

# Representing low-mass black hole seeds in cosmological simulations: A new sub-grid stochastic seed model

Aklant K. Bhowmick<sup>1</sup>,<sup>1</sup>★ Laura Blecha,<sup>1</sup> Paul Torrey<sup>1,2</sup>, Rainer Weinberger<sup>1,3</sup>, Luke Zoltan Kelley<sup>1,4</sup>, Mark Vogelsberger<sup>1,5</sup>, Lars Hernquist<sup>6</sup> and Rachel S. Somerville<sup>7,8</sup>

<sup>1</sup>Department of Physics, University of Florida, Gainesville, FL 32611, USA

<sup>2</sup>Department of Astronomy, University of Virginia, 530 McCormick Road, Charlottesville, VA 22903, USA

<sup>3</sup>Leibniz Institute for Astrophysics, An der Sternwarte 16, D-14482 Potsdam, Germany

<sup>4</sup>Department of Astronomy, University of California at Berkeley, 501 Campbell Hall, Berkeley, CA 94720, USA

<sup>5</sup>Department of Physics, Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

<sup>6</sup>Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

<sup>7</sup>Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA

<sup>8</sup>Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Rd, Piscataway, NJ 08854, USA

Accepted 2024 March 14. Received 2024 January 15; in original form 2023 September 26

## ABSTRACT

The physical origin of the seeds of supermassive black holes (SMBHs), with postulated initial masses ranging from  $\sim 10^5 M_\odot$  to as low as  $\sim 10^2 M_\odot$ , is currently unknown. Most existing cosmological hydrodynamic simulations adopt very simple, ad hoc prescriptions for BH seeding and seed at unphysically high masses  $\sim 10^5$ – $10^6 M_\odot$ . In this work, we introduce a novel sub-grid BH seeding model for cosmological simulations that is directly calibrated to high-resolution zoom simulations that explicitly resolve  $\sim 10^3 M_\odot$  seeds forming within haloes with pristine, dense gas. We trace the BH growth along galaxy merger trees until their descendants reach masses of  $\sim 10^4$  or  $10^5 M_\odot$ . The results are used to build a new stochastic seeding model that directly seeds these descendants in lower resolution versions of our zoom region. Remarkably, we find that by seeding the descendants simply based on total galaxy mass, redshift and an environmental richness parameter, we can reproduce the results of the detailed gas-based seeding model. The baryonic properties of the host galaxies are well reproduced by the mass-based seeding criterion. The redshift-dependence of the mass-based criterion captures the combined influence of halo growth, dense gas formation, and metal enrichment on the formation of  $\sim 10^3 M_\odot$  seeds. The environment-based seeding criterion seeds the descendants in rich environments with higher numbers of neighbouring galaxies. This accounts for the impact of unresolved merger dominated growth of BHs, which produces faster growth of descendants in richer environments with more extensive BH merger history. Our new seed model will be useful for representing a variety of low-mass seeding channels within next-generation larger volume uniform cosmological simulations.

**Key words:** methods: numerical – galaxies: evolution – galaxies: formation – quasars: supermassive black holes.

## 1 INTRODUCTION

The origin of supermassive black holes (SMBHs) is a key missing piece in our current understanding of galaxy formation. Several theoretical channels have been proposed for the first ‘seeds’ of SMBHs, predicting a wide range of postulated initial masses. At the lowest mass end of the initial seed mass function, we have the remnants of the first-generation Population III stars, a.k.a. Pop III seeds (Fryer, Woosley & Heger 2001; Madau & Rees 2001; Xu, Wise & Norman 2013; Smith et al. 2018) ranging from  $\sim 10^2$  to  $10^3 M_\odot$ . Next, we have seeds postulated at the ‘intermediate mass’ range of  $\sim 10^3$ – $10^4 M_\odot$  that can form via runaway stellar and black hole (BH) collisions within dense Nuclear Star Clusters, a.k.a NSC seeds (Davies, Miller & Bellovary 2011; Lupi et al. 2014; Kroupa

et al. 2020; Das et al. 2021a, b). Finally, we can have ‘high mass seeds’ formed via direct isothermal collapse of gas at sufficiently high temperatures ( $\gtrsim 10^4$  K), a.k.a direct collapse black hole or DCBH seeds (Bromm & Loeb 2003; Begelman, Volonteri & Rees 2006; Regan, Johansson & Wise 2014; Latif, Schleicher & Hartwig 2016; Luo et al. 2018, 2020; Wise et al. 2019; Begelman & Silk 2023). DCBHs masses are traditionally postulated to be ranging within  $\sim 10^4$ – $10^6 M_\odot$ , but recent works have suggested that they can also be as massive as  $\sim 10^8 M_\odot$  (Mayer et al. 2024).

The growing observed population of luminous quasars at  $z \sim 6$ –8 (Fan et al. 2001; Willott et al. 2010; Mortlock et al. 2011; Venemans et al. 2015; Bañados et al. 2016; Jiang et al. 2016; Reed et al. 2017; Bañados et al. 2018; Matsuoka et al. 2018; Wang et al. 2018, 2021; Matsuoka et al. 2019; Yang et al. 2019) tells us that  $\sim 10^9$ – $10^{10} M_\odot$  BHs already existed within the first billion after the Big Bang. These already pose a serious challenge to models of BH formation as well as BH growth. For example, light seeds may need to sustainably accrete

\* E-mail: [aklant.app@gmail.com](mailto:aklant.app@gmail.com)

gas at super-Eddington rates to grow by  $\sim 6\text{--}7$  orders of magnitude within such a short time. Alternatively, they can boost their seed mass via mergers, but it is unclear how efficiently these seeds sink and merge within the shallow potential wells of high redshift proto-galaxies (Volonteri 2007; Ma et al. 2021). Heavier seed masses such as DCBHs are substantially more conducive for assembling the high- $z$  quasars, but it is unclear if they form frequently enough to account for the observed number densities ( $1\text{ Gpc}^{-3}$ ).

Due to possible degeneracies in the impact of different BH formation versus BH growth models, it is challenging to constrain seed models solely using observations of luminous high- $z$  quasars. To that end, detections of lower mass BH populations at high- $z$  are going to be crucial for constraining seed models as these BHs are more likely to retain the memory of their initial seeds. The *JWST* (Gardner et al. 2006) is pushing the frontiers of SMBH studies by detecting lower luminosity active galactic nuclei (AGNs) at high redshifts. In addition to the first statistical sample of  $\sim 10^6\text{--}10^7\text{ M}_\odot$  AGN at  $z \sim 4\text{--}7$  (Harikane et al. 2023), *JWST* has also produced the first detections at  $z \gtrsim 8.3$  (Larson et al. 2023) and  $z \sim 10.6$  (Maiolino et al. 2023). Moreover, there is an exciting possibility of future detections of BHs as small as  $\sim 10^5\text{ M}_\odot$  using *JWST*, which would potentially enable us to probe the massive end of the seed population for the very first time (Natarajan et al. 2017; Cann et al. 2018; Inayoshi et al. 2022).

Even with *JWST* and proposed X-ray facilities like *ATHENA* (Barcons et al. 2017) and *Axis* (Mushotzky et al. 2019), low-mass seeds  $\sim 10^2\text{--}10^4\text{ M}_\odot$  are likely to be inaccessible to electromagnetic (EM) observations at high- $z$ . However, with the new observational window of gravitational waves (GW) opened for the first time by the Laser Interferometer Gravitational-Wave Observatory (LIGO; Abbott et al. 2009), we can close this gap. In addition to detecting numerous ( $\sim 80$ ) stellar mass BH mergers, LIGO has also started probing the elusive population of intermediate mass black holes (IMBH:  $\sim 10^2\text{--}10^5\text{ M}_\odot$ ) with GW190521 (Abbott et al. 2020) producing a  $\sim 142\text{ M}_\odot$  BH remnant. At the other end of BH mass spectrum, the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) have also detected the Hellings-Downs correlation expected from a stochastic GW background that most likely originates from populations of merging SMBHs (Agazie et al. 2023). But the strongest imprints of BH formation will likely be provided by the upcoming Laser Interferometer Space Antenna (LISA; Baker et al. 2019), which is expected to detect GWs from mergers of IMBHs as small as  $\sim 10^3\text{ M}_\odot$  up to  $z \sim 15$  (Amaro-Seoane et al. 2017).

Cosmological hydrodynamic simulations (Di Matteo et al. 2012; Vogelsberger et al. 2014b; Khandai et al. 2015; Schaye et al. 2015; Sijacki et al. 2015; Dubois et al. 2016; Volonteri et al. 2016, 2020; Kaviraj et al. 2017; Tremmel et al. 2017; Nelson et al. 2019a) have emerged as powerful tools for testing galaxy formation theories (see e.g. the review by Vogelsberger et al. 2020). However, most such simulations can resolve gas elements only down to  $\sim 10^5\text{--}10^7\text{ M}_\odot$ , depending on the simulation volume. This is particularly true for simulation volumes needed to produce statistical samples of galaxies and BHs that can be directly compared to observations. Therefore, most cosmological simulations only model BH seeds down to  $\sim 10^5\text{ M}_\odot$  (e.g. Vogelsberger et al. 2014b; Khandai et al. 2015; Tremmel et al. 2017). Notably, there are simulations that do attempt to capture seed masses down to  $\sim 10^4\text{ M}_\odot$  (Ni et al. 2022) and  $\sim 10^3\text{ M}_\odot$  (Taylor & Kobayashi 2014; Wang et al. 2019), but they do so without explicitly resolving the seed-forming gas to those masses. Overall, directly resolving the low mass seed population ( $\sim 10^2\text{--}10^4\text{ M}_\odot$  encompassing Pop III and NSC seeding channels) is completely

inaccessible within state-of-the-art cosmological simulations, and pushing beyond current resolution limits will require a substantial advancement in available computing power.

Given that BH seed formation is primarily governed by properties of the seed-forming gas, the insufficient resolution within cosmological simulations carries the additional liability of having poorly converged gas properties. For instance, Pop III and NSC seeds are supposed to be born out of star-forming and metal-poor gas. However, the rates of dense gas formation and metal enrichment may not be well converged in these simulations at their typical gas mass resolutions of  $\sim 10^5\text{--}10^7\text{ M}_\odot$  (for example, see fig. 19 of Bhowmick et al. 2021). As a result, many simulations (Di Matteo et al. 2012; Vogelsberger et al. 2014b; Nelson et al. 2018; Ni et al. 2022) simply use a host halo mass threshold to seed BHs. Several cosmological simulations have also used local gas properties for seeding (Taylor & Kobayashi 2014; Tremmel et al. 2017; Wang et al. 2019). These simulations produce seeds directly out of sufficiently dense and metal-poor gas cells, which is much more consistent with proposed theoretical seeding channels. But these approaches can lead to stronger resolution dependence in the simulated BH populations (see fig. 10 of Taylor & Kobayashi 2014). In any case, most of these seeding approaches have achieved significant success in generating satisfactory agreement with the observed SMBH populations at  $z \sim 0$  (Habouzit et al. 2020). However, it is important to note that they do not provide definitive discrimination among the potential seeding channels from which the simulated BHs may have originated.

A standard approach to achieve very high resolutions in cosmological simulations is to use the ‘zoom-in’ technique. In our previous work (Bhowmick et al. 2021, 2022a), we used cosmological zoom-in simulations with gas mass resolutions up to  $\sim 10^3\text{ M}_\odot$  to build a new set of gas-based seed models that placed seeds down to the lowest masses ( $2.3 \times 10^3\text{ M}_\odot$ ) within haloes containing sufficient amounts of dense and metal-poor gas. We systematically explored these gas-based seed models and found that the strongest constraints for seeding are expected within merger rates measurable with LISA. However, the predictions for these zoom simulations are subject to large cosmic variance, as they correspond to biased regions of the large-scale structure. In order to make observationally testable predictions with these gas-based seed models, we must find a way to represent them in cosmological simulations despite the lack of sufficient resolution.

In this work, we build a new sub-grid stochastic seed model that can represent low-mass seeds born out of dense and metal-poor gas, within lower-resolution and larger-volume simulations that cannot directly resolve them. To do this, we first run a suite of highest resolution zoom simulations that places  $\sim 10^3\text{ M}_\odot$  seeds within dense and metal-poor gas using the gas-based seed models from Bhowmick et al. (2021). We then study the growth of  $\sim 10^3\text{ M}_\odot$  seeds and the evolution of their formation environments. We particularly study the halo and galaxy properties wherein these seeds assemble higher mass ( $\sim 10^4$  &  $10^5\text{ M}_\odot$ ) descendants. We then use the results to build our stochastic seed model that directly seeds these descendants within lower resolution versions of the same zoom region. In the process, we determine the key ingredients required for these stochastic seed models to reproduce the results of the gas-based seed models in the lower resolution zooms.

Section 2 presents the basic methodology, which includes the simulation suite, the underlying galaxy formation model, as well as the BH seed models. In particular, our underlying galaxy formation model (except for BH seeding) is adopted from Illustris-TNG. Our main results are described in Sections 3 and 4. In Section 3, we present the results for the formation and growth of  $\sim 10^3\text{ M}_\odot$  seeds

within our highest resolution zoom simulations. In Section 4, we use the results from Section 3 to build our stochastic seed model. Finally, Section 5 summarizes our main results.

## 2 METHODS

### 2.1 AREPO cosmological code and the Illustris-TNG model

We use the AREPO gravity + magneto-hydrodynamics (MHD) solver (Springel 2010; Pakmor, Bauer & Springel 2011; Pakmor et al. 2016; Weinberger, Springel & Pakmor 2020) to run our simulations. The simulations use a  $\Lambda$  cold dark matter cosmology with parameters adopted from Planck Collaboration (2016): ( $\Omega_\Lambda = 0.6911$ ,  $\Omega_m = 0.3089$ ,  $\Omega_b = 0.0486$ ,  $H_0 = 67.74 \text{ km s}^{-1}\text{Mpc}^{-1}$ ,  $\sigma_8 = 0.8159$ ,  $n_s = 0.9667$ ). The gravity solver uses the PM Tree (Barnes & Hut 1986) method and the MHD solver for gas dynamics uses a quasi-Lagrangian description of the fluid within an unstructured grid generated via a Voronoi tessellation of the domain. Haloes are identified using the friends of friends (FOF) algorithm (Davis et al. 1985) with a linking length of 0.2 times the mean particle separation. Subhaloes are computed using the SUBFIND (Springel et al. 2001) algorithm for each simulation snapshot. Aside from our BH seed models, our underlying galaxy formation model is the same as the IllustrisTNG (TNG) simulation suite (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018, 2019a; Springel et al. 2018; Pillepich et al. 2018b; see also Genel et al. 2018; Weinberger et al. 2018; Donnari et al. 2019; Pillepich et al. 2019; Rodriguez-Gomez et al. 2019; Torrey et al. 2019; Nelson et al. 2019b; Habouzit et al. 2021; Übler et al. 2021). The TNG model includes a wide range of sub-grid physics for dense gas formation and evolution, metal enrichment, and feedback as detailed in Pillepich et al. (2018a) and also summarized in our earlier papers (Bhowmick et al. 2021, 2022a, b). Briefly, gas cells become star forming when they exceed a threshold of  $0.13 \text{ cm}^{-3}$ . Star particles represent single stellar populations (SSPs) with an assumed initial mass function of Chabrier (2003). These SSPs evolve and produce metals that are returned to the nearby ISM gas via feedback from SNIa, SNII, and AGB stars. The abundances of nine species (H, He, C, N, O, Ne, Mg, Si, and Fe) are individually tracked.

### 2.2 BH accretion, feedback, and dynamics

BH accretion rates are determined by the Eddington-limited Bondi–Hoyle formalism given by

$$\dot{M}_{\text{bh}} = \min(\dot{M}_{\text{Bondi}}, \dot{M}_{\text{Edd}}) \quad (1)$$

$$\dot{M}_{\text{Bondi}} = \frac{4\pi G^2 M_{\text{bh}}^2 \rho}{c_s^3} \quad (2)$$

$$\dot{M}_{\text{Edd}} = \frac{4\pi G M_{\text{bh}} m_p}{\epsilon_r \sigma_T c} \quad (3)$$

where  $G$  is the gravitational constant,  $\rho$  is the local gas density,  $M_{\text{bh}}$  is the BH mass,  $c_s$  is the local sound speed,  $m_p$  is the proton mass, and  $\sigma_T$  is the Thompson scattering cross-section. Accreting black holes radiate at bolometric luminosities given by,

$$L_{\text{bol}} = \epsilon_r \dot{M}_{\text{bh}} c^2, \quad (4)$$

where  $\epsilon_r = 0.2$  is the radiative efficiency.

IllustrisTNG implements a dual mode AGN feedback. ‘Thermal feedback’ is implemented for Eddington ratios ( $\eta \equiv \dot{M}_{\text{bh}}/\dot{M}_{\text{edd}}$ ) higher than a critical value of  $\eta_{\text{crit}} = \min[0.002(M_{\text{BH}}/10^8 M_\odot)^2, 0.1]$ . Here, thermal energy is deposited on to the neighbouring gas

at a rate of  $\epsilon_{f,\text{high}} \epsilon_r \dot{M}_{\text{BH}} c^2$  with  $\epsilon_{f,\text{high}} = 0.02$  where  $\epsilon_{f,\text{high}}$  is the ‘high accretion state’ coupling efficiency. ‘Kinetic feedback’ is implemented for Eddington ratios lower than the critical value. Here, kinetic energy is injected into the gas in a pulsed fashion whenever sufficient feedback energy is available, which manifests as a ‘wind’ oriented along a randomly chosen direction. The injected rate is  $\epsilon_{f,\text{low}} \dot{M}_{\text{BH}} c^2$  where  $\epsilon_{f,\text{low}}$  is called the ‘low accretion state’ coupling efficiency ( $\epsilon_{f,\text{low}} \lesssim 0.2$ ). For further details, we direct the interested readers to Weinberger et al. (2017). However, we demonstrated in Bhowmick et al. (2021) that accretion-driven BH growth is very small at these redshifts. As a result, the AGN feedback has a negligible impact on BH seeding and growth.

The limited mass resolution hinders our simulations from fully capturing the crucial BH dynamical friction force, especially for low masses. To stabilize the dynamics, BHs are relocated to the nearest potential minimum within their proximity, determined by the closest  $10^3$  neighbouring gas cells. When one BH enters the neighbourhood of another, prompt merger occurs.

### 2.3 Black hole seed models

#### 2.3.1 Gas-based seed model

We explore the formation and growth of the lowest mass  $2.3 \times 10^3 M_\odot$  seeds using the gas-based seeding prescriptions developed in Bhowmick et al. (2021). In order to contrast these seeds from those produced by the seed model discussed in the next subsection, we shall hereafter refer to them as *direct gas based* seeds or DGBs with mass  $M_{\text{seed}}^{\text{DGB}}$ . These seeding criteria are meant to broadly encompasses popular theoretical channels such as Pop III, NSC, and DCBH seeds, that are postulated to form in regions comprised of dense and metal-poor gas. We briefly summarize them as follows:

(i) *Dense and metal-poor gas mass criterion*: We place DGBs in haloes with a minimum threshold of dense (exceeding the star formation threshold of  $0.13 \text{ cm}^{-3}$ , the) & metal-poor ( $Z < 10^{-4} Z_\odot$ ) gas mass. Since our dense gas is also forming stars, to be consistent with our earlier works in this series (Bhowmick et al. 2021, 2022a, b), we denote this threshold by  $\tilde{M}_{\text{sfmt}}$  (in the units of  $M_{\text{seed}}^{\text{DGB}}$ ) where ‘sfmt’ stands for ‘star forming and metal-poor’. The values of  $\tilde{M}_{\text{sfmt}}$  are not constrained, but we expect it to be different for the various seeding channels. In this work, we consider models with  $\tilde{M}_{\text{sfmt}} = 5, 50, 150 \text{ \& } 1000$ .

(ii) *Halo mass criterion*: We place DGBs in haloes with a total mass exceeding a critical threshold, specified by  $\tilde{M}_h$  in the units of  $M_{\text{seed}}^{\text{DGB}}$ . In this work, we consider  $\tilde{M}_h = 3000 \text{ \& } 10000$ . While our seeding prescriptions are meant to be based on the gas properties within haloes, we still adopt this criterion to avoid seeding in haloes significantly below the atomic cooling threshold. This is because our simulations do not include the necessary physics (for e.g.  $H_2$  cooling) to self-consistently capture the collapse of gas and the formation of stars within these (mini)haloes. Additionally, these lowest mass haloes are also impacted by the finite simulation resolution, many of which are spuriously identified gas clumps with very little DM mass. (Please see Fig. B1 and Appendix B for further discussion about the foregoing points.) Another motivation for this criterion is that NSC seeds are anticipated to grow more efficiently within sufficiently deep gravitational potential wells where runaway BH merger remnants face difficulties escaping the cluster. Deeper gravitational potentials are expected in higher mass haloes.



As we simultaneously apply both of the above criteria, our gas-based seed models therefore contain three parameters, namely  $\tilde{M}_{\text{smp}}$ ,  $\tilde{M}_h$ , and  $M_{\text{seed}}^{\text{DGB}}$ . The simulation suite that will use these seed models will be referred to as `GAS_BASED`. The individual runs will be labelled as `SM*_FOF*` where the ‘\*’s correspond to the values of  $\tilde{M}_{\text{smp}}$  and  $\tilde{M}_h$  respectively. For example,  $\tilde{M}_{\text{smp}} = 5$  and  $\tilde{M}_h = 3000$  will correspond to `SM5_FOF3000`. As already mentioned, the seed masses in this suite will be  $M_{\text{seed}}^{\text{DGB}} = 2.3 \times 10^3 M_\odot$ . Lastly, we acknowledge that the use of halo-averaged properties of star-forming & metal-poor gas is less realistic than seeding based on local properties of individual dense and metal-poor gas cells (as done in Taylor & Kobayashi 2014; Tremmel et al. 2017; Bellovary et al. 2019). We do this partly because our interstellar medium is not explicitly resolved due to the effective equation-of-state description, which makes the local gas properties less reliable. Additionally, the halo averaged star-forming & metal-poor gas properties allows us to achieve excellent resolution convergence, as demonstrated in Bhowmick et al. (2021).

### 2.3.2 Stochastic seed model

As we mentioned, the key goal of this work is to build a new approach to represent low-mass seeds in larger-volume lower-resolution cosmological simulations that cannot directly resolve them. As we shall see in Section 4, this is achieved via a new stochastic seeding model. The complete details of this seed model are described in Section 4, where we thoroughly discuss their motivation and calibration using the results obtained from the `GAS_BASED` suite. Here, we briefly summarize key features so that the reader can contrast it against the gas-based seed models described in the previous subsection.

Since the simulations here will not fully resolve the  $2.3 \times 10^3 M_\odot$  DGBs, we will essentially seed their resolvable descendants. To distinguish them from the DGBs, we shall refer to these seeded descendants as *extrapolated seed descendants* or ESDs with masses (denoted by  $M_{\text{seed}}^{\text{ESD}}$ ) limited to the gas mass resolution of the simulations. In this work, we will largely explore ESD masses  $M_{\text{seed}}^{\text{ESD}} = 1.8 \times 10^4$  &  $1.5 \times 10^5 M_\odot$ , to be used for simulations with gas mass resolutions of  $\sim 10^4$  &  $10^5 M_\odot$ , respectively.

To seed the ESDs, we identify sites using the FOF algorithm, but with a shorter linking length (by factor of  $\sim 1/3$ ) compared to that used for identifying haloes. We shall refer to these short linking length FOFs as ‘best-Friends of Friends or bFOFs’. These bFOFs essentially correspond to galaxies or proto-galaxies residing inside the haloes. We do this to accommodate the formation of multiple ESDs per halo; this is because even if we seed one DGB per halo in the gas-based seed models, subsequent evolution of the hierarchical structure naturally leads to haloes occupying multiple higher mass descendants. This is because as these haloes merge with one another, their constituent BHs end up within the same halo. Notably, one could alternatively seed in subhaloes computed by SUBFIND. However, calling SUBFIND frequently enough for seeding BHs leads to a substantial slowdown in the simulations, making it unfeasible for larger cosmological runs. Hereafter, in most instances, we shall simply refer to these bFOFs as ‘galaxies’. Their properties are comprehensively studied in Section 4.1.

The ESDs will be stochastically placed in galaxies based on where the descendants of the  $2.3 \times 10^3 M_\odot$  DGBs end up within the `GAS_BASED` suite. Below we provide a brief summary of the seeding criteria

(i) *Galaxy mass criterion*: We will apply a galaxy mass (‘galaxy mass’ hereafter refers to the total mass including dark matter, gas

and stars) seeding threshold that will be stochastically drawn from galaxy mass distributions predicted for the assembly of ( $1.8 \times 10^4$  and  $10^5 M_\odot$ ) BHs that are descendants of  $2.3 \times 10^3 M_\odot$  DGBs within the `GAS_BASED` suite. As we explore further, it becomes evident that these distributions vary with redshift and exhibit significant scatter. The redshift dependence will capture the influence of halo growth, dense gas formation, and metal enrichment on seed formation in our gas-based seed models.

(ii) *Galaxy environment criterion*: In the context of a galaxy, we define its *environment* as the count of neighbouring haloes ( $N_{\text{ngb}}$ ) that exceed the mass of its host halo and are located within a specified distance (denoted by  $D_{\text{ngb}}$ ) from the host halo. We adopt this mass threshold for the neighbouring haloes because without it, the galaxy mass and galaxy environment would exhibit strong correlation; this would prevent us from independently applying the galaxy mass and galaxy environment criterion (see also Haas, Schaye & Jeesson-Daniel 2012). In this study, we determine  $N_{\text{ngb}}$  within a range of five times the virial radius ( $R_{\text{vir}}$ ) of the host halo, i.e.  $D_{\text{ngb}} = 5R_{\text{vir}}$ . This choice is suitable for investigating the immediate small-scale external surroundings of the galaxy, extending beyond its host halo. We then apply a seeding probability (less than unity) to suppress ESD formation in galaxies with  $\leq 1$  neighbouring haloes, thereby favouring seed formation in richer environments. By doing this, we account for the impact of unresolved hierarchical merger dominated growth from  $M_{\text{seed}}^{\text{DGB}}$  to  $M_{\text{seed}}^{\text{ESD}}$ , as it favours more rapid BH growth within galaxies in richer environments.

The simulations that use only the *galaxy mass criterion* will be referred to as the `STOCHASTIC_MASS_ONLY` suite. For simulations which use both *galaxy mass criterion* and *galaxy environment criterion*, we will refer to them as the `STOCHASTIC_MASS_ENV` suite. During the course of this paper, we will illustrate that the outcomes of each simulation of a specific region within the `GAS_BASED` suite, employing a distinct set of gas-based seeding parameters, can be reasonably well reproduced in a lower-resolution simulation of the same region within the `STOCHASTIC_MASS_ENV` suite. Note that by construction, the `STOCHASTIC_MASS_ONLY` and `STOCHASTIC_MASS_ENV` suite will not be able to probe those DGBs that do not undergo significant BH growth by  $z = 7$ . These DGBs that do not grow past the ESD mass can only be studied within the highest resolution `GAS_BASED` suite.

## 2.4 Simulation suite

Our simulation suite consists of zoom runs for the same overdense region as that used in Bhowmick et al. (2021) (referred to as `Zoom_REGION_z5`). At the starting redshift of  $z = 127$ , the zoom region was chosen as a  $3 \times 4 \times 3$  (Mpc/h)<sup>3</sup> initial volume from a parent uniform volume of (25 Mpc/h)<sup>3</sup>. It is targeted to produce a  $5.1 \times 10^{11} M_\odot$  halo at  $z = 5$ . The simulations were initialized using MUSIC (Hahn & Abel 2011) and run from  $z = 127$  to  $z = 7$ . The background grid’s resolution and the resolution of high-resolution zoom regions are determined by two key parameters:  $L_{\text{min}}$  (or levelmin) and  $L_{\text{max}}$  (or levelmax) respectively. These parameters define the resolution level, denoted as  $L$ , which is equivalent to the mass resolution produced by  $2^L$  number of dark matter (DM) particles per side in a uniform-resolution (25 Mpc/h)<sup>3</sup> box. Specifically, we set  $L_{\text{min}} = 7$  for the background grid, resulting in a DM mass resolution of  $7.8 \times 10^9 M_\odot$ . For the high-resolution zoom region, we explore  $L_{\text{max}}$  values of 10, 11, and 12. In addition, there is a buffer region that consists of DM particles with intermediate resolutions bridging the gap between the background grid and the zoom region. This buffer

region serves a crucial purpose of facilitating a smooth transition between the zoom region and the background grid. Our simulation suite is comprised of the following set of resolutions for the zoom regions:

(i) In our highest resolution  $L_{\max} = 12$  runs, we achieve a DM mass resolution of  $2.4 \times 10^4 M_{\odot}$  and a gas mass resolution of  $\sim 10^3 M_{\odot}$  (the gas cell masses are contingent upon the degree of refinement or derefinement of the Voronoi cells, thereby introducing some variability). These runs are used for the GAS\_BASED suite that seeds DGBs at  $2.3 \times 10^3 M_{\odot}$  using the gas-based seed models described in Section 2.3.1.

(ii) For our  $L_{\max} = 11$  & 10 runs, we achieve DM mass resolutions of  $1.9 \times 10^5$  &  $1.5 \times 10^6 M_{\odot}$  and gas mass resolutions of  $\sim 10^4$  &  $10^5 M_{\odot}$ , respectively. These runs will be used for the STOCHASTIC\_MASS\_ONLY and STOCHASTIC\_MASS\_ENV suite, that will seed ESDs at  $1.8 \times 10^4$  &  $1.5 \times 10^5 M_{\odot}$  for  $L_{\max} = 11$  & 10, respectively, using the stochastic seed models described in Section 2.3.2.

Further details of our full simulation suite are summarized in Table 1. It is important to note that our new stochastic seed models will be primarily designed for implementation within larger-volume uniform simulations. However, this paper specifically focuses on zoom simulations. In particular, we are using  $L_{\max} = 11$  & 10 zoom simulations for testing the stochastic seed models against the highest resolution  $L_{\max} = 12$  zooms that use the gas-based seed models. In a subsequent paper (Bhowmick et al., in preparation), we will be applying the stochastic seed models on uniform volume simulations of the same resolutions as the  $L_{\max} = 11$  & 10 zooms.

## 2.5 Tracing BH growth along merger trees: The SUBLINK algorithm

We use the GAS\_BASED suite to trace the growth of the lowest mass  $2.3 \times 10^3 M_{\odot}$  DGBs and study the evolution of their environments (halo and galaxy properties) as they assemble higher mass BHs. We do this by first constructing subhalo merger trees using the SUBLINK algorithm (Rodriguez-Gomez et al. 2015), which was designed for SUBFIND based subhaloes. Note that these SUBFIND based subhaloes, like bFOFs, also trace the substructure within haloes. Therefore, to avoid confusion, we shall refer to SUBFIND based subhaloes as ‘subfind-subhaloes’. It is also very common to interpret the subfind-subhaloes as ‘galaxies’. As we shall see however, in this work, we only use these subfind-subhaloes as an intermediate step to arrive at the FOF and bFOF merger trees. Therefore, there is no further mention of subfind-subhaloes after this subsection. On that note, recall again that any mention of ‘galaxy’ in our paper refers to the bFOFs.

SUBFIND was run on-the-fly to compute subfind-subhaloes within both FOF and bFOF catalogues. Therefore, for obtaining both FOF and bFOF merger trees, we first compute the merger trees of their corresponding subfind-subhaloes. The key steps in the construction of the subfind-subhalo merger tree:

- (i) For each progenitor subfind-subhalo at a given snapshot, SUBLINK determines a set of candidate descendant subfind-subhaloes from the next snapshot. Candidate descendants are those subfind-subhaloes who have common DM particles with the progenitor.
- (ii) Next, each candidate descendant is given a score based on the merit function  $\chi = \sum_i 1/R_i^{-1}$  where  $R_i$  is the binding energy rank of particle  $i$  within the progenitor. DM particles with higher binding

energy within the progenitor are given a lower rank.  $\sum_i$  denotes a sum for all the particles within the candidate descendant.

- (iii) Amongst all the candidate descendants, the final unique descendant is chosen to be the one with the highest score. This essentially ensures that the unique descendant has the highest likelihood of retaining the most bound DM particles that resided within the progenitor.

From the subfind-subhalo merger trees, we use the ones that only consist of central subfind-subhaloes (most massive within a FOF or bFOF) and construct the corresponding FOF/ halo merger trees and bFOF/galaxy merger trees. We then trace the growth of BHs along these merger trees, and the outcomes of this analysis are elaborated upon in the subsequent sections.

## 3 RESULTS I: BLACK HOLE MASS ASSEMBLY IN HIGH-RESOLUTION ZOOMS

We start our analysis by looking at the growth history of  $1.5 \times 10^3 M_{\odot}$  DGBs within the GAS\_BASED suite. We trace their growth along halo merger trees (see Section 2.5) from the time of their formation to when they assemble higher mass ( $1.8 \times 10^4$ ,  $1.5 \times 10^5$  &  $1.1 \times 10^6 M_{\odot}$ ) descendant BHs. We choose these descendant BH masses as they encompass the target gas mass resolutions of our lower resolution ( $L_{\max} = 11$  & 10) zooms. These are also comparable to typical gas mass resolutions of cosmological simulations in the existing literature. For example, the TNG100 (Nelson et al. 2018), Illustris (Vogelsberger et al. 2014a, b), EAGLE (Schaye et al. 2015), MassiveBlackII (Khandai et al. 2015), BlueTides (Feng et al. 2016) and HorizonAGN (Kaviraj et al. 2017) simulations have a gas mass resolution of  $\sim 10^6 M_{\odot}$  and similar values for the seed masses. The relatively smaller volume cosmological simulations such as ROMULUS25 (Tremmel et al. 2017) and TNG50 (Pillepich et al. 2019) have a gas mass resolution of  $\sim 10^5 M_{\odot}$  with a seed mass of  $10^6 M_{\odot}$ . Recall again that most of these simulations seed BHs simply based on either a constant halo mass threshold, or poorly resolved local gas properties. The results presented in this section will be used in Section 4 to calibrate the stochastic seed model that will represent the gas-based  $2.3 \times 10^3 M_{\odot}$  seeds in the lower-resolution zooms without resolving them directly.

### 3.1 Evolution of seed forming sites: Rapid metal enrichment after seed formation

Fig. 1 depicts the evolution of gas density, star formation rate (SFR) density, and gas metallicity at DGB forming sites from two distinct epochs ( $z = 20$  & 12). As dictated by our gas-based seed models, for each of the DGB forming sites there exists gas that is simultaneously forming stars but is also metal-poor (marked in yellow circles). However, we also find that metal enrichment has already commenced at the immediate vicinity of these DGB forming sites. In other words, DGB formation occurs in haloes where metal enrichment has already begun due to prior star formation and evolution, but it has not polluted the entire halo yet. But soon after DGB formation, i.e. within a few tens of million years, we find that the entirety of the regions becomes polluted with metals.

The rapid metal enrichment of DGB forming haloes is shown much more comprehensively and quantitatively in Fig. 2. Here we show the evolution of halo mass, dense gas mass, dense & metal-poor gas mass and gas metallicity from  $z \sim 25$ –7 for all DGB forming haloes along their respective merger trees (faded dotted lines). To avoid overcrowding of the plots, we select trees based on the most

**Table 1.** Spatial and mass resolutions within the zoom region of our simulations for various values of  $L_{\text{max}}$  (see Section 2.4 for the definition).  $M_{\text{dm}}$  is the mass of a dark matter particle,  $M_{\text{gas}}$  is the typical mass of a gas cell (note that gas cells can refine and de-refine depending on the local density), and  $\epsilon$  is the gravitational smoothing length. The 4th column represents the number of nearest gas cells that are assigned to be BH neighbours. The 5th and 6th columns correspond to the seed mass and seed model used at the different resolutions.

| $L_{\text{max}}$ | $M_{\text{dm}}$ ( $M_{\odot}$ ) | $M_{\text{gas}}$ ( $M_{\odot}$ ) | $\epsilon$ (kpc $h^{-1}$ ) | Black hole neighbours | Seed mass ( $M_{\odot}$ )                        | Seed model         |
|------------------|---------------------------------|----------------------------------|----------------------------|-----------------------|--|--------------------|
| 12               | $2.4 \times 10^4$               | $\sim 10^3$                      | 0.125                      | 256                   | $M_{\text{seed}}^{\text{DGB}} = 2.3 \times 10^3$ | Gas-based seeding  |
| 11               | $1.9 \times 10^5$               | $\sim 10^4$                      | 0.25                       | 128                   | $M_{\text{seed}}^{\text{ESD}} = 1.8 \times 10^4$ | Stochastic seeding |
| 10               | $1.5 \times 10^6$               | $\sim 10^5$                      | 0.5                        | 64                    | $M_{\text{seed}}^{\text{ESD}} = 1.5 \times 10^5$ | Stochastic seeding |

restrictive seeding criterion of  $\tilde{M}_{\text{sfrmp}} = 1000$  &  $\tilde{M}_{\text{h}} = 3000$ , but our general conclusions hold true for other seeding thresholds as well. Not surprisingly, the halo mass (1st row) and star forming gas mass (2nd row) tend to monotonically increase with decreasing redshift on average (thick solid black lines). Note that for individual trees, the halo mass can occasionally decrease with time due to tidal stripping. On more rare occasions, there may also be a sharp drop in the halo mass at a given snapshot followed by a sharp rise back to close to the original value. This is likely because the FOF finder ‘mistakenly’ splits a larger halo in two at that snapshot. The dense gas mass can also decrease with time due to the dense gas being converted to star particles.

Very importantly, the dense and metal-poor gas mass (3rd row of Fig. 2) increases initially and peaks at the time of DGB formation, following which it rapidly drops down. This happens independent of the formation redshift, and is due to the rapid metal enrichment depicted in Fig. 1. The rapid metal enrichment can be quantitatively seen in the average gas metallicity evolution (4th row of Fig. 2). We can see that even prior to the DGB formation, the average gas metallicities already start to increase from the pre-enrichment values ( $\sim 10^{-8} Z_{\odot}$ ), to  $\sim 10^{-3} Z_{\odot}$  at the time of formation. Therefore, even at the time of formation, the average metallicities of haloes are already greater than the maximum seeding threshold of  $10^{-4} Z_{\odot}$ ; however, there are still pockets of dense gas with metallicities  $\leq 10^{-4} Z_{\odot}$ , wherein DGBs form.

In Fig. 3, we select haloes that form DGBs at  $z = 15$  using gas-based seeding parameters  $\tilde{M}_{\text{sfrmp}} = 1000$  &  $\tilde{M}_{\text{h}} = 3000$ , and we show their evolution (orange circles) to  $z = 14, 13$  &  $12$  on the SFR versus halo mass plane (upper panels) and the gas metallicity versus halo mass plane (lower panels). We compare them to the full population of haloes at their respective redshifts (blue points). We investigate how biased these DGB forming haloes are compared to typical haloes of similar masses. On the SFR versus halo mass plane, the DGB forming haloes have similar SFRs compared to haloes of similar masses; not surprisingly, this continues to be so as they evolve to lower redshifts. On the metallicity versus halo mass plane, we find that DGB forming haloes have significantly lower metallicities compared to haloes of similar masses. This is a natural consequence of the requirement that the DGB forming haloes have sufficient amounts of metal-poor gas. However, due to the rapid metal enrichment of these haloes seen in Figs 1 and 2, their descendants at  $z = 14, 13$  &  $12$  end up having metallicities similar to haloes of comparable mass.

The picture that emerges from Figs 1–3 is one in which DGB-forming haloes are generally *not* a special subset of haloes (in terms of properties that persist to lower redshift), but rather they are fairly typical haloes that have the right conditions for DGB formation at a special moment in *time*. In other words, despite our seeding criterion favouring low-metallicity, star-forming haloes, their descendants still end up with similar SFRs and metallicities compared to the general population of similar-mass haloes. While

Fig. 3 only shows the evolution of DGB-forming haloes at  $z = 15$ , this general conclusion holds true for DGB-forming haloes at all redshifts. A key consequence is that the descendants of seed forming haloes can be well characterized by their halo mass distributions, largely because they are in this transient phase of rapid metal enrichment at the time of seed formation.

We utilize this characteristic of our gas-based seeding models to develop the new sub-grid seeding model for lower-resolution simulations in Section 4. Rather than requiring information about detailed properties of the descendant galaxies of these gas-based seeding sites, we show in Section 4.2 that most galaxy properties are well reproduced by simply matching the galaxy mass distribution. We then show in Section 4.3 that by additionally imposing a criterion on galaxy environment, we can robustly capture the evolved descendants of seeding sites from our high-resolution simulations.

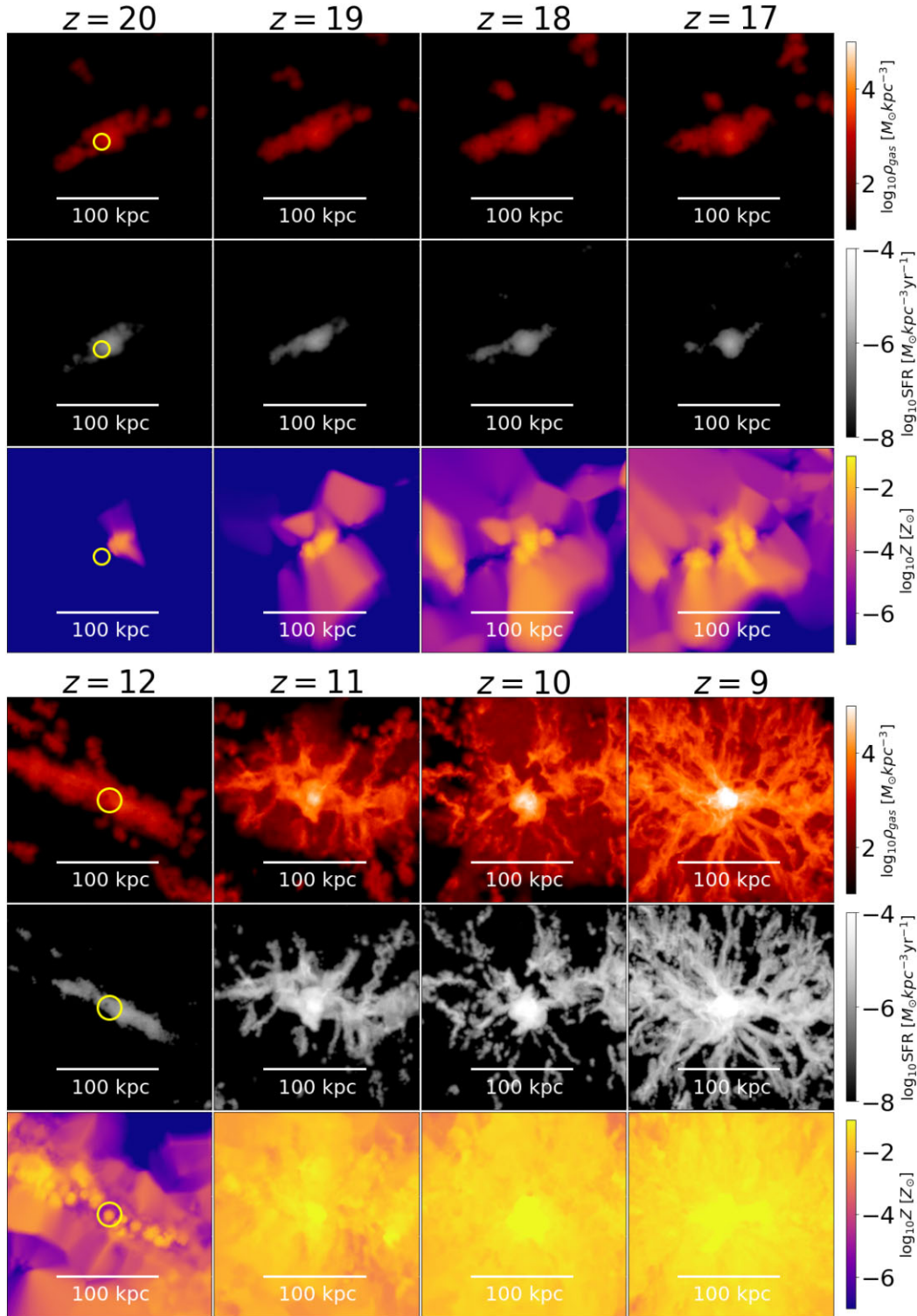
### 3.2 DGB formation and subsequent growth

We have thus far talked about the DGB forming haloes and their evolution. In this subsection, we will focus on the formation of the DGBs themselves, and their subsequent growth to assemble higher mass BHs.

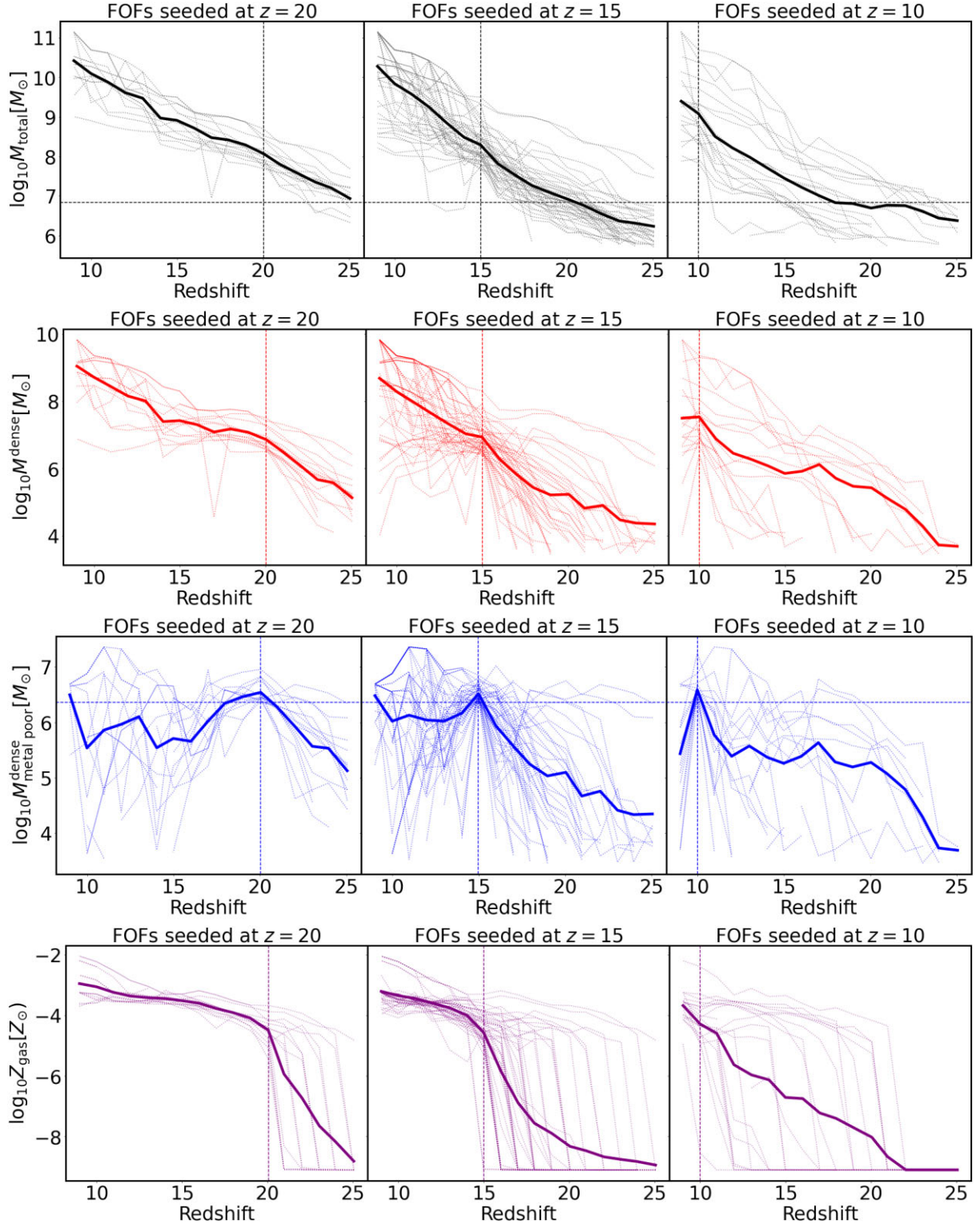
#### 3.2.1 Drivers of DGB formation: Halo growth, dense gas formation, and metal enrichment

Our gas-based seeding criteria identify three main physical processes that govern DGB formation in our simulations, i.e. halo growth, dense gas formation and metal enrichment. Halo growth and dense gas formation tend to promote DGB formation with time, whereas metal enrichment suppresses DGB formation with time. The overall rate of DGB formation at various redshifts is determined by the complex interplay between these three processes. We study this interplay in Fig. 4, wherein we show the number of haloes satisfying three different criteria:  $M_{\text{total}} > \tilde{M}_{\text{h}} \times M_{\text{seed}}^{\text{DGB}}$  (dotted line),  $M_{\text{dense}} > \tilde{M}_{\text{sfrmp}} \times M_{\text{seed}}^{\text{DGB}}$  (dashed line) and  $M_{\text{metal poor}}^{\text{dense}} > \tilde{M}_{\text{sfrmp}} \times M_{\text{seed}}^{\text{DGB}}$  (solid line).  $M_{\text{total}}$ ,  $M_{\text{dense}}$  and  $M_{\text{metal poor}}^{\text{dense}}$  correspond to the total halo mass, dense gas mass, and dense & metal-poor gas mass of haloes respectively. Amongst the above three criteria, the one that is most restrictive essentially determines the driving physical process for DGB formation at a given redshift. For example, in the rightmost panel of Fig. 4, the dotted lines have the lowest normalization from  $z \sim 25$  to  $10$ ; this implies that halo growth is primary driver and leads to the production of more DGBs with time. In the 3rd panel from the left, the solid and dashed lines have similar normalization, and both of them are lower than the dotted lines at the highest redshifts; this indicates that formation of dense gas is the key driver, which also enhances DGB formation with time. Lastly, in all of the panels, the solid lines have substantially lower



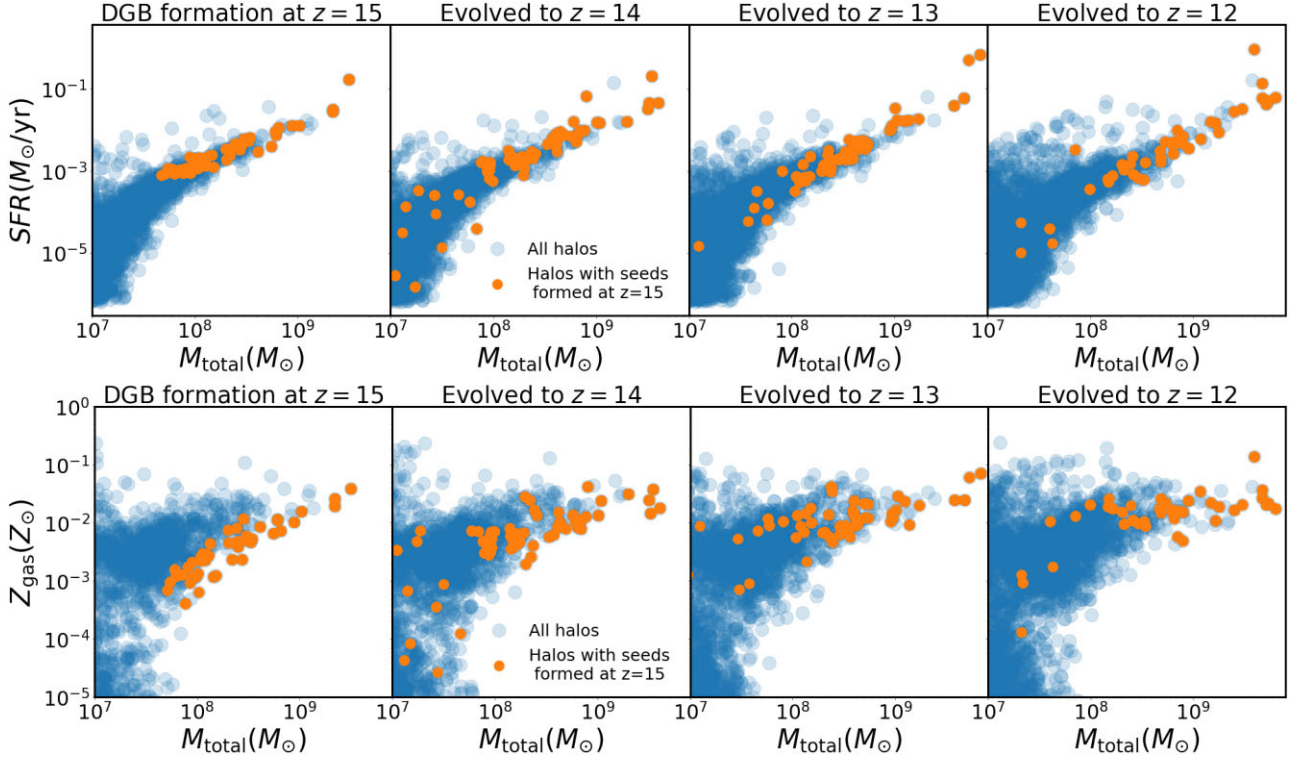


**Figure 1.** Evolution of gas density (red/orange), SFR density (greyscale), and gas metallicity (yellow/purple) of various seed forming sites in our zoom simulations that use the gas-based seed models described in Section 2.3.1. Hereafter, we shall refer to the seeds formed by the gas-based seed models as ‘Direct gas-based seeds’ or DGBs. The large panels correspond to DGB forming sites from two distinct epochs namely  $z = 20$  (top) and  $z = 12$  (bottom). Within each large panel, the leftmost sub-panel corresponds to the snapshot at the time of DGB formation, wherein the yellow circles mark the location of the formation site that contains the dense and metal-poor gas. The remaining subpanels from left to right show the evolution of that formation site along three subsequent snapshots. We can clearly see that at the time of DGB formation, the regions in the immediate vicinity of the formation site have already started the process of metal enrichment. As a result, these regions get completely polluted with metals within a very short time after DGB formation.

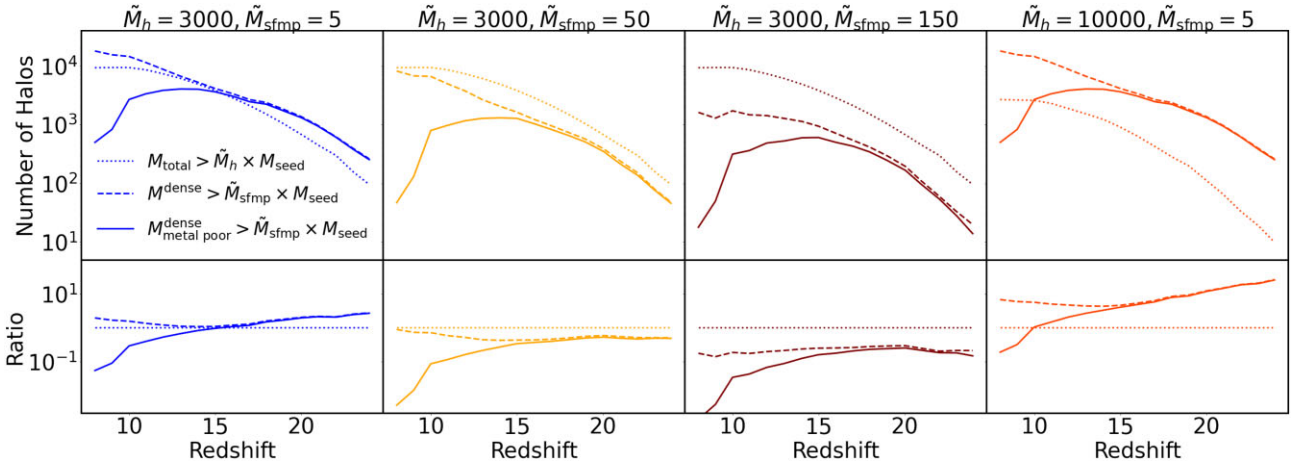


**Figure 2.** Assembly history of haloes forming  $2.3 \times 10^3 M_{\odot}$  DGBs using gas-based seed models. Top to bottom, the rows show the evolution of total halo mass ( $M_{\text{total}}$ ), dense gas mass ( $M^{\text{dense}}$ ), dense and metal-poor gas mass ( $M^{\text{dense}}_{\text{metal poor}}$ ), and gas metallicity ( $Z_{\text{gas}}$ ). Left, middle, and right panels show haloes seeded at  $z = 20$ ,  $z = 15$ , and  $z = 10$  (vertical dashed lines in each column) respectively, using the gas-based seeding criterion,  $\bar{M}_{\text{smp}} = 1000$  (horizontal dashed line in 3rd row) and  $\bar{M}_{\text{h}} = 3000$  (horizontal dashed line in 1st row). The faded dotted lines show the evolution of all DGB-forming haloes along their merger trees. The thick solid lines show the mean trend, i.e. logarithmic average of the values of all the faded dotted lines at each redshift. The dense and metal-poor gas masses tend to sharply drop soon after seeding, independent of the time of seeding. This is because the DGB forming haloes have already started to undergo rapid metal enrichment, which is shown in the fourth row by the rapid increase in gas metallicity prior to the seeding event.





**Figure 3.** The evolution of host star formation rates or SFR (top panels) and  $Z_{\text{gas}}$  (bottom panels) versus host mass is shown for  $2.3 \times 10^3 M_{\odot}$  DGBs formed at  $z = 15$ . In the leftmost panels, the filled orange circles indicate the haloes that form DGBs at  $z = 15$ . The filled orange circles in the subsequent panels (from left to right) show the same host haloes at  $z = 14, 13$  &  $12$ . The full population of haloes at each redshift is shown in blue. In other words, we select the orange circles at  $z = 15$  using our gas-based seeding criteria [ $\tilde{M}_h, \tilde{M}_{\text{sfrmp}} = 3000, 1000$ ] (assuming  $M_{\text{seed}}^{\text{DGB}} = 2.3 \times 10^3 M_{\odot}$ ), and follow their evolution on the halo merger tree. Comparing them to the full population of haloes at each redshift, we find that even though the DGB forming haloes at  $z = 15$  are biased towards lower gas metallicities at fixed halo mass (lower left panel), subsequent evolution of these haloes to lower redshifts causes them to become more unbiased at  $z = 14, 13$  &  $12$ . This is due to the rapid metal enrichment of these DGB forming haloes depicted in Fig. 2.



**Figure 4.** The upper panels show the number of haloes satisfying different cuts that were used in our gas-based seed models: dotted lines correspond to a total mass cut of  $\tilde{M}_h \times M_{\text{seed}}^{\text{DGB}}$ , dashed lines correspond to a dense gas mass cut of  $\tilde{M}_{\text{sfrmp}} \times M_{\text{seed}}^{\text{DGB}}$ , and solid lines show a dense and metal-poor gas mass cut of  $\tilde{M}_{\text{sfrmp}} \times M_{\text{seed}}^{\text{DGB}}$ . The lower panels show ratio of the normalizations w.r.t. the dotted lines from the top panel. The line with the smallest normalization determines which of the processes between halo growth versus dense gas formation versus metal enrichment is the key driver for DGB formation at a given epoch. For  $\tilde{M}_h = 3000$ , we find that metal enrichment becomes the key driver for (suppressing) DGB formation around  $z \sim 13$  for all  $\tilde{M}_{\text{sfrmp}}$  values between 5 and 150. However, when  $\tilde{M}_h = 10000$ , halo growth continues to be the primary regulator for DGB formation until  $z \sim 10$ , after which metal enrichment takes over.

normalization than both dashed and dotted lines at the lowest redshifts. In this case, metal enrichment is the primary driver, which leads to slow down and eventual suppression of DGB formation with time.

Comparing the different columns in Fig. 4, we note that the gas-based seeding parameters ( $\tilde{M}_h$  and  $\tilde{M}_{\text{sfrmp}}$ ) have a strong influence in determining which process dominantly drives DGB formation at various redshifts. For  $\tilde{M}_h = 3000$  and  $\tilde{M}_{\text{sfrmp}} = 5$  (leftmost panel),

halo growth is the key driver for DGB formation from  $z \sim 30$  to 15; at  $z \lesssim 15$ , metal enrichment becomes the primary driver and slows down DGB formation. When  $\tilde{M}_h$  is fixed at 3000 and  $\tilde{M}_{\text{smp}}$  is increased to 50 or 150 (2nd and 3rd panels, respectively), dense gas formation replaces halo growth to become the primary driver for DGB formation at  $z \sim 30$ –15; however, metal enrichment continues to be the main driver in slowing down DGB formation at  $z \lesssim 15$ . Finally, when  $\tilde{M}_{\text{smp}}$  is fixed at 5 and  $\tilde{M}_h$  is increased to 10000 (rightmost panels), halo growth becomes the key driver for DGB formation from  $z \sim 30$  to 10. In this case, metal enrichment takes the driving seat at a lower redshift of  $z \sim 10$  compared to the cases when  $\tilde{M}_h = 3000$ .

To further summarize the above findings from Fig. 4, we find that when  $\tilde{M}_h$  is 3000, DGB formation is ramped up by either dense gas formation or halo growth until  $z \sim 15$ . After  $z \sim 15$ , it is slowed down by metal enrichment. But when  $\tilde{M}_h = 10000$ , the halo mass criterion becomes much more restrictive and halo growth continues to ramp up DGB formation until  $z \sim 10$  before it is slowed down by metal enrichment. In the next subsection, we shall see the implications of the above on the rates of DGB formation at various redshifts.

### 3.2.2 Formation rates of $\sim 10^3 M_\odot$ DGBs

The leftmost panel of Fig. 5 shows the formation rates of  $2.3 \times 10^3 M_\odot$  DGBs for the different gas-based seed models. The interplay between halo growth, dense gas formation and metal enrichment discussed in the previous subsection is readily seen in the DGB formation rates. For  $\tilde{M}_h = 3000$  and  $\tilde{M}_{\text{smp}} = 5, 50, 150$  & 1000, we find that DGB formation ramps up as the redshift decreases from  $z \sim 30$  to 15, driven pre-dominantly either by halo growth (for  $\tilde{M}_{\text{smp}} = 5$ ) or dense gas formation (for  $\tilde{M}_{\text{smp}} = 50, 150$  & 1000). As the redshift decreases below  $z \sim 15$ , metal enrichment significantly slows down DGB formation. However, when  $\tilde{M}_h$  is increased to 10000 (red line), halo growth continues to ramp up DGB formation till  $z \sim 10$ , after which the suppression of DGB formation due to metal enrichment takes place. Note also that at  $z \lesssim 10$ , DGB formation is finally strongly suppressed due to metal pollution for all the seed models. This is because most of the regions of newly dense gas are already metal enriched by then, likely due to stellar feedback dispersing the metals throughout the simulation volume.

### 3.2.3 Assembly rates of $\sim 10^4$ – $10^6 M_\odot$ BHs from $\sim 10^3 M_\odot$ seeds

The assembly rates of  $1.8 \times 10^4$ ,  $1.5 \times 10^5$  &  $1.1 \times 10^6 M_\odot$  BHs are shown in 2nd, 3rd, and 4th panels of Fig. 5, respectively. As in Bhowmick et al. (2021), we find that nearly 100 per cent of the growth of these DGBs is happening via mergers. This is partly due to the  $M_{\text{BH}}^2$  scaling of Bondi Hoyle accretion rates, which leads to much slower accretion onto low mass DGBs, and it is consistent with the findings of Taylor & Kobayashi (2014; see fig. 3 in their paper).

Let us first focus on the impact of this merger dominated growth on the assembly of  $1.8 \times 10^4 M_\odot$  BHs (2nd panel of Fig. 5). They generally assemble at rates  $\sim 50$ –80 times lower than the rates at which  $2.3 \times 10^3 M_\odot$  DGBs form. Notably, the trends seen in the DGB formation rates directly reflect upon the rates at which  $1.8 \times 10^4 M_\odot$  BHs assemble. In particular, for  $\tilde{M}_h = 3000$  and  $\tilde{M}_{\text{smp}} = 5, 50$  & 150, we see an increase in the assembly rates as the redshift decreases from  $z \sim 25$  to 15 wherein DGB formation is driven by halo growth or dense gas formation. The assembly rates slow down at  $z \lesssim 15$  as metal enrichment slows down DGB formation. For a higher value of  $\tilde{M}_h = 10000$ , halo growth continues to increase the assembly rates until  $z \sim 10$ , before metal enrichment slows it down.

Overall, these results suggest that the interplay of halo growth, dense gas formation and metal enrichment processes that we witnessed on the formation rates of  $2.3 \times 10^3 M_\odot$  DGBs, are also retained in the assembly rates of their higher mass  $1.8 \times 10^4 M_\odot$  descendants.

We also see the assembly of a handful of  $1.5 \times 10^5$  and  $1.1 \times 10^6 M_\odot$  BHs (3rd and 4th panels of Fig. 5).  $1.5 \times 10^5 M_\odot$  BHs generally start assembling at  $z \lesssim 15$  and  $1.1 \times 10^6 M_\odot$  BHs assemble at  $z \lesssim 12$ . However, any potential trends similar to that identified in the previous paragraph for  $1.8 \times 10^4 M_\odot$  descendants, are difficult to discern for the  $1.5 \times 10^5$  and  $1.1 \times 10^6 M_\odot$  descendants due to very limited statistical power.

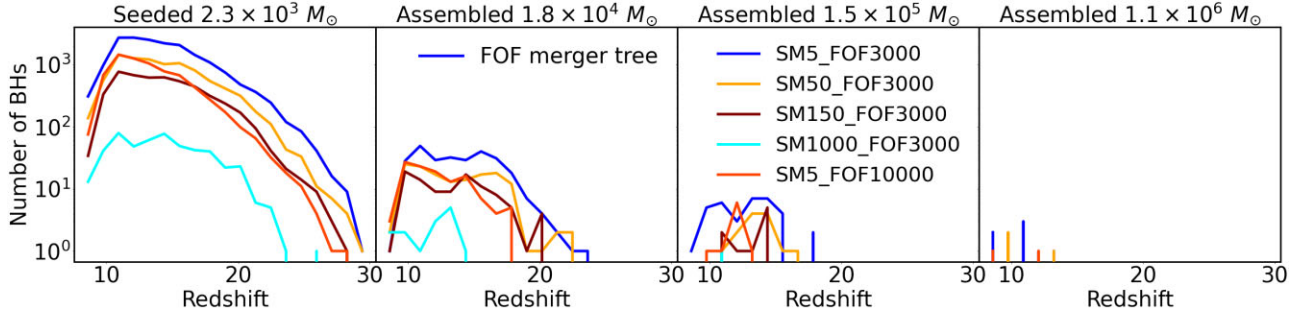
### 3.3 In which host haloes do the $\sim 10^4$ – $10^6 M_\odot$ descendant BHs assemble?

Fig. 6 shows the host halo masses (denoted by  $M_{\text{total}}^{\text{halo}}$ ) and redshifts at which  $2.3 \times 10^3 M_\odot$  DGBs form (leftmost panel), followed by the assembly of  $1.8 \times 10^4 M_\odot$  and  $1.5 \times 10^5 M_\odot$  BHs (middle and right-hand panels, respectively). Broadly speaking,  $2.3 \times 10^3 M_\odot$  DGBs form in  $\sim 10^{6.7}$ – $10^{7.7} M_\odot$  haloes,  $1.8 \times 10^4 M_\odot$  BHs assemble in  $\sim 10^{7.7}$ – $10^{8.7} M_\odot$  haloes, and  $1.5 \times 10^5 M_\odot$  BHs assemble in  $\sim 10^{8.7}$ – $10^{9.7} M_\odot$  haloes. Therefore, rates of BH growth versus halo growth are broadly similar. This is a natural expectation from merger-dominated BH growth, since the BH mergers crucially depend on the merging of their host haloes. Note, however, that in the absence of our currently imposed BH repositioning scheme that promptly merges close enough BH pairs, we could expect larger differences between the merger rates of BHs and their host haloes.

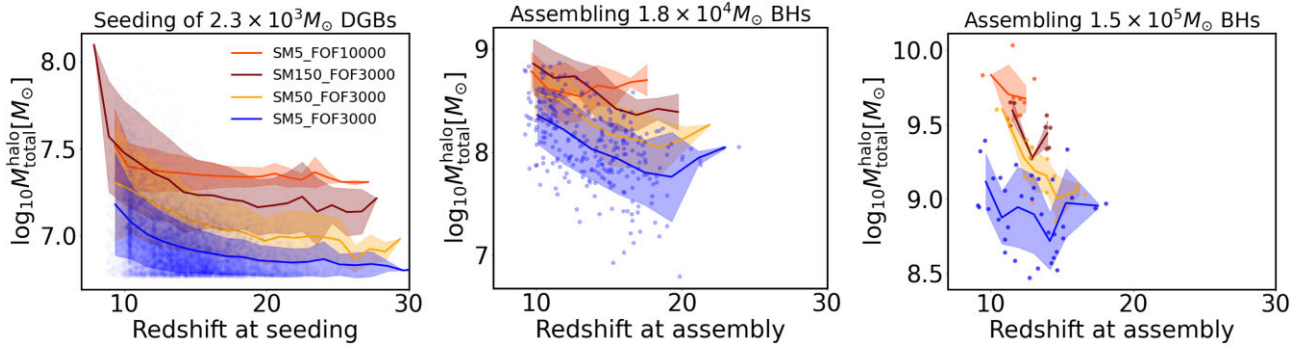
The interplay between halo growth, dense gas formation and metal enrichment at different redshifts (as noted in Section 3.2) profoundly influences the redshift evolution of the halo masses in which the seeding of  $2.3 \times 10^3 M_\odot$  DGBs and assembly of higher-mass BHs take place. Let us first focus on the seeding of  $2.3 \times 10^3 M_\odot$  DGBs (Fig. 6: left-hand panel).

We find for  $\tilde{M}_h = 3000$  &  $\tilde{M}_{\text{smp}} = 50, 150$  that the halo masses steadily increase with time as dense gas formation drives the formation of DGBs. As described in more detail in Appendix B, this is a simple consequence of cosmological expansion, which makes it more difficult for the gas to cool and form stars at later times within haloes of a fixed mass. Notably, as metal enrichment gradually takes over at  $z \lesssim 15$ , the redshift evolution becomes substantially steeper, pushing DGB formation towards even more massive haloes at later times. This may seem counterintuitive since we expect more massive haloes to have stronger metal enrichment, which should suppress DGB formation within them. However, more massive haloes also generally have higher overall dense gas mass, a portion of which may remain metal-poor since star-forming haloes are not fully metal enriched instantaneously. As it turns out in our simulations, when metal enrichment increases, it favours DGB formation in more massive haloes because they are more likely to have a sufficient amount of dense and metal-poor gas mass. For further details on this, the reader can refer to Appendix B. When  $\tilde{M}_h$  is increased to 10000, the redshift evolution of DGB forming halo mass is flat until  $z \sim 10$  since the seed formation is primarily driven by the *halo mass criterion*. It is only after  $z \sim 10$  that the DGB forming halo mass starts to steeply increase due to the full influence of metal enrichment.

The above trends directly impact the redshift evolution of the host halo masses in which  $1.8 \times 10^4 M_\odot$  assemble (middle panel of Fig. 6). For the model with a stricter halo mass criterion (i.e.  $\tilde{M}_h = 10000$  &  $\tilde{M}_{\text{smp}} = 5$ ), the transition in the slope of the  $M_{\text{total}}^{\text{halo}}$  versus redshift relation occurs much later (transition occurs between  $z \sim 12$  and 10) compared to models with more lenient halo mass



**Figure 5.** We trace the growth of  $2.3 \times 10^3 M_\odot$  DGBs (leftmost panels) along merger trees and show the redshifts when they assemble BHs of masses  $1.8 \times 10^4 M_\odot$ ,  $1.5 \times 10^5 M_\odot$ , and  $1.1 \times 10^6 M_\odot$  (2nd, 3rd, and 4th panels from the left). Different colours correspond to the different gas-based seed models with varying  $\tilde{M}_{\text{smp}} = 5, 50, 150$  &  $1000$ ,  $\tilde{M}_h = 3000$  and  $\tilde{M}_{\text{smp}} = 5$ ,  $\tilde{M}_h = 10000$ . We find that the impacts of increasing  $\tilde{M}_{\text{smp}}$  and  $\tilde{M}_h$  are qualitatively distinguishable. For  $\tilde{M}_h = 3000$  and  $\tilde{M}_{\text{smp}} = 5 - 1000$ , metal enrichment starts to slow down DGB formation around  $z \sim 15$ . In contrast, when  $\tilde{M}_h$  is increased from 3000 to 10 000, the slow down of DGB formation due to metal enrichment starts much later ( $z \lesssim 10$ ). Similar trends are seen in the assembly rates of higher mass descendants (particularly  $1.8 \times 10^4 M_\odot$  BHs).



**Figure 6.** The left panel shows the redshifts and the FOF total masses at which  $2.3 \times 10^3 M_\odot$  DGBs form. Middle and right panels show the redshifts and the FOF total masses at which  $1.8 \times 10^4 M_\odot$  and  $1.5 \times 10^5 M_\odot$  descendant BHs, respectively, assemble on the FOF merger tree. The different colors correspond to different gas-based seed models. Each data point corresponds to a single instance of assembly or seeding. We only show data points for a limited set of models to avoid overcrowding. Solid lines show the mean trend and the shaded regions show  $\pm 1\sigma$  standard deviations. We find that as metal enrichment takes over as the driving force and suppresses DGB formation at lower redshifts, DGBs form in increasingly massive haloes. This also drives a similar redshift dependence for the assembly of  $1.8 \times 10^4 M_\odot$  BHs.

criterion  $\tilde{M}_h = 3000$  &  $\tilde{M}_{\text{smp}} = 5 - 150$  ( $z \gtrsim 15$ ). This, again, is because metal enrichment starts to suppress DGB formation much later in the model with stricter halo mass criterion. Finally, for the assembly of  $1.5 \times 10^5 M_\odot$  BHs, the redshift evolution of the host halo masses cannot be robustly deciphered due to statistical uncertainties. But here too, we see hints of higher host halo masses at lower redshifts in regimes where metal enrichment is the primary driver for (the suppression of) DGB formation.

Overall, the impact of halo growth, dense gas formation and metal enrichment on DGB formation is well imprinted in the redshift evolution of the host halo masses within which their descendant BHs assemble. We shall see in later sections how this fact is going to be crucial in building the new seed model to represent (descendants of)  $2.3 \times 10^3 M_\odot$  DGBs in lower-resolution simulations.

#### 4 RESULTS II: A NEW STOCHASTIC SEED MODEL FOR LARGER SIMULATIONS

We have thus far traced the growth of low mass ( $2.3 \times 10^3 M_\odot$ ) DGBs born in regions with dense and metal-poor gas, in order to determine the host properties of their higher-mass ( $1.8 \times 10^4$  &  $1.5 \times 10^5 M_\odot$ ) descendant BHs. We will now use these results to build a new stochastic seed model that can represent these  $2.3 \times 10^3 M_\odot$  DGBs within simulations that cannot directly resolve them. In Section 2.3.2,

we gave a brief introduction of this seed model and mentioned that this model would rely on a *galaxy mass criterion* and a *galaxy environment criterion*. Here, we detail the motivation, construction, and calibration of both of these seeding criteria and demonstrate that the resulting model can reproduce reasonably well the high-resolution, gas-based seed model predictions in lower-resolution simulations.

Note that some of our gas-based seed parameter combinations do not produce enough descendant BHs in our zoom region to perform a robust calibration. These include  $\tilde{M}_h = 3000$ ;  $\tilde{M}_{\text{smp}} = 1000$  for the  $1.8 \times 10^4 M_\odot$  descendants and  $\tilde{M}_h = 3000$  &  $10000$ ;  $\tilde{M}_{\text{smp}} = 150$  &  $1000$  for the  $1.5 \times 10^5 M_\odot$  descendants. Therefore, we shall not consider these parameter values hereafter.

In the stochastic seed model, we will directly seed the descendants with initial masses set by the gas mass resolution ( $1.8 \times 10^4$  &  $1.5 \times 10^5 M_\odot$  in  $L_{\text{max}} = 11$  &  $10$  respectively). As already mentioned in Section 2.3.2, because these massive seeds are meant to represent descendants of  $2.3 \times 10^3 M_\odot$  DGBs that cannot be resolved directly, we refer to the former as ‘extrapolated seed descendants’ or ESDs with initial mass denoted by  $M_{\text{seed}}^{\text{ESD}}$ . In other words, our new stochastic seeding prescription will place ESDs with  $M_{\text{seed}}^{\text{ESD}}$  set by the gas mass resolution of  $1.8 \times 10^4$  or  $1.5 \times 10^5 M_\odot$ , but they are intended to represent our gas-based seed models with unresolvable  $2.3 \times 10^3 M_\odot$  DGBs. To that end, the next few subsections address the following



question: *How do we build a new seed model that can capture the unresolved growth phase from  $M_{\text{seed}}^{\text{DGB}} = 2.3 \times 10^3 M_{\odot}$  to  $M_{\text{seed}}^{\text{ESD}} = 1.8 \times 10^4$  or  $1.5 \times 10^5 M_{\odot}$ ?*

#### 4.1 Seeding sites for ESDs: ‘Best Friends of Friends (bFOF)’ galaxies

It is common practice in many (but not all) cosmological simulations to place one seed per halo at a given time step. The advantage to this is that the halo properties (particularly the total halo mass) show much better resolution convergence compared to the local gas properties. However, this is not quite realistic, as haloes typically have a significant amount of substructure and can therefore have multiple seeding sites at a given time. Despite this, subhaloes are not typically used to seed BHs, likely because on-the-fly subhalo finders like SUBFIND are much more computationally expensive compared to on-the-fly halo finders like the FOF finder.

Recall that in our gas-based seed model,  $2.3 \times 10^3 M_{\odot}$  DGBs were also seeded as ‘one seed per halo’. But even in this case, as these smaller seed-forming haloes and their BHs undergo mergers, configurations with multiple  $1.8 \times 10^4$  or  $1.5 \times 10^5 M_{\odot}$  BHs per halo tend to naturally emerge. We emulate this in our new seed model by seeding ESDs within bFOFs introduced in Section 2.3.2. The linking length for the bFOFs was chosen to be 1/3rd of the value adopted for standard FOF haloes (which is 0.2 times the mean particle separation). This value was chosen after exploring a number of possibilities. On one hand, a much larger linking length does not resolve the substructure adequately. On the other hand, if the linking length is much smaller, a significant number of FOFs end up not containing any bFOFs.

Fig. 7 summarizes the bFOF properties in relation to the familiar FOF haloes at  $z = 8$ . The leftmost panel shows the relationship between the masses of FOFs and bFOFs. Within a FOF, the most massive bFOF is assigned as the ‘central bFOF’ (blue circles) and the remaining bFOFs are assigned as the ‘satellite bFOFs’ (orange circles). The central bFOFs are about  $\sim 7$  times less massive than the host FOF. Not surprisingly, the satellite bFOFs span a much wider range of masses all the way down to the lowest possible masses at the bFOF/FOF identification limit ( $\geq 32$  DM particles). The middle panel of Fig. 7 shows the bFOF occupation statistics for FOFs of different masses. More massive FOFs tend to host a higher number of bFOFs; the most massive  $\sim 4 \times 10^{10} M_{\odot}$  FOF has about  $\sim 4 \times 10^3$  bFOFs. We can see that in addition to the central bFOF, the satellite bFOFs can also contain BHs (orange, green and maroon points in the middle panel). To that end, the right-hand panel of Fig. 7 shows the total BH occupations inside FOFs and bFOFs as a function of their respective masses. We can clearly see that while individual FOFs can contain multiple BHs (up to a few tens), the vast majority of individual bFOFs contain 0 or 1 BHs. In fact, amongst the  $\sim 30\,000$  bFOFs at  $z = 8$ , only 12 of them have more than 1 BH. These results generally hold true at all redshifts.

By building our seed model based on bFOFs instead of FOFs (i.e. one ESD per bFOF), we expect to naturally place multiple  $1.8 \times 10^4 M_{\odot}$  or  $1.5 \times 10^5 M_{\odot}$  ESDs in individual haloes. As a result, we will successfully capture situations where multiple  $1.8 \times 10^4 M_{\odot}$  or  $1.5 \times 10^5 M_{\odot}$  descendant BHs assemble from  $2.3 \times 10^3 M_{\odot}$  DGBs in a single halo within close succession. As mentioned in Section 2.3.2, these bFOFs are essentially the sites where high- $z$  (proto)galaxies reside; we therefore use the phrase ‘galaxies’ to refer to these bFOFs.

#### 4.2 Building the galaxy mass criterion

Recall from Section 3.1 that because DGB formation in our gas-based seeding model occurs during a transient phase of rapid metal enrichment, their descendants have metallicities (and SFRs) that are typical for haloes with similar total masses. This motivates us to first explore low-resolution simulations with seeding criterion that simply matches the galaxy mass distribution of seeding sites in our high-resolution, gas based models. We refer to this seeding criterion as the *galaxy mass criterion*; notably, this differs from typical halo-mass-based seeding models in the use of a distribution of host mass thresholds rather than a single value. The corresponding simulations are referred to as STOCHASTIC\_MASS\_ONLY.

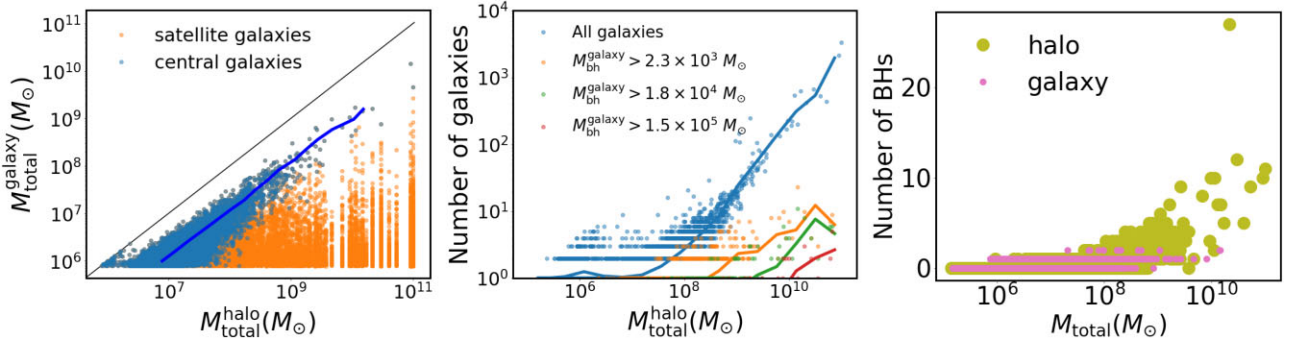
##### 4.2.1 Galaxy masses at assembly of $\sim 10^4$ & $10^5 M_{\odot}$ BHs from $\sim 10^3 M_{\odot}$ seeds

To calibrate our seed models, we first determine the galaxy masses ( $M_{\text{total}}^{\text{galaxy}}$ ) in which  $1.8 \times 10^4 M_{\odot}$  and  $1.5 \times 10^5 M_{\odot}$  BHs assemble from  $2.3 \times 10^3 M_{\odot}$  DGBs within our GAS\_BASED simulations; these are shown in Fig. 8. Let us first focus on the assembly of  $1.8 \times 10^4 M_{\odot}$  descendants (Fig. 8, top panels). Similar to that of  $M_{\text{total}}^{\text{halo}}$  versus redshift relations (Fig. 6, middle panel), the  $M_{\text{total}}^{\text{galaxy}}$  versus redshift relations show features that reflect the interplay between halo growth, dense gas formation, and metal enrichment in influencing DGB formation. For  $\tilde{M}_h = 3000$ ,  $\tilde{M}_{\text{sfrp}} = 50$  &  $150$ , we see that the slope of redshift evolution of the mean (denoted by  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  and shown as solid lines) undergoes a gradual transition between  $z \sim 13$  and  $15$ . This corresponds to the slow down of DGB formation due to metal enrichment. When  $\tilde{M}_h = 10000$  &  $\tilde{M}_{\text{sfrp}} = 5$ , this transition occurs at comparatively lower redshifts ( $z \sim 12$ – $10$ ) as the influence of metal enrichment starts later due to the higher  $\tilde{M}_h$ . We then fit the mean trend by a double power law (dashed lines in Fig. 8, upper panels) given by

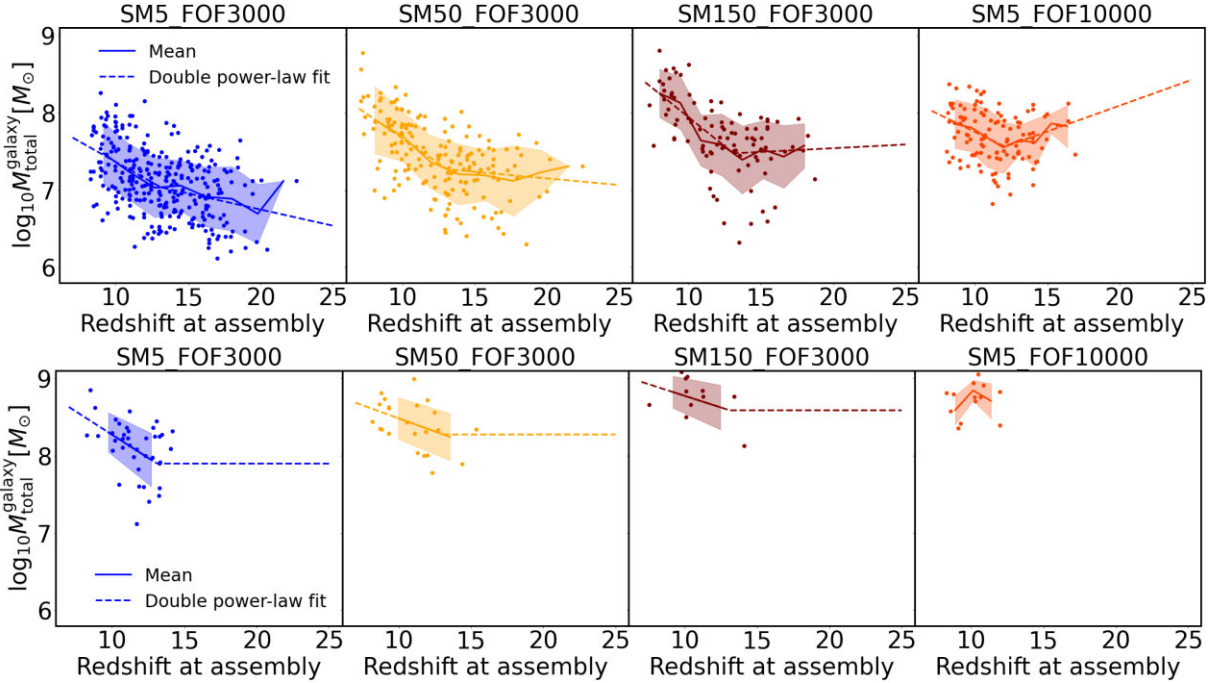
$$\log_{10} \langle M_{\text{total}}^{\text{galaxy}} \rangle = \begin{cases} (z - z_{\text{trans}}) \times \alpha + \log_{10} M_{\text{trans}}, & \text{if } z \geq z_{\text{trans}} \\ (z - z_{\text{trans}}) \times \beta + \log_{10} M_{\text{trans}}, & \text{if } z < z_{\text{trans}} \end{cases} \quad (5)$$

$z_{\text{trans}}$  roughly marks the transition in the driving physical process for DGB formation. For  $z > z_{\text{trans}}$ , halo growth or dense gas formation primarily drives DGB formation; for  $z < z_{\text{trans}}$ , metal enrichment takes over as the primary driver to suppress DGB formation.  $M_{\text{trans}}$  is the value of  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  at the transition redshift. Finally,  $\alpha$  and  $\beta$  are the slopes of the  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  versus redshift relation at  $z > z_{\text{trans}}$  and  $z < z_{\text{trans}}$  respectively. To simplify our fitting procedure, we first select  $z_{\text{trans}}$  for each of the cases via visual inspection and determine  $M_{\text{trans}}$  by interpolating the  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  versus redshift relation. We then fit for  $\alpha$  and  $\beta$  using the `scipy.optimize.curve_fit` python package. Note that the double power-law function assumes a sharp transition in the  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  versus redshift relation at  $z = z_{\text{trans}}$ . However, as we can see in Fig. 8, this transition occurs much more gradually as metal enrichment starts to slow down and eventually suppresses DGB formation. Nevertheless, the double power-law model offers a simple (albeit approximate) framework to capture the intricate convolution of the impact of halo growth, dense gas formation and metal enrichment that leads to the initial rise and eventual suppression of DGB formation.

The values of  $z_{\text{trans}}$ ,  $M_{\text{trans}}$ ,  $\alpha$  and  $\beta$  for the different gas-based seed models are listed in the top four rows of Table 2. We choose  $z_{\text{trans}} = 13.1$  for  $\tilde{M}_h = 3000$ ,  $\tilde{M}_{\text{sfrp}} = 5, 50$  &  $150$ .  $z_{\text{trans}}$  is the same for all three  $\tilde{M}_{\text{sfrp}}$  values to encode that the slow down of seed formation due to metal enrichment starts at similar redshifts for all these models.



**Figure 7.** Introduction to best friends of friends (bFOF) galaxies, which are identified using the FOF algorithm but with one-third of the linking length used for identifying haloes: Lefthand panel shows the relation between halo mass and the mass ( $M_{\text{total}}^{\text{galaxy}}$ ) of the central or most massive bFOF in blue, and satellite bFOF in orange. On an average, the central bFOFs are  $\sim 7$  times less massive than their host FOFs, but with substantial scatter ( $\gtrsim 1$  dex) for fixed FOF mass ( $M_{\text{total}}^{\text{halo}}$ ). The middle panel shows the number of bFOFs for FOFs of different total masses. The plots are shown at  $z = 8$  and for the gas-based seed model [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 3000, 5$ ]. Blue color shows all bFOFs (with or without BHs); orange, green and maroon lines show bFOFs with a total BH mass of  $2.3 \times 10^3$ ,  $1.8 \times 10^4$ , and  $1.5 \times 10^5 M_{\odot}$ , respectively. Right-hand panel shows the number of BHs occupied by FOFs and bFOFs. While  $\gtrsim 12$  per cent of FOFs contain multiple BHs (up to  $\sim 30$ ), only  $\sim 1$  per cent of bFOFs contain multiple BHs. All this motivates us to use bFOFs as seeding sites (instead of FOFs) in our new stochastic seed models that would be able to represent the lowest mass ( $\sim 10^3 M_{\odot}$ ) DGBs in lower resolution simulations that cannot directly resolve them. These bFOFs are essentially sites of (proto)galaxies residing within the high- $z$  haloes. We hereafter refer to these bFOFs as ‘galaxies’.



**Figure 8.** Top and bottom rows show the redshifts and the galaxy total masses ( $M_{\text{total}}^{\text{galaxy}}$  that includes DM, gas and stars) at which  $1.8 \times 10^4$  and  $1.5 \times 10^5 M_{\odot}$  BHs, respectively, assemble from  $2.3 \times 10^3 M_{\odot}$  DGBs when the BH growth is traced along the galaxy merger tree. The 1st, 2nd and 3rd columns show different gas-based seeding models with  $\tilde{M}_h = 3000$  and  $\tilde{M}_{\text{smp}} = 5, 50$  &  $150$ . The 4th column shows  $\tilde{M}_h = 10000$  and  $\tilde{M}_{\text{smp}} = 5$ . Solid lines show the mean trend and the shaded regions show  $\pm 1\sigma$  standard deviations. We find that for all the models, there is a transition in the slope of the mean trend at redshift  $z \equiv z_{\text{trans}} \sim 12$ – $13$ , which is driven by the suppression of seed formation by metal enrichment. The trends are reasonably well fit by a double power law (dashed lines). These fits are used in our stochastic seed models that directly seed the descendants (referred to as ‘extrapolated seed descendants or ESDs’) at  $1.8 \times 10^4 M_{\odot}$  or  $1.5 \times 10^5 M_{\odot}$  within the lower resolution  $L_{\text{max}} = 11$  &  $10$  zooms, respectively. To obtain fits in the top row, we first assumed  $z_{\text{trans}} = 13.1$  for  $\tilde{M}_h = 3000, \tilde{M}_{\text{smp}} = 5, 50$  &  $150$ , and  $z_{\text{trans}} = 12.1$  for  $\tilde{M}_h = 10000, \tilde{M}_{\text{smp}} = 5$  via a visual inspection. The fits were then performed to obtain the slopes at  $z < z_{\text{trans}}$  and  $z > z_{\text{trans}}$  using `scipy.optimize.curve_fit`. The final fitted parameters are shown in Table 2.

For  $\tilde{M}_h = 10000, \tilde{M}_{\text{smp}} = 5$ , we choose a lower transition redshift of  $z_{\text{trans}} = 12.1$  as halo growth continues to drive up seed formation up to lower redshifts compared to the models with  $\tilde{M}_h = 3000$ .

The impact of  $\tilde{M}_h$  and  $\tilde{M}_{\text{smp}}$  on  $M_{\text{trans}}$ ,  $\alpha$  and  $\beta$  is noteworthy. As  $\tilde{M}_h$  or  $\tilde{M}_{\text{smp}}$  increases, the value of  $M_{\text{trans}}$  also increases to generally

reflect the fact that descendant BHs of a fixed mass are assembling in more massive haloes.  $\alpha$  is significantly more sensitive to  $\tilde{M}_{\text{smp}}$  compared to  $\tilde{M}_h$ ; this is not surprising as  $\alpha$  corresponds to the regime where metal enrichment primarily governs seed formation. A higher value of  $\tilde{M}_{\text{smp}}$  produces a steeper  $\alpha$ , as it leads to stronger suppression

**Table 2.** Fiducial model parameters for the stochastic seed model, calibrated for each of the gas-based seeding parameters. Columns 1 and 2 show the gas-based seeding parameters  $\tilde{M}_h$  and  $\tilde{M}_{\text{sfrmp}}$ . For each set of  $\tilde{M}_h$  and  $\tilde{M}_{\text{sfrmp}}$  values, the remaining columns list the parameters of the stochastic seed model. Columns 3–7 show the parameter values used for the *galaxy mass criterion*, which are derived from gas-based seed model predictions of the  $M_{\text{total}}^{\text{galaxy}}$  versus redshift relations (Fig. 8).  $z_{\text{trans}}$ ,  $M_{\text{trans}}$ ,  $\alpha$ , &  $\beta$  are obtained by fitting the mean trends using the double power-law function shown in equation (5).  $\sigma$  is the standard deviation. Columns 8–10 show the parameter values for the *galaxy environment criterion* (i.e.  $p_0$ ,  $p_1$ , and  $\gamma$ ). These are obtained by exploring a range of possible values to find the best match with the small-scale BH clustering and overall BH counts predicted by the gas-based seed model. Note that these calibrations may be different for other simulations with different sets of sub-grid models for dense gas formation, metal enrichment and stellar feedback. They also show some dependence on the overdensity of the zoom region i.e. they are subject to cosmic variance.

| $\tilde{M}_{\text{sfrmp}}$   | $\tilde{M}_h$ | $z_{\text{trans}}$ | $\log_{10} M_{\text{trans}} [\text{M}_{\odot}]$ | $\alpha$ | $\beta$ | $\sigma$ | $p_0$ | $p_1$ | $\gamma$ |
|--|---------------|--------------------|---|----------|---------|----------|-------|-------|----------|
| $M_{\text{seed}}^{\text{ESD}} = 1.8 \times 10^4 \text{ M}_{\odot}$ |               |                    |   |          |         |          |       |       |          |
| 5  | 3000          | 13.1               | 7.03  | −0.105   | −0.041  | 0.330    | NA    | NA    | NA       |
| 50   | 3000          | 13.1               | 7.26  | −0.128   | −0.017  | 0.319    | 0.1   | 0.3   | 1.6      |
| 150  | 3000          | 13.1               | 7.47  | −0.151   | 0.009   | 0.360    | 0.1   | 0.3   | 1.6      |
| 5  | 10 000        | 12.1               | 7.56  | −0.091   | 0.067   | 0.278    | 0.2   | 0.4   | 1.2      |
| $M_{\text{seed}}^{\text{ESD}} = 1.5 \times 10^5 \text{ M}_{\odot}$ |               |                    |   |          |         |          |       |       |          |
| 5  | 3000          | 13.1               | 7.89  | −0.120   | 0       | 0.246    | 0.2   | 0.4   | 1.2      |
| 50   | 3000          | 13.1               | 8.27  | −0.067   | 0       | 0.286    | 0.2   | 0.4   | 1.2      |
| 150  | 3000          | 13.1               | 8.58  | −0.060   | 0       | 0.298    | 0.2   | 0.4   | 1.2      |

of DGB formation by metal enrichment. Lastly,  $\beta$  is impacted by both  $\tilde{M}_{\text{sfrmp}}$  and  $\tilde{M}_h$ . This also makes sense because  $\beta$  corresponds to the regime where either dense gas formation or halo growth can drive seed formation. Increasing  $\tilde{M}_{\text{sfrmp}}$  enhances the role of dense gas formation, and increasing  $\tilde{M}_h$  enhances the role of halo growth. Generally, we see that as the number of DGBs forming at the highest redshifts is decreased due to increase in  $\tilde{M}_h$  or  $\tilde{M}_{\text{sfrmp}}$ ,  $\beta$  tends to go from negative to positive values thereby favouring higher  $M_{\text{total}}^{\text{galaxy}}$  at higher redshifts. This is likely because when BHs are very few, merger driven growth is slow and galaxies have more time to grow via DM accretion between successive mergers. As a result, galaxy growth is slightly faster than merger dominated BH growth at these highest redshifts where there are very few BHs.

We now turn our attention to the assembly of  $1.5 \times 10^5 \text{ M}_{\odot}$  descendant BHs (bottom panels of Fig. 8). In this case, we do not have adequate statistics to robustly determine the  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  versus redshift relations. We can see that data points only exist at  $z \lesssim 13$ , wherein  $\langle M_{\text{total}}^{\text{galaxy}} \rangle$  tends to increase with decreasing redshift, except for  $\tilde{M}_h = 10000$ ,  $\tilde{M}_{\text{sfrmp}} = 5$ , where statistics are too poor to reveal any useful trends). Here, we only fit for  $\alpha$  after assuming the same values of  $z_{\text{trans}}$  that were used for the assembly of  $1.8 \times 10^4 \text{ M}_{\odot}$  BHs (dashed lines in Fig. 8, lower panels). The best-fitting values are shown in the bottom three rows of Table 2. Overall, we should still keep in mind that there are very few  $1.5 \times 10^5 \text{ M}_{\odot}$  descendants. Therefore, these fits are not very statistically robust. Nevertheless, they will still be useful to test our stochastic seed models in the next subsection.

In addition to the mean trends, the  $M_{\text{total}}^{\text{galaxy}}$  versus redshift relations show a significant amount of scatter ( $\sigma$ ). This is defined to be the  $1\sigma$  standard deviation shown by the shaded regions in Fig. 8. Generally, we see that the scatter does not have a strong redshift evolution. The overall mean scatter (averaged over the entire redshift range) for the different gas-based seed models is shown in the seventh column of Table 2. The scatter decreases slightly as we make the gas-based seeding criterion more restrictive by increasing  $\tilde{M}_h$  or  $\tilde{M}_{\text{sfrmp}}$ . This is likely because for more restrictive seed models, assembly of higher-mass BHs occurs in more massive galaxies for which the underlying galaxy mass function is steeper. For the same reason, the scatter is also smaller for the assembly of  $1.5 \times 10^5 \text{ M}_{\odot}$  BHs compared to that of  $1.8 \times 10^4 \text{ M}_{\odot}$  BHs.

#### 4.2.2 Properties of galaxies that form ESDs: Comparison with gas-based seed model predictions

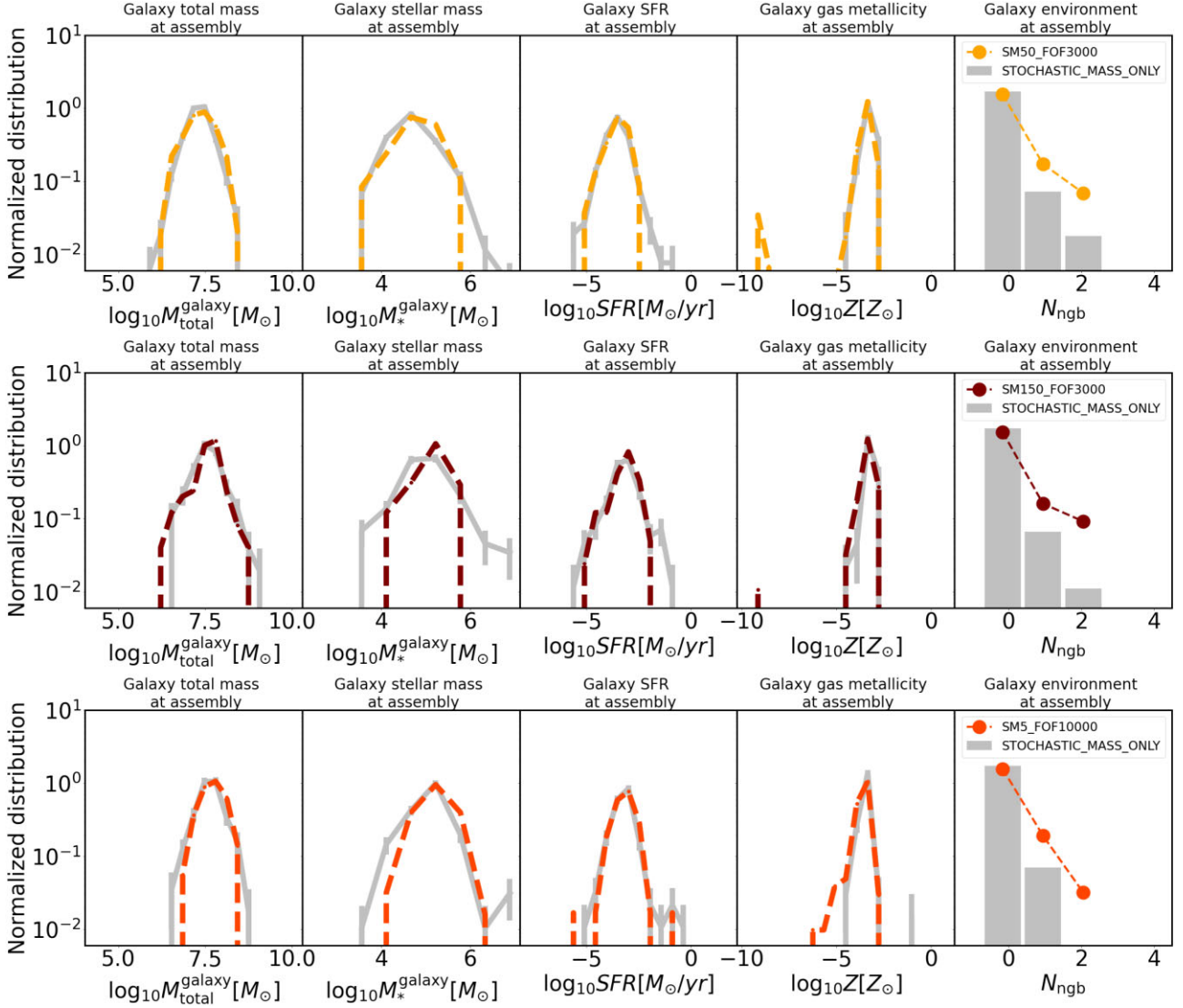
We finally use the  $M_{\text{total}}^{\text{galaxy}}$  versus redshift relations to formulate our *galaxy mass criterion*. More specifically, we place ESDs of mass  $1.8 \times 10^4$  and  $1.5 \times 10^5 \text{ M}_{\odot}$  based on minimum galaxy mass thresholds. The threshold value ( $M_{\text{th}}$ ) is stochastically drawn from redshift dependent distributions described by a log-normal function, i.e.  $\propto \exp[-\frac{1}{2}(\log_{10} M_{\text{th}}^2 - \mu^2)/\sigma^2]$ , with mean  $\mu \equiv \langle M_{\text{total}}^{\text{galaxy}} \rangle(z)$  described by the double power-law fits shown in Fig. 8 and Table 2. The standard deviation  $\sigma$  is shown in Table 2 (column 7).

In Fig. 9, we show the 1D distributions (marginalized over all redshifts until  $z = 7$ ) of the various galaxy properties wherein  $1.8 \times 10^4 \text{ M}_{\odot}$  descendants assemble (i.e. total mass, stellar mass, SFRs, gas metallicities and environments). We compare the predictions for the GAS\_BASED simulations that assemble the  $1.8 \times 10^4 \text{ M}_{\odot}$  descendants from  $2.3 \times 10^3 \text{ M}_{\odot}$  DGBs (colored lines), and the STOCHASTIC\_MASS\_ONLY simulations that directly seed the  $1.8 \times 10^4 \text{ M}_{\odot}$  ESDs (grey lines). We can clearly see that after calibrating the STOCHASTIC\_MASS\_ONLY simulations to reproduce the total galaxy masses (1st panels from the left) predicted by the GAS\_BASED simulation, it also broadly reproduces the baryonic properties of the galaxies such as stellar masses, SFRs and metallicities (2nd, 3rd and 4th panels). This further solidifies our findings from Figs 1 to 3, that the galaxies wherein the  $1.8 \times 10^4 \text{ M}_{\odot}$  descendants assemble are reasonably well characterized by their total mass alone. Recall that this is attributed to the transience of the rapid metal enrichment phase in which haloes form  $2.3 \times 10^3 \text{ M}_{\odot}$  DGBs in the GAS\_BASED suite.

However, we see that the *galaxy mass criterion* places the ESDs in sparser environments (hosts with fewer neighbouring haloes) compared to the GAS\_BASED simulation predictions (rightmost panels in Fig. 9). This reflects the fact that when the low-mass DGBs assemble higher-mass BHs through merger-dominated BH growth, their descendants naturally grow faster in regions with more frequent major halo and galaxy mergers. Therefore, for a given distribution of total galaxy masses, those living in richer environments are more likely to contain higher-mass descendant BHs.

These results for the assembly of  $1.8 \times 10^4 \text{ M}_{\odot}$  BHs also hold true for the assembly of  $1.5 \times 10^5 \text{ M}_{\odot}$  BHs, as shown in Fig. 10. In the next section, we develop an additional seeding criterion to account





**Figure 9.** Colored dashed lines show 1D distributions of galaxy properties in which  $1.8 \times 10^4 M_\odot$  BHs assemble from  $2.3 \times 10^3 M_\odot$  DGBs within GAS\_BASED simulations. From left to right, the panels in each row show the total galaxy masses ( $M_{\text{total}}^{\text{galaxy}}$ ), stellar masses ( $M_*^{\text{galaxy}}$ ), SFRs, gas metallicities ( $Z$ ), and environments ( $N_{\text{ngb}}$  i.e. the number of neighbouring haloes around the galaxy as defined in Section 2.3.2). Top, middle and bottom rows correspond to different sets of gas-based seed parameters:  $[\tilde{M}_h, \tilde{M}_{\text{sfrmp}} = 3000, 50]$ ,  $[\tilde{M}_h, \tilde{M}_{\text{sfrmp}} = 3000, 150]$  and  $[\tilde{M}_h, \tilde{M}_{\text{sfrmp}} = 10000, 5]$  respectively. In each panel, the light grey lines show host properties for the  $1.8 \times 10^4 M_\odot$  ESDs in the corresponding STOCHASTIC\_MASS\_ONLY simulation. Note that unlike the rest of the paper, here the STOCHASTIC\_MASS\_ONLY simulations are run at the highest resolution of  $L_{\text{max}} = 12$  for a fair comparison of their predicted galaxy baryonic properties with the GAS\_BASED simulations run at the same resolution. The total galaxy masses of BH hosts in the STOCHASTIC\_MASS\_ONLY simulations are calibrated to match the GAS\_BASED simulations, but no other calibration is performed. The agreement of the distributions of baryonic properties ( $M_*$ , SFR, &  $Z$ ) between the two types of simulations results naturally from matching the  $M_{\text{total}}^{\text{galaxy}}$  distribution. However, the STOCHASTIC\_MASS\_ONLY simulations do end up placing the ESDs in significantly less rich environments (smaller  $N_{\text{ngb}}$ ) compared to what is required by the GAS\_BASED simulations.

for this small-scale clustering of the assembly sites of higher mass descendants in our GAS\_BASED models.

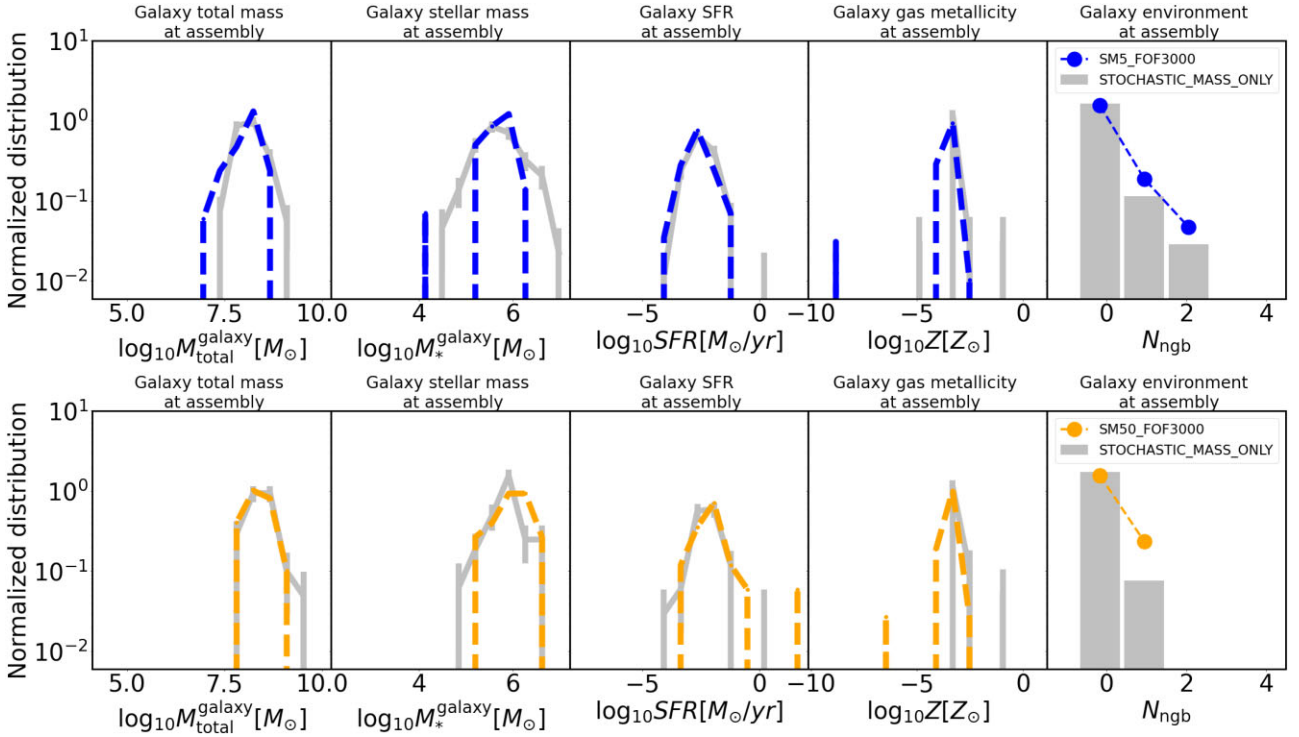
#### 4.3 Building the galaxy environment criterion

In this section, we describe an additional *galaxy environment criterion* to favour the placement of ESDs in galaxies in richer environments (at fixed galaxy mass). We then explore its implications on their two-point clustering and the overall BH population.

First, we assume that any potential seeding site with two or more neighbours ( $N_{\text{ngb}} \geq 1$ ) will always seed an ESD. Potential seeding sites with zero or one neighbours will seed an ESD with

a probability  $0 \leq P_{\text{seed}}^{\text{env}} \leq 1$ . For these cases, we assign a different linear dependence of  $P_{\text{seed}}^{\text{env}}$  on the galaxy mass  $M_{\text{total}}^{\text{galaxy}}$ , such that the probability for any potential seeding site to actually form an ESD is given by

$$P_{\text{seed}}^{\text{env}} = \begin{cases} \left( M_{\text{total}}^{\text{galaxy}} - \langle M_{\text{total}}^{\text{galaxy}} \rangle \right) \gamma + p_0, & \text{if } N_{\text{ngb}} = 0 \\ \left( M_{\text{total}}^{\text{galaxy}} - \langle M_{\text{total}}^{\text{galaxy}} \rangle \right) \gamma + p_1, & \text{if } N_{\text{ngb}} = 1 \\ 1, & \text{if } N_{\text{ngb}} \geq 1 \end{cases}. \quad (6)$$



**Figure 10.** Similar to Fig. 9, but for the assembly of  $1.5 \times 10^5 M_\odot$  BHs from  $2.3 \times 10^3 M_\odot$  DGBs. Here, the top and bottom rows correspond to  $[M_h, \bar{M}_{\text{smp}} = 3000, 5]$  and  $[M_h, \bar{M}_{\text{smp}} = 3000, 50]$ .

Here,  $p_0$  and  $p_1$  denote the seeding probability in galaxies with 0 and 1 neighbours, respectively, at the mean ( $\langle M_{\text{total}}^{\text{galaxy}} \rangle$ ) of the total mass distributions of galaxies wherein the descendant BHs assemble.

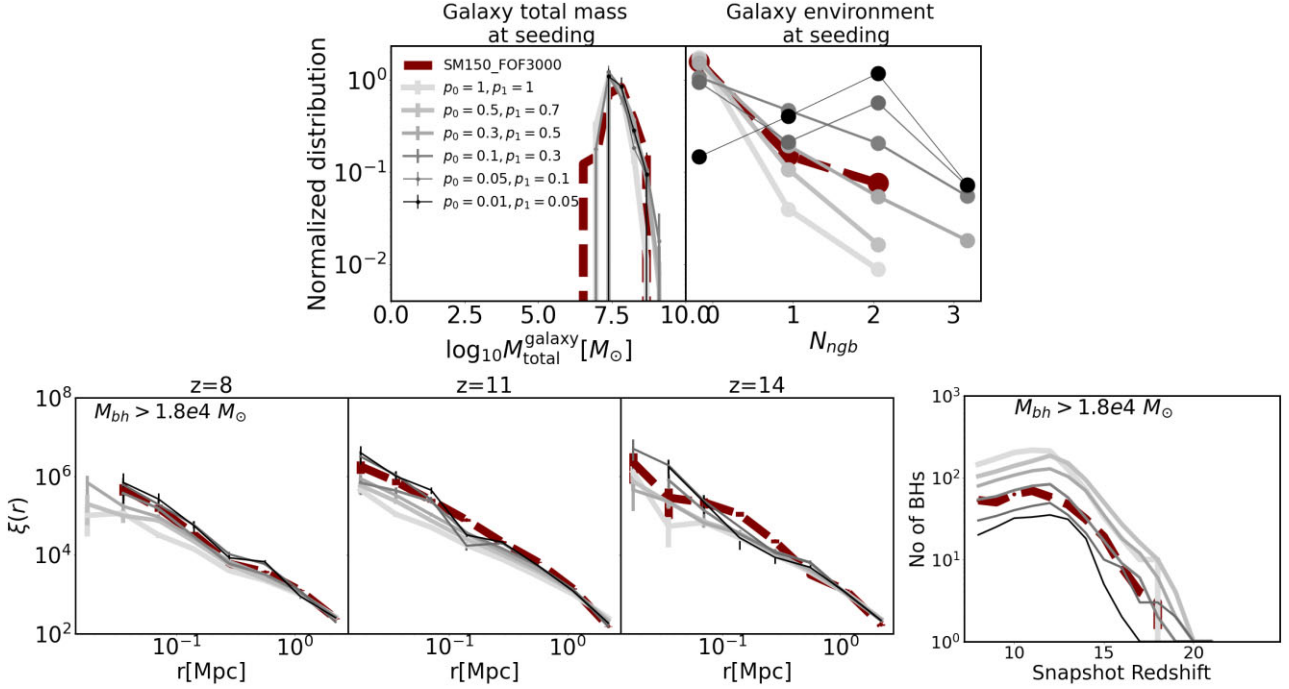
The parameter  $\gamma$  defines the slope for the linear dependence of  $P_{\text{seed}}^{\text{env}}$  on the galaxy mass; it varies slightly between the underlying gas-based seed models used for calibration, as listed in Table 2. The motivation for this linear dependence and the adopted  $\gamma$  values are described in Appendix A. But to briefly summarize the main physical motivation, we use a  $\gamma > 0$  to encode the natural expectation that for fixed  $N_{\text{ngb}}$ , descendants will grow faster within galaxies with higher total mass. This is because  $N_{\text{ngb}}$ , by definition, counts the number of haloes with masses *higher* than the host halo mass of the galaxy that are within  $5R_{\text{vir}}$ . As a result, a higher-mass galaxy with  $N_{\text{ngb}}$  neighbours is in a more overdense region than a lower-mass galaxy with the same  $N_{\text{ngb}}$  neighbours.

We add the *galaxy environment criterion* to the already applied *galaxy mass criterion*. We shall refer to the resulting suite of simulations as *STOCHASTIC.MASS.ENV*. In Fig. 11, we systematically compare the *GAS.BASED* simulations (maroon lines) to the *STOCHASTIC.MASS.ENV* simulations that trace  $1.8 \times 10^4 M_\odot$  descendants (grey lines) for a range of parameter values for  $p_0$  and  $p_1$ . We start with  $p_0 = 1, p_1 = 1$ , which is basically the *STOCHASTIC.MASS.ONLY* simulation (lightest grey lines), and find that it significantly underestimates the two point clustering (by factors up to  $\sim 5$ ) of the  $\geq 1.8 \times 10^4 M_\odot$  BHs compared to the *GAS.BASED* simulations (lower left three panels). At the same time, the *STOCHASTIC.MASS.ONLY* simulation also over-estimates the overall counts of the  $\geq 1.8 \times 10^4 M_\odot$  BHs (lower right most panel). Upon decreasing the probabilities as  $p_0 < p_1 < 1$ , we can see that the two-point clustering starts to increase while the overall BH counts simultaneously decrease. For  $p_0 = 0.1$  &  $p_1 = 0.3$ , we produce the best agreement of the two-point clustering as well as

the overall BH counts. Further decreasing  $p_0$  and  $p_1$  mildly enhances the two-point clustering, but leads to too much suppression of the BH counts compared to *GAS.BASED* simulations. Therefore, we identify  $p_0 = 0.1$  &  $p_1 = 0.3$  as the best set of parameter values for the gas-based seeding parameters  $[M_h, \bar{M}_{\text{smp}} = 3000, 150]$ .

As a caveat, we must also note in Fig. 11 that while  $p_0 = 0.1$  &  $p_1 = 0.3$  produces the best agreement with the two point correlation function between *GAS.BASED* and *STOCHASTIC.MASS.ENV* simulations, it does place the ESDs in galaxies with somewhat higher  $N_{\text{ngb}}$  compared to the *GAS.BASED* simulations (upper right panels). To that end, recall that  $N_{\text{ngb}}$  only measures the galaxy environment at a fixed separation scale of  $D_{\text{ngb}} = 5 R_{\text{vir}}$  (revisit Section 2.3.2). Therefore, we cannot expect  $N_{\text{ngb}}$  to fully determine the two-point correlation profile, which measures the environment over a wide range of separation scales ( $\sim 0.01$ – $1 \text{ Mpc}/h$  in our case). In other words, one could come up with alternative set of *galaxy environment criteria* (for example, using  $N_{\text{ngb}}$  within a different  $D_{\text{ngb}} \neq 5 R_{\text{vir}}$  or even multiple set of  $N_{\text{ngb}}$  values within different multiple  $D_{\text{ngb}}$  values) and still be able simultaneously reproduce the two-point correlation function as well as the BH counts. Finding all these different possibilities of *galaxy environment criteria* is not the focus of this work. Instead, our objective here is simply to demonstrate that to reproduce the *GAS.BASED* simulation predictions, we need a *galaxy environment criterion* to favour the placing of ESDs in galaxies with richer environments. Furthermore, we showed that by applying a *galaxy environment criterion* that brings the two point correlation function into agreement with the *GAS.BASED* simulations, our *STOCHASTIC.MASS.ENV* simulations achieve the primary goal for our sub-grid seeding model: faithfully representing the descendants of  $2.3 \times 10^3 M_\odot$  seeds produced in the *GAS.BASED* simulations.

Thus far we have calibrated a *STOCHASTIC.MASS.ENV* simulation to reproduce the  $1.8 \times 10^4 M_\odot$  descendant BH population



**Figure 11.** Impact of *galaxy environment criterion* on the two-point clustering and the overall counts of  $>1.8 \times 10^4 M_{\odot}$  BHs. The dashed maroon lines show a simulation that uses the gas-based seed model [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 3000, 150$ ] with  $M_{\text{seed}}^{\text{DGB}} = 2.3 \times 10^3 M_{\odot}$ . The grey solid lines correspond to simulations that use the stochastic seed model, and directly place ESDs of mass  $1.8 \times 10^4 M_{\odot}$  based on both the *galaxy mass criterion* and *galaxy environment criterion*. For the *galaxy environment criterion*, we systematically decrease  $p_0$  and  $p_1$  as the shade gets darker (see legend). *Upper panels:* The total galaxy mass (left panel) and galaxy environment (right panel) during the initial assembly of  $1.8 \times 10^4 M_{\odot}$  BHs. *Lower panels:* The left three panels show the two point clustering of  $>1.8 \times 10^4 M_{\odot}$  BHs at  $z = 8, 11$  &  $14$ , respectively, and the rightmost panel shows the overall number of  $>1.8 \times 10^4 M_{\odot}$  BHs in each snapshot. We find that the STOCHASTIC.MASS.ONLY simulation ( $p_0 = 1$  and  $p_1 = 1$ ) significantly underestimates the small-scale clustering and overestimates the BH counts compared to the GAS\_BASED simulations. As we introduce the *galaxy environment criterion* (STOCHASTIC.MASS.ENV) and decrease  $p_0$  and  $p_1$  to favour seeding in richer environments, we find that the small-scale clustering is enhanced and the BH counts decrease. The model with  $p_0, p_1 = 0.1, 0.3$  produces the best match for the small-scale clustering as well as the BH counts.

from a gas-based seed model with [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 3000, 150$ ] and  $M_{\text{seed}} = 2.3 \times 10^3 M_{\odot}$ . We can perform the same calibration for the remaining gas-based seed models in our suite, and for the assembly of  $1.5 \times 10^5 M_{\odot}$  descendant BHs in addition to  $1.8 \times 10^4 M_{\odot}$  descendants. The resulting  $p_0$  and  $p_1$  values for all the gas-based seeding parameters are listed in Table 2. Broadly speaking, we require  $p_0 \sim 0.1$ – $0.2$  and  $p_1 \sim 0.3$ – $0.4$  to simultaneously reproduce the gas-based seed model predictions for the small-scale clustering and BH counts of the descendant BHs. Slightly higher  $p_0$  and  $p_1$  values are favoured for more restrictive gas based criteria and for higher-mass galaxies, possibly because in both cases the descendant BHs assemble in higher-mass galaxies. Note that higher-mass galaxies tend to be more strongly clustered than lower mass galaxies. As a result, during the calibration of the STOCHASTIC.MASS.ENV simulations, the *galaxy mass criterion* alone will already produce a slightly stronger clustering for the ESDs. This lessens the burden on the *galaxy environment criterion* to achieve the desired clustering predicted by the gas-based seed models.

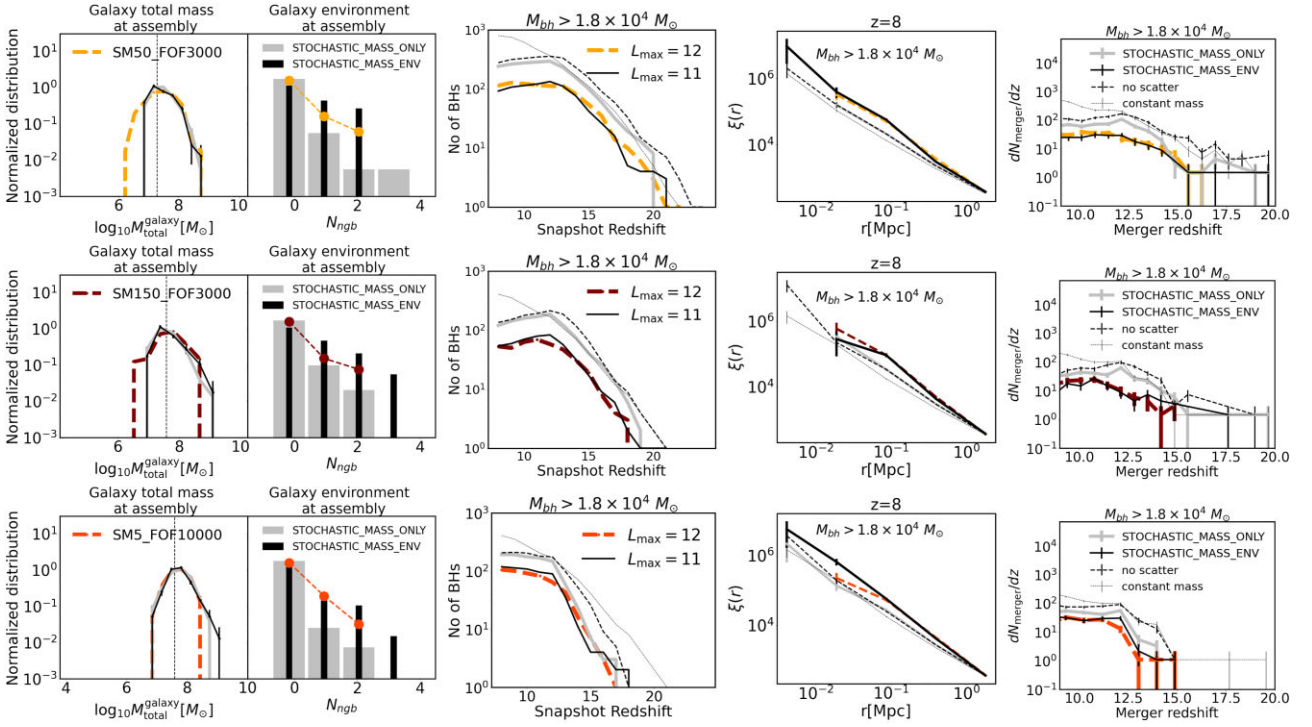
In Figs 12 and 13, we show the STOCHASTIC.MASS.ENV (solid black lines) versus GAS\_BASED (colored dashed lines) seed model predictions. For  $M_{\text{seed}}^{\text{ESD}} = 1.8 \times 10^4 M_{\odot}/h$  (Fig. 12), we calibrate models corresponding to [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 3000, 50$  &  $3000, 150$ ] and [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 10000, 5$ ]. We exclude the most lenient gas-based seed parameters of [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 3000, 5$ ], since it leads to a significant portion of  $1.8 \times 10^4 M_{\odot}/h$  descendants to assemble in galaxies that cannot be resolved in the  $L_{\text{max}} = 11$  runs. For the remaining gas-based

seed parameters, the STOCHASTIC.MASS.ENV simulations well reproduce the GAS\_BASED simulation predictions for the BH counts, two-point correlation functions and merger rates of  $>1.8 \times 10^4 M_{\odot}$  BHs.

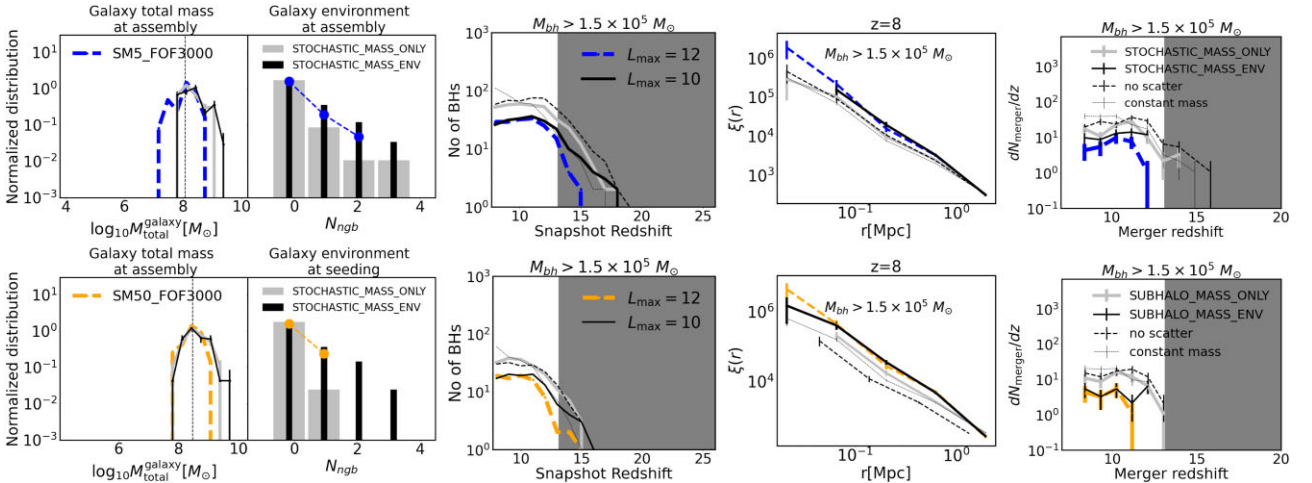
For  $M_{\text{seed}}^{\text{ESD}} = 1.5 \times 10^5 M_{\odot}$  (Fig. 13), we only do this exercise for the most lenient gas-based seed models i.e. [ $\tilde{M}_h, \tilde{M}_{\text{smp}} = 3000, 5$  &  $3000, 50$ ]. This is because for the stricter gas-based seed models, there are too few BHs produced overall. Here, the STOCHASTIC.MASS.ENV simulations well reproduce the counts of  $>1.5 \times 10^5 M_{\odot}$  BHs at  $z < 13.1$  (wherein there is enough data to calibrate the slope  $\alpha$ ; revisit Fig. 8, bottom row). For  $z > 13.1$ ,  $\beta = 0$  is assumed due to the absence of enough data points to perform any fitting; here, the STOCHASTIC.MASS.ENV seed model overestimates the number of  $>1.5 \times 10^5 M_{\odot}$  BHs and their high- $z$  merger rates. Regardless, where enough data exist for robust calibration, these results imply that with a calibrated combination of *galaxy mass criterion* and *galaxy environment criterion*, the STOCHASTIC.MASS.ENV simulations can well reproduce the GAS\_BASED simulation predictions for a wide range of gas-based seeding parameters.

Figs 12 and 13 also disentangle the impact of the various components of our final stochastic seed model, and they highlight the importance of each component in the successful representation of the gas-based seed models. As seen previously, the STOCHASTIC.MASS.ONLY seed model overestimates the BH counts and merger rates by factors between  $\sim 2$  and  $5$ . Next, when we assume





**Figure 12.** Here we demonstrate the ability of different  $L_{\max} = 11$  stochastic seed models to represent the  $1.8 \times 10^4 M_{\odot}$  descendants of  $2.3 \times 10^3 M_{\odot}$  DGBs formed in  $L_{\max} = 12$  gas-based seed models. The leftmost two panels show the total galaxy mass and galaxy environment at the time of assembly of  $1.8 \times 10^4 M_{\odot}$  BHs. The remaining three panels on the right show the statistics of  $>1.8 \times 10^4 M_{\odot}$  BHs, namely the total BH counts versus redshift, the two-point clustering at  $z = 8$ , and the merger rates. The colored dashed lines show the GAS\_BASED simulations wherein  $2.3 \times 10^3 M_{\odot}$  DGBs form and eventually grow to assemble  $1.8 \times 10^4 M_{\odot}$  BHs. The different rows correspond to different values of  $M_{\text{sfmp}}$  and  $\bar{M}_h$  (see legend). The remaining lines correspond to simulations using stochastic seed models that place ESDs directly at  $1.8 \times 10^4 M_{\odot}$ . The thick and solid silver and black lines and histograms show the STOCHASTIC\_MASS\_ONLY and STOCHASTIC\_MASS\_ENV simulations respectively; they use the fiducial seeding parameters calibrated for each set of gas-based seeding parameters listed in Table 2. The thin black dashed lines in the right three panels show STOCHASTIC\_MASS\_ONLY simulations that assume zero scatter in the *galaxy mass criterion* i.e.  $\sigma = 0$ . The thinnest black solid line in the same panels show simulations that assume a constant galaxy mass threshold fixed at the mean of the distributions from the leftmost panels (see vertical line). Amongst all the simulations that use stochastic seeding, only the STOCHASTIC\_MASS\_ENV simulations are able to successfully capture the GAS\_BASED simulation predictions.



**Figure 13.** Same as Fig. 12, but for the assembly of  $1.5 \times 10^5 M_{\odot}$  BHs. The statistics are more limited compared to the previous figure. The shaded grey regions correspond to  $z > 13.1$ , wherein we could not calibrate the *galaxy mass criterion* due to lack of data points in Fig. 8. But at  $z < 13.1$  where calibration was possible, we find that the STOCHASTIC\_MASS\_ENV simulations (at a resolution of  $L_{\max} = 10$ ) do reasonably match with the BH counts predicted by the  $L_{\max} = 12$  GAS\_BASED simulations.

zero scatter in the *galaxy mass criterion* ( $\Sigma = 0$ , black dashed lines), it further overestimates the BH counts and merger rates up to factors of  $\sim 1.5$  (grey solid versus black dashed lines). Finally, if we remove the redshift dependence in the *galaxy mass criterion* and instead assume a constant threshold value (thin dotted lines), the BH counts and merger rates monotonically increase with time. Not surprisingly, this is because such a model cannot capture the suppression of seed formation due to metal enrichment.

Overall, we can clearly see that in order to represent our  $L_{\max} = 12$  gas-based seed models forming  $2.3 \times 10^3 M_{\odot}$  BH seeds in lower-resolution, larger-volume simulations, we need a stochastic seed model that places their resolvable descendant BHs (ESDs) using the following two criteria

(i) A *galaxy mass criterion* with a galaxy mass seeding threshold that is drawn from a distribution that evolves with redshift. The redshift evolution encodes the impact of dense gas formation, halo growth and metal enrichment on seed formation.

(ii) A *galaxy environment criterion* that favours seeding within galaxies living in rich environments. This encodes the impact of the unresolved, hierarchical-merger-dominated growth of these seeds from  $M_{\text{seed}}^{\text{DGB}}$  to  $M_{\text{seed}}^{\text{ESD}}$ .

Lastly, we note that while Table 2 presents the calibrated parameter values for the stochastic seed models, these values are specific to the underlying TNG galaxy formation model and zoom region used in this work. In general, they are subject to cosmic variance. In the next paper (in preparation) where we apply the stochastic seed models on uniform boxes, we re-calibrate our seeding parameters by running our gas-based seed models on high-resolution uniform simulations. Additionally, they may also be different for other simulations with different sub-grid models for dense gas formation, metal enrichment and stellar feedback. Therefore, before applying these seed models to other simulations, care should be taken to re-calibrate the seed parameters in order to ensure self-consistency.

#### 4.4 Accounting for unresolved minor mergers

We have thus far successfully built a new stochastic BH seed model that places ESDs which represent the  $\sim 10^4$ – $10^5 M_{\odot}$  descendants of  $\sim 10^3 M_{\odot}$  DGBs in simulations that cannot directly resolve these lowest-mass BHs. In this section, we model the subsequent growth of these ESDs. To do so, we must first account for one additional contribution to their growth: unresolved minor mergers.

Recall from Bhowmick et al. (2021) that the earliest growth of these  $\sim 10^3 M_{\odot}$  DGBs is completely driven by BH mergers, with negligible contribution from gas accretion. For our present purposes, these BH mergers can be classified into three types:

(i) *Heavy mergers*: In these mergers, both the primary and secondary black holes (with masses  $M_1$  and  $M_2$ , respectively) are greater than the mass of the ESDs ( $M_1 > M_2 > M_{\text{seed}}^{\text{ESD}}$ ). Therefore, these mergers will be fully resolvable within STOCHASTIC\_MASS\_ENV simulations.

(ii) *Light major mergers*: In these mergers, both the primary and secondary black holes are less massive than the ESDs ( $M_{\text{seed}}^{\text{DGB}} < M_2 < M_1 < M_{\text{seed}}^{\text{ESD}}$ ). These mergers cannot be resolved in STOCHASTIC\_MASS\_ENV simulations. However, these are the mergers that lead to the initial assembly of the descendants represented by the ESDs, such that their contribution to BH assembly is already implicitly captured within the stochastic seed model.

(iii) *Light minor mergers*: In these mergers, the primary black hole is more massive than the ESD mass, but the secondary black hole is not ( $M_1 > M_{\text{seed}}^{\text{ESD}}$  &  $M_{\text{seed}}^{\text{DGB}} < M_2 < M_{\text{seed}}^{\text{ESD}}$ ). These mergers

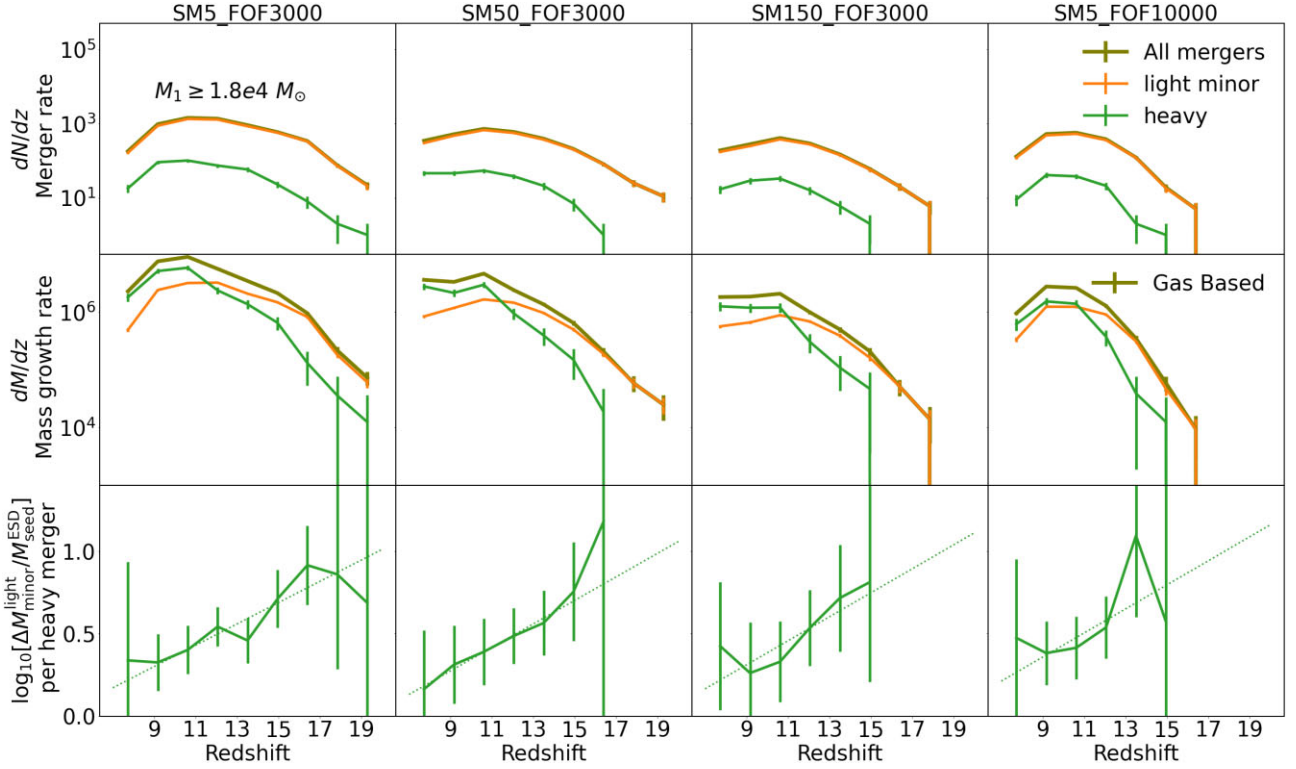
cannot be resolved in STOCHASTIC\_MASS\_ENV simulations, and their contributions to BH mass assembly cannot be captured by the *galaxy mass criterion* or the *galaxy environment criterion*. Therefore, we must modify our prescription to explicitly add their contribution to the growth of the ESDs.

We first determine the contribution of light minor mergers within the GAS\_BASED simulations. Here we only show the results for  $M_{\text{seed}}^{\text{ESD}} = 1.8 \times 10^4 M_{\odot}$ , since there are too few  $1.5 \times 10^5 M_{\odot}$  BHs formed in the GAS\_BASED simulations to robustly perform this analysis for the latter. The light minor mergers are thus defined to have  $M_1 > 1.8 \times 10^4 M_{\odot}$  and  $2.3 \times 10^3 < M_2 < 1.8 \times 10^4 M_{\odot}$ , and heavy mergers are defined to be those with  $M_1 > M_2 > 1.8 \times 10^4 M_{\odot}$ . In Fig. 14, we compare the contributions of the light minor mergers and heavy mergers to the growth of  $> 1.8 \times 10^4 M_{\odot}$  BHs in the GAS\_BASED simulations. The light minor mergers are  $\sim 30$  times more frequent than the heavy mergers (top row); this is simply due to higher overall number of  $M_{\text{BH}} < 1.8 \times 10^4 M_{\odot}$  BHs compared to  $M_{\text{bh}} > 1.8 \times 10^4 M_{\odot}$  BHs. When we compare the mass growth contributed by light minor mergers versus heavy mergers (middle row), we find that the light minor mergers dominate at the highest redshifts ( $z \sim 15$ – $19$ ). As BH growth proceeds over time, the mass growth contributed by heavy mergers increases and eventually exceeds that of the light minor mergers at  $z \lesssim 12$ , even though the overall merger rates are still dominated by light minor mergers. This is because the masses of the BHs involved in the heavy mergers continue to increase with time. Eventually, when new DGB formation is strongly suppressed by metal enrichment, the mass growth due to the light minor mergers becomes small. We clearly see these trends in the third row of Fig. 14 which shows  $\Delta M_{\text{minor}}^{\text{light}}$  defined as the amount of mass growth due to light minor mergers between successive *heavy merger* events.  $\Delta M_{\text{minor}}^{\text{light}}$  monotonically decreases with redshift and its evolution is reasonably well fit by power laws.

We use the power law fits of  $\Delta M_{\text{minor}}^{\text{light}}$  (shown in the last row of Fig. 14) to determine the missing BH growth contribution from light minor mergers. More specifically, for each heavy merger event in a STOCHASTIC\_MASS\_ENV simulation, we added extra mass growth of  $\Delta M_{\text{minor}}^{\text{light}}$  due to light minor mergers, calculated based on these power law fits. Fig. 15 shows that it is only after the inclusion of these unresolved light minor mergers, we achieve reasonable agreement between the BH mass functions predicted by the GAS\_BASED and the STOCHASTIC\_MASS\_ENV simulations (colored dashed lines versus solid black lines). Note that at masses between  $M_{\text{seed}}^{\text{ESD}}$  and  $2M_{\text{seed}}^{\text{ESD}}$ , the STOCHASTIC\_MASS\_ENV simulations will inevitably continue to slightly underpredict the mass functions. This is because within our prescription, the contribution from light minor mergers does not occur until the first heavy merger event between the ESDs.

## 5 SUMMARY AND CONCLUSIONS

In this work, we tackle one of the longstanding challenges in modelling BH seeds in cosmological hydrodynamic simulations: how do we simulate low-mass ( $\lesssim 10^3 M_{\odot}$ ) seeds in simulations that cannot directly resolve them? We address this challenge by building a new sub-grid seed model that can stochastically seed the smallest resolvable descendants of low-mass seeds in lower-resolution simulations (hereafter referred to as ‘stochastic seed model’). Our new seed model is motivated and calibrated based on highest resolution simulations that directly resolve the low-mass seeds. With this new tool, we have bridged a critical gap between high-resolution simulations that directly resolves low-mass seeds, and larger-volume simulations that can generate sufficient numbers



**Figure 14.** Comparing the contributions of heavy mergers versus light minor mergers to the merger driven BH growth within the GAS\_BASED suite. The green lines show heavy mergers where the masses of both primary and secondary BHs are  $\geq 1.8 \times 10^4 M_\odot$ . The orange lines show the light minor mergers where the secondary BH mass is  $< 1.8 \times 10^4 M_\odot$  but the primary BH mass is  $\geq 1.8 \times 10^4 M_\odot$ . The olive lines show the total contribution from both types of mergers i.e. all mergers with primary BHs  $\geq 1.8 \times 10^4 M_\odot$ . The different columns show different gas-based seed models. Middle panels show the mass growth rate due to mergers as a function of redshift, which is defined as the total mass of all merging secondary BHs per unit redshift. The light minor mergers show a dominant contribution at  $z \gtrsim 11$ , whereas heavy mergers tend to be more prevalent at  $z \lesssim 11$ . The bottom panels show the mass growth ( $\Delta M_{\text{light minor}}^{\text{light}}$ ) due to the light minor mergers between successive heavy mergers. This contribution needs to be explicitly included in simulations that use the stochastic seed models, to produce BH growth consistent with the GAS\_BASED simulations.

of BHs to compare against observational measurements. This paves the way for making statistically robust predictions for signatures of low-mass seeds using cosmological hydrodynamic simulations, which is a crucial step in preparation for the wealth of observations with ongoing *JWST*, as well as upcoming facilities such as LISA.

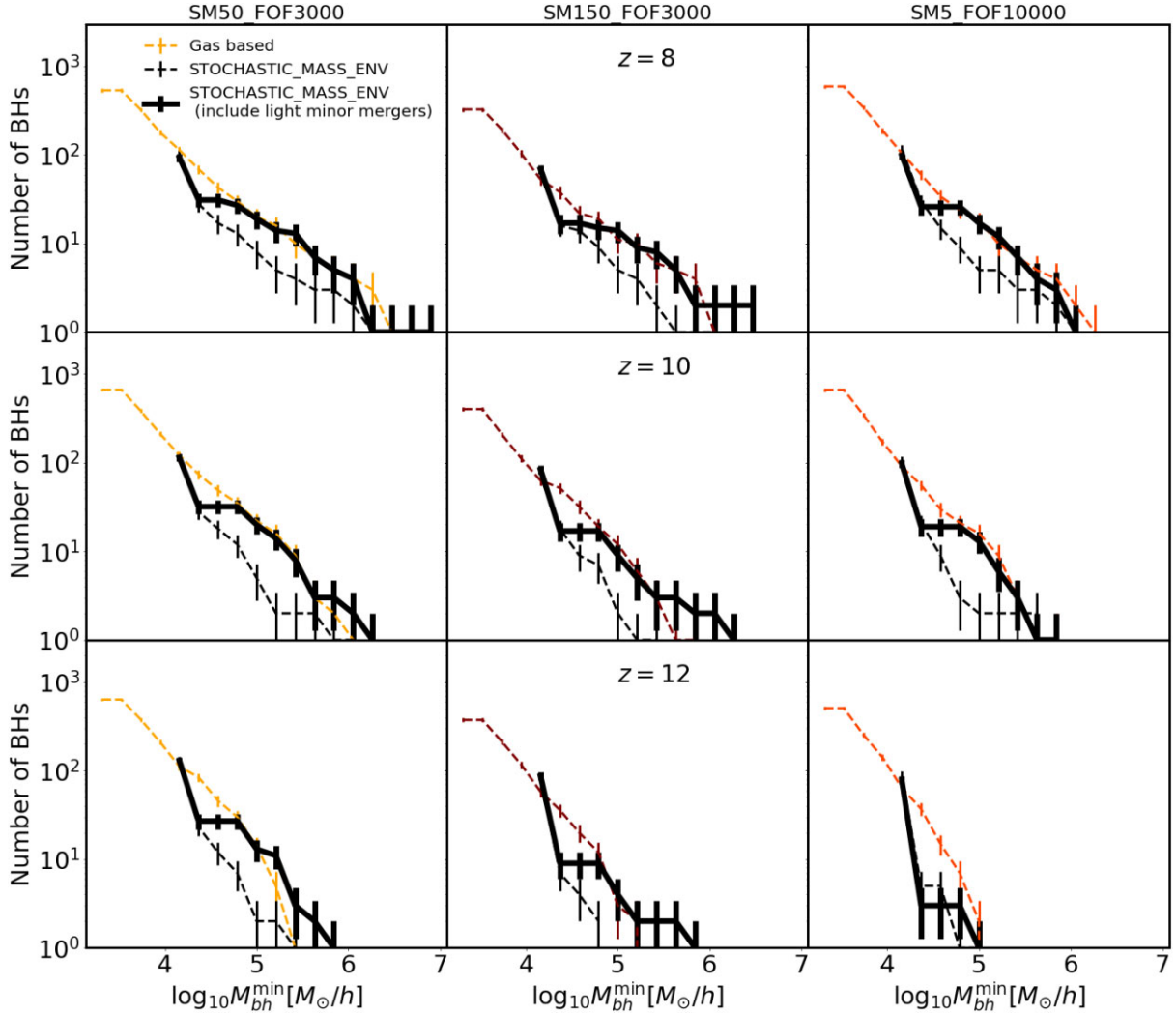
The core objective of this work has been to determine the key ingredients needed to construct such a seed model. To do this, we study the growth of the lowest mass  $2.3 \times 10^3 M_\odot$  seeds that were fully resolved using highest resolution zoom simulations. These seeds are placed in haloes containing gas that is simultaneously dense as well as metal-poor ( $< 10^{-4} Z_\odot$ ), consistent with proposed low mass seeding candidates such as Pop III stellar remnants. We trace the growth of these  $2.3 \times 10^3 M_\odot$  seeds until they assemble descendants with masses that are close to different possible gas mass resolutions ( $\sim 10^4$ – $10^6 M_\odot$ ) expected in larger cosmological volumes. We characterize the environments in which these descendants assemble; for e.g. they assemble in haloes with masses ranging from  $\sim 10^7$  to  $10^9 M_\odot$ . The results are used to build our stochastic seed model that directly seeds these descendants in lower resolution simulations. To distinguish against the *actual*  $2.3 \times 10^3 M_\odot$  seeds, we refer to the ‘seeds’ formed by the stochastic seed model as ‘extrapolated seed descendants’ or ESDs (with mass  $M_{\text{seed}}^{\text{ESD}}$ ). We consider  $1.8 \times 10^4$  &  $1.5 \times 10^5 M_\odot$  ESDs that are aimed at faithfully representing the descendants of  $2.3 \times 10^3 M_\odot$  seeds born out of dense and metal-poor gas. Specifically, we explore a wide range of stochastic seed models on lower resolution versions of our zoom

region, and determine the crucial ingredients required to reproduce the results of the highest resolution zoom simulations that explicitly resolve the  $2.3 \times 10^3 M_\odot$  seeds. Following are the key features of our new seed model:

(i) We seed the ESDs in high- $z$  (proto)galaxies which are bound substructures within high- $z$  haloes. Since haloes can contain multiple galaxies, this naturally allows the placement of multiple ESDs per halo. This is important because even if  $2.3 \times 10^3 M_\odot$  seeds are placed as one seed per halo, their subsequent hierarchical growth inevitably assembles multiple higher mass descendants within individual haloes.

(ii) We introduce a *galaxy mass criterion* which places the ESDs based on galaxy mass thresholds. These thresholds are stochastically drawn from galaxy mass (including DM, stars and gas) distributions wherein  $1.8 \times 10^4$  &  $1.5 \times 10^5 M_\odot$  BHs assemble from  $2.3 \times 10^3 M_\odot$  seeds. We find that the *galaxy mass criterion* effortlessly also replicates the baryonic properties of the galaxies at the time of assembly of the seed descendants, including stellar mass, SFRs, and gas metallicities. This is because, although  $2.3 \times 10^3 M_\odot$  seeds form within haloes exhibiting a bias towards lower metallicities in comparison to typical haloes of similar masses, they undergo a transient phase characterized by rapid metal enrichment. As a result, the higher mass  $1.8 \times 10^4$  &  $1.5 \times 10^5 M_\odot$  descendants end up in unbiased haloes with metallicities similar to haloes with similar masses. The redshift dependence of the distributions underlying the galaxy mass thresholds capture the





**Figure 15.** Comparison of the cumulative mass functions (i.e. the number of BHs above a minimum BH mass threshold  $M_{bh}^{\min}$ ) between the GAS\_BASED (coloured lines) and STOCHASTIC\_MASS\_ENV (black lines) simulations. The top, middle, and bottom rows show  $z = 8, 10$ , and  $12$ , respectively. The black dashed and solid lines show the STOCHASTIC\_MASS\_ENV predictions with and without the explicit inclusion of the contribution from the unresolved light minor mergers. Without the light minor mergers, the STOCHASTIC\_MASS\_ENV BH mass functions are significantly steeper than in the GAS\_BASED simulations. After including the contribution from the *unresolved light mergers*, the STOCHASTIC\_MASS\_ENV simulations are able to achieve reasonable agreement with the BH mass functions predicted by the GAS\_BASED simulations.

complex influence of processes such as halo growth, dense gas formation and metal enrichment, on the formation of  $2.3 \times 10^3 M_{\odot}$  seeds.

(iii) However, if our stochastic seed model only contains the *galaxy mass criterion*, it underestimates the two-point clustering (at scales of  $0.01$ – $0.1$  Mpc/h) of  $\geq 1.8 \times 10^4$  &  $1.5 \times 10^5 M_{\odot}$  BHs by factors of  $\sim 5$ . At the same time, it overestimates the BH abundances and merger rates of  $\geq 1.8 \times 10^4$  &  $1.5 \times 10^5 M_{\odot}$  BHs by factors up to  $\sim 5$ . This is a direct consequence of the fact that in our highest resolution zooms, the  $2.3 \times 10^3 M_{\odot}$  seeds grow primarily via BH-BH mergers. As a result, the assembly of the higher mass descendants is more efficient in galaxies with richer environments (higher number of neighbouring haloes) with a more extensive merger history. This cannot be captured solely by the *galaxy mass criterion*.

(iv) To successfully capture the two-point clustering of the  $\geq 1.8 \times 10^4$  &  $1.5 \times 10^5 M_{\odot}$  descendant BHs, we introduce a *galaxy environment criterion*, where we assign seeding probabilities less

than unity for galaxies with  $\leq 1$  neighbours. By doing this, we preferentially place ESDs in richer environments, which enhances the two-point clustering. We demonstrate that by adding a *galaxy-environment criterion* that is calibrated to produce the correct two-point clustering, our stochastic seed models can simultaneously also reproduce the BH abundances and merger rates of the  $\geq 1.8 \times 10^4$  &  $1.5 \times 10^5 M_{\odot}$  BHs.

(v) Lastly, the BH growth in our stochastic seed models is underestimated due to the absence of light minor mergers, defined as those involving a resolved primary ( $M_1 > M_{\text{seed}}^{\text{ESD}}$ ) but an unresolved secondary ( $M_2 < M_{\text{seed}}^{\text{ESD}}$ ). We compute the contribution of these mergers from the highest resolution zooms that resolve the  $2.3 \times 10^3 M_{\odot}$  seeds, and explicitly add them to the simulations that use the stochastic seed models. It is only after adding the contribution from light minor mergers, do our stochastic seed models achieve success in accurately reproducing the BH mass functions predicted by the highest resolution zooms.

Overall, our stochastic seed model requires three main seeding components to successfully represent low-mass seeds in lower resolution-larger volume simulations: (1) a *galaxy mass criterion*, (2) *galaxy environment criterion*, and (3) inclusion of unresolved light minor mergers. In our upcoming companion paper (Bhowmick et al., in preparation), we apply these stochastic seed models to uniform volume cosmological simulations, and thereby make predictions that would be directly comparable to facilities such as *JWST* and *LISA* for different seeding scenarios.

The construction of our stochastic seed model essentially rests only on two important aspects of the formation of low-mass seeds. First, these seeds are forming in regions which are already in the process of rapid metal enrichment, which is a natural consequence of seeding within dense and metal-poor gas. Secondly, the BH growth is dominantly driven by BH-BH mergers. Therefore, our stochastic seed model could be tuned to represent *any* low-mass seeding scenario for which the foregoing assumptions hold true. These include scenarios beyond the ones we consider in this work. Furthermore, we can calibrate our stochastic seed model against any high resolution simulation run with different galaxy formation models or using different state-of-the-art numerical solvers such as GADGET-4 (Springel et al. 2021), GIZMO (Hopkins 2015) etc. Lastly, a key advantage of our seed model is that it depends only on galaxy total mass (which is dark matter dominated) and galaxy environment. Therefore, it can also be readily applied to DM only simulations as well as semi-analytic models that are typically much less expensive compared to full hydrodynamic simulations.

In the near future, we shall test the ability of our stochastic seed model to represent low-mass seeds when coupled with alternate models for star formation, metal enrichment, stellar feedback, BH accretion and dynamics. In particular, the recipes for star formation, metal enrichment and stellar feedback can substantially influence our results by impacting the abundance of seed formation sites. For example, stronger stellar feedback can significantly enhance the dispersion of metals and prevent the formation of dense and metal-poor pockets. Supernova feedback at high redshifts can significantly enhance the availability of gas to fuel BH accretion (see also Habouzit, Volonteri & Dubois 2017). In terms of the accretion prescription itself, having a smaller scaling exponent between BH accretion rate and BH mass (such as  $\alpha = 1/6$  for gravitational torque driven accretion model) may also enhance the accretion rates of low-mass seeds. Similarly, having a more physically motivated BH dynamics prescription will likely impact the merger rates and change the relative importance of accretion versus mergers in driving BH growth. In such a case, we can envision requiring additional ingredient(s) in our stochastic seed model to capture the impact of unresolved accretion driven growth of low-mass seeds, similar to how the galaxy environment criterion was needed to account for the impact of unresolved merger dominated BH growth.

Nevertheless, our new stochastic seed model offers a substantial improvement over existing cosmological simulations that have either relied on a threshold halo/stellar mass, or on poorly resolved gas properties for seeding. Unlike most of these currently used seed models, our models will allow us to represent low-mass seeds in cosmological simulations without the need to either explicitly resolve the seeds, or seed below the gas mass resolution of the simulation. Overall, this work is an important step towards the next generation of cosmological hydrodynamic simulations in terms of improved modeling of high redshift SMBHs, to finally understand their role in shaping high redshift galaxy evolution in the ongoing *JWST* and upcoming *LISA* era.

## ACKNOWLEDGEMENTS

LB acknowledges support from NSF award AST-1909933 and Cottrell Scholar Award #27553 from the Research Corporation for Science Advancement. PT acknowledges support from NSF-AST 2008490. RW acknowledges funding of a Leibniz Junior Research Group (project number J131/2022).

## DATA AVAILABILITY

The underlying data used in this work shall be made available upon reasonable request to the corresponding author.

## REFERENCES

- Abbott B. P. et al., 2009, *Rep. Prog. Phys.*, 72, 076901  
 Abbott R. et al., 2020, *ApJ*, 900, L13  
 Agazie G. et al., 2023, *ApJ*, 951, L8  
 Amaro-Seoane P. et al., 2017, preprint (arXiv:1702.00786)  
 Bañados E. et al., 2018, *Nature*, 553, 473  
 Baker J. et al., 2019, *BAAS*, 51, 7  
 Barcons X. et al., 2017, *Astron. Nachr.*, 338, 153  
 Barnes J., Hut P., 1986, *Nature*, 324, 446  
 Bañados E. et al., 2016, *ApJS*, 227, 11  
 Begelman M. C., Silk J., 2023, *MNRAS*, 526, L94  
 Begelman M. C., Volonteri M., Rees M. J., 2006, *MNRAS*, 370, 289  
 Bellovary J. M., Cleary C. E., Munshi F., Tremmel M., Christensen C. R., Brooks A., Quinn T. R., 2019, *MNRAS*, 482, 2913  
 Bhowmick A. K. et al., 2021, *MNRAS*, 507, 2012  
 Bhowmick A. K., Blecha L., Torrey P., Kelley L. Z., Vogelsberger M., Nelson D., Weinberger R., Hernquist L., 2022a, *MNRAS*, 510, 177  
 Bhowmick A. K. et al., 2022b, *MNRAS*, 516, 138  
 Bromm V., Loeb A., 2003, *ApJ*, 596, 34  
 Cann J. M., Satyapal S., Abel N. P., Ricci C., Secrest N. J., Blecha L., Gliozzi M., 2018, *ApJ*, 861, 142  
 Chabrier G., 2003, *PASP*, 115, 763  
 Das A., Schleicher D. R. G., Basu S., Boekholt T. C. N., 2021a, *MNRAS*, 505, 2186  
 Das A., Schleicher D. R. G., Leigh N. W. C., Boekholt T. C. N., 2021b, *MNRAS*, 503, 1051  
 Davies M. B., Miller M. C., Bellovary J. M., 2011, *ApJ*, 740, L42  
 Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, *ApJ*, 292, 371  
 Di Matteo T., Khandai N., DeGraf C., Feng Y., Croft R. A. C., Lopez J., Springel V., 2012, *ApJ*, 745, L29  
 Donnari M. et al., 2019, *MNRAS*, 485, 4817  
 Dubois Y., Peirani S., Pichon C., Devriendt J., Gavazzi R., Welker C., Volonteri M., 2016, *MNRAS*, 463, 3948  
 Fan X. et al., 2001, *AJ*, 122, 2833  
 Feng Y., Di-Matteo T., Croft R. A., Bird S., Battaglia N., Wilkins S., 2016, *MNRAS*, 455, 2778  
 Fryer C. L., Woosley S. E., Heger A., 2001, *ApJ*, 550, 372  
 Gardner J. P. et al., 2006, *Space Sci. Rev.*, 123, 485  
 Genel S. et al., 2018, *MNRAS*, 474, 3976  
 Haas M. R., Schaye J., Jeason-Daniel A., 2012, *MNRAS*, 419, 2133  
 Habouzit M., Volonteri M., Dubois Y., 2017, *MNRAS*, 468, 3935  
 Habouzit M., Pisani A., Goulding A., Dubois Y., Somerville R. S., Greene J. E., 2020, *MNRAS*, 493, 899  
 Habouzit M. et al., 2021, *MNRAS*, 503, 1940  
 Hahn O., Abel T., 2011, *MNRAS*, 415, 2101  
 Harikane Y. et al., 2023, *ApJ*, 959, 39  
 Hopkins P. F., 2015, *MNRAS*, 450, 53  
 Inayoshi K., Onoue M., Sugahara Y., Inoue A. K., Ho L. C., 2022, *ApJ*, 931, L25  
 Jiang L. et al., 2016, *ApJ*, 833, 222  
 Kaviraj S. et al., 2017, *MNRAS*, 467, 4739  
 Khandai N., Di Matteo T., Croft R., Wilkins S., Feng Y., Tucker E., DeGraf C., Liu M.-S., 2015, *MNRAS*, 450, 1349

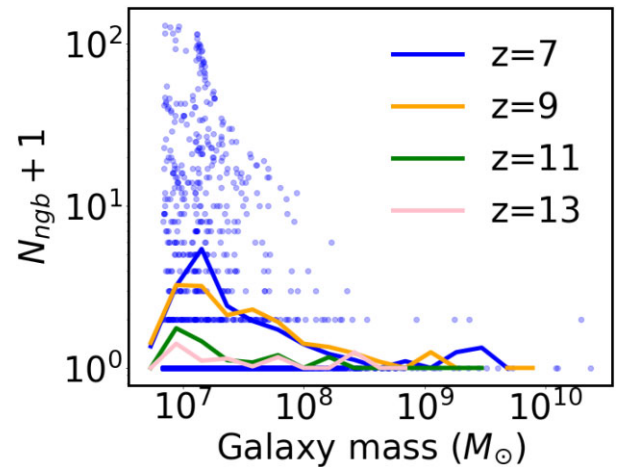
Kroupa P., Subr L., Jerabkova T., Wang L., 2020, *MNRAS*, 498, 5652  
 Larson R. L. et al., 2023, *ApJ*, 953, L29  
 Latif M. A., Schleicher D. R. G., Hartwig T., 2016, *MNRAS*, 458, 233  
 Luo Y., Ardanesh K., Shlosman I., Nagamine K., Wise J. H., Begelman M. C., 2018, *MNRAS*, 476, 3523  
 Luo Y., Shlosman I., Nagamine K., Fang T., 2020, *MNRAS*, 492, 4917  
 Lupi A., Colpi M., Devecchi B., Galanti G., Volonteri M., 2014, *MNRAS*, 442, 3616  
 Ma L., Hopkins P. F., Ma X., Anglés-Alcázar D., Faucher-Giguère C.-A., Kelley L. Z., 2021, *MNRAS*, 508, 1973  
 Madau P., Rees M. J., 2001, *ApJ*, 551, L27  
 Maiolino R. et al., 2023, preprint (arXiv:2305.12492)  
 Marinacci F. et al., 2018, *MNRAS*, 480, 5113  
 Matsuoka Y. et al., 2018, *ApJS*, 237, 5  
 Matsuoka Y. et al., 2019, *ApJ*, 872, L2  
 Mayer L., Capelo P. R., Zwick L., Di Matteo T., 2024, *ApJ*, 961, 76  
 Mortlock D. J. et al., 2011, *Nature*, 474, 616  
 Mushotzky R. et al., 2019, *BAAS*, 51, 107  
 Naiman J. P. et al., 2018, *MNRAS*, 477, 1206  
 Natarajan P., Pacucci F., Ferrara A., Agarwal B., Ricarte A., Zackrisson E., Cappelluti N., 2017, *ApJ*, 838, 117  
 Nelson D. et al., 2018, *MNRAS*, 475, 624  
 Nelson D. et al., 2019a, *Comput. Astrophys Cosmol.*, 6, 2  
 Nelson D. et al., 2019b, *MNRAS*, 490, 3234  
 Ni Y. et al., 2022, *MNRAS*, 513, 670  
 Pakmor R., Bauer A., Springel V., 2011, *MNRAS*, 418, 1392  
 Pakmor R., Pfrommer C., Simpson C. M., Kannan R., Springel V., 2016, *MNRAS*, 462, 2603  
 Pillepich A. et al., 2018a, *MNRAS*, 473, 4077  
 Pillepich A. et al., 2018b, *MNRAS*, 475, 648  
 Pillepich A. et al., 2019, *MNRAS*, 490, 3196  
 Planck Collaboration, 2016, *A&A*, 594, A13  
 Reed S. L. et al., 2017, *MNRAS*, 468, 4702  
 Regan J. A., Johansson P. H., Wise J. H., 2014, *ApJ*, 795, 137  
 Rodriguez-Gomez V. et al., 2015, *MNRAS*, 449, 49  
 Rodriguez-Gomez V. et al., 2019, *MNRAS*, 483, 4140  
 Schaye J. et al., 2015, *MNRAS*, 446, 521  
 Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, *MNRAS*, 452, 575  
 Smith B. D., Regan J. A., Downes T. P., Norman M. L., O'Shea B. W., Wise J. H., 2018, *MNRAS*, 480, 3762  
 Springel V., 2010, *MNRAS*, 401, 791  
 Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726  
 Springel V. et al., 2018, *MNRAS*, 475, 676  
 Springel V., Pakmor R., Zier O., Reinecke M., 2021, *MNRAS*, 506, 2871  
 Taylor P., Kobayashi C., 2014, *MNRAS*, 442, 2751  
 Torrey P. et al., 2019, *MNRAS*, 484, 5587  
 Tremmel M., Karcher M., Governato F., Volonteri M., Quinn T. R., Pontzen A., Anderson L., Bellovary J., 2017, *MNRAS*, 470, 1121  
 Übler H. et al., 2021, *MNRAS*, 500, 4597  
 Venemans B. P. et al., 2015, *MNRAS*, 453, 2259  
 Vogelsberger M. et al., 2014a, *MNRAS*, 444, 1518  
 Vogelsberger M. et al., 2014b, *Nature*, 509, 177  
 Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020, *Nat. Rev. Phys.*, 2, 42  
 Volonteri M., 2007, *ApJ*, 663, L5  
 Volonteri M., Dubois Y., Pichon C., Devriendt J., 2016, *MNRAS*, 460, 2979  
 Volonteri M. et al., 2020, *MNRAS*, 498, 2219  
 Wang F. et al., 2018, *ApJ*, 869, L9  
 Wang E. X., Taylor P., Federrath C., Kobayashi C., 2019, *MNRAS*, 483, 4640  
 Wang F. et al., 2021, *ApJ*, 907, L1  
 Weinberger R. et al., 2017, *MNRAS*, 465, 3291  
 Weinberger R. et al., 2018, *MNRAS*, 479, 4056  
 Weinberger R., Springel V., Pakmor R., 2020, *ApJS*, 248, 32  
 Willott C. J. et al., 2010, *AJ*, 139, 906  
 Wise J. H., Regan J. A., O'Shea B. W., Norman M. L., Downes T. P., Xu H., 2019, *Nature*, 566, 85

Xu H., Wise J. H., Norman M. L., 2013, *ApJ*, 773, 83  
 Yang J. et al., 2019, *AJ*, 157, 236

## APPENDIX A: RELATIONSHIP BETWEEN SUBHALO ENVIRONMENT AND SUBHALO MASS

While our stochastic seed models apply seeding criteria based on galaxy mass and galaxy environment (number of neighbouring haloes  $N_{\text{ngb}}$ ), these two galaxy properties are not completely independent of each other. In Fig. A1, we can clearly see that the galaxy with lower masses tend to have higher number of neighbouring haloes. This is not surprising given the precise definition of  $N_{\text{ngb}}$  described in Section 2.3.2, which only counts neighbouring haloes that exceed the host halo mass of the galaxy. In other words, higher mass galaxies are typically hosted by higher mass haloes. Therefore, for a higher mass galaxy, there are going to be fewer neighbouring haloes that have enough mass to be counted in the  $N_{\text{ngb}}$  calculation. Notably, galaxies of a fixed mass tend to have higher  $N_{\text{ngb}}$  at lower redshifts; this is simply due to higher number of haloes at lower redshifts in general.

Due to this negative correlation between galaxy mass and galaxy environment, applying a *galaxy environment criterion* (that favours seeding in richer environments) can cause the ESDs to form more favourably in lower mass galaxies. This can alter our desired calibration for the *galaxy mass criterion* that we apply prior to the *galaxy environment criterion*. To prevent this from happening, we impose the environment based seeding probabilities  $p_0$  and  $p_1$  to linearly increase with the galaxy mass with a slope  $\gamma > 0$  [see equation (6)]. Depending on the gas-based seed parameters,  $\gamma$  values of  $\sim 1.2$ – $1.6$  (quoted in Table 2) are the ones found to maintain the calibration of the *galaxy mass criterion*. For values significantly higher or lower than  $\sim 1.2$ – $1.6$ , the *galaxy environment criterion* starts to skew the galaxy mass distributions (wherein ESDs are



**Figure A1.** Relationship between galaxy mass and galaxy environment (number of neighbouring haloes  $N_{\text{ngb}}$  as defined in Section 2.3.2) for the galaxy populations in our GAS-BASED simulations. We plot ' $N_{\text{ngb}} + 1$ ' on the y-axis in order to also show galaxies with no neighbours on the log scale. The circles show data points at  $z = 7$ , and the solid lines show the mean trends at  $z = 7, 9, 11$  &  $13$ . We can see that smaller mass galaxies generally have higher number of neighbours. This is not unexpected, given the fact that  $N_{\text{ngb}}$  counts only those neighbours which exceed the host halo mass of the galaxy. And as expected from hierarchical structure formation, galaxies of a given mass have fewer number of neighbours at higher redshifts.

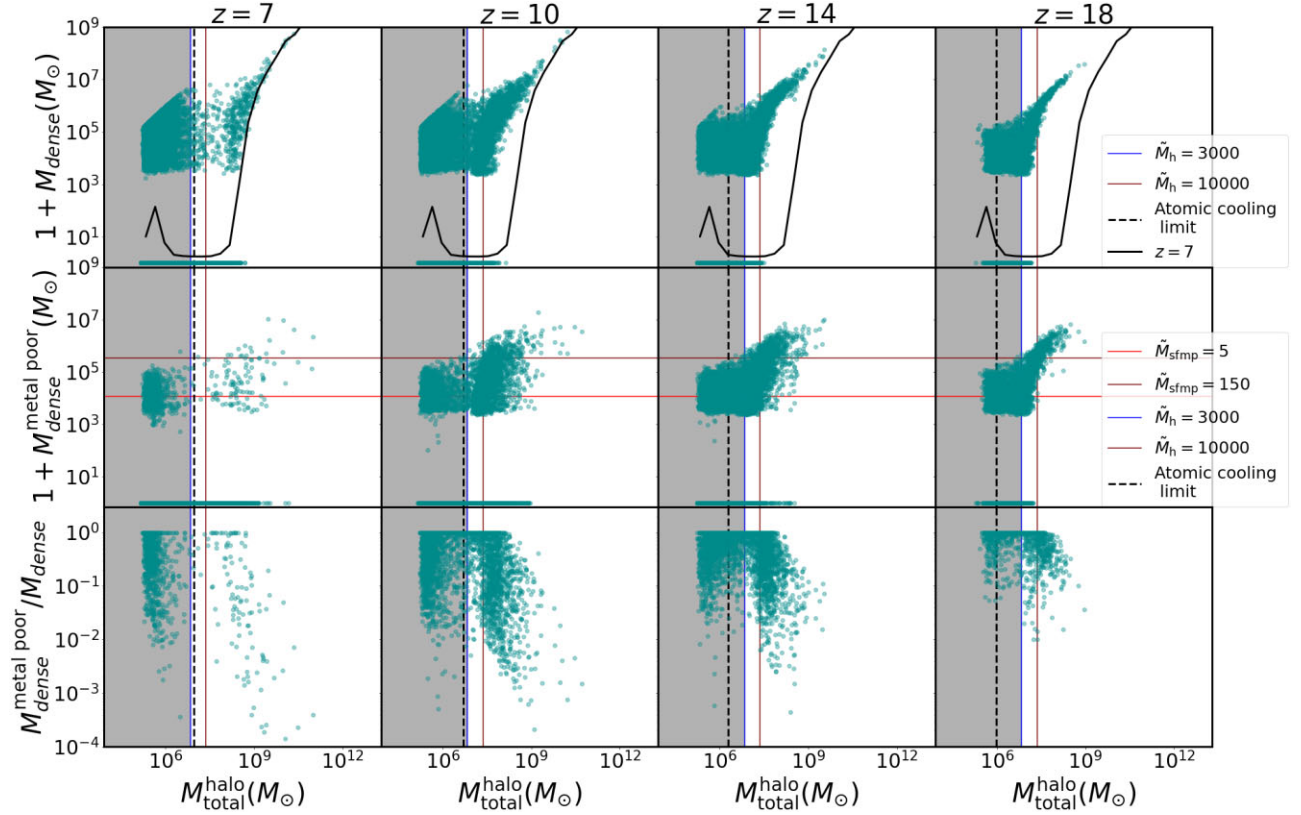


formed) towards higher or lower masses respectively, compared to our desired calibration. Lastly, incorporating this linear dependence with  $\gamma > 0$  is also physically motivated. This is because it captures the notion that, for a given value of  $N_{\text{ngb}}$ , seeding should be favoured in a galaxy with higher mass because it exists in a more extreme environment compared to a lower mass galaxy with the same  $N_{\text{ngb}}$ .

## APPENDIX B: EVOLUTION OF DENSE AND METAL-POOR GAS IN HALOES

In Fig. B1, we show scatter plots of the dense gas mass ( $M_{\text{dense}}$ ) and dense & metal-poor gas mass ( $M_{\text{dense}}^{\text{metal poor}}$ ) versus the total halo mass ( $M_{\text{total}}^{\text{halo}}$ ) at different redshifts. The top row shows that there is a straightforward positive correlation between the halo mass and dense gas mass, except at the lowest halo masses wherein

the results are likely impacted by the finite simulation resolution. Notably, several of these lowest mass objects are spuriously identified gas clumps with very little DM mass. In addition, these haloes are also significantly below the atomic cooling threshold (virial Temperature of  $10^4$  K, dashed black vertical lines), which we do not self-consistently simulate due to the absence of  $H_2$  cooling. With our adopted halo mass thresholds ( $\tilde{M}_h = 3000$  &  $10000$ ), we avoid seeding in these lowest mass haloes (marked as shaded grey region). Hereafter, we shall focus only on haloes with reasonably well converged stellar and gas properties (outside the grey region). The top row also shows that at fixed halo mass, the dense gas mass (top row) steadily decreases with time (green circles vs. black solid line). This is a simple consequence of cosmological expansion, which increases the atomic cooling threshold with time. As a result, at later times, haloes of a given mass have lower ability to contain



**Figure B1.** Dense gas masses ( $M_{\text{dense}}$ , top panels), dense and metal-poor gas masses ( $M_{\text{dense}}^{\text{metal poor}}$ , middle panels) and their ratios (bottom panels) are plotted versus the total mass ( $M_{\text{total}}$ ) for haloes in different snapshots within the GAS-BASED suite that explicitly resolves the  $2.3 \times 10^3 M_{\odot}$  DGBs. The different columns show different redshift snapshots; however, the mean trend at  $z = 7$  is plotted as solid black line in all the top panels to clearly see the redshift evolution. We added 1 to the y-axis in order to include haloes with no dense gas on the log-scale. The black dashed vertical lines correspond to the atomic cooling limit (halo virial temperature  $T_{\text{vir}} = 10^4$  K). The red and orange horizontal lines correspond to the seeding thresholds of  $\tilde{M}_{\text{simp}} = 5$  &  $150$ , respectively. The blue and brown vertical lines correspond to the seeding thresholds of  $\tilde{M}_h = 3000$  &  $10000$  respectively. Shaded regions correspond to the lowest mass objects below the  $\tilde{M}_h = 3000$  limit, which are also below the atomic cooling threshold. We avoid seeding in these haloes since they are impacted by the limited mass resolution and lack of sufficient physics (absence of  $H_2$  cooling). Top panels show that at fixed halo mass, dense gas mass decreases with time due to cosmological expansion. Middle and bottom panels show that despite stronger metal enrichment in more massive haloes, the dense and metal-poor gas mass is still positively correlated with halo mass. As a result, DGB formation is favoured in more massive haloes when the primary driver is metal enrichment.

gas and collapse it to high enough densities to form stars. This is overall responsible for the steady increase in DGB forming halo masses with time in epochs where dense gas formation is the primary driver of DGB formation (seen in Section 3.3). In the bottom row, the fraction of dense gas mass that is also metal-poor ( $<10^{-4} Z_{\odot}$ ), sharply decreases with halo mass at fixed redshift. This is not surprising because metal enrichment is expected to be more prevalent in massive haloes. Regardless, the middle row shows that the overall dense and metal-poor gas mass continues to be positively

correlated with halo mass. This is simply due to more massive haloes having higher overall dense gas mass. As a result, whenever metal enrichment becomes the primary driver of DGB formation, it leads to a more rapid increase in the DGB forming halo mass with time, compared to that of simple cosmological expansion (see again Section 3.3).

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.