2.05$ for the single satellite, and $W_c/K_c \sim 0.08, 5.59$ for the sequence of satellites. This implies that typical low-mass, pre-compaction satellites are not expected to heat up the host cusps haloes while the massive, post-compaction satellites are expected to significantly heat up the host cusp over a non-negligible fraction of a virial crossing time. Using the satellite total-mass profiles, the three quantities are slightly larger, but the qualitative result remains the same. For the steep-cusp host, Fig. G1 (available as supplementary material online), the satellite stripping is stronger due to the steeper cusp, but the heating by compact satellites is still significant during half a virial time.

$s = 2$ (and $\epsilon = 0.5$ and $B = 2$ as before). The same for the steep-cusp host is shown in Fig. G1 (available as supplementary material online). The contours $\log W_c/K_c = 1$ mark the boundaries for significant cusp heating. The fiducial diffuse and compact satellites, as evaluated from the DM in the VELA simulations, §2, are marked by open and filled circles at $(c, a) = (3, 0.5)$ and $(7, 1)$, respectively. We read for the NFW host $W_c/K_c \simeq 2.05$ and 0.03 for the fiducial compact and diffuse satellite, respectively. For the fits to the more compact total satellite mass, we obtain $W_c/K_c \simeq 7.5$ and 0.2, respectively. This is more heating, but with a similar difference between the compact and diffuse satellites. For the steep-cusp host, and the DM satellite profiles, the respective results are $W_c/K_c \simeq 0.56$ and 0.001, namely less heating, as expected. Overall, we learn quite robustly that a single compact satellite of $m_v \sim 0.1 M_v$ is clearly capable of significantly heating up the cusp in an NFW host, and is expected to do so also in a steep-cusp host. In contrast, a fiducial diffuse satellite is expected to have only a negligible effect on the host cusp in the two types of hosts, even if the satellite compactness is evaluated via its total mass profile. According to eq. (17), the energy

deposited by a more massive satellite, $m_v = 0.1\,\mu\,M_v$, is expected to provide more heating by a factor μ^2.

A note of caution is that the deposited kinetic energy may not all be in the form of "heat", e.g., it could partly be in the form of bulk motion and in particular rotational energy. In this case the analysis should involve angular momentum, both in the estimation of the effect of DF on the DM cusp and in the evaluation of the DM response to outflows (§6). For satellite entry to the cusp on a rather tangential orbit, it can be shown that the angular-momentum deposit by DF with respect to the maximum angular momentum of the cusp, had it been a disk with circular orbits, is comparable to the relative energy deposit of eq. (17). This implies a likely significant effect of the compact satellite on the cusp angular momentum. We defer an analysis involving angular momentum to future work.

3.2.2 A cosmological sequence of satellites

For a cosmological sequence of satellites, we assume that all satellites of mass $m_v > 10^{11.3} M_\odot$ are post-compaction (Zolotov et al. 2015; Tomassetti et al. 2016), with a high concentration and a cusp (§2), which make

then penetrate to the host cusp with a non-negligible mass fraction f. For a host halo of $M_v = 10^{12.3} M_\odot$, these are mergers of mass ratio $m_v/M_v \geqslant 1:10$. We read from Figure 7 of Neistein & Dekel (2008), based on EPS theory, that $\sim 60\%$ of the total accretion rate is in such satellites (at all redshifts). For the average total accretion rate we adopt the analytic estimate of Dekel et al. (2013) in the EdS cosmological regime (a good approximation at $z > 1$),

$$\dot{M} = 0.47\ \mathrm{Gyr}^{-1}\ M_v\ (1+z)_3^{5/2}. \quad (18)$$

The rate of total energy deposited in the cusp by dynamical friction involving all these satellites can be computed by integration over their EPS conditional satellite mass function normalized to a total that equals $0.6\dot{M}$, convolved with the energy per satellite as given by eq. (17). The mass function of first-order subhaloes at infall, could be approximated by $dn/dm \propto m^{-1.5}$ at $m \simeq 0.02 - 0.2 M_v$, with a truncation at $m_{max} \lesssim 0.5 M_v$ (Jiang & van den Bosch 2014). For a crude conservative estimate, we assume here that each of the relevant satellites above $0.1 M_v$ is of a given mass m_v somewhat above $0.1 M_v$, and write the accreted mass rate in these satellites as $\dot{N} m_v = 0.6\dot{M}$, where $\dot{N}$ is the number of relevant satellites that merge during one Gigayear. This gives

$$\dot{N} = 1.4\,\mathrm{Gyr}^{-1}\ (m_v/M_v)_{0.2}^{-1}\ (1+z)_3^{5/2}. \quad (19)$$

Multiplying the result for a single satellite from eq. (17) by $\dot{N}$, we obtain an approximation for the rate of relative work done on the host cusp by all the compact satellites

$$\frac{\dot{W}}{K_c} = 3\ \mathrm{Gyr}^{-1}\ f_{0.2}^2\ s_2\ (m_v/M_v)_{0.2}\ (1+z)_3^{5/2}. \quad (20)$$

In τ halo virial crossing times, each given by $t_v \simeq 0.5\,\mathrm{Gyr}\,(1+z)_3^{-3/2}$ (Dekel et al. 2013), the relative DF work becomes

$$\frac{W_c(\Delta t = \tau t_v)}{K_c} = 1.5\,\tau\ f_{0.2}^2\ s_2\ (m_v/M_v)_{0.2}\ (1+z)_3. \quad (21)$$

Based on Figure 7 of Neistein & Dekel (2008), we adopt $m_v/M_v = 0.2$ as our fiducial value.

An upper limit for the total energy deposited could be estimated by integrating over a whole Hubble time, $t \simeq 6.6 t_v \simeq 3.3\,\mathrm{Gyr}(1+z)_3^{-3/2}$, namely about a factor of 6.6 times the energy in eq. (21). During a time window of $0.5 t_v$, the timescale for the "heat" deposited by dynamical friction to "dissipate" as estimated below in §5.2 from an N-body simulation, we estimate $W/K_c \sim 0.75\,(1+z)_3$.

The right column of Fig. 3 shows the toy-model estimates for W_c/K_c for a cosmological sequence of satellites as a function of the satellite's profile parameters (c, α), for the NFW host, based on eq. (21). This is assuming $m_v = 0.2 M_v$, at $z = 2$, and a duration of one virial time $\sim 0.5\,\mathrm{Gyr}$, namely $\tau = 1$ (with $s = 2$, $B = 2$, and $\epsilon = 0.5$ when solving for f_t in eq. (9), as before). The same for the steep-cusp host is shown in Fig. G1 (available as supplementary material online). We read

for the NFW host, using the satellite DM mass profiles, $W_c/K_c \simeq 5.6\,\tau$ and 0.08τ for the fiducial compact and diffuse satellite, respectively. When using the satellite total mass, the results are $W_c/K_c \simeq 20\,\tau$ and $0.5,\tau$. For the steep-cusp host, with the satellite DM profiles, the respective results are $W_c/K_c \simeq 1.5\,\tau$ and $0.003\,\tau$. We learn that the median energy deposited by a sequence of satellites in one virial time is crudely estimated to be comparable to and slightly larger than the energy deposited by a single satellite of $m_v \sim 0.1 M_v$. During a period of $\sim 0.5 t_v$, the satellites are expected to significantly heat up the cusp in an NFW host, and they are estimated to do so also in a steep-cusp host. In contrast, a sequence of diffuse satellites is expected to have only a negligible effect on the host cusp for the two types of hosts during a period of order a virial time.

Based on the proportionality to $(1 + z)$ in eq. (21), a somewhat higher satellite compactness is needed for heating the cusp to the same level over a similar period at lower redshifts, while a somewhat lower compactness would be sufficient at higher redshifts.

Figure I1 (available as supplementary material online) complements Fig. 3 by showing the same toy-model estimates for satellite penetration and the energy deposited in the host cusp by dynamical friction, but in the plane of the alternative DZ parameters (c_2, s_1) instead of the natural DZ parameters (c, α).

We note that, with a satellite initial mass function that declines flatter than m_v^{-2} up to a drop at m_{max}, the dependence on m_v^2 of the energy deposited in the cusp, eq. (17), indicates that the total heating of the cusp is dominated by the more massive satellites. One could comment that this dominance of massive satellites may be strengthened because, beyond the crude tidal-stripping approximation of eq. (6), the actual satellite mass loss also depends on the time the satellite spends in the strong-tide regime before entering the cusp. In particular, satellites that make it to the cusp in the first or second pericenter of their orbit are expected to bring in more mass than satellites that hang around and are subject to stripping for a longer time (see Fig. 4 below). One can estimate that, for a given satellite density profile, the rate of decay of the satellite orbit is increasing with m_v/M_v. This can be obtained by relating the decay rate to either the ratio of AM loss by the DF torque to the orbital AM of the satellite, or the ratio of work done on the satellite to its energy. These ratios are both proportional to m/M, because the DF acceleration is proportional to m, while the specific angular momentum or energy are proportional to M, and the tidal stripping is not strongly dependent on either.

4 DYNAMICAL-FRICTION HEATING: SEMI-ANALYTIC SATGEN SIMULATIONS

In order to explore more accurately the satellite properties required for efficient penetration and strong DM heating by dynamical friction, we utilize the dedicated Semi-Analytic Satellite Generator, SatGen (Jiang et al.

2020),[5] (described in Appendix §C, available as supplementary material online). It simulates tidal stripping while the satellite of an assumed initial density profile orbits in the potential well of the host halo and the central galaxy and penetrates to the inner halo subject to dynamical friction by the host dark matter. The code utilizes EPS merger trees and physical prescriptions for the relevant mechanisms. The tidal mass loss and the structural evolution of the satellites depend on the initial structure of the satellites, following Errani, Peñarrubia & Walker (2018), such that a dense, cuspy satellite at infall is more resistant to tidal disruption than a cored satellite. In the version used here, we model the satellite with non-dissipative particles, without explicitly treating the gas component.

The energy deposited in the radius shell $(r \pm dr)$ of the host is computed either by the work done by dynamical friction on the satellite along its orbit, or by the orbital energy loss of the satellite, with similar results. Dynamical friction is modeled using eq. (12) with $\ln \Lambda = \ln(M_v/m)$. In both cases, rather crudely, the energy is assumed to be deposited locally.

4.1 A single satellite

Figure 4 summarizes the results of running SatGen with a single satellite, our fiducial compact or diffuse, penetrating into an NFW host of $c_{\mathrm{NFW}} = 5$. The DZ mass profile parameters of the initial satellites at the host virial radius are the fiducial $(c, \alpha) = (7, 1)$ and $(3, 0.5)$, respectively, consistent with the DM VELA profiles after and before the compaction in Fig. 1 and Fig. 2. The initial satellite mass at R_v is $m_v = 10^{11.5} M_\odot$ in a host of $M_v = 10^{12.5} M_\odot$. The orbit has a typical circularity at R_v of $\epsilon = 0.4$.

The orbit (top) shows that the compact satellite enters the 10 kpc cusp near its second pericenter, while the diffuse satellite enters only near its third pericenter. The satellite bound mass fraction at the half-mass radius of the core (7 kpc) is $m/m_v \simeq 0.32$ and 0.09 respectively. These are to be compared to the toy model predictions of $m/m_v \sim 0.39$ and 0.046. The energy deposited in the cusp is $W_c/K_c \simeq 2.0$ and 0.32 for the compact and diffuse satellite respectively. In comparison, substituting the mass fractions from SatGen in eq. (17) of the toy model yields $W_c/K_c \simeq 2.05$ and 0.029, emphasizing the qualitative difference between the effects of the compact and diffuse satellites. The semi-analytic simulation thus confirms the qualitative estimates from the toy model that the fiducial compact satellite is expected to significantly heat up the host cusp, while the fiducial diffuse satellite is not expected to generate major heating.

In order to evaluate the effect of including the central baryons in the satellites, we re-ran SatGen with our fiducial NFW host halo but with the satellites following the more compact DZ-profile fits to the VELA simulated galaxies using the *total* mass rather than the dark matter alone. The DZ profiles are here $(c, \alpha) = (5, 0.5)$ and $(3, 1.8)$ for the diffuse and compact satellites respectively, following the VELA pre and post compaction galaxies from Fig. 1. We learn from Fig. F1 (available as supplementary material online), which summarizes this SatGen run compared to the fiducial case of Fig. 4, that the more compact satellites indeed penetrate to the host cusp with a higher mass and deposit more energy there accordingly, as expected. However, the difference is rather small, with $m/m_v \simeq 0.5$ compared to 0.4, and $W_c/K_c \simeq 2.5$ compared to 2, for the compact satellites. This is a smaller difference than expected based on the toy model of §3. These results indicate that our main analysis using the fiducial satellite profiles as derived from the DM mass in the VELA simulated galaxies should provide good, conservative estimates for the cusp heating by dynamical friction.

In order to evaluate the effect of a *steep-cusp* host halo, we re-ran SatGen with the fiducial DM satellite profiles but with the steep-cusp host instead of the fiducial NFW. We read from Fig. G2 (available as supplementary material online) that for the steep-cusp host the same satellites penetrate with $m/m_v \simeq 0.30$ and 0.075 for the compact and diffuse satellite, respectively. These values are only slightly smaller than for the NFW host, probably because the two host haloes are not very different outside the cusp. The satellite mass becomes significantly smaller only in the inner cusp, inside ~ 5 kpc, where the tidal stripping is more efficient due to the higher density of the steep-cusp halo. The corresponding toy-model estimates are $m/m_v \simeq 0.23$ and 0.01. The energy deposited in the steep cusp according to SatGen is $W_c/K_c \simeq 2.0$ and 0.32. These values are rather similar to those in the NFW host. Apparently, the lower satellite mass that has penetrated into the cusp is balanced by the higher density in the cusp, making the work by dynamical friction almost the same in the two cases. The toy-model estimates are $W_c/K_c \simeq 0.56$ and 0.001, indicating a stronger dependence on the host-cusp properties than is actually produced by SatGen, but emphasizing again that the effects of the compact and diffuse satellites are very different, and that the compact satellite is expected to provide non-negligible heating also in the steep-cusp halo.

4.2 A cosmological sequence of merging satellites

We next run SatGen with a cosmological distribution of satellites over time τt_v, where $\tau \sim 1$, from $z = 2.3$ to 2. We actually generate random merger trees for a target halo of of a given mass at $z = 2$ starting at $z = 20$, evolve these cosmological sequences of satellites in time, and keep track of the DF heating induced by them over the last virial time at each given halo-centric radius. The satellite masses are drawn from the EPS conditional mass function (e.g., Lacey & Cole 1993; Neistein & Dekel 2008; Parkinson, Cole & Helly 2008; Benson 2017). We assume that the orbital energy at infall through R_v is the same as that of a circular orbit of radius $R_v(t)$, and draw the circularity ϵ at random from the distribution $\mathrm{dP}/\mathrm{d}\epsilon = \pi \sin(\pi\epsilon)/2$, which approxi-

---

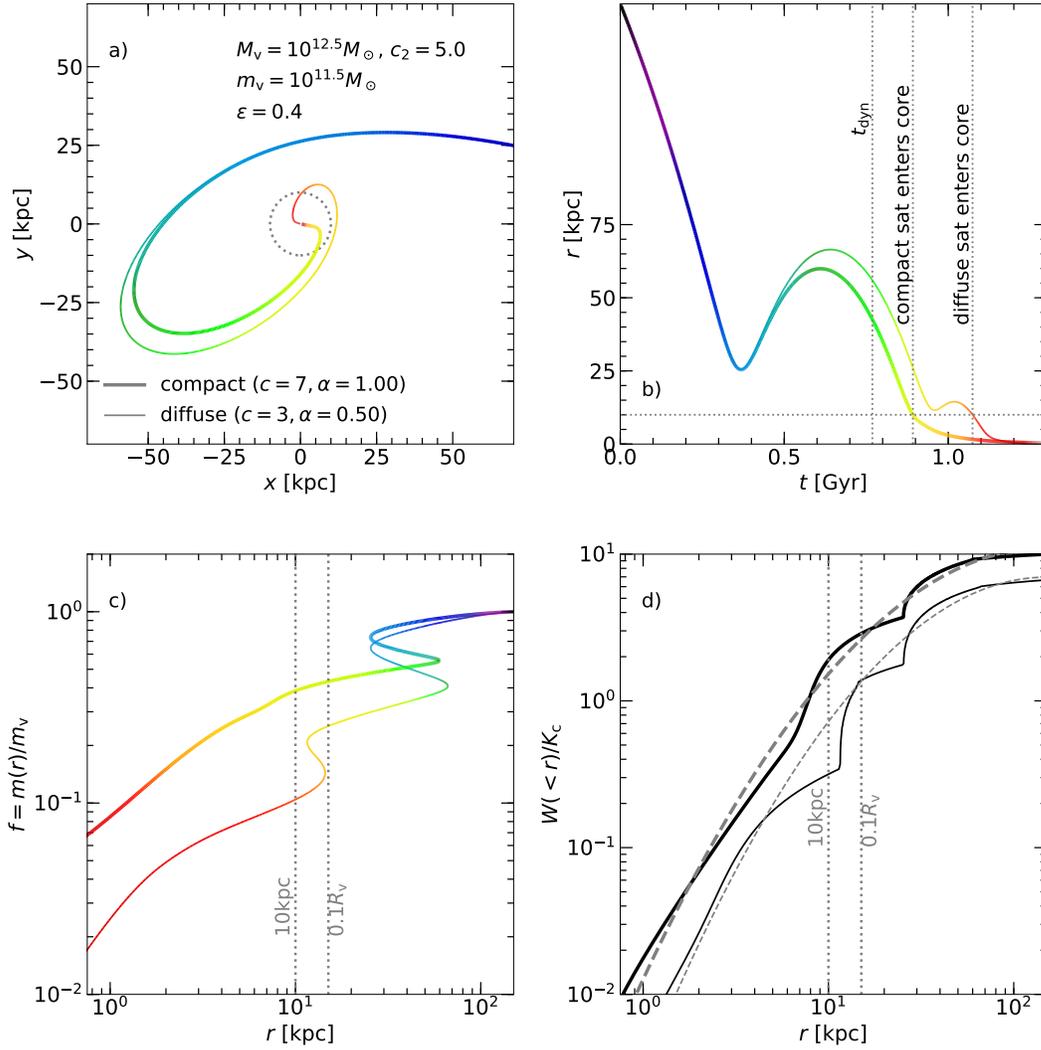[5] Available for implementation in https://github.com/shergreen/SatGen .

Figure 4. Semi-analytic SatGen simulations of single satellites, diffuse and compact, marked by thin lines and thick lines respectively. The Dekel-Zhao mass profile parameters of the satellites are $(c,\alpha)=(3,0.5)$ and $(7,1)$, respectively, consistent with the VELA profile pre and post compaction in Fig. 1 and Fig. 2. The initial satellite mass at $R_{\rm v}$ is $m_{\rm v}=10^{11.5}M_{\odot}$ in a host of $M_{\rm v}=10^{12.5}M_{\odot}$, with an NFW profile of $c_{\rm NFW}=5$. The same for a steep-cusp halo is shown in Fig. G2 (available as supplementary material online). The orbit has a typical circularity at $R_{\rm v}$ of $\epsilon=0.4$. **Top:** The satellite orbits, a face on projection (left) and radius within the host halo as a function of time, with the times of entry into the $10\,{\rm kpc}$ cusp marked. **Bottom left:** The satellite bound mass fraction $f=m_{\rm t}/m_{\rm v}$ as a function of radius. **Bottom right:** The energy deposited by dynamical friction in the host, interior to radius r, with respect to the kinetic energy within the cusp ($K_{\rm c}=0.5M(r<10\,{\rm kpc})V_{\rm v}^2$). The dashed curves refer to the fits used in the CuspCore analysis of §6, listed in Table 1. We read $m_{\rm c}/m_{\rm v}\simeq0.1,0.4$ and $W_{\rm c}/K_{\rm c}\simeq0.3,1.8$ for the diffuse and compact satellites respectively. We learn that the bound gas fraction of the compact satellite is ~10 times larger than for the diffuse satellite. As a result the DF on the compact satellite generates a significant energy change in the host cusp, while the effect of the diffuse satellite is negligible, consistent with the analytic estimate of §3.

mates the distribution measured in cosmological simulations (e.g., Wetzel 2011; van den Bosch 2017), where $\langle\epsilon\rangle\simeq0.45$ at $z=2$. We consider the infall locations to be isotropically distributed on the virial sphere. In the fiducial run, the host halo is again of mass $M_{\rm v}=10^{12.5}M_{\odot}$ with an NFW profile and $c_{\rm NFW}=5$. Every satellite with $m_{\rm v}>10^{11}M_{\odot}$ is assumed to be post-compaction, with the fiducial compact DZ profile of $(c,a)=(7,1)$, while all the less massive satellites are assumed to be diffuse, pre-compaction, with $(c,a)=(3,0.5)$. In a reference run, with the same distribution of masses and initial orbits, all the satellites are assumed to be diffuse. This refer-

ence case is supposed to approximate the real situation for less massive hosts, say $M_{\rm v}<10^{11.5}M_{\odot}$, where no compact satellites (of $m_{\rm v}>10^{11}M_{\odot}$) are expected.

Figure 5 shows the energy deposited within the sphere of radius r about the halo center, during a period of one virial time at $z=2$, $t_{\rm v}\simeq0.5\,{\rm Gyr}$, corresponding to a redshift interval $\Delta z\simeq0.3$.[6] We read for the NFW host halo $W_{\rm c}/K_{\rm c}\simeq1.8$ and 0.35 for the run with and without post-compaction satellites, respectively. This is

---

[6] In the EdS regime, approximately valid at $z>1$, the relation is $\Delta t\simeq-1.68\,{\rm Gyr}\,(1+z)_3^{-5/2}\Delta z$.
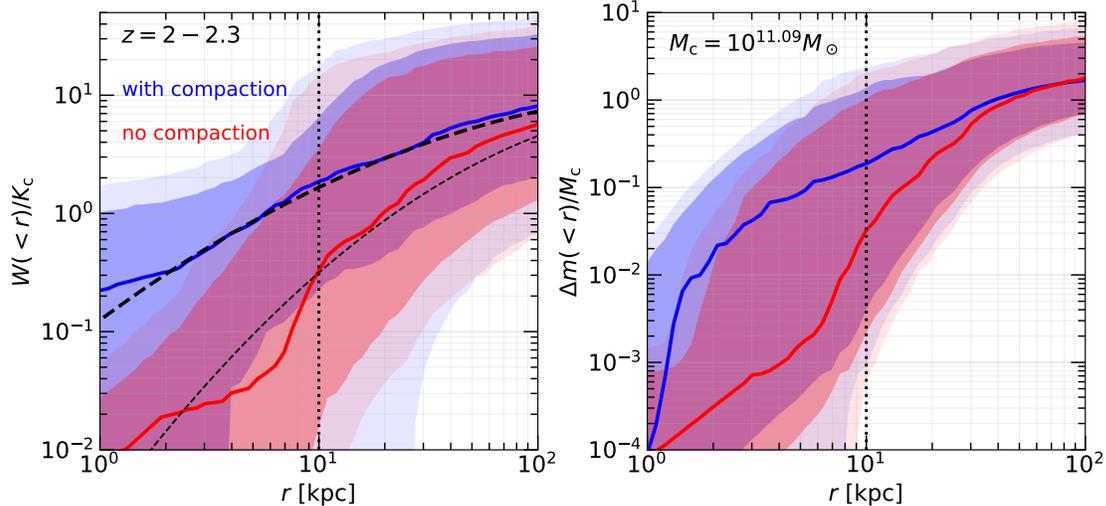
Figure 5. SatGen simulations of a cosmological sequence of satellites during one halo virial time at $z \sim 2$, $t_{\rm v} \simeq 0.5$ Gyr, corresponding to a redshift interval $\Delta z \simeq 0.3$. **Left:** the energy deposited within radius r with respect to the kinetic energy in the 10 kpc cusp, $W(<r)/K_{\rm c}$ (left), with the dashed lines representing the fit used in §6 and listed in Table 1. **Right:** the satellite mass deposited inside r with respect to the host cusp mass within 10 kpc, $\Delta m(<r)/M_{\rm c}$. The host halo mass is $M_{\rm v} = 10^{12.5} M_\odot$, with an NFW profile of $c_{\rm NFW} = 5$. The same for the steep-cusp profile is shown in Fig. G3 (available as supplementary material online). The fiducial run (blue) assumes compact satellites of $m_{\rm v} > 10^{11} M_\odot$, while the reference run (red) assumes that all satellites are diffuse. Shown are the medians and the 68% and 95% percentiles over 600 random merger trees. The DZ profiles for the compact and diffuse satellites have the fiducial parameters $(c, \alpha) = (7, 1)$ and $(3, 0.5)$, respectively. We see significant mass and energy deposit in the cusp for the compact satellites (valid for hosts of $M_{\rm v} > 10^{12} M_\odot$ at $z > 1$), and low mass and energy deposit for the case of diffuse satellites (approximating the situation in host haloes of $M_{\rm v} < 10^{11.5} M_\odot$ or at low redshifts).

compared to the toy model predictions of $W_{\rm c}/K_{\rm c} \simeq 5.6$ and 0.08. They both point to a similar qualitative difference where the post-compaction satellites generate significant cusp heating while the diffuse satellites provide only minor heating during a time interval comparable to the virial time at $z \sim 2$.

For the steep-cusp halo run of SatGen with a sequence of satellites, shown in Fig. G3 available as supplementary material online), we obtain $W_{\rm c}/K_{\rm c} \simeq 1.8$ and 0.35, respectively. This is very similar to the SatGen results for the NFW host, consistent with the similarity of the results obtained for a single satellite. The corresponding toy-model predictions are $W_{\rm c}/K_{\rm c} \simeq 1.5$ and 0.003, consistent with the SatGen result for the compact satellite, though estimating less energy deposit in the steep-cusp halo compared to the NFW halo. We learn that the cosmological sequence of compact satellites during half a virial time is expected to provide significant cusp heating in the steep-cusp host halo as well.

5   POST-HEATING RELAXATION

Before we proceed to the response of the heated DM to AGN outflows in §6, we address the behavior of the cusp after the DF heating. In the spirit of our simple toy modeling, we first assume that the work done by dynamical friction on the satellite is instantaneously deposited in the dark matter as kinetic energy ("heating"), in terms of increased velocity dispersion, while the potential energy and the associated density profile

remain unchanged at that instant. In the case of a single satellite, after it has disrupted or settled in the center, we expect the host halo to relax over a certain time window $\tau t_{\rm v}$ into a new Jeans equilibrium. In this process the dark matter "cools" while it expands, namely the potential energy grows (becoming less negative) at the expense of the kinetic energy, and the cuspy density profile partly flattens. We first address this process in §5.1 via simplified toy models, and then in §5.2 using an N-body simulation.

5.1   Toy models

For a first crude qualitative impression of this process, we consider the cusp to be an isolated shell of mass M in virial equilibrium, where the kinetic and potential energies are related via $K = -U/2 = GM^2/(2R)$, both at the initial radius $R_{\rm i}$ and at the final radius $R_{\rm f}$, after a deposit of energy W and a relaxation process. Conservation of energy implies

$$-\frac{GM^2}{2R_{\rm i}} + W = -\frac{GM^2}{2R_{\rm f}}, \qquad (22)$$

so

$$\frac{K_{\rm f}}{K_{\rm i}} = \frac{R_{\rm i}}{R_{\rm f}} = 1 - \frac{W}{K_{\rm i}}. \qquad (23)$$

The system expands, while it cools, from a peak kinetic energy of $K_{\rm peak} = K_{\rm i} + W$, immediately after the action of dynamical friction, to below the initial kinetic energy, $K_{\rm f} = K_{\rm i} - W$. In the relaxation process, the kinetic and potential energy changes are $\Delta K = -\Delta U = -2W$ Thus,
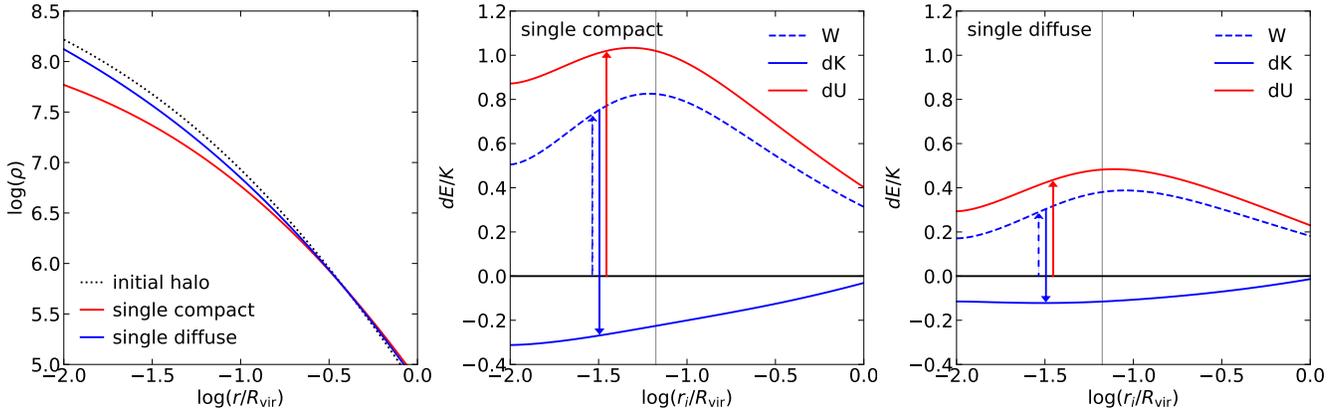
Figure 6. The relaxation of a DM halo from an initial to a final configuration in Jeans equilibrium, based on the CuspCore model, after a deposit of kinetic energy that mimics dynamical friction by our fiducial compact and a diffuse satellites. The initial halo is NFW with a central point mass of $8 \times 10^{10} M_\odot$ resenting baryons. The energy deposited as a function of radius in the host is adopted from the SatGen simulations, the bottom-left panel of Fig. 4. Shown are the density profile (left) and the changes in the kinetic and potential energies interior to radius r (dK and dU, blue and red, respectively), normalized by the kinetic energy $K(<r)$. We see that after the initial heating of K by the work of dynamical friction W (dashed), the changes are the same in amplitude, $\Delta K \simeq -\Delta U$, as the cusp cools and flattens. The compact satellite causes a much stronger effect.

after an energy deposit due to DF by a single satellite, the system may be found in one of two configurations. During the time window for expansion and cooling, τt_v (to be estimated below), the system is "hotter" then before with the cuspy profile practically unchanged. After this period, the system has a flatter density profile but it is "cooler" than in its initial state.

For a more quantitative estimate, we assume for the host halo a Dekel-Zhao density profile both at the initial and at the final configurations, both in Jeans equilibrium, as described in section 3.3 of Freundlich et al. (2020a). The transition between the two states is assumed to start with an instantaneous addition of kinetic energy to the dark matter at every radius according to an assumed input profile $W(<r)$, mimicking the work by dynamical friction, while the mass distribution and therefore the potential energy are fixed. This is followed by a relaxation to a new equilibrium while shells enclosing given DM masses are assumed to conserve energy. This is analogous to the CuspCore model used in §6 and described in Appendix §E (available as supplementary material online) for the response to outflows, except that the first instantaneous change is in the kinetic energy rather than in the potential energy. We assume here an initial host halo of $M_v = 10^{12.5} M_\odot$ at $z = 2$, with an NFW profile of $c_{NFW} = 5$, and adopt the corresponding input energy-deposit profile from the results of the SatGen simulations of a single satellite, the medians in the bottom-left panel of Fig. 4. This is for the fiducial compact and diffuse satellites, with $m_v/M_v = 0.1$ and $\epsilon = 0.4$.

Figure 6 shows the resultant changes in the density profile and in the kinetic and potential energies, dK and dU, interior to r, normalized by $K(<r)$, with respect to the initial configuration. The initial heating of K through the work W by dynamical friction (dashed blue line) shows for the compact satellite maximum relative heating near the cusp radius of $10\,\mathrm{kpc} \simeq 0.067 R_v$. The

subsequent relaxation to a new Jeans equilibrium (solid lines) is associated with cooling and expansion such that $\Delta K \sim -\Delta U$, in qualitative agreement with the result from the crude shell toy model. The effect of a compact satellite is naturally much stronger than the effect of a diffuse satellite.

The N-body simulation described in §5.2, of the effect of a single compact satellite on the cusp, confirms the expected two-stage behavior of first heating the cusp without affecting the density profile, and then relaxing to a cooler and extended configuration with a flatter density profile. We learn from this simulation that the time window for the "hot" phase is $t_{hot} \sim 0.5 t_v$. Substituting $\tau = 0.5$ in eq. (21), we expect a total heating of $W_c/K_c \simeq 0.75\,(1+z)_3$ during this hot phase. This being of order unity at $z \sim 2$ indicates that, at these high halo masses and redshifts, the cusp can be maintained in the "hot" phase, and thus be more susceptible to core formation by AGN outflows (§6). We will study in §6 the DM response to AGN-driven outflows when starting from either the "hot" or the "cold" configuration.

5.2 Heating and flattening in N-body simulations

We realized in §5.1 that a halo cusp, after being heated by a single satellite, would relax into a new Jeans equilibrium in an energy-conserving process that involves expansion and cooling, yielding a cold cusp with a somewhat flatter density profile. Here, we use an N-body simulation in order to verify these heating and cooling processes. We wish in particular to estimate the time it takes for the heated cusp to cool down. This will be used as the effective duration for evaluating the heating effect of a cosmological sequence of satellites, expressed as τ in eq. (21) of the toy model and as the proper duration of the SatGen simulation in §4.2. According to these estimates, if the cooling process takes a significant fraction of a virial time or more, namely τ is not much
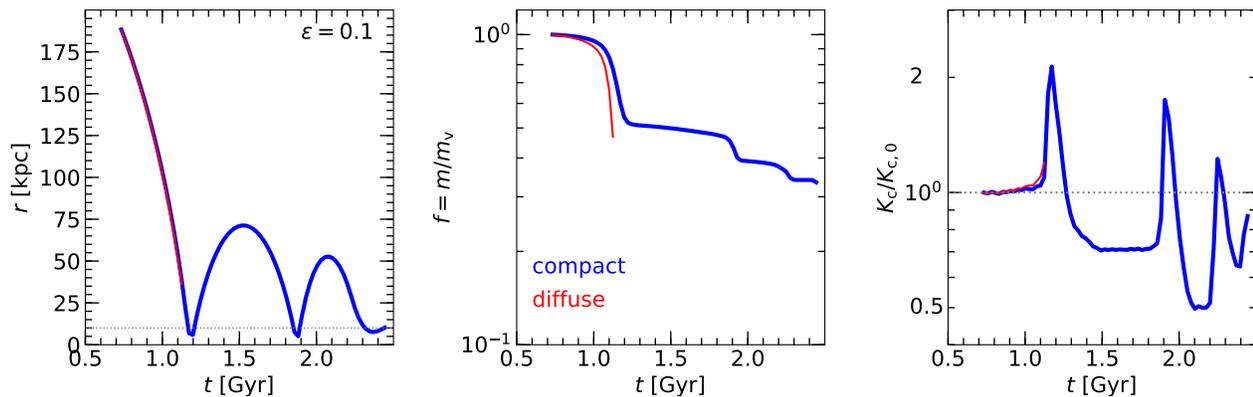
Figure 7. Time evolution in an N-body simulation of a compact satellite (blue) and a diffuse satellite (red) of $m_v = 10^{11.5} M_\odot$ on a rather radial orbit in a host halo of $M_v = 10^{12.5} M_\odot$ at $z = 2$. **Left:** the satellite distance from the host centre. **Middle:** the bound mass of the satellite. **Right:** the kinetic energy within the host cusp with respect to its initial value. We read that the heated episodes near the pericenters last a total of ~ 0.25 Gyr, which is $\sim 0.5 t_v$.
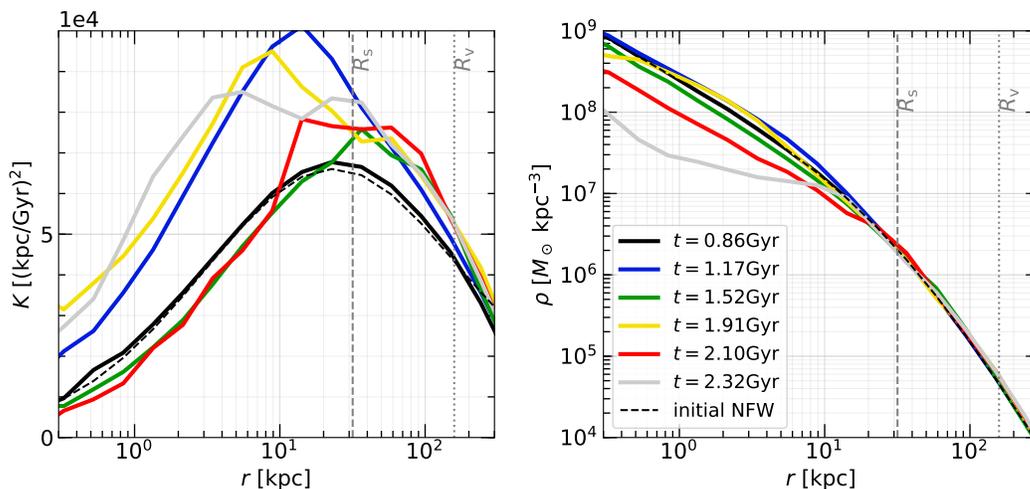

Figure 8. Evolution of the profiles of kinetic energy (left) and density (right) in an N-body simulation of a compact satellite in a massive halo, the same as Fig. 7. The initial NFW profile is shown (dashed). The times shown are: (a) $t = 0.86$ Gyr (black), virial crossing, cold, cuspy. (b) $t = 1.17$ Gyr (blue), first pericenter, hot, cuspy. (c) $t = 1.52$ Gyr (green), first apocenter, cold, cuspy. (d) $t = 1.91$ Gyr (yellow), second pericenter, hot, cuspy. (e) $t = 2.10$ Gyr (red), second apocenter, slightly hot, slightly flattened. (f) $t = 2.32$ Gyr (grey), third pericenter, hot, more flattened.

smaller than unity, then we can expect the cusp to be maintained hot by the sequence of mergers with compact satellites. If, on the other hand, $\tau \ll 1$, the cusp will drop to a cool phase between the individual merger episodes. The different responses of a hot cusp and a cooled-flattened cusp to AGN-driven outflows will be discussed in §6.

The N-body simulation uses the public GADGET-2 code (Springel 2005), with a particle mass of $m_p = 3 \times 10^5 M_\odot$ and a softening length of $r_{ref} = 0.0003 (N_v/10^6)^{-1/3} \ell_v$, where $N_v = m_v/m_p$ is the number of particles within ℓ_v (following van Kampen 2000; Ogiya et al. 2019), namely $r_{ref} = 20$ pc. We consider a satellite of $m_v = 10^{11.5} M_\odot$ orbiting a host of $M_v = 10^{12.5} M_\odot$ near $z = 2$ along a rather eccentric orbit of circularity $\epsilon = 0.1$ at the halo virial radius and a velocity virial velocity at that radius. We compare the cases of a compact and a diffuse satellite, where the compact satel-

lite is initialized with an NFW profile of $c_{NFW} = 50$, and the diffuse satellite is initialized with a Burkert profile (Burkert 1995) of a scale radius $r_s = 12$ kpc, the same as the inner scale radius of an NFW profile with $c_{NFW} = 6$. These profiles are not the same as the fiducial profiles used in the rest of the paper, but they are rather close, and serve us for our qualitative purpose here. More details on this N-body simulation are provided in Appendix §D (available as supplementary material online).

Figure 7 shows the time evolution of the relative excess of kinetic energy of the dark matter within the cusp with respect to its initial kinetic energy. It also shows the distance of the satellite from the host center, indicating the successive pericenters and apocenters with the orbit decay into the cusp, and the mass of the satellite at that time. We see hot-phase periods near the first three orbit pericenters, relaxing to cold-phase pe-

riods in between as well as after the coalescence. The overall time in the hot phase is $t_{\mathrm{hot}} \sim 0.5 t_{\mathrm{v}}$.

Figure 8 shows the density profile and kinetic-energy profile of the host halo at different times through the merger process compared to the initial cold NFW profile. In the kinetic energy, we see hot phases in the cusp during periods about the first, second and third pericenters, cooling down in between. The maximum heating is near and interior to the cups radius of $\sim 10\,\mathrm{kpc}$. The density profiles show that the initial profile hardly changes until after the first pericenter. Very minor flattening starts appearing after the first pericenter, and a more noticeable but still partial flattening develops after the second pericenter, becoming more significant after the third pericenter. The latter may be subject to uncertainties due to difficulties in identifying a center for the halo at these stages of satellite coalescence with the host halo center.

Substituting $\tau = 0.5$ in the toy-model estimate eq. (21), we expect a total heating of $W_{\mathrm{c}}/K_{\mathrm{c}} \simeq 0.75(1+z)_3$ during this hot phase. This being of order unity at $z \sim 2$ indicates that, at these high halo masses and redshifts, the cusp can be maintained for most of the time in the "hot" phase, and thus be more susceptible to core formation by AGN outflows, to be demonstrated in §6.

6 DM RESPONSE TO AGN OUTFLOW

6.1 Modeling the response to outflows

In order to model the dark-matter response in the inner halo to an episode of AGN-driven gas outflow, we use the CuspCore analytic model presented in Freundlich et al. (2020a), that has been originally implemented for supernova-driven outflows in low-mass galaxies. The model, as summarized in Appendix §E[7] (available as supplementary material online), consists of a two-stage response, starting from an instantaneous change of potential due to an assumed central mass removal while the velocities are frozen, and following by an energy-conserving relaxation to a new Jeans equilibrium. The model assumes a spherical halo, isotropic velocities, Jeans equilibrium in the initial and the final stages, and energy conservation for shells enclosing a given DM mass. This model proved successful when compared to cosmological simulations (NIHAO) except in situations of major mergers. It uses the DZ density profile with flexible concentration and inner slope, utilizing its analytic expressions for the potential and kinetic energies, as described in Appendix §A and Freundlich et al. (2020b).

We assume here as before a halo of mass $M_{\mathrm{v}} = 10^{12.5} M_{\odot}$ with a given initial density profile. We add a central point mass representing baryons of $M_{\mathrm{g}} = M_{\mathrm{s}} = 4 \times 10^{10} M_{\odot}$, following the average stellar-to-halo mass ratio for such a halo mass at $z = 2$ (Behroozi et al. 2019) and assuming a gas fraction of 0.5 for star-forming

---

[7] The simplest version of CuspCore is available online for application in https://github.com/Jonathanfreundlich/CuspCore.

Fitting function for the energy deposited by DF

Host	Satellite	Compactness	A	B	R_{max}
NFW	single	compact	16.81	0.62	R_{v}
NFW	single	diffuse	16.61	0.71	R_{v}
NFW	cosmo	compact	16.80	0.24	692
NFW	cosmo	diffuse	16.70	0.45	610
steep	single	compact	16.81	0.58	R_{v}
steep	single	diffuse	16.57	0.65	R_{v}
steep	cosmo	compact	16.70	0.70	R_{v}
steep	cosmo	diffuse	16.50	1.00	R_{v}

Table 1. We use as input to the outflow model CuspCore the energy deposited in the host cusp by the dynamical friction exerted on satellites as obtained from the corresponding SatGen simulations. We actually use a functional fit (shown in the SatGen figures) of the form $\log_{10} W(<r) = A - B\,[\log_{10}(r/R_{\mathrm{max}})]^2$, with the parameters listed in this table, where A and B are log energy in units of $M_{\odot}\,\mathrm{kpc}^2\,\mathrm{Gyr}^{-2}$, and R_{max} is in kpc, with $R_{\mathrm{v}} = 150\,\mathrm{kpc}$.

galaxies at that redshift (Tacconi et al. 2018). The corresponding dark-matter velocity-dispersion profile is determined by requiring Jeans equilibrium (Freundlich et al. 2020b).

The outflowing mass is assumed to be a fraction η of the gas mass, ranging from zero to unity, namely a fraction 0.5η of the central baryonic mass. We estimate in §7.2 that the typical gas mass in the galaxy of $M_{\mathrm{gas}} \sim 4 \times 10^{10} M_{\odot}$ is comparable to the typical mass of the fresh cosmologically accreted gas during $\sim 0.5 t_{\mathrm{v}} \sim 0.25\,\mathrm{Gyr}$. Taking into account accreting recycled gas that is three times more massive than the accreting fresh gas, the gas available for ejection by AGN feedback during that period is $\sim 1.2 \times 10^{11} M_{\odot}$. Recall that we learned from the N-body simulation of §5.2 that $\sim 0.5 t_{\mathrm{v}}$ is the "cooling" time for the cusp heated by dynamical friction, and we will show below that a pre-heated cusp is most responsive to outflows, so this is the time interval within which AGN feedback is most effective in producing a core. We model here only a single episode of $\eta = 0.5$ or $\eta = 1$, and consider it to provide a lower limit for the effect of core formation.

6.2 A single satellite

Figure 9 shows the change in the DM density profile due to an instantaneous outflow episode based on the CuspCore model. The initial DM halo is our standard, with $M_{\mathrm{v}} = 10^{12.5} M_{\odot}$ and an NFW profile of $c_{\mathrm{NFW}} = 5$ at $z = 2$. The initial kinetic energy obeys Jeans equilibrium (following Freundlich et al. 2020a), with an additional presence of a $10^{10.9} M_{\odot}$ central point mass representing a baryonic component, half gas and half stars. The pre-outflow configuration is the result of DF heating by a single satellite of mass m_{v}, either the compact or the diffuse fiducial cases deduced from the cosmological simulations described in §2. The energy deposited by dynamical friction is read from the SatGen simulation for a satellite of $m_{\mathrm{v}} = 0.1 M_{\mathrm{v}}$, as shown in Fig. 4. We actually use the functional fit that is shown in the figure and presented in Table 1. A satellite mass of $0.1\mu m_{\mathrm{v}}$ is mim-
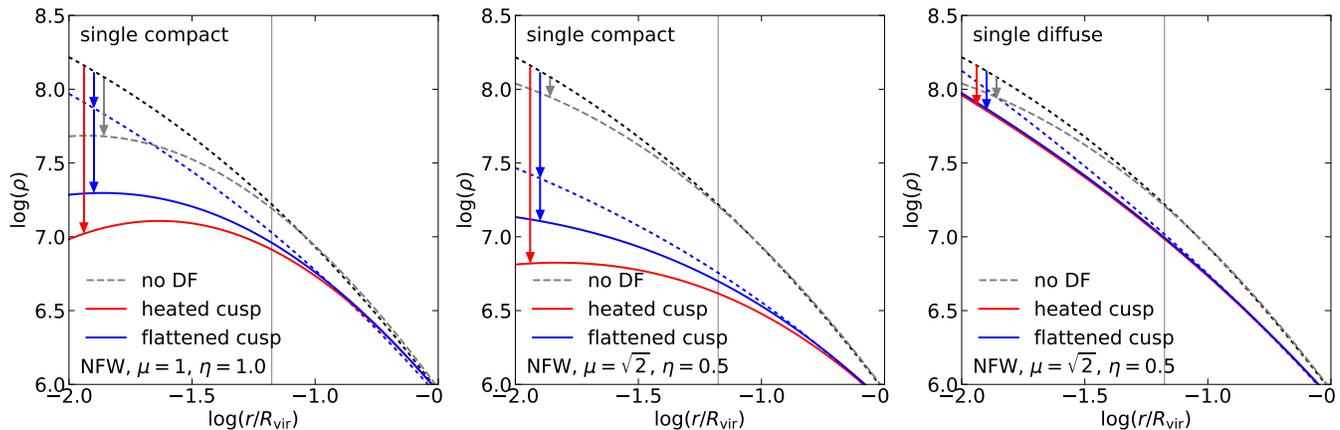
Figure 9. Dark-matter density-profile response to an outflow according to the CuspCore model. The initial halo is NFW with $M_{\rm v} = 10^{12.5} M_\odot$ and $c_{\rm NFW} = 5$ with $R_{\rm v} = 150\,{\rm kpc}$ at $z = 2$. The DF heating is by a single fiducial satellite of initial mass $m_{\rm v}$, compact or diffuse, based on the energy-deposit profile as revealed by SatGen, using the fit shown in the bottom-left panel of Fig. 4 and Table 1. The instantaneous outflow involves a fraction η of the central gas mass of $10^{10.6} M_\odot$. **Left:** a compact satellite $m_{\rm v} = 0.1 M_{\rm v}$ and $\eta = 1$. **Middle:** a compact satellite and $m_{\rm v} = 0.1\sqrt{2} M_{\rm v}$ (obtained by doubling the energy deposited by DF according to Fig. 4) and $\eta = 0.5$. **Right:** the same but for a diffuse satellite. Three cases are shown in each panel as follows. **Grey:** the dotted line is the pre-DF initial cuspy DM profile, cold NFW in Jeans equilibrium. It is assumed to be the pre-outflow profile, with the dashed grey line referring to the corresponding post-outflow profile. **Red:** the pre-outflow density profile is the same grey dotted line, but in this case the halo has been heated by dynamical friction, and the solid red line is the corresponding post-outflow profile, showing a flat core extending to the core radius of $\sim 10\,{\rm kpc}$ (vertical line). **Blue:** the dotted line is the somewhat flatter density profile late after the heating by dynamical friction, after relaxation involving expansion and cooling. With this cold pre-outflow density profile, the solid blue line refers to the post-outflow profile, showing a core but not as flat as in the pre-heated case. Results for a steep-cusp halo are shown in Fig. G4 (available as supplementary material online). We learn that the diffuse satellite causes a much weaker effect than the compact satellite, as predicted by the toy model. We also learn that a heated cusp is more responsive to an outflow than a flattened, cold cusp, though both can end up as extended cores under pre-heating by the compact satellite.

icked by multiplying the deposited energy from Fig. 4 by μ^2, following eq. (17). The mass removed by the outflow is assumed to be a fraction η of the $10^{10.6} M_\odot$ gas (a fraction 0.5η of the baryonic mass), and the halo expands differentially while conserving energy into its final Jeans equilibrium according to the CuspCore model. The two cases shown, both ending up with an extended core, are for $\mu = 1$ with $\eta = 1$ and for $\mu = \sqrt{2}$ with $\eta = 0.5$.

Three cases are shown in each panel of Fig. 9, as follows. The first (black) is a reference case of an outflow ignoring any pre-heating by dynamical friction, starting from the initial cold NFW halo (short-dashed black). One can see that the post-outflow cusp (long-dashed grey) is somewhat flatter than the original cusp, but even with $\eta = 1$ it is not as extended and as flat as deduced from observations.

In the second case in each panel (red), the pre-outflow configuration is set to mimic the hot phase soon after the energy deposit by dynamical friction, namely the density profile is the same initial NFW profile while all the input energy is deposited as kinetic energy. We see that in this case the cusp heated by the compact satellite generates an extended flat core, as desired (solid red). On the other hand, the cusp heated by the diffuse satellite shows only a minor flattening. The unrealistic positive slope at small radii is an artifact of the CuspCore model, and should be interpreted as a flat inner core.

In the third case in each panel (blue), the pre-outflow configuration is set to mimic the cooled, slightly

flattened phase after relaxation, which may in principle be reached a while ($\sim 0.5 t_{\rm v}$) after the heating by dynamical friction (short dashed blue). Now, in the case of a compact satellite, the outflow generates a flattened profile (solid blue), but not as flat and extended as in the case of a hot pre-outflow cusp. We learn that the response of a pre-heated inner DM halo to central mass ejection is indeed much stronger than that of a standard cold NFW halo, and stronger than that of a cooled flattened cusp that may in principle be reached late after the heating by dynamical friction.

6.3 A cosmological sequence of satellites

Figure 10 shows the same as Fig. 9 but now starting with the DF energy deposit from the SatGen simulation of a cosmological sequence of satellites, the fiducial compact and diffuse, based on Fig. 5, which refers to a duration of $\tau t_{\rm v}$ with $\tau = 1$. The functional fits used here and shown in the figure are listed in Table 1. Figure 10 assumes $\tau = 0.5$, namely half the energy deposited by DF at each radius in Fig. 5, to match the duration of $\sim 0.5 t_{\rm v}$ for a heated cusp, as discussed in §5, based on the estimates from the N-body simulation described in §5.2, Fig. 7. The positive slope at small radii is an artifact of the CuspCore model, and should be interpreted as a flat inner core.

We learn that the overall effect of an outflow after heating by a sequence of satellites over half a virial time is comparable to, and slightly stronger than, the
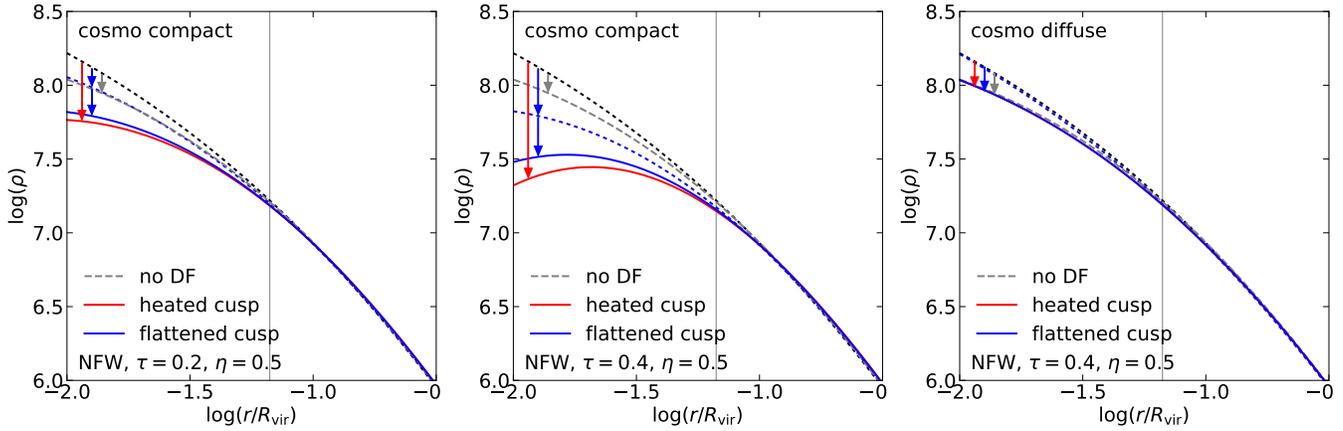
Figure 10. Similar to Fig. 9, with the same NFW initial host halo, but for DF energy deposit by a cosmological sequence of satellites. The energy deposited during $\tau t_{\rm v}$ is taken to be τ times the fit to the median from the SatGen runs for $t_{\rm v}$ shown in Fig. 5 and Table 1. **Left:** compact satellites with $\tau = 0.2$. **Middle:** compact satellites with $\tau = 0.4$, mimicking the typical duration of the hot phase discussed in §5. **Right:** the same but for diffuse satellites. The outflow is with a fraction $\eta = 0.5$ of the gas. We learn that the effect of a typical sequence of satellites during $\lesssim 0.5 t_{\rm v}$ is comparable to the effect of a single $m_{\rm v} = 0.14 M_{\rm v}$ satellite. This implies that the case of a pre-heated cusp (red) is realistic, generating a flat core extending to $\sim 10\,{\rm kpc}$ once the halo mass is above the golden mass, such that the satellites are above the mass-threshold for compaction, and the halo allows effective AGN feedback. The diffuse satellites have a negligible effect, as predicted. The corresponding results for a steep-cusp halo are shown in Fig. G5 (available as supplementary material online).

effect of an outflow after heating by a single satellite of $m_{\rm v}/M_{\rm v} \sim 0.14$ based on Fig. 9. This implies that the case of a pre-heated cusp (red) is realistic for the median galaxy, generating a flat core extending to $\sim 10\,{\rm kpc}$ once the halo mass is above the golden mass, such that the satellites are above the mass-threshold for compaction, and where the halo allows effective AGN feedback.

6.4   Other cases

Figures G4 and G5 (available as supplementary material online) are analogous of Figs. 9 and 10 but for an initial host halo with a cusp steeper than NFW, a DZ profile with $(2h, s_{1h}) = (5, 1.5)$, the same as studied in Figs. G1, G2, and G3. We learn that core formation in response to AGN-driven outflow is more challenging in the steep-cusp halo than in the NFW halo, both in terms of less heating by dynamical friction and a weaker response to an outflow. In order to demonstrate the formation of an extended core by compact satellites in this case, one needs a larger energy deposit by dynamical friction and/or a stronger outflow. For a single satellite, Figs. G4, we model a satellite of $m_{\rm v} = 0.1 \mu M_{\rm v}$ with $\mu = \sqrt{3}$ and an outflow that involves all the gas mass of $10^{10.6} M_\odot$, $\eta = 1$. In the case of DF heating by a cosmological sequence of compact satellites, Fig. G5, we learn that an outflow of $\eta = 2$, namely involving at least two successive episodes of outflow, is sufficient for producing a moderate core without DF heating. A flatter core is produced with $\tau = 1$ and $\eta = 2$ if the slope of $K(<r)$ within the cusp is artificially flattened in order to make a better fit to the DZ profile in CuspCore, and thus allow convergence with $\tau = 1$. The high τ can be interpreted as representing the top 1/3 of the random realizations of the satellites in the sequence (in terms of mass and circularity) during $\sim 0.5 t_{\rm v}$. Alternatively it

can be interpreted as the median during a longer period of $\sim t_{\rm v}$. Thus, we learn that an extended core can be reproduced even in an initially steep cusp, at least in a fraction of the galaxies.

We explored the general dependence of the formation of a core on various parameters of the model. This was done using the fiducial case of an NFW host of $M_{\rm v} = 10^{12.5} M_\odot$ and a single satellite of $m_{\rm v} = 0.1 M_{\rm v}$ with the fiducial compact or diffuse DM profile on a typical orbit, and with an outflow of $\eta = 0.5$ of the gas. The general conclusions are as follows. (1) While a compact satellite with an outflow of $\eta = 0.5$ is needed for an extended core, even a diffuse satellite can lead to a core with a stronger outflow of $\eta \sim 1$, but it is at best a small core of a few kiloparsecs. (2) While a gas mass of $\sim 10^{10.6} M_\odot$ leads to a core by an $\eta = 0.5$ outflow even if it occurs after post-heating relaxation, §5, a gas mass smaller than $\sim 10^{10} M_\odot$ may be enough for generating a core provided that the outflow occurs during the hot phase of the DM. It should be noted that the direction of the dependence on the baryonic mass is not obvious a priori, because a lower baryonic mass is associated with both a shallower potential well and a weaker outflow, which work to strengthen and weaken the formation of a core, respectively. (3) Spreading the baryons from a central point mass to an extended configuration, out to $0.2 R_{\rm v}$, makes only a small difference to the core formation. This is due to the competing effects mentioned above, where a more extended baryon distribution is associated with both, a shallower potential and a weaker effect of the outflow.

We also tested multiple episodes of outflow, each with $\eta = 0.2$, following one event of DF heating and no further supply of central baryons during the sequence of outflows. The total fractional outflown mass in $N = 3$ such episodes is comparable to the outflow mass in a

single episode of $\eta = 0.5$. We learn that under these conditions, the successive three weaker episodes and the stronger single episode lead to a similar core. This is line with the toy model of shell response to an outflow (Dutton et al. 2016, Section 4), which predicts that the expansion of the shell radius is approximately by a factor $1 + Nf$, where $f \ll 1$ is the fraction of the total mass ejected in each episode, such that Nf is the same for all N. However, if each outflow is preceded by an inflow of the same gas mass, the net expansion factor (after certain contraction) becomes $1 + Nf^2$. If the outflow events are preceded by a cosmological inflow of the same total gas mass $f_{\rm inf}$, then $f = f_{\rm inf}/N$, such that the single outflow is expected to be more effective than a sequence. If, on the other hand, each outflow is preceded by an inflow of recycled gas, then f is the same for all episodes, and a sequence of weak outflows is expected to be more effective than a single outflow.

7 DISCUSSION

7.1 Cores at large masses and high redshifts

We showed that the observed preferred appearance of cores in haloes of $M_{\rm v} > 10^{12} M_\odot$ is a natural outcome of the characteristic threshold for compaction events at $m_{\rm v} > 10^{11.3} M_\odot$, as seen in simulations (Zolotov et al. 2015; Tomassetti et al. 2016; Tacchella et al. 2016a) and observations (Huertas-Company et al. 2018). Once the host halo is above $10^{12} M_\odot$, satellites above $10^{11.3} M_\odot$ constitute most of the accretion into the host halo, and these are the post-compaction compact satellites that allow 20% or more of their mass to penetrate intact into the host cusp and significantly heat up the cusp by dynamical friction. The transition in the satellite compactness as a function of mass translates to a threshold for haloes that can generate extended DM cores.

The preference of cores at redshifts above $z \sim 1$ arises from several different sources. First, as estimated by our toy model, the relative energy deposited by dynamical friction in the cusp in a Gyr is $\propto (1 + z)^{5/2}$, eq. (20), and in a halo crossing time it is $\propto (1 + z)$, eq. (21). With a typical bound fraction $f \sim 0.2$ surviving stripping for a post-compaction satellite, the deposited energy is indeed comparable to the cusp energy near $z \sim 2$, where the majority of the observed haloes show cores. Second, the compaction events, which tend to occur near the characteristic mass, tend to be deeper at higher redshifts due to the higher gas fraction which allows more dissipation and angular-momentum loss and therefore more compact satellites. Third, the satellite orbits tend to be more radial at higher redshifts, where a given halo mass represents a higher-sigma peak. According to Wetzel (2011), the mean orbit circularity is $\epsilon \sim 0.6, 0.5, 0.4$ at $z \sim 0, 2, 4$ respectively. The more radial orbits at higher redshift give rise to less tidal stripping, deeper satellite penetration and larger energy deposit by DF in the host cusp. Further discussion of the redshift and mass dependence of the orbit circularity is in Appendix §I1 (available as supplementary material online).

The second-stage AGN-driven core formation gives rise to a similar mass threshold for cores. This is because the central black-hole growth is suppressed by supernova feedback when the halo is less massive than the golden mass $\sim 10^{12} M_\odot$ (Dubois et al. 2015; Bower et al. 2017; Anglés-Alcázar et al. 2017b), and it turns into a rapid growth triggered by a compaction event (Dekel, Lapiner & Dubois 2019; Lapiner, Dekel & Dubois 2020) in the hot-CGM phase once above the golden mass, consistent with observations (Kocevski et al. 2017; Förster Schreiber et al. 2019). In turn, the AGN-driven winds also contribute to the preference of cores at high redshifts, because the winds tend to be more effective at high redshifts due to the higher gas fraction, providing more gas for activating the AGN and as a substance to be ejected.

7.2 Gas availability for AGN-driven outflows

For a halo of mass $M_{\rm v} = 10^{12.5} M_\odot$, we assume a galaxy gas mass of $M_{\rm g} \simeq 4 \times 10^{10} M_\odot$, following the average stellar-to-halo mass ratio for such a halo mass at $z = 2$ (Behroozi et al. 2019) and assuming a gas fraction of 0.5 (Tacconi et al. 2018). The outflowing mass in each ejection episode is assumed to be a fraction η of the gas mass in the galaxy, between zero and unity, namely a fraction 0.5η of the central baryonic mass at $z \sim 2$. An upper limit on η is provided by the gas mass available for being ejected, which we estimate from cosmological accretion plus recycling during a period of, e.g., $0.5t_{\rm v}$, the typical time for the hot phase as estimated in §5.2. The typical cosmological specific accretion rate in the EdS regime ($z > 1$) can be approximated by eq. (18), $\dot{M}/M \simeq 0.47 \, {\rm Gyr}^{-1} (1 + z)_3^{5/2}$, where the mass can refer successfully either to the total mass or to the baryonic mass (Dekel et al. 2013). Assuming a cosmological baryon fraction $f_{\rm b} = 0.17$, with an accreting gas fraction with respect to the total accreting baryons of $f_{\rm g} = 0.5$ at $z = 2$, and assuming that the accretion rate of fresh and recycled gas is a factor $f_{\rm rec}$ times the fresh gas accretion rate, the total gas accretion rate into the galaxy inside a halo of $M_{\rm v} = 10^{12.5} M_\odot$ is

$$\dot{M}_{\rm g} \simeq f_{\rm rec} \, 1.3 \times 10^{11} M_\odot \, {\rm Gyr}^{-1} . \qquad (24)$$

This provides $M_{\rm g} \sim f_{\rm rec} \, 3.2 \times 10^{10} M_\odot$ of gas that is available for ejection during $0.5t_{\rm v} \simeq 0.25 \, {\rm Gyr}$. Given the comparable gas mass of $\sim 4 \times 10^{10} M_\odot$ in the galaxy, one can expect about $f_{\rm rec}$ outflow episodes of as much as $\eta \sim 1$ each during that period. Based on FIRE simulations, the recycling factor at $z \sim 2$ is estimated to be as high as $f_{\rm rec} \sim 3$ (Anglés-Alcázar et al. 2017a, Fig. 3). We therefore assume that from the point of view of gas availability for ejection, by modeling the response to one outflow episode of $\eta = 1$ we obtain a conservative lower limit to the effect of outflows during $0.5t_{\rm v}$. This is provided that the duty cycle of AGN outbursts is appropriate for the purpose, and that their energy/momentum and coupling to the available gas are sufficient for a removal of a significant fraction of the available gas.

7.3 Central stars, gas and dark matter

For the proposed scenario to produce DM cores with a low fraction of dark matter and with a stellar system that remains intact, several conditions have to be fulfilled. For example, (1) the accretion rate of dark matter to the cusp by satellites should not overcome the expansion rate induced by DF heating and/or AGN-driven outflows, (2) the stellar system should be more compact than the gas that is to be removed from the galaxy, or it has to be kinematically cooler than the dark matter, and (3) the dark matter that is to be pushed away from the cusp has to originally be at radii that are comparable to (or slightly larger than) those of the gas to be removed.

7.3.1 Dark-matter accretion

The galaxies with DM cores are observed to have a low fraction of dark matter with respect to the total core mass, $f_{DM} \sim 0-0.3$ within the effective radius R_e (Genzel et al. 2020, Fig. 10). The compact merging satellites that provide the DF heating bring into the cusp a mass that is comparable to the original cusp mass. For example, according to Fig. 4, for a typical single compact satellite of $m_v \sim 0.1 M_v$, the mass that enters the 10 kpc host cusp is $m_c \sim 0.4 m_v = 0.04 M_v$, indeed comparable to the mass inside the cusp. For our scenario of core formation to work, the dark matter that dominates this incoming mass has to be pushed away from the cusp, together with the dark matter that has already been in the cusp, either by the post-heating expansion discussed in §5, and/or by the response to AGN-driven outflows discussed in §6. In order to compute the timescale for doubling the cusp mass by incoming satellites, we recall that the specific accretion rate into the halo is $\dot{M}_v/M_v \simeq 0.5 \, \mathrm{Gyr}^{-1}(1 + z)_3^{5/2}$ (Dekel et al. 2013). If about 60% of this input mass is in compact satellites of $m_v > 0.1 M_v$ (Neistein & Dekel 2008), and if the satellite mass that penetrates to the cusp is $m_c \sim 0.4 m_v$ while the cusp mass in an NFW host is $M_c \sim 0.04 M_v$, we obtain a specific DM accretion rate into the cusp of $\dot{m}_c/M_c \sim 3 \, \mathrm{Gyr}^{-1}(1+z)_3^{5/2}$, namely a DM cusp doubling time of

$$t_c \sim 0.3 \, \mathrm{Gyr} \, (1 + z)_3^{-5/2} \sim 0.6(1 + z)_3^{-1} t_v \, . \qquad (25)$$

According to Fig. 4, this DM mass is mostly deposited in the outer cusp, at $3-10$ kpc. This timescale turns out to be comparable to the timescale for post-heating relaxation (§5) and to the timescale for fresh gas supply for an AGN-driven outflow that generates a core (§7.2). This implies that the DM accretion rate into the cusp is slow enough for the dark matter to be maintained hot by dynamical friction, and that the gas accretion rate (especially when adding recycling) is more than needed for producing AGN-driven outflows that push away the additional dark matter, once the AGN duty cycle, energy/momentum and coupling to the gas are appropriate.

Figure H1 (available as supplementary material online) show the CuspCore results with the DM satellite mass that has been added to the host according to Sat-Gen, Fig. 4 and Fig. 5, prior to the outflow event. Shown are cases with the same parameters as in Fig. 9 and Fig. 10, confirming core formation in the presence of the added mass. The core is slightly less flat than without the additional mass for a single satellite, but it is as flat for a sequence of satellites, where the total added mass is only 0.2 of the cusp mass. This justifies ignoring the added mass in our main analysis.

7.3.2 The stellar system

What may lead to the initial compactness and the associated coldness of the stellar systems? The cored galaxies tend to have massive compact stellar bulges, with masses $10^{10.5} - 10^{11} M_\odot$ (Genzel et al. 2020, Fig. 11). These bulges have to be compact enough such that they are not significantly heated by the merging satellites, and especially such that their compactness compared to the gas and the (possibly heated) dark matter, and their self-gravity, allow them to survive intact while the more extended dark matter expands in response to the AGN-driven outflows. These bulges are likely to have partly formed by earlier dissipative compaction events that occurred when the host halo crossed the threshold mass, generating a stellar nugget, blue that turned red, inside ~ 1 kpc (Zolotov et al. 2015; Tacchella et al. 2016a).

Furthermore, the stars that come in with the merging post-compaction satellites are expected to be more compact than the gas and DM at the centers of these satellites, and thus to penetrate by DF deep into the host halo cusp with little stripping. The stellar-to-halo mass ratio at $z \sim 2$, for our dominant post-compaction satellite mass of $m_v \sim 10^{11.5} M_\odot$, is expected to be $m_s/m_v \sim 0.004$ (Behroozi et al. 2019). According to Fig. 4, the bound mass of a typical satellite of $m_v = 0.1 M_v$ that enters the inner 1 kpc of the host is $m_1 \sim 0.09 m_v = 0.009 M_v$. We learn that if the satellite stars are indeed confined to and dominate the central regions of the satellites, they are all expected to penetrate intact into the inner 1 kpc of the host halo, while the mass that is stripped within the outer cusp is primarily dark matter.

The compact stellar nugget is not expected to heat up due to dynamical friction. The gravitational potential of the stars in the host central ~ 1 kpc nugget is deeper by a factor of a few than that of the dark matter at several kiloparsecs. According to Fig. 4, the energy deposited by dynamical friction inside the inner 1 kpc of the host is two order of magnitude less than at a a radius of a few kiloparsecs. These indicate that the heating effect of dynamical friction on the inner nugget is negligible.

7.3.3 The gas distribution

In order for the stellar nugget to remain intact while the dark matter is expanding in response to AGN-driven outflows, the gas to be expelled has to lie mostly outside the compact stellar system. We learn from the VELA

hydro cosmological simulations that in post-compaction galaxies, even without AGN feedback, the gas half-mass radius is typically ~ 3 times the stellar half-mass radius (Tacchella et al. 2016a; Kretschmer et al. 2021)[8] With $R_{\rm e,stars} \sim 1\,{\rm kpc}$, we estimate $R_{\rm e,gas} \sim 3\,{\rm kpc}$, which is outside most of the stellar system, but inside most of the dark-matter cusp of $R_{\rm e,dm} \sim 7\,{\rm kpc}$ that is to be pushed away, as required. The gas is more diffuse than the stars in these simulations because of the depletion by central star formation and partial effects of supernova feedback when near the golden mass, and because the gas might have been segregated from the stars already in the merging satellites due to similar effects as well as ram-pressure stripping. Considering the effect of earlier AGN-driven outflows may make the gas even more diffuse.

Clearly, the above discussion is based on crude estimates. A quantitative analysis of the stellar system, the gas distribution and the dark-matter distribution in the presence of baryons is deferred to a future study that will properly address the baryonic components of the satellites and the host halo. In particular, the CuspCore model will be generalized to deal with a three-component system.

7.4 Why do cosmological simulations fail?

We can see several reasons for the failure of DF to heat up the host cusps (and/or produce cores) in current cosmological simulations (e.g. Wang et al. 2020; Lazar et al. 2020, TNG and FIRE-2 respectively). First, the simulation resolution does not allow a full treatment of the compaction of satellites of $\sim 10^{11.3}\,M_\odot$. While the $\sim 10\,{\rm kpc}$ environment in the satellite is resolved, the inner $\sim 1\,{\rm kpc}$ is not properly resolved in most simulations (e.g. EAGLE, Illustris-TNG). The compaction on these scales is important as it suppresses tidal heating in the inner satellites and reduces the overall tidal stripping of the satellites, which allows better penetration and stronger DF heating of the host cusp.

Second, the compaction process in the satellites may be suppressed due to feedback that is too strong. This may be the case in some of the popular cosmological simulations (e.g. FIRE, Illustris-TNG).

Third, if the satellites are not properly resolved by more than 10^6 particles and an optimal scale-dependent force softening length, there is an artificial tidal disruption of the satellites (van den Bosch & Ogiya 2018; Errani & Navarro 2020). This mostly affects satellites in the inner host halo, where the satellite abundance could be artificially suppressed by a factor of two or more (Green, van den Bosch & Jiang 2021), thus suppressing the DF heating of the host cusp.

The fact that the AGN environment is unresolved may suppress its efficiency in driving winds that could

participate in generating DM cores. First, the simulations do not resolve any clumpiness in the accretion onto the black hole. Such clumpiness should boost the black-hole growth rate and AGN activity, and it can be incorporated as a subgrid model (DeGraf et al. 2017). Second, the simulations may fail to resolve the coupling of the AGN with the ISM around the black hole. For example, if the ISM is clumpy, the push of the dense clumps may boost the generated wind.

The key test for the proposed scenario for core formation in massive galaxies should eventually be cosmological simulations that reproduce such cores. A maximum resolution of $\sim 20\,{\rm pc}$ will enable the formation of compact blue nuggets as merging satellites, and $\sim 10^7\,M_\odot$ particles in their haloes will avoid artificial tidal disruption. Improvements of the black-hole and AGN subgrid recipes could help generating the desired bursty duty cycle. First, by introducing subgrid clumpiness in the accretion onto the black holes (e.g. following DeGraf et al. 2017), and Second, by limiting the AGN-feedback push to cold clumps in the ISM/CGM. In addition, positive AGN feedback (Silk 2013) may trigger SN feedback to assist the driving of the wind.

7.5 Caveats

In the current absence of proper simulations, the analysis performed above is approximate, meant to provide a feasibility test for the proposed hybrid scenario of massive core formation at high redshift, so its quantitative predictions should be taken with a grain of salt. It is yet to be verified with proper cosmological simulations that resolve the satellite compactness and the processes of tidal stripping and dynamical friction, and model realistic black-hole growth and AGN feedback. We mention below certain caveats of our analysis.

We derived fiducial satellite profiles from cosmological simulations (VELA), which indicate a compaction-driven transition from diffuse to compact when the galaxy is near a golden halo mass of $M_{\rm v} \sim 10^{11.3}\,M_\odot$. While this transition is rather robust for most galaxies, in simulations and observations, we adopted in our simplified analysis representative fiducial profiles below and above the golden mass, and ignored the large scatter in these profiles and in the threshold mass. This scatter may weaken the systematic mass dependence of core formation.

The analytic toy model for DF heating is clearly very approximate, meant to motivate our qualitative expectations. The SatGen semi-analytic simulations are a step forward in approximating the evolution of satellites, single and in a cosmological sequence, under tidal stripping and dynamical friction. It has been successfully calibrated against cosmological simulations. However, the way we translate the satellite evolution to local heating of the dark matter in the host halo is very crude, as dynamical friction is not a local effect (Tremaine & Weinberg 1984).

The CuspCore modeling of the dark-matter response to an outflow episode is only an approximation. In particular, the assumptions of a pre-outflow hot halo

---

[8] For example, we read from Table A1 of Kretschmer et al. (2021) typical values of $R_{\rm d}/R_{\rm e,stars} \sim 5$, and obtain the desired ratio of effective radii using $R_{\rm e,gas}/R_{\rm d} \sim 0.5$.

as the outcome of earlier DF heating, and of an instantaneous mass loss from a central point mass, may be oversimplified. Furthermore, the implementation of energy conservation in the subsequent relaxation process is not formally justified in the case of shell crossing. Here, the proof is in the pudding, in the sense that the model has been demonstrated to approximate the response of dark matter to outflow episodes in hydro cosmological simulations (NIHAO), except under merger situations. Finally, the CuspCore model as implemented here treats a single instantaneous mass loss, while realistic AGN feedback may involve a duty cycle of mass ejection episodes with mutual cross-talk.

In most of the above analysis, the focus was on the dark matter, ignoring certain effects that may involve the baryons. In particular, for the scenario to work, the baryons in the satellites should penetrate into the halo center such that the gas is available for ejection by AGN feedback while the stars, incoming or forming in-situ, remain intact at the center as a visible galaxy. Clearly, a more sophisticated analysis that better addresses the baryons is required.

8 CONCLUSION

This work has been inspired by the puzzling observational indication for a dearth of dark matter and flat-density cores extending to $\sim 10\,\mathrm{kpc}$ in about one third of the DM haloes of mass $M_\mathrm{v} \geqslant 10^{12} M_\odot$ at $z \sim 2$ (Genzel et al. 2020). This is puzzling because such cores, which are not detected at low redshifts, are not reproduced by any current cosmological simulation, and because supernova feedback, the common process assumed responsible for core formation in low-mass galaxies, is not expected to be energetic enough for producing cores in massive galaxies (Dekel & Silk 1986). Despite potential uncertainties in the non-trivial interpretation of the observations, we take it as a non-trivial theoretical challenge.

We study a hybrid scenario where two processes combine to generate such massive cores at high redshifts. These are "heating" of the inner host-halo cusp by dynamical friction acting on *compact* merging satellites, followed by AGN-driven outflows that flatten the hot cusp into a core. Each of these processes by itself seems to be incapable of forming extended-enough cores (El-Zant, Shlosman & Hoffman 2001; El-Zant et al. 2004; El-Zant 2008; Martizzi et al. 2012; Martizzi, Teyssier & Moore 2013). For the scenario to work, they should act in concert, and for reproducing the desired combined effects in simulations, each should be treated properly.

The key for efficient dynamical-friction heating is the compactness of the incoming satellites. This turns out to be a natural outcome of the wet compaction process that galaxies typically undergo when their haloes are near or above a golden mass of $m_\mathrm{v} \sim 10^{11.3} M_\odot$ (Zolotov et al. 2015; Barro et al. 2017; Huertas-Company et al. 2018). This is indeed the mass range for the dominant merging satellites in host haloes of $M_\mathrm{v} \geqslant 10^{12} M_\odot$, in which the cores are observed.

We use analytic toy modeling and dedicated semi-analytic simulations (SatGen, Jiang et al. 2020) to explore the energy deposit by dynamical friction as a function of satellite compactness, for a single satellite on a typical orbit and for a cosmological sequence of satellites during a period comparable to the halo virial crossing time at $z \sim 2$.

Zoom-in hydro cosmological simulations (VELA) allow us to determine fiducial density profiles for the post-compaction high-mass satellites in contrast to the pre-compaction low-mass satellites. The Dekel-Zhao functional form (Dekel et al. 2017; Freundlich et al. 2020b), with a flexible concentration and inner slope and analytic expressions for the potential and kinetic energies, best-fit these with the DZ parameter values $(c,\alpha) = (7,1)$ and $(3,0.5)$ for the compact and diffuse satellites. These correspond to the equivalent pairs of parameters $(c_2, s_1) = (15, 1.5)$ and $(3,1)$, respectively, where c_2 is the concentration where the density-profile slope is -2 and s_1 is the density slope at $0.01 R_\mathrm{v}$.

For a massive host halo of an initial NFW cusp, we find using the toy model and SatGen simulations that the fiducial compact satellites, relevant for massive haloes, significantly heat up the host cusp. They typically penetrate into the cusp with one to a few tens of percents of their original mass, and by dynamical friction they deposit there an energy comparable to the initial kinetic energy in the cusp during one half of the virial time at $z \sim 2$, about $0.25\,\mathrm{Gyr}$. On the other hand, the diffuse satellites, relevant for less massive haloes, penetrate with only a small fraction of their original mass, and deposit energy that is a similarly small fraction of the cusp energy, causing a negligible effect.

We find, using toy models and N-body simulations, that the "hot" phase of the cusp, driven by dynamical friction acting on a single compact satellite, is temporary, lasting for about half a virial crossing time. Later, the cusp would have relaxed to a new Jeans equilibrium, by expanding and cooling into a cold and somewhat flatter cusp. However, at $z \sim 2$ and for $M_\mathrm{v} \geqslant 10^{12} M_\odot$, given the energy deposited by dynamical friction in half a virial time, the hot phase is expected to be maintained by the cumulative heating from a cosmological sequence of merging satellites.

In the second stage, using an analytic model for outflow-driven core formation (CuspCore, Freundlich et al. 2020a), we demonstrate that the heated or flattened cusps respond strongly to removal of about half the central gas mass. This is a small fraction of the gas available, given the accretion and recycling in the available time of half a virial time at $z \sim 2$. A pre-heated cusp develops into a flat core extending to $\sim 10\,\mathrm{kpc}$, while a cooled and somewhat flattened cusp generates a partly flattened core. AGN feedback is indeed expected to become effective in haloes of $M_\mathrm{v} \geqslant 10^{12} M_\odot$, where the black-hole growth is no longer suppressed by supernova feedback and its rapid growth, triggered by a wet compaction event, is maintained by a hot CGM (Bower et al. 2017; Dekel, Lapiner & Dubois 2019; Lapiner, Dekel & Dubois 2020).

Repeating the analysis for a host halo with a steeper

initial cusp of slope $s_1 = 1.5$, we find that core formation is more demanding, but plausible for about one third of the galaxies. For dynamical friction to sufficiently heat up the cusp in half a virial time, the actual sequence of accreting satellites should include larger masses and/or more radial orbits than typical, in the top third of the distribution. For the AGN feedback to then generate a core, the gas mass removed should be comparable to the instantaneous gas mass in the galaxy, which is still only a fraction of the gas supplied by accretion and recycling during half a virial time.

The preference for cores in haloes of mass above the golden mass of $M_v \sim 10^{12} M_\odot$ is primarily due to the threshold mass for compaction events that (a) generate sufficiently compact satellites for significant DF cusp heating, and (b) trigger strong AGN-driven outflows in the host galaxies (Dekel, Lapiner & Dubois 2019; Lapiner, Dekel & Dubois 2020). The preference for cores at $z > 1$ arises from the more frequent mergers that drive the DF heating, and the stronger compaction events due to the higher gas fraction. The preferred mass and redshift are related to each other through the fact that the typical haloes at the Press-Schechter nonlinear mass scale are $\sim 10^{12} M_\odot$ at $z \sim 2$.

We estimate that the accretion rate of dark matter into the cusp via the merging satellites is slow compared to the hot-cusp relaxation rate and the rate of gas supply for AGN-driven outflows. This allows the formation of a core with a low central DM fraction, as observed. We anticipate that the central stellar nugget is likely to be more compact than the gas distribution such that the stellar system will not be significantly heated by dynamical friction and will not be strongly affected by the AGN-driven gas removal that generates the DM core. These should be verified via simulations including baryons.

For cosmological simulations to reproduce the required DF heating, they should resolve the wet compaction of the satellites of $m_v > 10^{11.3} M_\odot$ on scales below $\sim 100\,\mathrm{pc}$, and have at least $\sim 10^7$ particles in each subhalo in order to avoid artificial tidal disruption. AGN feedback could be boosted by resolving clumpy black-hole accretion and clumpy response to outflows. Simulations of this kind are not beyond reach in the near future, and they should be performed for a more realistic test of the proposed scenario.

REFERENCES

Agertz O., Kravtsov A. V., Leitner S. N., Gnedin N. Y., 2013, ApJ, 770, 25

Anglés-Alcázar D., Faucher-Giguère C.-A., Kereš D., Hopkins P. F., Quataert E., Murray N., 2017a, MNRAS, 470, 4698

Anglés-Alcázar D., Faucher-Giguère C.-A., Quataert E., Hopkins P. F., Feldmann R., Torrey P., Wetzel A., Kereš D., 2017b, MNRAS, 472, L109

Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15

Barro G. et al., 2017, ApJ, 840, 47

Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143

Behroozi P. S., Wechsler R. H., Conroy C., 2013, ApJ, 762, L31

Benson A. J., 2017, MNRAS, 467, 3454

Birnboim Y., Dekel A., 2003, MNRAS, 345, 349

Bower R. G., Schaye J., Frenk C. S., Theuns T., Schaller M., Crain R. A., McAlpine S., 2017, MNRAS, 465, 32

Bryan G. L., Norman M. L., 1998, ApJ, 495, 80

Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001, MNRAS, 321, 559

Burkert A., 1995, ApJ, 447, L25

Burkert A., Genzel R., Bouché N., Cresci G., Khochfar S., Sommer-Larsen J., Sternberg A., et al.,, 2010, ApJ, 725, 2324

Ceverino D., Dekel A., Bournaud F., 2010, MNRAS, 404, 2151

Ceverino D., Dekel A., Mandelker N., Bournaud F., Burkert A., Genzel R., Primack J., 2012, MNRAS,

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

Ceverino D., Klypin A., 2009, ApJ, 695, 292

Ceverino D., Klypin A., Klimek E. S., Trujillo-Gomez S., Churchill C. W., Primack J., Dekel A., 2014, MNRAS, 442, 1545

Ceverino D., Primack J., Dekel A., 2015, MNRAS, 453, 408

Chandrasekhar S., 1943, ApJ, 97, 255

Danovich M., Dekel A., Hahn O., Ceverino D., Primack J., 2015, MNRAS, 449, 2087

de Blok W. J. G., McGaugh S. S., Bosma A., Rubin V. C., 2001, ApJ, 552, L23

de Blok W. J. G., Walter F., Brinks E., Trachternach C., Oh S. H., Kennicutt, R. C. J., 2008, AJ, 136, 2648

DeGraf C., Dekel A., Gabor J., Bournaud F., 2017, MNRAS, 466, 1462

Dekel A., Birnboim Y., 2006, MNRAS, 368, 2

Dekel A., Devor J., Hetzroni G., 2003, MNRAS, 341, 326

Dekel A., Ginzburg O., Jiang F., Freundlich J., Lapiner S., Ceverino D., Primack J., 2020a, arXiv e-prints

Dekel A., Ishai G., Dutton A. A., Maccio A. V., 2017, MNRAS, 468, 1005

Dekel A., Krumholz M. R., 2013, MNRAS, 432, 455

Dekel A., Lapiner S., Dubois Y., 2019, arXiv e-prints

Dekel A., Lapiner S., Ginzburg O., Jiang F., Ceverino D., Primack J., 2020b, arXiv e-prints

Dekel A., Sarkar K. C., Jiang F., Bournaud F., Krumholz M. R., Ceverino D., Primack J. R., 2019, arXiv e-prints

Dekel A., Silk J., 1986, ApJ, 303, 39

Dekel A., Zolotov A., Tweed D., Cacciato M., Ceverino D., Primack J. R., 2013, MNRAS, 435, 999

Dubois Y., Volonteri M., Silk J., Devriendt J., Slyz A., Teyssier R., 2015, MNRAS, 452, 1502

Dutton A. A. et al., 2016, MNRAS, 461, 2658

Eddington A. S., 1916, Monthly Notices of the Royal Astronomical Society, 76, 572

El-Zant A., Shlosman I., Hoffman Y., 2001, ApJ, 560, 636

El-Zant A. A., 2008, ApJ, 681, 1058

El-Zant A. A., Hoffman Y., Primack J., Combes F., Shlosman I., 2004, ApJ, 607, L75

Errani R., Navarro J. F., 2020, arXiv e-prints, arXiv:2011.07077

Errani R., Peñarrubia J., Walker M. G., 2018, MNRAS, 481, 5073

Ferland G. J., Korista K. T., Verner D. A., Ferguson J. W., Kingdon J. B., Verner E. M., 1998, PASP, 110, 761

Flores R. A., Primack J. R., 1994, ApJ, 427, L1

Förster Schreiber N. M. et al., 2019, ApJ, 875, 21

Freundlich J., Dekel A., Jiang F., 2019, in SF2A-2019: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, Di Matteo P., Creevey O., Crida A., Kordopatis G., Malzac J., Marquette J. B., N'Diaye M., Venot O., eds., p. Di

Freundlich J., Dekel A., Jiang F., Ishai G., Cornuault N., Lapiner S., Dutton A. A., Macciò A. V., 2020a, MNRAS, 491, 4523

Freundlich J. et al., 2020b, MNRAS, 499, 2912

Genzel R. et al., 2020, arXiv e-prints, arXiv:2006.03046

Green S. B., van den Bosch F. C., Jiang F., 2021, Monthly Notices of the Royal Astronomical Society, 503, 4075

Haardt F., Madau P., 1996, ApJ, 461, 20

Hopkins P. F., Quataert E., Murray N., 2012, MNRAS, 421, 3522

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

Inoue S., Dekel A., Mandelker N., Ceverino D., Bournaud F., Primack J., 2016, MNRAS, 456, 2052

Jiang F., Dekel A., Freundlich J., van den Bosch F. C., Green S. B., Hopkins P. F., Benson A., Du X., 2020, arXiv e-prints, arXiv:2005.05974

Jiang F. et al., 2019, MNRAS, 488, 4801

Jiang F., van den Bosch F. C., 2014, MNRAS, 440, 193

Kazantzidis S., Zentner A. R., Kravtsov A. V., 2006, Astrophys. J., 641, 647

King I., 1962, AJ, 67, 471

Kocevski D. D., Barro G., Faber S. M., Dekel A., Somerville R. S., Young J. A., Williams C. C., et al., 2017, ApJ, 846, 112

Komatsu E., Dunkley J., Nolta M. R., Bennett C. L., Gold B., Hinshaw G., Jarosik N., et al., 2009, ApJS, 180, 330

Kravtsov A. V., 2003, ApJ, 590, L1

Kravtsov A. V., Klypin A. A., Khokhlov A. M., 1997, ApJS, 111, 73

Kretschmer M., Dekel A., Freundlich J., Lapiner S., Ceverino D., Primack J., 2021, MNRAS, 503, 5238

Krumholz M. R., Thompson T. A., 2013, MNRAS, 434, 2329

Lacey C., Cole S., 1993, MNRAS, 262, 627

Lapiner S., Dekel A., Dubois Y., 2020, arXiv e-prints, arXiv:2012.09186

Lazar A. et al., 2020, MNRAS, 497, 2393

Macciò A. V., Crespi S., Blank M., Kang X., 2020, MNRAS, 495, L46

Mandelker N., Dekel A., Ceverino D., DeGraf C., Guo Y., Primack J., 2017, MNRAS, 464, 635

Mandelker N., Dekel A., Ceverino D., Tweed D., Moody C. E., Primack J., 2014, MNRAS, 443, 3675

Martizzi D., Teyssier R., Moore B., 2013, MNRAS, 432, 1947

Martizzi D., Teyssier R., Moore B., Wentz T., 2012, MNRAS, 422, 3081

Moody C. E., Guo Y., Mandelker N., Ceverino D., Mozena M., Koo D. C., Dekel A., Primack J., 2014, MNRAS, 444, 1389

Moster B. P., Naab T., White S. D. M., 2018, MNRAS, 477, 1822

Murray N., Quataert E., Thompson T. A., 2010, ApJ, 709, 191

Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493

Neistein E., Dekel A., 2008, MNRAS, 388, 1792

Ogiya G., van den Bosch F. C., Hahn O., Green S. B., Miller T. B., Burkert A., 2019, MNRAS, 485, 189

Oh S.-H., Brook C., Governato F., Brinks E., Mayer L., de Blok W. J. G., Brooks A., Walter F., 2011a, AJ, 142, 24

Oh S.-H., de Blok W. J. G., Brinks E., Walter F., Kennicutt, Robert C. J., 2011b, AJ, 141, 193

Oh S.-H. et al., 2015, AJ, 149, 180

Parkinson H., Cole S., Helly J., 2008, MNRAS, 383, 557

Peirani S. et al., 2017, MNRAS, 472, 2153

Peirani S. et al., 2019, MNRAS, 483, 4615

Penarrubia J., Benson A. J., Walker M. G., Gilmore G., McConnachie A. W., Mayer L., 2010, MNRAS, 406, 1290

Pontzen A., Governato F., 2012, MNRAS, 421, 3464

Roca-Fàbrega S. et al., 2018, arXiv e-prints

Rodríguez-Puebla A., Primack J. R., Avila-Reese V., Faber S. M., 2017, MNRAS, 470, 651

Sharma G., Salucci P., van de Ven G., 2021, arXiv e-prints, arXiv:2105.13684

Silk J., 2013, ApJ, 772, 112

Snyder G. F., Lotz J., Moody C., Peth M., Freeman P., Ceverino D., Primack J., Dekel A., 2015, MNRAS, 451, 4290

Springel V., 2005, MNRAS, 364, 1105

Springel V., Hernquist L., 2005, ApJ, 622, L9

Strawn C. et al., 2020, MNRAS

Tacchella S., Dekel A., Carollo C. M., Ceverino D., DeGraf C., Lapiner S., Mandelker N., Primack J. R., 2016a, MNRAS, 458, 242

Tacchella S., Dekel A., Carollo C. M., Ceverino D., DeGraf C., Lapiner S., Mandelker N., Primack Joel R., 2016b, MNRAS, 457, 2790

Tacconi L. J., Genzel R., Saintonge A., Combes F., García-Burillo S., Neri R., Bolatto A., et al., 2018, ApJ, 853, 179

Tomassetti M. et al., 2016, MNRAS, 458, 4477

Tremaine S., Weinberg M. D., 1984, MNRAS, 209, 729

van den Bosch F. C., 2017, MNRAS, 468, 885

van den Bosch F. C., Ogiya G., 2018, MNRAS, 475, 4066

van Kampen E., 2000, arXiv

Wang Y. et al., 2020, MNRAS, 491, 5188

Weinberger R. et al., 2018, MNRAS, 479, 4056

Wetzel A. R., 2011, MNRAS, 412, 49

Wuyts S. et al., 2016, ApJ, 831, 149

Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2009, ApJ, 707, 354

Zhao H., 1996, MNRAS, 278, 488

Zolotov A. et al., 2015, MNRAS, 450, 2327

APPENDIX A: THE DEKEL-ZHAO PROFILE

The Dekel-Zhao halo profile (Dekel et al. 2017; Freundlich et al. 2020b),[9] following a more general mathematical analysis by Zhao (1996), is a functional form for dark-matter haloes with two free shape parameters, a concentration c and an inner slope α, allowing the central region to range continuously from a steep cusp to a flat core. It has been found to fit dark-matter haloes in cosmological hydro simulations better than other two-parameter profiles such as the Einasto and the generalized-NFW profiles with a flexible inner slope. A unique feature is that it has analytic expressions not only for the density and mass-velocity profiles but also for the potential and kinetic energy profiles (as well as for gravitational lensing properties). Freundlich et al. (2020b) also provide the typical profile parameters as a function of mass.

The profile of *mean* density within a sphere of radius r is given by

$$\bar{\rho}(r) = \frac{\bar{\rho}_{\rm c}}{x^\alpha (1 + x^{1/2})^{2(3-\alpha)}}, \quad x = \frac{r}{R_{\rm v}} c, \quad \text{(A1)}$$

$$\bar{\rho}_{\rm c} = c^3 \mu(c, \alpha) \bar{\rho}_{\rm v}, \quad \bar{\rho}_{\rm v} = \frac{M_{\rm v}}{(4\pi/3)R_{\rm v}^3}, \quad \text{(A2)}$$

[9] Available for implementation in https://github.com/JonathanFreundlich/Dekel_profile .

$$\mu(c, \alpha) = c^{\alpha-3} (1 + c^{1/2})^{2(3-\alpha)}. \quad \text{(A3)}$$

For completeness, the local density profile is

$$\rho(r) = \frac{(1 - \alpha/3) \bar{\rho}_{\rm c}}{x^\alpha (1 + x^{1/2})^{2(3.5-\alpha)}}. \quad \text{(A4)}$$

The associated mass profile is

$$\frac{M(x)}{M_{\rm v}} = \frac{1}{c^3 \bar{\rho}_{\rm v}} x^3 \bar{\rho}(x) = \frac{\mu}{\bar{\rho}_{\rm c}} x^3 \bar{\rho}(x). \quad \text{(A5)}$$

The log slope of the mass profile is

$$\nu(r) = \frac{3 - \alpha}{1 + x^{1/2}}. \quad \text{(A6)}$$

The negative log slope of the density profile is

$$s(r) = \frac{\alpha + 3.5x^{1/2}}{1 + x^{1/2}}. \quad \text{(A7)}$$

In order to obtain the mass fraction $f = M(< r)/M_{\rm v}$ within a sphere of radius r, we express it using eq. (A5) as

$$f(r) = \mu \bar{\rho}_{\rm c}^{-1} x^3 \bar{\rho}(x). \quad \text{(A8)}$$

Combining eq. (A8) and eq. (A1), we get

$$f(r) = \mu x^{3-\alpha} (1 + x^{1/2})^{-2(3-\alpha)}, \quad \text{(A9)}$$

from which we obtain

$$x = [(f/\mu)^{-1/[2(3-\alpha)]} - 1]^{-2}. \quad \text{(A10)}$$

Inserting eq. (A10) in eq. (A8), we obtain an equation for $f(r)$,

$$(f/\mu) [(f/\mu)^{-1/[2(3-\alpha)]} - 1]^6 = \bar{\rho}(r_{\rm t})/\bar{\rho}_{\rm c}. \quad \text{(A11)}$$

We recall from equations 11-14 of Freundlich et al. (2020b) that a more physical pair of shape parameters, that refer to $\rho(r)$ rather to $\bar{\rho}(r)$, may be (c_2, s_1). The concentration c_2 refers to the virial radius with respect to the radius where the log density slope is -2,

$$c_2 = c \left(\frac{1.5}{2 - \alpha}\right)^2, \quad \text{(A12)}$$

valid for $\alpha < 2$. The inner slope s_1 is minus the log slope of the local density profile $\rho(r)$ at a given radius r_1 (say $r_1 = 0.01R_{\rm v}$),

$$s_1 = \frac{\alpha + 3.5x_1^{1/2}}{1 + x_1^{1/2}}, \quad x_1 = \frac{r_1}{R_{\rm v}} c. \quad \text{(A13)}$$

For completeness, the inverse relations are

$$c = \left(\frac{s_1 - 2}{(3.5 - s_1)(r_1/R_{\rm v})^{1/2} - 1.5c_2^{-1/2}}\right)^2, \quad \text{(A14)}$$

$$\alpha = \frac{1.5s_1 - 2(3.5 - s_1)x_{2,1}^{1/2}}{1.5 - (3.5 - s_1)x_{2,1}^{1/2}}, \quad x_{2,1} = \frac{r_1}{R_{\rm v}} c_2. \quad \text{(A15)}$$

We note that a valid DZ solution is not guaranteed for any arbitrary pair of values (c_2, s_1), e.g., there is no

solution where the denominator in either eq. (A14) or eq. (A15) vanishes.

A best fit for an NFW profile with a given concentration is obtained, e.g., by minimizing residuals in uniformly spaced log radii in the range $\log(r/R_{\rm v}) = (-2, 0)$. For $c_{\rm NFW} = 5$ we obtain for the best-fit DZ parameters $(c, \alpha) = (7.126, 0.2156)$ or $(c_2, s_1) = (5.035, 0.9076)$. A slightly better fit near $0.01 R_{\rm v}$ can be obtained with $(c_2, s_1) \simeq (4.8, 1.0)$, but this is at the expense of slightly larger deviations at large radii.

The gravitational potential as given in eq. 19 of Freundlich et al. (2020b) is

$$U(r) = -V_{\rm v}^2 \left[1 + 2c\mu \left(\frac{\chi_c^{\tilde{\alpha}} - \chi^{\tilde{\alpha}}}{\tilde{\alpha}} - \frac{\chi_c^{\tilde{\alpha}+1} - \chi^{\tilde{\alpha}+1}}{\tilde{\alpha}+1} \right) \right], \tag{A16}$$

$$\chi = \frac{x^{1/2}}{1+x^{1/2}}, \quad \chi_c = \frac{c^{1/2}}{1+c^{1/2}}, \quad \tilde{\alpha} = 2(2-\alpha). \tag{A17}$$

The velocity dispersion that stems from the Jeans equation (Freundlich et al. 2020b, eq. 22), providing the kinetic energy per unit mass, is

$$\sigma_r^2(r) = 2c\mu V_{\rm v}^2 \frac{\rho_c}{\rho(r)} \Big[\mathcal{B}(4 - 4\alpha, \, 9, \, \zeta) \Big]_\chi^{\chi_c}, \tag{A18}$$

where $\mathcal{B}(a, b, x) = \int_0^x t^{a-1}(1 - t)^{b-1}dt$ is the incomplete beta function, the brackets denote the difference of the enclosed function between 1 and χ, i.e., $[f(\zeta)]_\chi^{\chi_c} \equiv f(\chi_c) - f(\chi)$, and $\rho_c = (1 - \alpha/3)\bar{\rho}_c$. The definition of the incomplete beta function has been extended to negative parameters since the bracketed term is well-defined. Equations B1 and B3 of Freundlich et al. (2020b) are the equivalent expressions in terms of finite series.

Adding an additional point mass $M_{\rm b}$ at the halo center adds the velocity dispersion term (Freundlich et al. 2020b, eq. C15)

$$\sigma_{M_{\rm b}}^2(r) = 2c \frac{GM_{\rm b}}{R_{\rm v}} \frac{\rho_c}{\rho(r)} \Big[\mathcal{B}(-2 - 2\alpha, \, 9, \, \zeta) \Big]_\chi^1. \tag{A19}$$

For comparison, the NFW mass profile, with an inner cusp of negative log density slope $\alpha = 1$ and a free NFW concentration parameter $c_{\rm h}$, is

$$\frac{M(r)}{M_{\rm v}} = \frac{A(x)}{A(c_{\rm h})}, \quad \frac{\bar{\rho}(r)}{\bar{\rho}_{\rm v}} = \frac{A(x)}{A(c_{\rm h})} \frac{c_{\rm h}^3}{x^3}, \tag{A20}$$

$$A(x) = \ln(1+x) - \frac{x}{1+x}, \quad x = \frac{r}{R_{\rm v}} c_{\rm h}, \tag{A21}$$

The profile of the negative log slope of the mass profile is

$$\nu(r) = \frac{x^2}{(1+x)^2 A(x)}. \tag{A22}$$

APPENDIX B: THE VELA SIMULATIONS

The VELA suite consists of hydro-cosmological simulations zooming-in on 34 moderately massive galaxies,

presented in more detail in Ceverino et al. (2014) and Zolotov et al. (2015). This suite has been used to study central issues in the evolution of galaxies at high redshifts, including, e.g., compaction to blue nuggets and the trigger of quenching (Zolotov et al. 2015; Tacchella et al. 2016b,a), evolution of global shape (Ceverino, Primack & Dekel 2015; Tomassetti et al. 2016), violent disc instability (Mandelker et al. 2014, 2017; Inoue et al. 2016), the SFR-density relation by supernova feedback (Dekel et al. 2019), post-compaction formation of discs and rings (Dekel et al. 2020a,b), OVI in the CGM (Roca-Fàbrega et al. 2018; Strawn et al. 2020), and angular momentum and galaxy size (Jiang et al. 2019). Additional analysis of the same suite of simulations are discussed in Moody et al. (2014); Snyder et al. (2015). This appendix provides an overview of the relevant features of these simulations.

The VELA simulations make use of the Adaptive Refinement Tree (ART) code (Kravtsov, Klypin & Khokhlov 1997; Kravtsov 2003; Ceverino & Klypin 2009), which follows the evolution of a gravitating N-body system and the Eulerian gas dynamics using an adaptive mesh refinement. The maximum spatial resolution is $17 - 35\,$pc at all times. The code incorporates subgrid recipes for physical process that are relevant for galaxy formation, such as gas cooling by atomic hydrogen and helium, metal and molecular hydrogen cooling, photoionization heating by the UV background with partial self-shielding, star formation, stellar mass loss, metal enrichment of the ISM and stellar feedback. Supernovae and stellar winds are implemented by local injection of thermal energy as described in Ceverino & Klypin (2009); Ceverino, Dekel & Bournaud (2010) and Ceverino et al. (2012). Radiation-pressure stellar feedback is implemented at a moderate level, following Dekel et al. (2013), as described in Ceverino et al. (2014).

Cooling and heating rates are based on the CLOUDY code (Ferland et al. 1998). A uniform UV background based on the redshift-dependent Haardt & Madau (1996) model is assumed, except at gas densities higher than $0.1\,{\rm cm}^{-3}$, where partial self-shielding allows dense gas to cool down to ~ 300K. The assumed equation of state is that of an ideal mono-atomic gas. Artificial fragmentation on the cell size is prevented by introducing a pressure floor, which ensures that the Jeans scale is resolved by at least 7 cells (see Ceverino, Dekel & Bournaud 2010). Star particles form in timesteps of $5\,$Myr in cells where the gas density exceeds $1\,{\rm cm}^{-3}$ and the temperatures is below 10^4K. The code implements a stochastic star formation where a star particle with a mass of 42% of the gas mass forms with a probability $P = (\rho_{\rm g}/10^3\,{\rm cm}^{-3})^{1/2}$ but not higher than 0.2.

Thermal feedback that mimics the energy release from stellar winds and supernova explosions is incorporated as a constant heating rate over the 40 Myr following star formation. A velocity kick of $\sim 10\,{\rm km\,s}^{-1}$ is applied to 30 % of the newly formed stellar particles – this enables SN explosions in lower density regions where the cooling may not overcome the heating without implementing an artificial shutdown of cooling

(Ceverino & Klypin 2009). The code also incorporates the later effects of Type Ia supernova and stellar mass loss, and it follows the metal enrichment of the ISM. Radiation pressure is incorporated through the addition of a non-thermal pressure term to the total gas pressure in regions where ionizing photons from massive stars are produced and may be trapped. This ionizing radiation injects momentum in the cells neighbouring massive star particles younger than $5\,\mathrm{Myr}$, and whose column density exceeds $10^{21}\,\mathrm{cm}^{-2}$, isotropically pressurizing the star-forming regions (see more details in Agertz et al. 2013; Ceverino et al. 2014).

The initial conditions for the simulations are based on DM haloes that were drawn from dissipationless N-body simulations at lower resolution in cosmological boxes of $15 - 60\,\mathrm{Mpc}$. The ΛCDM cosmological model was assumed with the WMAP5 values of the cosmological parameters, $\Omega_{\mathrm{m}} = 0.27$, $\Omega_{\Lambda} = 0.73$, $\Omega_{\mathrm{b}} = 0.045$, $h = 0.7$ and $\sigma_8 = 0.82$ (Komatsu et al. 2009). Each halo was selected to have a given virial mass at $z = 1$ and no ongoing major merger at $z = 1$. This latter criterion eliminated less than 10 % of the haloes, those that tend to be in a dense, proto-cluster environment at $z \sim 1$. The virial masses at $z = 1$ were chosen to be in the range $M_{\mathrm{v}} = 2 \times 10^{11} - 2 \times 10^{12}\,M_{\odot}$, about a median of $4.6 \times 10^{11}\,M_{\odot}$. If left in isolation, the median mass at $z = 0$ was intended to be $\sim 10^{12}\,M_{\odot}$.

The VELA cosmological simulations are state-of-the-art in terms of high-resolution adaptive mesh refinement hydrodynamics and the treatment of key physical processes at the subgrid level. In particular, they trace the cosmological streams that feed galaxies at high redshift, including mergers and smooth flows, and they resolve the violent disc instability that governs high-z disc evolution and bulge formation (Ceverino, Dekel & Bournaud 2010; Ceverino et al. 2012, 2015; Mandelker et al. 2014). To mention a few limitations, like in other simulations, the treatments of star formation and feedback processes are rather simplified. The code may assume a realistic SFR efficiency per free fall time on the grid scale but it does not follow in detail the formation of molecules and the effect of metallicity on SFR. The feedback is treated in a crude way, where the resolution does not allow the capture of the Sedov-Taylor phase of supernova bubbles. The radiative stellar feedback assumed no infrared trapping, in the spirit of low trapping advocated by Dekel & Krumholz (2013) based on Krumholz & Thompson (2013), which makes the radiative feedback weaker than in other simulations that assume more significant trapping (Murray, Quataert & Thompson 2010; Hopkins, Quataert & Murray 2012). AGN feedback, and feedback associated with cosmic rays and magnetic fields, are not yet implemented. Nevertheless, as shown in Ceverino et al. (2014), the star-formation rates, gas fractions, and stellar-to-halo mass ratio are all in the ballpark of the estimates deduced from observations.

The virial and stellar properties of the galaxies are listed for example in Table 1 of Dekel et al. (2020b). The virial mass M_{v} is the total mass within a sphere of radius R_{v} that encompasses an overdensity of $\Delta(z) =$

$[18\pi^2 - 82\Omega_{\Lambda}(z) - 39\Omega_{\Lambda}(z)^2]/\Omega_{\mathrm{m}}(z)$, where $\Omega_{\Lambda}(z)$ and $\Omega_{\mathrm{m}}(z)$ are the cosmological parameters at z (Bryan & Norman 1998; Dekel & Birnboim 2006). The stellar mass M_{s} is the instantaneous mass in stars within a radius of $0.2R_{\mathrm{v}}$, accounting for past stellar mass loss. We start the analysis at the cosmological time corresponding to expansion factor $a = 0.125$ (redshift $z = 7$), and most galaxies reach $a = 0.50$ ($z = 1$). Each galaxy is analyzed at output times separated by a constant interval in a, $\Delta a = 0.01$, corresponding at $z = 2$ to ~ 100 Myr (roughly half an orbital time at the disc edge). The sample consists of totally ~ 1000 snapshots in the redshift range $z = 7 - 0.8$ from 35 galaxies that at $z = 2$ span the stellar mass range $(0.2 - 6.4) \times 10^{11}\,M_{\odot}$. The half-mass sizes R_{e} range $R_{\mathrm{e}} \simeq 0.4 - 3.2\,\mathrm{kpc}$ at $z = 2$. The determination of the centre of the galaxy is outlined in detail in appendix B of Mandelker et al. (2014). Briefly, starting form the most bound star, the centre is refined iteratively by calculating the centre of mass of stellar particles in spheres of decreasing radii down to $130\,\mathrm{pc}$ or when the number of stellar particles in the sphere drops below 20.

We identify the major event of wet compaction to a blue nugget for each galaxy. This is the one that leads to a significant central gas depletion and SFR quenching, and marks the transition from dark-matter to baryon dominance within R_{e}. Following Zolotov et al. (2015) and Tacchella et al. (2016a), the most physical way to identify the compaction and blue nugget is by the steep rise of gas density (and SFR) within the inner $1\,\mathrm{kpc}$ to the highest peak, as long as it is followed by a significant, long-term decline in central gas mass density (and SFR). The onset of compaction can be identified as the start of the steep rise of central gas density prior to the blue-nugget peak. An alternative identification is using the shoulder of the stellar mass density within $1\,\mathrm{kpc}$ where its rise due to the starburst associated with the compaction turns into a plateau of maximum long-term compactness slightly after the blue-nugget peak of gas density. This is a more practical way to identify blue nuggets in observations (e.g. Barro et al. 2017).

APPENDIX C: SATGEN - A SEMI-ANALYTIC SATELLITE GENERATOR

The semi-analytic model for satellite galaxies SatGen is presented in Jiang et al. (2020).[10] It can generate statistical samples of satellite populations for a host halo of desired mass, redshift, and cosmological parameters. The model combines DM halo merger trees, empirical relations for the galaxy-halo connection, and simple analytical prescriptions for tidal effects, dynamical friction, and ram pressure stripping (if the satellites contains gas). SatGen emulates cosmological zoom-in simulations in certain aspects. Satellites can reside in subhaloes of desired density profiles, with cores or cusps, depending on the subhalo response to baryonic physics

---

[10] Available for implementation in https://github.com/shergreen/SatGen .

that are formulated from hydro-simulations or physical modeling. The host potential can be composed of a DM halo and baryonic components, such as a disc and a bulge, each described by a density profile that allows analytic integration of the satellite orbits. The subhalo profile and the stellar mass and structure of a satellite evolves due to tidal heating and tidal mass loss, which depend on its initial structure. SatGen complements simulations by propagating the effect of halo response found in simulated central galaxies to satellites (which are typically not properly resolved in simulations). It outperforms simulations by capturing the halo-to-halo variance of satellite statistics and overcoming artificial disruption due to insufficient resolution (van den Bosch & Ogiya 2018; Green, van den Bosch & Jiang 2021). Certain features of SatGen that are relevant for our current study are elaborated on below.

SatGen generates halo merger trees using the algorithm of Parkinson, Cole & Helly (2008) as re-calibrated by Benson (2017). Merger trees are constructed using the time-stepping advocated in Appendix A of Parkinson, Cole & Helly (2008), which corresponds to $\Delta z \simeq 0.001$, but for book keeping the temporal resolution is down-sampled to timesteps of $\Delta t = 0.1 t_{\rm dyn}(z)$, where $t_{\rm dyn} = \sqrt{3\pi/[16 G \Delta \rho_{\rm crit}(z)]}$ is the instantaneous virial time of DM haloes. In the EdS regime, approximately valid at $z > 1$, $\Delta \simeq 200$ and the mean universal density approaches the critical cosmological density.

The structure of the host potential is determined in the following way. First, the virial mass of the system $M_{\rm v}(t)$ is given by following the main progenitor along the main branch of the merger tree. The stellar mass $M_{\rm s}(t)$ is assigned according to the abundance matching relations of Rodríguez-Puebla et al. (2017). Second, we determine the DZ profile of the halo, including the effect of baryonic, as follows. The concentration parameter in a DM-only scenario, $c_{2,{\rm DMO}}(M_{\rm v}, t)$, is obtained from the empirical relation of Zhao et al. (2009). We then consider the halo response to baryons following Freundlich et al. (2020a), which provides empirically the ratio of the baryon-affected concentration and the DM-only concentration, $c_2/c_{2,{\rm DMO}}$, as a function of the stellar-to-halo-mass ratio $M_{\rm s}/M_{\rm v}$, and the inner logarithmic slope of the system $s_1 = {\rm d}\ln\rho/{\rm d}\ln r$ at $r = 0.01 R_{\rm v}$. Finally, we compute the DZ-profile parameters (c, α) using c_2 and s_1. This procedure applies to both the host halo and the progenitors of satellites prior to infall.

The orbits of incoming satellites are initialized as follows. We consider the infall locations to be isotropically distributed on the virial sphere of the host halo, for which we randomly draw an azimuthal angle ϕ from $[0, 2\pi]$ and a cosine polar angle $(\cos\theta)$ from $[0, 1]$. We assume that the orbital energy is the same as that of a circular orbit of of velocity $V_{\rm v}(t)$ at radius $R_{\rm v}(t)$, and randomly assign a circularity ϵ from a distribution, ${\rm dP}/{\rm d}\epsilon = \pi\sin(\pi\epsilon)/2$, which approximates the ϵ distribution of infalling satellites measured in cosmological simulations (Wetzel 2011; van den Bosch 2017).

We follow the orbits by treating satellites as point masses. At each timestep, SatGen solves the equations of motion

$$\ddot{\boldsymbol{r}} = -\nabla\Phi + a_{\rm DF}, \qquad (C1)$$

where $\boldsymbol{r}$ is the position vector, Φ is the gravitational potential, and $a_{\rm DF}$ is the acceleration due to dynamical friction, modeled using the Chandrasekhar (1943) formula as given in eq. (12) and eq. (13).

We model the tidal mass loss using

$$\dot{m} = -A\frac{m(> \ell_{\rm t})}{t_{\rm dyn}(r)}, \qquad (C2)$$

where we have introduced a fudge parameter A as the stripping efficiency to encapsulate uncertainties in the definition of the tidal radius. That is, the timescale on which stripping occurs is the local dynamical time $t_{\rm dyn}(r) = \sqrt{3\pi/16G\bar{\rho}(r)}$ divided by A, with $\bar{\rho}(r)$ the average density of the host system within radius r. We use $A = 0.55$ following the calibration by Green, van den Bosch & Jiang (2021) from simulations. The mass loss over a timestep Δt is then given by $\Delta m = \dot{m}\,\Delta t$.

To keep track of DF heating, we register the work done by DF on a satellite at each step, or equivalently the orbital energy change at each step,

$$\Delta W(t + \Delta t) = E(t) - E(t + \Delta t), \qquad (C3)$$

Note that the orbital energy E at time $t + \Delta t$ includes the contribution from the stripped mass Δm, which is assumed to be on the same orbit of the satellite that it used to belong to.

The structural evolution of satellites in response to tidal mass loss, heating, and re-virialization, is modeled using the empirical tidal tracks from simulations (Penarrubia et al. 2010). Note that the tidal track is conditioned on the initial structure of the satellites, which is important for capturing the difference in DF heating due to a compact satellite versus a diffuse one. In the current study we do not explicitly include baryons within the satellites.

Jiang et al. (2020) used the model to study satellites of Milky-Way sized hosts, making it emulate simulations of bursty or smooth star formation and experimenting with a disc potential in the host halo. They found that the model reproduces the observed satellite statistics in the Milky Way and M31 reasonably well. Different physical recipes make a difference in satellite abundance and spatial distribution at the 25% level, not large enough to be distinguished by current observations given the halo-to-halo variance. The MW/M31 disc depletes satellites by ~20% and has only a subtle effect of diversifying the internal structure of satellites, which may be important for alleviating certain small-scale problems. We do not explicitly include in the current study a central baryonic component.

APPENDIX D: N-BODY SIMULATIONS OF CUSP HEATING AND ITS RELAXATION

To test the impact of cusp heating by satellites and the following relaxation, we run idealized N-body simulations, each with a host halo and a single merging satel-

lite. At the beginning of the simulation, the two haloes are set in equilibrium, spherically symmetric and with an isotropic velocity dispersion. In this case, the phase-space density of the particles is determined by the particle specific energy E and its radial distance from the halo center r. The halo density profile can be written as

$$\rho(r) = m_{\mathrm{p}} \int f(E) d^3 \mathbf{v} = 4\pi\sqrt{2}m_{\mathrm{p}} \int_0^{\Psi} \sqrt{\Psi - \mathcal{E}} f(\mathcal{E}) d\mathcal{E},$$
(D1)

where $\Psi = \Phi_0 - \Phi$, with Φ the gravitational potential and Φ_0 its value at the boundary of the system, which we set at $4R_{\mathrm{v}}$. The energy per unit mass is $\mathcal{E} = \Psi - (1/2)v^2$. Given the density profile, the gravitational potential can be derived from the Poisson equation. For a realistic stationary halo, Ψ is a monotonically decreasing function of r, so ρ can be written as a function of Ψ. Taking the derivative of both sides of eq. (D1) with respect to Ψ, one gets

$$\frac{d\rho}{d\Psi} = \sqrt{8}\pi m_{\mathrm{p}} \int_0^{\Psi} \frac{f(\mathcal{E})}{\sqrt{\Psi - \mathcal{E}}} \, d\mathcal{E},$$
(D2)

where m_{p} is the particle mass. The above equation can be solved to give the Eddington's inversion formula (Eddington 1916)

$$f(\mathcal{E}) = \frac{1}{\sqrt{8}\pi^2 m_{\mathrm{p}}} \frac{\mathrm{d}}{\mathrm{d}\mathcal{E}} \int_0^{\mathcal{E}} \frac{\mathrm{d}\Psi}{\sqrt{\mathcal{E} - \Psi}} \frac{\mathrm{d}\rho}{\mathrm{d}\Psi} \mathrm{d}\Psi.$$
(D3)

The particle positions and velocities are randomly drawn from the density profile, eq. (D1) and the velocity (energy) distribution, eq. (D3). For the host halo, an NFW profile is used with a sharp truncation at $4R_{\mathrm{v}}$. For the satellites, the density profile, either NFW profile (compact satellite) or Burkert profile (diffuse satellite), is truncated exponentially at the virial radius following Kazantzidis, Zentner & Kravtsov (2006) to roughly account for tidal truncation.

After generating the stationary halos, the satellite is put at the apocenter of an orbit with specified orbital parameters, e.g., the circularity and total energy. The system is then evolved using the public N-body SPH code GADGET-2 (Springel & Hernquist 2005).

The centers of the host and satellite are identified by searching for the most-bound particle, the particle that has the most negative total energy $E_i = \Phi(r_i) + (1/2)\,v_i^2$ within the corresponding halo. For the satellite, the bound mass is computed at each snapshot using an iterative un-binding algorithm (van den Bosch & Ogiya 2018).

APPENDIX E: CUSPCORE - AN ANALYTIC MODEL FOR DM RESPONSE TO OUTFLOWS

Freundlich et al. (2020a)[11] presents a simple analytic

[11] Available for application in
https://github.com/Jonathanfreundlich/CuspCore .

model for the response of a dissipationless spherical system to an instantaneous mass change at its center. It has been applied there to the formation of flat cores in low-mass dark-matter haloes and the origin of ultra-diffuse galaxies (UDGs) from outflow episodes driven by supernova feedback, but it is applicable for any rapid changes in the central mass. Here we use it for the dark-matter response to AGN-driven central gas ejection. This model generalizes an earlier simplified analysis of an isolated shell (Dutton et al. 2016) into a system with continuous density, velocity and potential profiles.

The DM response is divided into two steps: an instantaneous change of potential at constant velocities due to a given rapid mass loss (or mass gain), followed by energy-conserving relaxation to a new Jeans equilibrium. The halo profile is modeled by the two-parameter Dekel-Zhao profile described in §A, using the analytic expressions for the associated potential and kinetic energies at equilibrium. The way energy conservation is applied in the second stage of this model is not formally justified in the case of shell crossing, so its validity as an approximation should be based on testing against simulations. In Freundlich et al. (2020a), the model has been tested against NIHAO cosmological zoom-in simulations, where it successfully predicts the evolution of the inner DM profile between successive snapshots in about 75% of the cases, failing mainly in merger situations when the system strongly deviates from Jeans equilibrium.

The energy per unit mass of a shell at radius r_{i} in the initial halo at Jeans equilibrium is the sum

$$E_{\mathrm{i}}(r_{\mathrm{i}}) = U(r_{\mathrm{i}}; p_{\mathrm{i}}) + K(r_{\mathrm{i}}; p_{\mathrm{i}}),$$
(E1)

where $U(r_{\mathrm{i}}; p_{\mathrm{i}})$ and $K(r_{\mathrm{i}}; p_{\mathrm{i}})$ are functional forms for the potential and kinetic energies per unit mass, which depend on the parameters p_{i} that characterize the initial halo density profile. We use the DZ profile with the parameters c and α, for which the potential $U(r_{\mathrm{i}}; p_{\mathrm{i}})$ is given by eq. (A16), and the kinetic energy $K(r_{\mathrm{i}}; p_{\mathrm{i}})$ derives from eq. (A19), stemming from Jeans equilibrium. For the two energies we may consider an additional baryonic component, characterized by additional parameters. In the temporary state immediately after the instantaneous mass change by m (where $m < 0$ for an outflow and $m > 0$ for an inflow), the energy becomes

$$E_{\mathrm{t}}(r_{\mathrm{i}}) = U(r_{\mathrm{i}}; p_{\mathrm{i}}) - \frac{Gm}{r_{\mathrm{i}}} + K(r_{\mathrm{i}}; p_{\mathrm{i}}).$$
(E2)

After relaxation to the final Jeans equilibrium state of the halo, whose profile is described by the parameters p_{f}, the shell encompassing a given mass has moved to a final radius r_{f} and its energy is

$$E_{\mathrm{f}}(r_{\mathrm{f}}) = U(r_{\mathrm{f}}; p_{\mathrm{f}}) - \frac{Gm}{r_{\mathrm{f}}} + K(r_{\mathrm{f}}; p_{\mathrm{f}}, m),$$
(E3)

where the kinetic energy is again set by the Jeans equation but it also depends on the mass change m. The radius r_{f} is itself a function of the final parameters p_{f}, given that the enclosed mass is constant, $M(r_{\mathrm{f}}; p_{\mathrm{f}}) = M(r_{\mathrm{i}}; p_{\mathrm{i}})$. The assumed energy conservation during the relaxation phase corresponds to $E_{\mathrm{f}}(r_{\mathrm{f}}) =$

$E_t(r_i)$, which is solved numerically to obtain the final halo parameters p_f. In practice, we minimize the difference $E_f(r_f) - E_t(r_i)$ for hundred shells equally spaced in $\log(r/R_v)$ from -2 to 0 (thus giving more weight to central regions than linearly-spaced shells). The assumed energy conservation per shell that encompasses a given mass is not formally justified in the case of shell crossing, and our use of it is based on the success of this model in reproducing the results of simulations. We refer to Freundlich et al. (2020a) for more details, and to Freundlich, Dekel & Jiang (2019) for a brief presentation of the model.

APPENDIX F: A SATELLITE WITH CENTRAL BARYONS

Complementing the main text, Fig. F1 is the analog of Fig. 4, showing the result of a SatGen run with our fiducial NFW host halo, but with the satellites following the more compact DZ-profile fits to the VELA simulated galaxies using the *total* mass including the baryons rather than the dark matter alone. We learn that for the more compact satellites, as expected, the penetrating mass to the host cusp is higher, and the energy deposited in the cusp is higher accordingly. However, the difference is rather small, with $m/m_v \simeq 0.5$ compared to 0.4, and $W_c/K_c \simeq 2.5$ compared to 2, for the compact satellites.

APPENDIX G: A STEEP-CUSP HOST HALO

Here we show the same results that have been shown in the main text for an NFW host halo with a moderately steep cusp, but for a steep-cusp host of a DZ profile with $s_1 = 1.5$ and $c_2 = 5$.

Figure G1, same as Fig. 3, shows the toy model predictions as a function of the satellite compactness. Figure G2, same as Fig. 4, shows the results of a SatGen run with a single satellite. Figure G3, same as Fig. 5, refers to SatGen runs with a cosmological sequence of satellites. Figure G4 and Fig. G5, same as Fig. 9 and Fig. 10, show the results of CuspCore for a single satellite and a cosmological sequence of satellites, respectively. The results for the steep-cusp host are discussed in comparison to the results for the NFW host in the main text.

APPENDIX H: OUTFLOW WITH SATELLITE MASS ADDED TO HOST

APPENDIX I: TOY-MODEL S_1 AND C_2

Figure I1 is the analog of Fig. 3, showing the toy-model estimates for the satellite mass in the host cusp and the energy deposited there by dynamical friction as a function of the satellite initial profile, but here for the more accessible parameters (c_2, s_1) instead of the natural DZ parameters (c, α). The conclusion is the same as in Fig. 3.

I1   Orbit circularity as a function of redshift and halo mass

One may elaborate on the redshift and mass dependence of the satellite orbit circularity as one of the factors in the tendency of DF heating to be more effective at higher masses and higher redshifts.

The orbit is characterized at R_v by two parameters, e.g., energy and angular momentum, or the velocity magnitude V_{in} and the circularity $\epsilon = V_{tan}/V_v$. For reference, the orbit eccentricity is $e^2 = 1 - \epsilon^2$, and the corresponding spin parameter is

$$\lambda = \frac{V_{tan}R_v}{\sqrt{2}V_v R_v} = \frac{\epsilon}{\sqrt{2}}, \qquad (I1)$$

independent of V_{in}. The orbit, and the effects of dynamical friction and tidal stripping, depend in addition on V_{in}, which for $M_v \sim 10^{12}M_\odot$ at $z \sim 2$ is roughly $V_{in}/V_v \simeq 1.15 \pm 0.15$ (Wetzel 2011, Figs. 2,5,9).

According to the cosmological N-body simulations of Wetzel (2011), Figs. 5 and 8, for minor mergers of $m_v/M_v \sim 0.02$, ϵ tends to decrease with increasing redshift, where the average is $\langle \epsilon \rangle \simeq 0.55$ and 0.45 at $z = 0$ and 2.5, respectively. The corresponding median of the pericenter of the orbit is roughly $\langle r_{peri} \rangle/R_v \simeq 0.24$ and 0.17, respectively, namely a deeper penetration at higher redshifts. The distribution of ϵ is found to be approximately universal for a given host halo mass when measured with respect to the non-linear Press-Schechter mass $M_{ps}(z)$. For a given satellite mass, ϵ and r_{peri} tend to decrease with increasing host halo mass. These redshift and mass dependencies are in the desired sense, but they are rather mild, possibly not sufficient by themselves for explaining the redshift and mass dependencies of the DM core phenomenon.

A qualitatively similar redshift dependence is obtained from hydro cosmological simulations, via the analysis of the angular momentum carried by the cosmic-web cold streams that build the galaxies at high redshift, and contain the incoming satellites (Danovich et al. 2015). According to their figure 15, the dominant stream carries on average 84% of the angular-momentum inflow rate, and 64% of the mass influx. In order to relate the measured spin parameter to measured eccentricity, we consider for an upper limit only one dominant stream, and obtain $\lambda \sim \epsilon/\sqrt{2} \sim 0.35$ for $\epsilon = 0.5$. For a lower limit, we consider three comparable streams with random orientation and impact parameter, and obtain $\lambda \sim (\epsilon/\sqrt{2})/\sqrt{3} \sim 0.2$ for $\epsilon = 0.5$. We can therefore assume that the mean eccentricity measured by Wetzel (2011), $\epsilon \sim 0.5$, would typically correspond to $\lambda \sim 0.3$. However, from Danovich et al. (2015, Fig. 1), at R_v, we read for the dark matter that $\lambda \simeq 0.13$ and it is not varying with redshift, while for the cold gas $\lambda \simeq 0.3, 0.2, 0.15$ at $z = 1.5, 2.5, 3.5$, respectively. Smaller λ values are measured at higher redshifts also in their Fig. 7. Similar results are obtained for cold gas at $z = 1.6 - 3$ in their Fig. 14, where $\lambda \sim 0.2$. This indicates that at $z \sim 2.5$ one should assume $\epsilon \sim 0.25$ for the dark matter and $\epsilon \sim 0.33$ for the cold gas. These values are lower than the average value obtained for satellites by
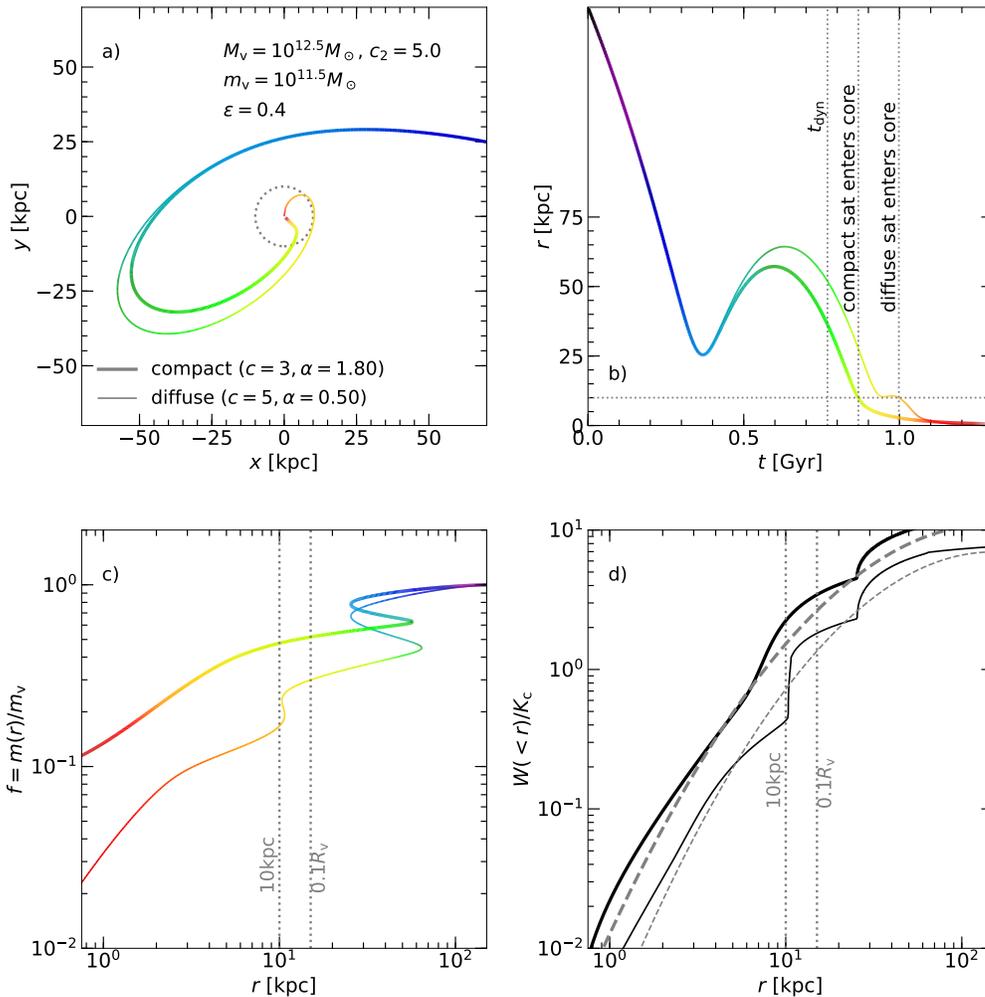
Figure F1. Semi-analytic SatGen simulations of single satellites, diffuse and compact, similar to Fig. 4, but where the satellite profiles are the best DZ fits to the *total* mass in the VELA pre and post compaction galaxies from Fig. 1, including the baryons, with DZ parameters $(c, \alpha) = (5, 0.5)$ and $(3, 1.8)$, respectively.

Wetzel (2011), indicating more radial orbits and thus stronger dynamical friction. The redshift dependence in Danovich et al. (2015) for the cold gas is stronger than in Wetzel (2011), but this may be balanced by the weaker redshift dependence for the dark matter.

Qualitatively similar redshift and mass dependencies can be deduced from the analysis of random Gaussian fluctuation fields by Bardeen et al. (1986), who predict that higher-sigma density peaks have lower λ values and more radial orbits. This is consistent with the trends found in Wetzel (2011) and Danovich et al. (2015), and with the core phenomenon being more pronounced at higher redshifts and masses, being higher-sigma peaks.

Figure G1. Same as Fig. 3 but for the fiducial steep-cusp host, with $s_1 = 1.5$ and $c_2 = 5$ in the DZ profile. Shown are the toy-model estimates for satellite penetration and energy deposited in the host cusp by dynamical friction, as a function of the satellite compactness via the Dekel-Zhao profile parameters of concentration and inner slope (c, α). For the steep-cusp host we read for the diffuse and compact satellites respectively $m/m_v \sim 0.01, 0.23$, $W_c/K_c \text{(single)} \sim 0.001, 0.56$ and $W_c/K_c \text{(cosmo)} \sim 0.003, 1.53$. The satellite stripping is stronger due to the steeper cusp, but the heating by compact satellites is still significant during half a virial time.

Figure G2. Same as Fig. 4, for an $m_v/M_v = 0.1$ single satellite, but for a steep-cusp host halo with DZ parameters $(c_2, s_1) = (5, 1.5)$. The penetrating satellite mass and the DF energy deposited in the steep cusp are similar to the case of an NFW halo, indicating significant heating by the compact satellite and only partial heating by the diffuse satellite. The functional fits (dashed) to be used by CuspCore are listed in Table 1.
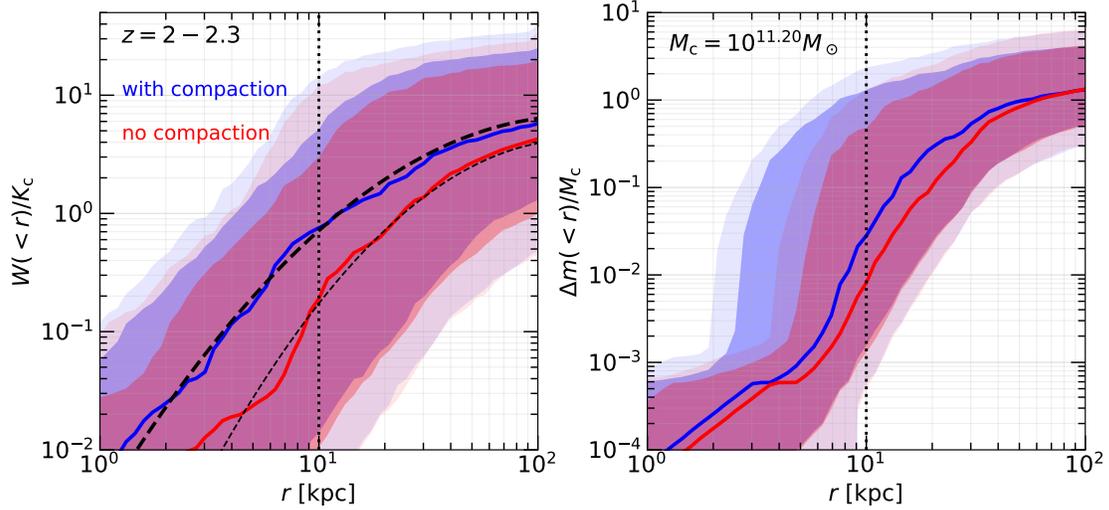
Figure G3. Same as Fig. 5 but for the steep-cusp host. Shown are the results of a SatGen simulation of a cosmological sequence of satellites during one halo virial time at $z \sim 2$, where $t_v \simeq 0.5\,\mathrm{Gyr}$. The host halo starts with a DZ steep-cusp profile of $(c_2, s_1) = (5, 1.5)$. The fits for the deposited energy by DF to be used in CuspCore is marked (dashed black), and listed in Table 1. Most of the satellite mass is deposited near the outer edge of the cusp, with only a small fraction of the mass penetrating to the inner cusp. The energy deposited in the steep cusp by DF on a sequence of compact satellites is lower than that deposited in the NFW cusp (Fig. 5) by a factor of ~ 2.5. However, it is 75% of the cusp kinetic energy, implying heating also in the steep-cusp host. The heating by diffuse satellites is weaker, only 20% of the cusp energy.
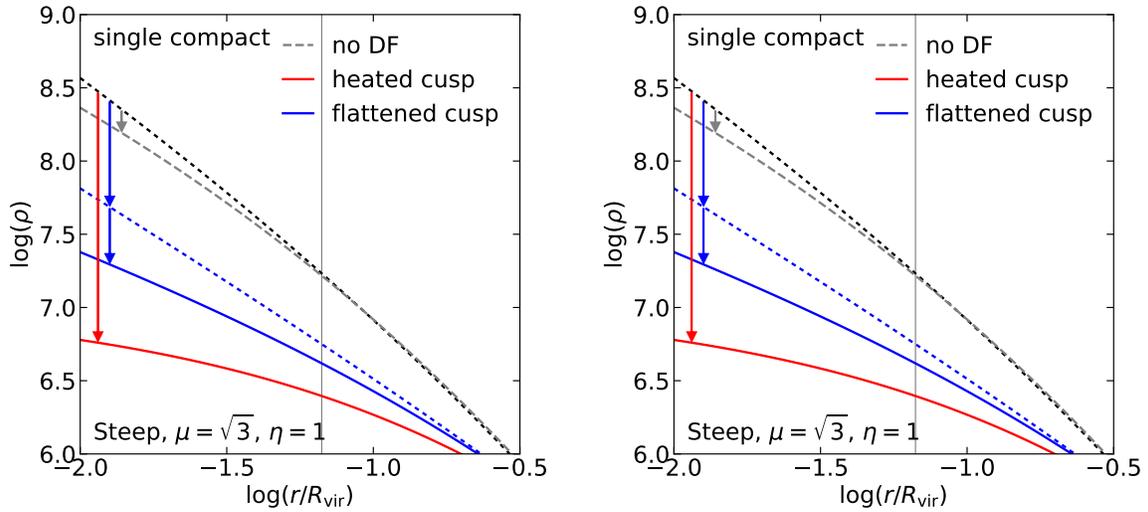

Figure G4. Density profiles as in Fig. 9, for DF heating by single satellite, but for a steep-cusp initial host halo with $(c_{2h}, s_{1h}) = (5, 1.5)$ instead of NFW. Here, in order to obtain a significant effect, the initial satellite mass is $m_v = 0.1\mu M_v$ with $\mu = \sqrt{3}$ and the outflow is with $\eta = 1$, namely involving all the available gas of $10^{10.6} M_\odot$ (compared to $\mu = \sqrt{2}$ and $\eta = 0.5$ in Fig. 9). The DF heating is based on the energy deposit profile by SatGen, bottom-left panel of Fig. G2, which turns out to be comparable to the energy deposit of the NFW host. We learn that the steep cusp is more resilient than the NFW cusp both to DF heating and to outflows, requiring more massive satellite and outflow for generating an extended core.
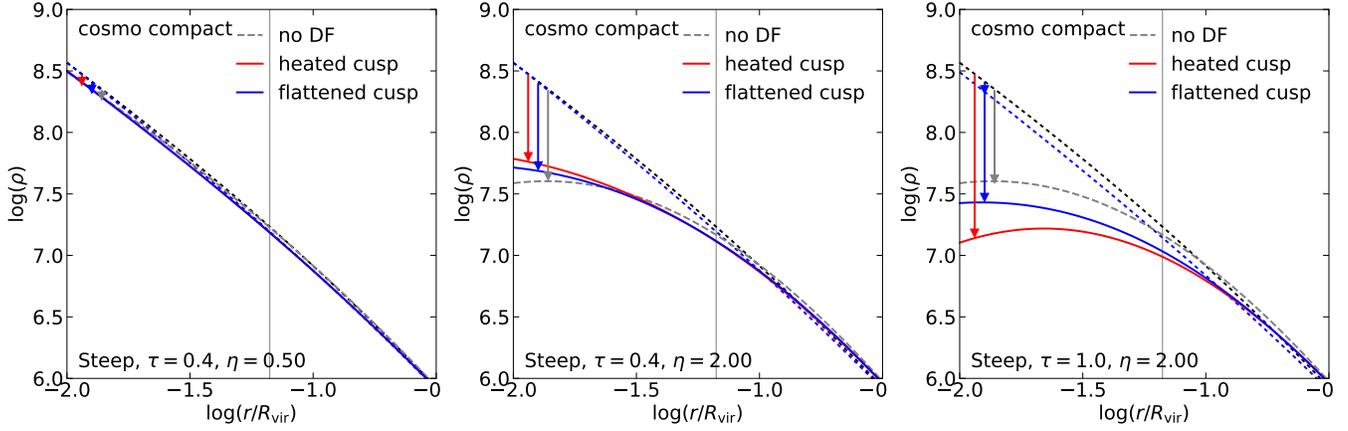
Figure G5. Same as Fig. 10, for DF heating by a cosmological sequence of satellites, but for a steep-cusp host halo with $(2h, s_{1h}) = (5, 1.5)$ instead of NFW. The energy deposited in a virial time is based on the fit to the SatGen run shown in Fig. G3. **Left:** Compact satellites and outflows with the parameters that produced a core in the NFW cusp cause a negligible effect on the steep cusp. **Middle:** An outflow of $\eta = 2$ is sufficient for forming a moderate core without DF heating. In this case, the DF actualy steepens the profile, making it a little harder to produce a core by inflow. **Right:** Similar but using a slightly flatter slope of $K(<r)$ than produced by SatGen within the cusp, which enables a better fit to the DZ profile and thus a convergence of CuspCore, leading to a flat core after DF heating with $\tau = 1$. The higher value of τ can be interpreted as roughly representing the top $1/3$ of the random realizations drawn from the mass function and the circularity distribution of satellites during $0.4\,t_{\rm v}$. Alternatively it can be interpreted as the median energy during $t_{\rm v}$ or or a longer duration.
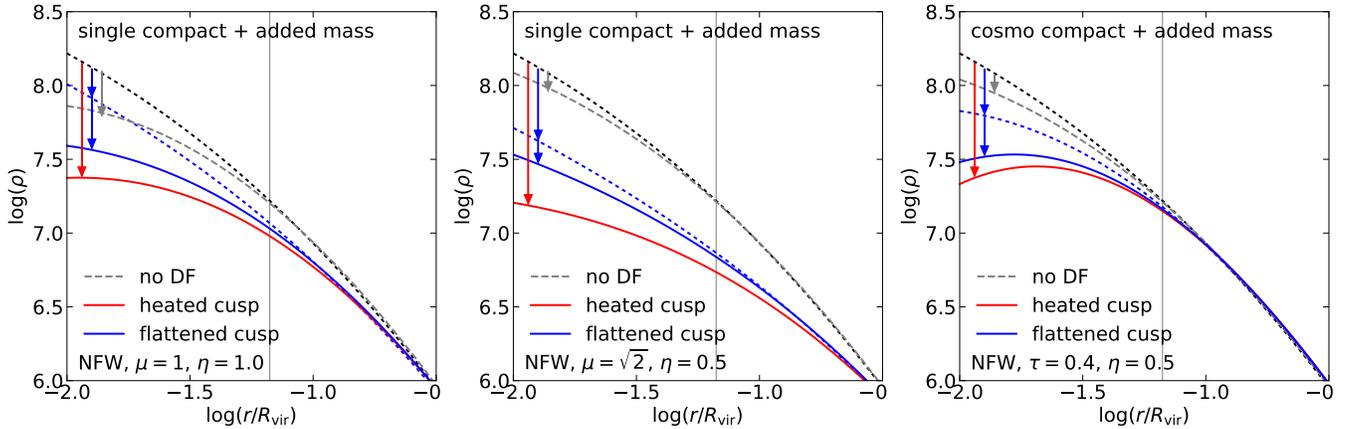

Figure H1. The effect of added satellite mass to the cusp. Same as Fig. 9 (left and middle) and Fig. 10, for DF heating by a single compact satellite or a sequence of compact satellites, in an NFW host, but with the mass of the satellite added to the host cusp where it is stripped or at the center, based on the bottom-left panel of Fig. 4 and the right panel of Fig. 5. The difference from the results obtained without this additional mass in Fig. 9 and Fig. 10 is small. For the single satellite, the satellite mass deposited in the cusp is comparable to the cusp mass, slightly steepening the fonal core. For the sequence of satellites, the satellite mass is only $0.2 M_{\rm c}$, causing almost no change to the final core. This justifies ignoring the added mass in our main analysis.
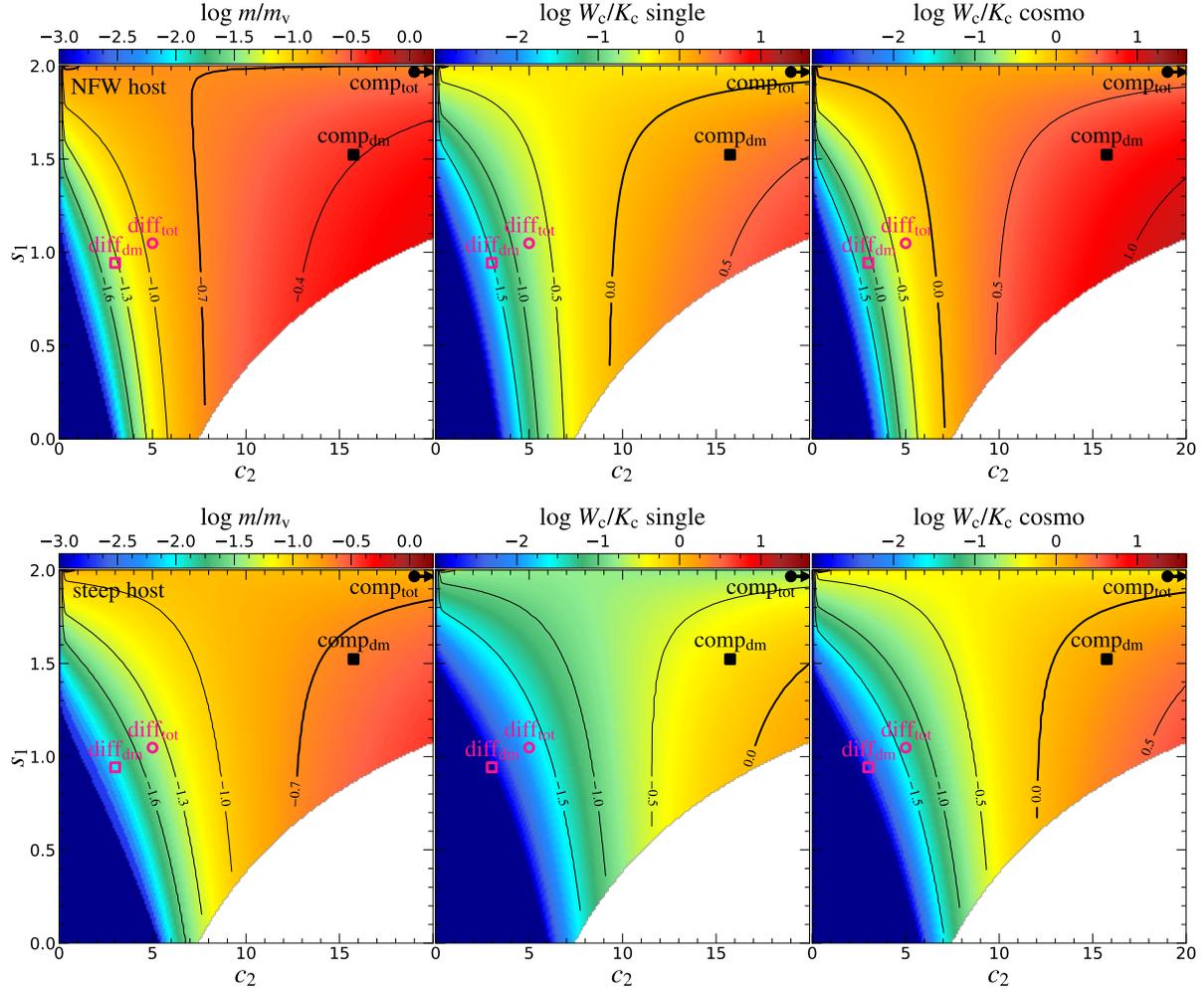
Figure I1. Toy-model estimates for satellite penetration and energy deposited in the host cusp by dynamical friction, similar to Fig. 3, but for the more accessible parameters (c_2, s_1) instead of the natural Dekel-Zhao parameters (c, α). The concentration parameter c_2 refers to the radius where the local log slope of the density profile is -2 (as in the concentration of the NFW profile), and s_1 is the minus the inner local log slope at $r = 0.01 R_{\rm v}$. The transformation between the two alternative pairs of parameters is given in eq. (A12) to eq. (A15). While there is a valid DZ profile for any values of c (>0) and α (<3), a valid profile is not guaranteed for arbitrary values of c_2 and s_1. For example, $c_2 \to \infty$ for $\alpha = 2$. The case of compact total $(c, \alpha) = (3, 1.8)$ has $(c_2, s_1) = (170, 2.05)$, which is outside the box of this figure. We truncate the plot where $c > 100$ (or where $\alpha \geqslant 2$).