Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **493,** 1178–1196 (2020)
Advance Access publication 2020 February 4

doi:10.1093/mnras/staa324

# Clustering with JWST: Constraining galaxy host halo masses, satellite quenching efficiencies, and merger rates at $z = 4{-}10$

Ryan Endsley,[1★] Peter Behroozi,[1] Daniel P. Stark,[1] Christina C. Williams,[1†] Brant E. Robertson,[2,3] Marcia Rieke,[1] Stefan Gottlöber[4] and Gustavo Yepes[5,6]

[1] *Steward Observatory, University of Arizona, 933 N Cherry Ave, Tucson, AZ 85721, USA*
[2] *Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA*
[3] *Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ 08540, USA*
[4] *Leibniz-Institut für Astrophysik, D-14482 Potsdam, Germany*
[5] *Departamento de Física Teórica, Módulo 8, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*
[6] *CIAFF, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*

Accepted 2020 January 31. Received 2020 January 28; in original form 2019 July 4

## ABSTRACT

Galaxy clustering measurements can be used to constrain many aspects of galaxy evolution, including galaxy host halo masses, satellite quenching efficiencies, and merger rates. We simulate *JWST* galaxy clustering measurements at z ∼ 4–10 by utilizing mock galaxy samples produced by an empirical model, the UNIVERSEMACHINE. We also adopt the survey footprints and typical depths of the planned joint NIRCam and NIRSpec Guaranteed Time Observation program planned for Cycle 1 to generate realistic *JWST* survey realizations and to model high-redshift galaxy selection completeness. We find that galaxy clustering will be measured with $\gtrsim 5\sigma$ significance at $z \sim 4$–10. Halo mass precisions resulting from Cycle 1 angular clustering measurements will be ∼0.2 dex for faint ($-18 \gtrsim M_{\rm UV} \gtrsim -19$) galaxies at $z \sim 4$–10 as well as ∼0.3 dex for bright ($M_{\rm UV} \sim -20$) galaxies at $z \sim 4$–7. Dedicated spectroscopic follow-up over $\sim 150\,{\rm arcmin}^2$ would improve these precisions by ∼0.1 dex by removing chance projections and low-redshift contaminants. Future *JWST* observations will therefore provide the first constraints on the stellar–halo mass relation in the epoch of reionization and substantially clarify how this relation evolves at $z > 4$. We also find that ∼1000 individual satellites will be identifiable at $z \sim 4$–8 with *JWST*, enabling strong tests of satellite quenching evolution beyond currently available data ($z \lesssim 2$). Finally, we find that *JWST* observations can measure the evolution of galaxy major merger pair fractions at $z \sim 4$–8 with ∼0.1–0.2 dex uncertainties. Such measurements would help determine the relative role of mergers to the build-up of stellar mass into the epoch of reionization.

**Key words:** galaxies: high-redshift – dark ages, reionization, first stars – cosmology: large-scale structure of Universe.

## 1 INTRODUCTION

Over the past decade, multiple observational programs have opened a window into the first billion years following the big bang (see Stark 2016 for a review). Ground and space-based photometric campaigns have revealed over one thousand galaxies at $z \gtrsim 6$, enabling constraints on luminosity functions and star formation rate densities out to $z \sim 10$ (e.g. Atek et al. 2015; Bouwens et al. 2015; Finkelstein et al. 2015a; Livermore, Finkelstein & Lotz 2017; Oesch et al. 2018; Ono et al. 2018) as well as galaxy stellar mass functions out to $z \sim 8$ (e.g. Stark et al. 2009; González et al. 2011; Grazian et al. 2015; Song et al. 2016). *JWST* promises to advance luminosity and stellar mass function constraints at high redshifts via detailed spectra and extremely deep ($m \sim 29$–30) photometry in the near to mid-infrared. However, less appreciated is *JWST*'s potential to place new constraints on early galaxy evolution via high-redshift galaxy clustering measurements.

Galaxy clustering measurements are commonly used to infer halo masses by exploiting the strong relation between halo mass and clustering strength (e.g. Mo & White 1996; Tinker et al. 2010). These halo masses can then be used to infer the stellar–halo mass relation, a fundamental constraint on the connection between galaxies and their host haloes (see Wechsler & Tinker 2018 for a review). Significant effort has been devoted to measuring $z \gtrsim 4$

★ E-mail: rendsley@email.arizona.edu
† NSF Fellow.

galaxy clustering, revealing that the most UV luminous galaxies at $z \sim 4$–7 tend to reside in the most massive haloes (Barone-Nugent et al. 2014; Harikane et al. 2016, 2018; Ishikawa et al. 2017; Hatfield et al. 2018). Combined with stellar mass constraints, high-redshift clustering results may already be signalling evolution in gas cooling efficiency, feedback efficiency, and merger rates from $z \sim 4$ to $z \sim 7$ (Harikane et al. 2016, 2018). However, stringent ($\lesssim$0.5 dex error) halo mass constraints at $z \geq 4$ are primarily limited to bright ($M_{\rm UV} \sim -20$) galaxies at $z \lesssim 5$ (Barone-Nugent et al. 2014; Harikane et al. 2018). It therefore remains unclear how the stellar–halo mass relation evolves at $z \gtrsim 6$, leaving a gap in our understanding of galaxy evolution in the early Universe and, in particular, the epoch of reionization.

*JWST*'s greatly increased sensitivity is expected to significantly improve the precision of high-redshift clustering measurements. Deep ($m \gtrsim 29$–30) NIRCam photometry from Cycle 1 programs will enable the detection of $\sim$5000 galaxies at $z > 6$ and $\sim$300 at $z > 9$ (Mason, Trenti & Treu 2015; Tacchella et al. 2018; Williams et al. 2018). Such increased numbers will highly benefit clustering measurements because Poisson noise falls as the linear inverse of galaxy number density (Peebles 1980; Landy & Szalay 1993). Deeper imaging also enables the selection of lower mass haloes which are less susceptible to cosmic variance (e.g. Somerville et al. 2004; Trenti & Stiavelli 2008).

*JWST*'s NIRSpec will also enable the first spatial galaxy clustering measurements at $z > 4$ thanks to its multiplex capabilities and sensitivity to strong rest-optical lines at these redshifts (e.g. H$\alpha$ and [O III]; Chevallard et al. 2019; De Barros et al. 2019). These spatial clustering measurements will avoid chance projections as well as low-redshift contaminants present in angular clustering measurements based on broad-band photometric redshifts.

All of these advancements should improve constraints on galaxy halo masses, and consequently, the stellar–halo mass relation at $z > 4$. Recent studies have predicted the clustering of $z \sim 8$–10 haloes expected to host galaxies detectable with *JWST* (Bhowmick et al. 2018; Zhang et al. 2019b). Adopting idealized assumptions, Zhang et al. (2019b) concluded that galaxy clustering should be measurable to $\sim$4–5$\sigma$ significance at $z \sim 10$. Here, we simulate *JWST* galaxy clustering measurements at $z \sim 4$–10, noting that our methodology provides more observationally realistic predictions compared to Zhang et al. (2019b) in the following ways. First, we model high-redshift galaxy selection completeness by adopting the typical depths of a planned Cycle 1 program, the *JWST* Advanced Deep Extragalactic Survey (JADES; Williams et al. 2018), and utilizing an empirical model, the UNIVERSEMACHINE (Behroozi et al. 2019), to assign galaxy properties to haloes from a dark matter simulation. Secondly, we simulate clustering measurements over the exact survey footprint of the JADES program, including detector and pointing gaps, to accurately account for boundary effects. Thirdly, we account for peculiar motion distortions in spatial clustering measurements. Finally, we include satellite galaxies in our simulated measurements, which can significantly impact the recovered clustering signal because clustering strengths are density dependent. We furthermore simulate the process of inferring halo masses from these clustering measurements to provide the first predictions of $z \sim 4$–10 halo mass precisions resulting from future *JWST* surveys.

We also investigate how *JWST*'s improved sensitivity and spectroscopic capabilities will enable studies of satellite quenching and merger rates in the early Universe. Results from ground-based facilities have shown that satellites out to $z \sim 2$ are systematically more quenched than field galaxies at fixed stellar mass (e.g. van den Bosch et al. 2008; Kawinwanichakij et al. 2016). These findings imply a strong environmental dependence on star-formation properties within the last $\sim$10 Gyr, possibly driven by the presence of a hot circumgalactic medium surrounding the host halo. We therefore test how well *JWST* will facilitate the identification of high-redshift satellites and thereby push satellite quenching efficiency measurements to earlier epochs. We also investigate how *JWST* will improve galaxy pair fraction measurements at $z > 4$. Galaxy pair fractions can be used to infer galaxy merger rates and thus the relative contribution of mergers to stellar mass buildup throughout cosmic time (e.g. Lotz et al. 2011; Mundy et al. 2017) as well as the AGN-merger connection (Kocevski et al. 2012).

In Section 2, we describe our methods for generating realistic mock *JWST* survey realizations (Section 2.1), modelling high-redshift selection completeness (Section 2.2), calculating clustering strengths via the two-point correlation function (Section 2.3), and inferring halo masses from our simulated clustering measurements (Section 2.4). Our results are presented in Section 3 where we discuss the quality of our simulated $z \sim 4$–10 galaxy clustering measurements (Section 3.1) and the halo mass uncertainties expected to result from these measurements (Section 3.2), concluding with a discussion of how *JWST* will improve constraints on the stellar–halo mass relation at $z \gtrsim 4$. In Section 4, we discuss how well *JWST* observations will enable the identification of individual satellites (Section 4.1) as well as measurements of galaxy major merger pair fractions (Section 4.2) at $z \gtrsim 4$. Our main conclusions are listed in Section 5.

All magnitudes are quoted in the AB system (Oke & Gunn 1983). Luminosity to stellar mass conversions assume a Chabrier (2003) stellar initial mass function. We adopt a flat, $\Lambda$CDM cosmology with parameters ($\Omega_{\rm M} = 0.307$, $\Omega_\Lambda = 0.693$, $\Omega_{\rm B} = 0.048$, $h = 0.678$, $\sigma_8 = 0.823$, $n_s = 0.96$) consistent with *Planck* results (Planck Collaboration XIII 2016). Halo masses follow the Bryan & Norman (1998) spherical overdensity definition and refer to peak historical halo masses extracted from the merger tree except where otherwise specified. All distances are quoted in comoving coordinates unless otherwise stated.

## 2 METHODS

In this section, we describe our methods for simulating $z \sim 4$–10 *JWST* galaxy clustering measurements and inferring halo masses from those measurements. We begin by detailing our procedure for generating realistic *JWST* survey realizations based on a planned Cycle 1 GTO program, JADES (Williams et al. 2018), in Section 2.1. We then model the selection of high-redshift galaxies to account for completeness, as well as low-redshift interlopers in the angular clustering measurements, as described in Section 2.2. Our methods for simulating galaxy clustering measurements are then detailed in Section 2.3, followed by a description of how we infer halo masses from these measurements in Section 2.4.

### 2.1 Generating mock *JWST* survey realizations

Our approach to generating mock *JWST* survey realizations begins by taking halo positions from a dark matter simulation so that predicted clustering strengths arise from $\Lambda$CDM theory. Here, we use the Very Small MultiDark Planck (VSMDPL[1]) dark matter

[1] Details of VSMDPL can be found at https://www.cosmosim.org/cms/simulations/vsmdpl/. The series of MultiDark simulations is summarized at https://www.cosmosim.org/cms/simulations/simulations-overview/.

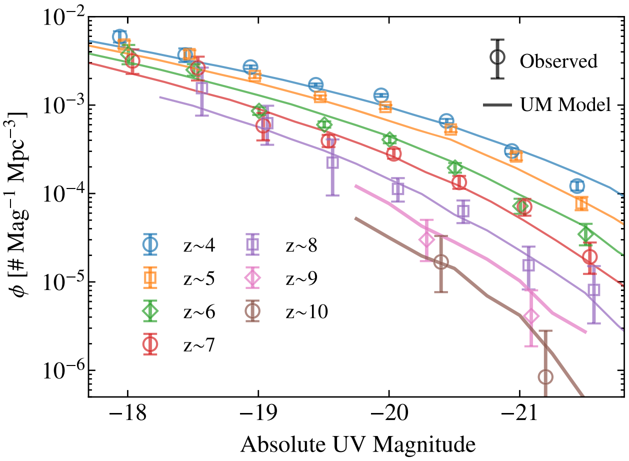

**Figure 1.** Comparison of the $z \sim 4$–10 UV luminosity functions from the entire mock galaxy catalogue output by the UNIVERSEMACHINE model (UM; Behroozi et al. 2019) used in this work (lines) versus the observational constraints input into that model. These observational data are taken from Finkelstein et al. (2015a) at $z \sim 4$–8 and Bouwens et al. (2016a) at $z \sim 9$–10. Overall, the UV luminosity functions output by the model are well matched to input observations as expected.

simulation, which continues the series described in Klypin et al. (2016) to higher resolution. VSMDPL was run using GADGET-2 (Springel 2005) with a box size of (160 Mpc $h^{-1}$)$^3$, containing $3840^3$ particles with a high mass ($6.2 \times 10^6$ M$_\odot$ $h^{-1}$) and force (1 kpc $h^{-1}$) resolution. The cosmological parameters adopted in VSMDPL are from Planck Collaboration XVI (2014) and listed in Section 1. Haloes and sub-haloes within VSMDPL were identified using the ROCKSTAR algorithm (Behroozi, Wechsler & Wu 2013a), and merger trees were constructed using the CONSISTENT TREES algorithm (Behroozi et al. 2013b).

We assign galaxy properties (including UV luminosities) to the VSMDPL haloes using an empirical model, the UNIVERSEMACHINE (hereafter referred to as the UM model; Behroozi et al. 2019). Briefly, the UM model utilizes observational constraints spanning $z =$ 0–10 to empirically infer halo star formation histories as a function of potential well depth, assembly history, and redshift. This procedure is performed using a Markov Chain Monte Carlo algorithm where each chain results in a mock catalogue of galaxies paired with haloes. In this work, we use the best-fitting catalogue, hereafter referred to as the UM-VSMDPL mock catalogue. Within the UM model, UV luminosities are computed from the star formation histories assigned to each halo using FSPS v3.0 (Conroy, Gunn & White 2009; Conroy & Gunn 2010) and a Chabrier (2003) stellar initial mass function, with dust attenuation scaled to match high-redshift observations (Bouwens et al. 2016b).

At high redshifts, the galaxy–halo connection set by the UM model is primarily governed by the input empirical UV luminosity functions. These are taken from Finkelstein et al. (2015a) and Bouwens et al. (2016a) at $z \sim$ 4–8 and $z \sim$ 9–10, respectively. As shown in Fig. 1, the UV luminosity functions from the UM-VSMDPL mock catalogue used in this work match the input constraints as expected. Additional high-redshift empirical constraints input into the UM model include UV–stellar mass relations ($z$ = 4–8), specific SFRs ($z$ = 0–8), and cosmic SFRs ($z$ = 0–9). See Behroozi et al. (2019) for additional details of the UM model including the calculation of UV luminosities.

In Appendix A, we test how our results change if we instead adopt an abundance matching approach to assign UV luminosities to haloes, finding no strong dependence. We also check whether our conclusions are dependent on the choice of empirical $z \sim 4$–10 luminosity functions used to set the galaxy–halo connection at high redshifts. The only substantial difference occurs when adopting published $z \sim 10$ luminosity functions that are ∼0.3 dex lower than those adopted by the UM model. This choice leads to significantly larger cosmic and Poisson variance in pair counts at this redshift, discussed further in Appendix B. Finally, we also test whether our results are dependent on the size of the dark matter simulation (and hence the modelled cosmic variance) input into the UM model. As discussed in Appendix C, our results are largely similar after adopting a ∼4 × larger dark matter simulation. Again, only at $z \sim 10$ is a significant difference found, though to a smaller degree than from adopting lower published $z \sim 10$ luminosity functions as described above.

To incorporate realistic sample variance into our simulated clustering measurements, we seek to only use mock galaxies that fall within the footprints of a planned Cycle 1 GTO program, the *JWST* Advanced Deep Extragalactic Survey (JADES; Williams et al. 2018). Briefly, JADES will observe the GOODS-S and GOODS-N *HST* Legacy fields with both photometry via NIRCam over 236 arcmin$^2$ and multi-object spectroscopy via NIRSpec over 142 arcmin$^2$. We therefore overlay the JADES footprints (including detector and pointing gaps) on mock survey volumes (i.e. lightcones) extracted from the UM-VSMDPL catalogue.

We choose to use only mock survey volumes that best match observed high-redshift galaxy number densities within the two GOODS fields to be observed by JADES. Our reasons for doing so are twofold. First, this procedure ensures our simulated clustering measurements will possess realistic Poisson variance which is of particular importance for the brightest and highest redshift samples. Secondly, it anchors the intrinsic clustering strengths of high-redshift galaxies within the GOODS fields because it has been shown that clustering strengths correlate with number density (i.e. environment; Croton, Gao & White 2007; Zehavi et al. 2018). We generated 500 mock survey volumes in total by first extracting 100 lightcones from the UM-VSMDPL catalogue, each spanning $z =$ 0–20 and subtending 50 arcmin in right-ascension by 25 arcmin in declination. Because each lightcone is large enough to encompass multiple JADES footprints for either field, we choose to extract five mock survey volumes from each lightcone centred at five evenly spaced RA positions within the lightcone. These mock volumes span an area of 0.4 × 0.4 deg$^2$ to ensure that each can host a full footprint for either field.

We select the best matching mock survey volumes by first calculating the UV luminosity functions within each mock volume and then performing $\chi^2$ fits against observed luminosity functions within the two GOODS fields. Specifically, we adopt the luminosity functions reported by Finkelstein et al. (2015a) at $z \sim 4$–8 and Oesch et al. (2014) at $z \sim$ 10. To avoid fitting incomplete data, we only fit to the bright-end ($M_{\rm UV} \leq -20$) luminosity function data points which are $\gtrsim$50 per cent complete (see fig. 9 in Finkelstein et al. 2015a). We perform the fitting procedure for each GOODS field and each redshift interval individually. We also only use the central 0.2 × 0.2 deg$^2$ of each mock volume as that better approximates the area covered by *HST* in each GOODS field.[2]

One of the primary goals of this work is to determine 68 per cent confidence interval statistics for both clustering strength measurements and inferred halo masses. We find that generating 100

[2] Each mock survey volume has an area of 0.4 × 0.4 deg$^2$ because parallel observations with JADES lead to more extended coverage relative to *HST*.

JADES realizations is sufficient for this purpose. Generating more realizations would require us to use mock survey volumes that are a poorer match to the empirical luminosity functions. To generate 100 realizations of the JADES program, we take the 10 best-fitting mock survey volumes (out of 500) for each GOODS field and generate one realization for every possible pair of GOODS-N and GOODS-S volumes. The ten best-fitting volumes for each field and redshift interval have reduced $\chi^2$ values less than three in all cases and are often $\lesssim$1.

We note that because we anchor to the GOODS luminosity functions, our simulated clustering strength measurements will not necessarily be representative of the average galaxy population at a given redshift. If the GOODS fields are under(over)-dense at $z \sim 4$–10, then our simulated clustering strengths will be higher (lower) than that in an average field.[3] Current *HST* measurements (Bouwens et al. 2015) suggest that the GOODS fields, when combined, have roughly average number densities at $z \sim 4$, $z \sim 6$, and $z \sim 7$.[4] However, both GOODS fields are slightly underdense (∼0.8–0.9 × the average number density) at $z \sim 6$ and GOODS-S is very underdense (∼0.5 × the average number density) at $z \sim 8$. Therefore, it is likely that the clustering measurements from the JADES program will have boosted clustering strengths at $z \sim 6$ and $z \sim 8$ relative to the general galaxy population. It will be possible to model the strength of this boost using improved luminosity function constraints within each field from *JWST*.

## 2.2 Modelling selection of high-redshift galaxies

We now determine which mock galaxies would be used for high-redshift clustering measurements by simulating the selection of $z \sim 4$–10 galaxies. In doing so, we adopt typical depths of the planned Cycle 1 program, JADES, and utilize empirically calibrated mock galaxy photometry and spectra produced by the JAdes extraGalactic Ultradeep Artificial Realizations (JAGUAR) package[5] (Williams et al. 2018). The overall goal here is to determine, as a function of redshift and UV magnitude, the photometric and spectroscopic selection completeness for angular and spatial clustering measurements, respectively.

For the angular clustering measurements, we assume that high-redshift galaxies will be photometrically selected using colour cuts as commonly done with *HST* imaging (e.g. Stark, Ellis & Ouchi 2011; González et al. 2012; Oesch et al. 2014; Bouwens et al. 2015, 2019). The colour cuts adopted in this work are detailed in Appendix D1 and are designed to separate galaxies into photometric redshift intervals of $z$ = [3.7,4.4], [4.4,5.5], [5.5,6.7], [6.7,7.7], [7.7,9.0], and [9.0,11.0]. We will hereafter refer to these as the $z \sim 4$, 5, 6, 7, 8, and 10 intervals, respectively.

The photometric selection completeness, $\mathscr{C}_P(\mathrm{z}, M_{\mathrm{UV}})$, is calculated as the fraction of mock JAGUAR galaxies (as a function of redshift and UV magnitude) that are selected to lie within the redshift interval of interest. Prior to this mock selection process, we add noise to the photometry of each JAGUAR galaxy 50 times with noise set equal to the medium JADES/NIRCam depths (Williams et al. 2018) for *JWST* photometry and deep GOODS/ACS depths (Bouwens et al. 2015) for optical *HST* photometry.[6] We find that >50 per cent of $z \sim 4$, 5, 6, 7, 8, and 10 galaxies with apparent rest-UV magnitudes of $m_{\mathrm{UV}}$ = 28.0, 28.5, 28.8, 29.2, 29.2, and 29.5, respectively, are selected to lie within the correct photometric redshift interval. We further find that only using galaxies brighter than these magnitudes restricts the low-redshift ($z < 3$) contamination fraction to reasonably small values (≤15 per cent) within the context of JAGUAR. Therefore, mock galaxies in the UM-VSMDPL catalogue fainter than the above magnitudes are ignored while brighter galaxies are selected at random for the simulated angular clustering measurements with probability equal to $\mathscr{C}_P(\mathrm{z}, M_{\mathrm{UV}})$.

**Table 1.** The limiting rest-UV apparent magnitudes used to select high-redshift galaxies when simulating *JWST* clustering measurements. All of the angular apparent magnitudes correspond to an approximate absolute magnitude of $M_{\mathrm{UV}} = -18$. These magnitudes were determined by modelling high-redshift selection completeness using typical depths of the planned Cycle 1 JADES program as described in Section 2.2. We do not simulate spatial clustering measurements at $z \sim 10$ because of the low spectroscopic completeness expected in this regime.

| | Adopted limiting rest-UV apparent magnitudes | | |
|---|---|---|---|
| | Targeted redshift interval | Angular clustering | Spatial clustering |
| $z \sim 4$ | 3.7–4.4 | 28.0 | 28.0 |
| $z \sim 5$ | 4.4–5.5 | 28.5 | 28.5 |
| $z \sim 6$ | 5.5–6.7 | 28.8 | 28.8 |
| $z \sim 7$ | 6.7–7.7 | 29.2 | 29.0 |
| $z \sim 8$ | 7.7–9.0 | 29.2 | 28.7 |
| $z \sim 10$ | 9.0–11.0 | 29.5 | – |

Because NIRSpec microshutter assembly (MSA) targets must first be photometrically identified, the overall spectroscopic completeness, $\mathscr{C}_S(\mathrm{z}, M_{\mathrm{UV}})$, is determined by the spectroscopic redshift completeness, the photometric selection completeness, and the availability of MSA slits for targets. We assume that spectroscopic redshifts can be obtained from galaxies with at least one emission line detected at $\geq 5\sigma$. Therefore, spectroscopic redshift completeness is calculated as the fraction of JAGUAR mock galaxies (as a function of redshift and UV luminosity) that have at least one emission line brighter than the $5\sigma$ flux limit set by the planned medium JADES/NIRSpec depths. For the low-resolution ($R \sim 100$) prism, the $5\sigma$ limiting flux is $1.4 \times 10^{-18}$ erg s$^{-1}$ cm$^{-2}$ at 2.5 µm. The medium-resolution ($R \sim 1000$) G235M and G395M grisms are approximately twice as sensitive with $5\sigma$ flux limits of 0.8 and $0.5 \times 10^{-18}$ erg s$^{-1}$ cm$^{-2}$ at 2.5 and 4.5 µm, respectively. Flux limits at other wavelengths are computed using the online sensitivity curves[7] for each disperser. We note that [O III]+H$\beta$ equivalent widths in the JAGUAR catalogue are consistent with inferences from observations (e.g. Labbé et al. 2013).

We assume that MSA targets will only include $z \gtrsim 4$ photometric candidates that have an estimated >50 per cent spectroscopic redshift completeness and are brighter than the limiting photometric magnitudes summarized in Table 1. As detailed in Appendix E, we find that the number of $z \gtrsim 4$ photometric candidates will not

[3] An underdense field translates to higher clustering strengths because haloes at fixed mass assemble from stronger peaks within the underlying density field and clustering scales with the peak strength (e.g. Mo & White 1996). Conversely, in overdense regions, haloes at fixed mass form from weaker peaks and therefore have lower clustering strengths.

[4] At $z \sim 10$, Poisson uncertainties are too large to determine whether either field is over- or underdense.

[5] We use the JAGUAR package because the UM model currently does not directly calibrate mock galaxy photometry and spectra.

[6] Such *HST* coverage is expected over a large fraction of the two GOODS fields to be imaged by JADES.

[7] https://jwst-docs.stsci.edu/display/JTI/NIRSpec + Sensitivity

exceed the number of available MSA slits assuming that the $z\sim4$–6 candidates will be targeted with the $R\sim100$ prism while the $z\gtrsim7$ candidates will be targeted with the $R\sim1000$ G235M and G395M grisms. Adopting those dispersers, we infer that $z\sim4$, 5, 6, 7, and 8 galaxies with $m_{\rm UV}=28.0$, 28.5, 28.8, 29.0, and 28.7 are >50 per cent likely to have at least one emission line detection with NIRSpec. We therefore adopt these as the limiting magnitudes for simulating spatial clustering measurements where brighter galaxies are selected with probability equal to $\mathscr{C}_S(z, M_{\rm UV})$. This net spectroscopic completeness is calculated as the convolution of the spectroscopic detection completeness and photometric selection completeness, both as a function of redshift and UV luminosity. This assumes that every $z\gtrsim4$ photometric candidate in a given NIRSpec pointing can eventually be placed on an MSA mask. In practice, it is likely that only ~50 per cent of such sources can be placed on MSA masks in the Cycle 1 JADES program alone to avoid overlapping spectra. We discuss how a 50 per cent slit placement efficiency would impact our simulated spatial clustering measurements in Appendix E.

We do not consider spectroscopic samples at $z\sim10$ because the very strong rest-optical lines such as [O III] and H$\alpha$ are no longer accessible to NIRSpec at $z>9$, suggesting low spectroscopic completeness in this regime. While moderately strong lines such as [O II]$\lambda$3729 and [Ne III]$\lambda$3869 are still accessible to NIRSpec out to $z\sim12$, these lines are expected to be $\gtrsim3\times$ weaker than [O III] at these redshifts (Tang et al. 2019). As such, we do not expect a sufficient number of $z\sim10$ galaxies to be spectroscopically detectable for spatial clustering measurements, at least within the medium JADES/NIRSpec survey. We discuss ways to address this challenge in Section 3.1.

When modelling the photometric selection completeness, we do not account for high-redshift sources missed due to bright foreground objects (e.g. low-redshift galaxies or stars). Using deep *HST* imaging, we estimate that only ~5–15 per cent of $z\sim4$–10 galaxies will be missed for this reason (Appendix D3). Because Poisson noise in clustering measurements scale with the number of sources, this suggests that our predicted clustering measurement significances would decrease by an equivalent ~10 per cent if we did account for this. However, this is much less than the ~30–40 per cent increase that would result from including the intrahalo clustering measurements (Section 3.1) and we therefore ignore the impact of bright foreground sources.

### 2.3 Measuring clustering strengths via the two-point correlation function

We simulate galaxy clustering measurements using the two-point correlation function as it allows us to consider the full scale-dependent clustering of galaxies. This approach is similar to recent studies (e.g. Harikane et al. 2016, 2018; Hatfield et al. 2018; Zhang et al. 2019b) and goes beyond past studies that have focused on a single-variable measurement of, e.g. the correlation length. The spatial two-point correlation function, $\xi(r)$, quantifies the excess pair counts at a given 3D separation distance, $r$, compared to pair counts for randomly distributed galaxies (e.g. Peebles 1980):

$$dP = n\,[1+\xi(r)]\,dV, \tag{1}$$

where d$P$ is the number of excess pairs, $n$ is the galaxy number density, and d$V$ is the volume element. Because peculiar motions distort the observed line-of-sight distance, we instead measure the projected spatial correlation function,

$$w_{\rm p}(r_{\rm p}) = \int_{-\pi_{\rm max}}^{\pi_{\rm max}} \xi(r_{\rm p}, \pi)\, d\pi. \tag{2}$$

Here, the 3D separation distance, $r$, has been split into a projected 2D distance, $r_{\rm p}$, and a line-of-sight distance, $\pi$, that includes peculiar motion distortions. We adopt $\pi_{\rm max}$ values of 10, 7, 5, 5, and 3 Mpc $h^{-1}$ for the $z\sim4$, 5, 6, 7, and 8 bins, respectively. These values were chosen to encompass the majority of peculiar motion distortions, calculated via the halo velocity information in the UM-VSMDPL mock catalogue. Our adopted $\pi_{\rm max}$ values also satisfy the resolution limitations of each NIRSpec disperser (see Section 2.2), assuming that emission line wavelength measurements can be made to within ~1/4 of the resolution element.

We calculate $w_{\rm p}(r_{\rm p})$ using the Landy & Szalay (1993) estimator:

$$w\,(D1, D2) = \frac{1}{R1R2}\,[D1D2 - 2\times D1R2 + R1R2]\,, \tag{3}$$

where $D1D2$ is the number of pairs between two galaxy samples ($D1$ and $D2$) at a given projected separation distance ($r_{\rm p}$) and $D1R2$ ($R1R2$) are the number of data–random (random–random) pairs, appropriately normalized (see Landy & Szalay 1993). To ensure that our random samples contribute negligible Poisson noise, we generate them using a surface density of $10^5$ objects per arcmin$^2$ which is more than 1000 × that of our most populated mock galaxy sample.

We simulate angular clustering measurements using the angular two-point correlation function, w($\theta$). For this, we again use the Landy & Szalay (1993) estimator and a random sample with surface density of $10^5$ objects per arcmin$^2$. Because low-redshift galaxies contaminate high-redshift photometrically selected samples, we account for the resulting reduction in measured angular two-point correlation functions (at most 50 per cent) via the methods described in Appendix D2.

We assume that galaxy clustering will be observationally measured in samples binned by redshift and threshold apparent rest-UV magnitudes, $m_{\rm UV}^{\rm th}$ (e.g. $m_{\rm UV}^{\rm th}=29$ means that only $m_{\rm UV}<29$ galaxies are included in the sample). Such binning is common in observational clustering studies at high redshifts (e.g. Harikane et al. 2016, 2018; Hatfield et al. 2018). First, we adopt a threshold magnitude equal to the limiting magnitude at each redshift (see Table 1). These mock galaxy samples are autocorrelated ($D1=D2$) and simply correspond to the entire sample used for our simulated clustering measurements at their respective redshift. We also consider galaxy samples with threshold magnitudes one or two magnitudes brighter than the limiting magnitude for redshifts where sufficient numbers of such bright galaxies are expected to lie within our adopted survey area. These brighter samples are cross-correlated with the entire galaxy sample at their respective redshift to minimize Poisson noise in pair counts. Table 2 shows the full list of our adopted threshold magnitudes.

We simulate correlation function measurements for each *JWST* survey realization described in Section 2.1. Because Poisson error can be significant in any single mock clustering measurement, we also simulate correlation function measurements using all twenty mock survey volumes adopted for each redshift (see Section 2.1). These 'true' correlation functions are used to assess the accuracy of simulated JADES clustering measurements via the $\chi^2$ statistic,

$$\chi^2 = \sum_i \frac{\left(w_{i,\rm meas} - w_{i,\rm true}\right)^2}{\sigma_i^2}. \tag{4}$$

**Table 2.** Estimates of the true halo mass precisions, $\sigma_{\rm halo}$, (in dex) expected to result from future *JWST* clustering measurements at $z \sim 4$–10. Each row shows the predicted precision for a given galaxy sample selected by redshift and threshold rest-UV magnitude, $m_{\rm UV}^{\rm th}$. These precisions are defined to reflect true 68 per cent confidence intervals (see Section 2.4) and result from adopting the footprints and typical depths of the *JWST* Cycle 1 GTO program JADES. We do not consider spatial clustering measurements at $z \sim 10$ due to the low spectroscopic completeness expected in this regime (Section 2.2).

| | Predicted true halo mass precisions with JWST | | | |
|---|---|---|---|---|
| | Angular clustering | | Spatial clustering | |
| Redshift | $m_{\rm UV}^{\rm th}$ | $\sigma_{\rm halo}$ | $m_{\rm UV}^{\rm th}$ | $\sigma_{\rm halo}$ |
| $z \sim 4$ | 28.0 | 0.21 | 28.0 | 0.10 |
| | 27.0 | 0.21 | 27.0 | 0.15 |
| | 26.0 | 0.31 | 26.0 | 0.21 |
| $z \sim 5$ | 28.5 | 0.23 | 28.5 | 0.10 |
| | 27.5 | 0.15 | 27.5 | 0.13 |
| | 26.5 | 0.14 | 26.5 | 0.15 |
| $z \sim 6$ | 28.8 | 0.12 | 28.8 | 0.10 |
| | 27.8 | 0.17 | 27.8 | 0.17 |
| | 26.8 | 0.38 | 26.8 | 0.29 |
| $z \sim 7$ | 29.2 | 0.16 | 29.0 | 0.24 |
| | 28.2 | 0.21 | 28.0 | 0.30 |
| | 27.2 | 0.21 | 27.0 | 0.39 |
| $z \sim 8$ | 29.2 | 0.22 | 28.7 | 0.36 |
| | 28.2 | 0.21 | 27.7 | 0.38 |
| $z \sim 10$ | 29.5 | 0.22 | – | – |
| | 28.5 | 0.29 | – | – |

Here, $w_{i,\,\rm meas}$ is the simulated two-point correlation function measurement for a separation bin $i$, $w_{i,\,\rm true}$ is the 'true' two-point correlation function at that separation, and $\sigma_i$ is the simulated jackknife error for $w_{i,\,\rm meas}$. We use jackknife errors because they are commonly adopted in observational clustering studies (e.g. Harikane et al. 2016, 2018; Coil et al. 2017) and we find that they will reasonably approximate a Gaussian error distribution (Appendix G). These errors are calculated from 10 jackknife samples of roughly equal area split by constant right-ascension over the JADES survey footprint.

We also assess the significance of each simulated clustering measurement by comparing it to the null result ($w = 0$). When calculating accuracy and significance, we exclude the innermost distance bin (i.e. the intrahalo term) which is dominated by satellite clustering. Satellite clustering, particularly for faint samples at high redshifts, remains highly uncertain in dark matter simulations due to the artificial loss of low-mass satellites (e.g. van den Bosch et al. 2018; van den Bosch & Ogiya 2018). While the UM model compensates for these effects (see Appendix F), we choose to exclude the intrahalo term to reduce any potential bias introduced by this correction method.

We also report how the significances of our simulated clustering measurements change when instead utilizing a covariance matrix. As in Zhang et al. (2019b), we calculate covariance matrices by computing correlation function measurements across 100 randomly selected mock survey volumes. While the covariance matrix provides a better representation of the true significance, it will likely not be possible to quantify this with *JWST* observations as doing so requires observing a large number (>10) of independent fields. Here, we report both significances so that those obtained from a jackknife approach can be compared with those derived from a covariance matrix.

To avoid imposing artificial clustering signals, we ensure that the number densities of random samples, $n_R$, used for spatial clustering measurements follow the redshift and luminosity evolution of galaxy number densities and their selection completeness:

$$n_R(z) \propto \int_{-\infty}^{M_{\rm UV}^{\rm th}} \phi(z, M_{\rm UV})\, \mathscr{C}_S(z, M_{\rm UV})\, {\rm d}M_{\rm UV}. \qquad (5)$$

Here, $\phi(z, M_{\rm UV})$ is the luminosity function taken from the best-fitting redshift evolution reported by Finkelstein et al. (2015a), $\mathscr{C}_S(z, M_{\rm UV})$ is the spectroscopic completeness as a function of redshift and luminosity (see Section 2.2), and $M_{\rm UV}^{\rm th}$ is the corresponding value of $m_{\rm UV}^{\rm th}$ at a given redshift, z.

### 2.4 Inferring halo masses

We now simulate the process of inferring halo masses from our simulated galaxy clustering measurements. By comparing these inferred halo masses to the true values in the UM-VSMDPL mock catalogue, we are able to predict the precisions that will result from future *JWST* clustering measurements. Specifically, we infer halo masses by fitting our simulated galaxy two-point correlation function measurements to a grid of correlation functions expected from haloes at different masses. The inferred halo mass is then taken as the median value from the likelihood function, $\mathcal{L} \propto \exp(-\chi^2/2)$, where this process is performed for each mock galaxy sample and each survey realization separately.

We first generate a grid of autocorrelation functions for haloes of different masses. Here, we use halo mass grid points between $\log(M_{\rm halo}/{\rm M}_\odot) = 9.50$–12.75 with a spacing of 0.01 dex. To calculate correlation functions for each halo mass grid point, we autocorrelate samples of haloes drawn from the twenty mock survey volumes used to simulate galaxy clustering measurements at the redshift of interest (see Section 2.1). Haloes are selected such that the typical (i.e. median) halo mass is equal to the grid point value of interest. They are also selected such that the resulting mass distribution is lognormal with 0.25 dex scatter, as this well-approximates the true halo mass distribution shape of all the $z \sim 4$–10 mock galaxy samples considered in this work (see Fig. 2 for examples). To account for the partially random selection necessary to force this desired distribution, we perform the halo selection process ten times for each grid point and average the correlation functions from the ten selections. Haloes are only selected if they lie within the redshift interval of interest (see Table 1). For each survey realization, we infer halo masses for the faintest galaxy sample in each redshift interval by fitting its simulated autocorrelation function with these gridded halo autocorrelation functions.

To infer halo masses for the brighter galaxy samples, we generate a grid of halo cross-correlation functions to match the procedure for simulating correlation function measurements of these brighter samples (Section 2.3). In this case, the halo samples $D1$ and $D2$ are selected to have different median masses $M_1$ and $M_2$, with the same 0.25 dex scatter as for the halo autocorrelation functions. $M_2$ is fixed to the inferred halo mass of the faintest galaxy sample for the redshift interval and survey realization of interest. The $M_1$ values (representing the brighter sample) are binned between $M_2$ minus 0.25 dex and $\log(M_{\rm halo}/{\rm M}_\odot) = 12.75$, again with a spacing of 0.01 dex. As in the autocorrelation grid case, we average the correlation functions resulting from ten selections. We then fit for

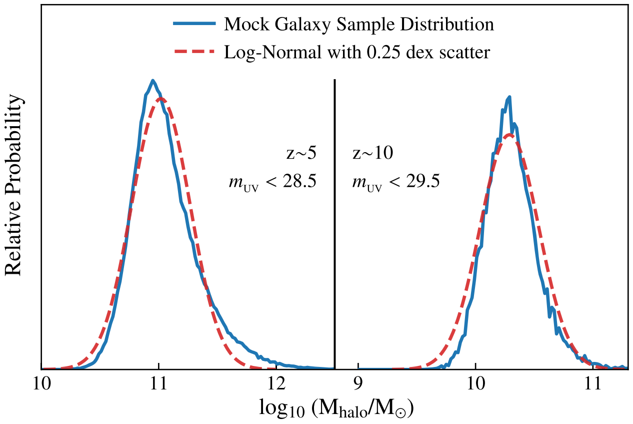

**Figure 2.** Comparison of the true halo mass distribution of mock galaxy samples used to simulate *JWST* clustering measurements (blue) and our adopted approximation to that distribution (red). As examples, we show this comparison for $z \sim 5$ and 10 galaxy samples with $m_{\rm UV}$ < 28.5 and 29.5, respectively.

the value of $M_1$, which is taken as the inferred halo mass of the brighter sample.

Finally, we determine the true 68 per cent confidence interval halo mass precisions for each galaxy sample resulting from this simulated procedure. Specifically, the true halo mass precision, $\sigma_{\rm halo}$, is defined such that the inferred halo mass is within $\sigma_{\rm halo}$ dex of the true typical halo mass for each mock galaxy sample (taken from the UM-VSMDPL catalogue) in 68 of the 100 mock survey realizations. These halo mass precisions are therefore not necessarily equal to the halo mass uncertainties that would be inferred directly from observations. We find that observationally inferred halo mass uncertainties (68 per cent confidence intervals from the $\chi^2$ distribution using jackknife errors) can underestimate the true precision by up to a factor of ∼2 in dex. This is largely due to the systematic uncertainty imposed by the requisite assumption of the true halo mass distribution for the observed galaxy sample (see Section 3.2) since we find that jackknife errors reasonably well capture the 68 per cent confidence interval from statistical fluctuations in the galaxy correlation function (Appendix G).

As motivated in Section 2.3, we use simulated jackknife errors when computing $\chi^2$ values because it will not be possible to derive accurate covariance matrices from observations alone. We do not calculate errors in our modelled halo mass-selected clustering strengths because the area from which these haloes are selected is $>10\times$ larger than the mock observed fields thereby substantially reducing Poisson error. Furthermore, we exclude the intrahalo term when calculating $\chi^2$ values due to the model dependence of satellite clustering strengths (see Section 2.3).

Finally, throughout this work, we use the peak halo mass (i.e. the maximum halo mass throughout that halo's lifetime prior to the redshift of interest) rather than the current halo mass.[8] We do so because peak halo masses better predict halo clustering strengths (e.g. Reddick et al. 2013; Guo et al. 2016). As such, when calculating projected spatial halo correlation functions, we use random samples ($R1$ and $R2$) with number densities that follow the redshift evolution of the $z = 4$–10 peak halo mass function from the UM-VSMDPL catalogue (Appendix H). The peak halo mass is on average <0.05 dex higher than the current halo mass for the $z \sim 4$–10 mock galaxy samples used in this work.

## 3 RESULTS

We now present $z \sim 4$–10 galaxy clustering measurements and halo mass precisions expected to result from future *JWST* observations. In Section 3.1, we discuss the quality of simulated $z \sim 4$–10 galaxy clustering measurements which adopt the footprints and typical depths of a planned Cycle 1 GTO program, JADES. Halo mass precisions resulting from these simulated clustering measurements are presented in Section 3.2, followed with a discussion of how *JWST* will improve constraints on the evolution of the stellar–halo mass relation at $z > 4$.

### 3.1 $z \sim$ 4–10 galaxy clustering measurements with JWST

We begin by investigating the quality of $z \sim 4$–10 galaxy clustering measurements expected from future *JWST* surveys. Fig. 3 shows simulated angular (left-hand panel) and projected spatial (right-hand panel) two-point correlation function measurements utilizing the footprint and depths of the planned Cycle 1 GTO program JADES (Williams et al. 2018). The median measurement significance from 100 survey realizations is shown in the lower left of each panel, one for each redshift bin. We find that Cycle 1 observations will enable $\gtrsim$5 sigma clustering measurements at $z \sim 4$–10.

Fig. 3 shows examples of typical measurements expected with the JADES program. Specifically, we plot measurements from realizations with significances within 0.5 of the median significance and $\chi^2$ values (relative to the true clustering strengths; see Section 2.3) within unity of the median $\chi^2$. The median $\chi^2$ lies between 2.4 and 3.9 for the angular measurements and 2.7 and 7.7 for the spatial measurements, resulting in median reduced $\chi^2$ values that are less than two in all cases.

The high quality of $z \sim 4$–10 angular clustering measurements expected from Cycle 1 observations are the result of NIRCam's ∼50 × improved sensitivity relative to WFC3/*HST*. Such capabilities will enable surveys that cover $\gtrsim$100 arcmin$^2$ regions with depths similar to *Hubble* ultradeep fields ($m \sim 29.5$; Bouwens et al. 2015). Access to galaxies ∼2 magnitudes further down the luminosity function will significantly reduce Poisson noise due to the steep faint-end slope (e.g. Bouwens et al. 2015; Finkelstein et al. 2015a), and weaken cosmic variance because fainter galaxies reside in lower mass haloes (e.g. Somerville et al. 2004; Trenti & Stiavelli 2008; Moster et al. 2011).

As shown in Fig. 3, we find that the precision of Cycle 1 angular clustering measurements will gradually evolve with redshift. Specifically, the typical measurement significance will increase from $\sim 7\sigma$ at $z \sim 4$ to $\sim 9\sigma$ at $z \sim 6$, then decrease at higher redshifts to $\sim 5\sigma$ at $z \sim 10$. This evolution is caused by two sources of noise that have opposite trends with redshift. In general, Poisson noise increases with redshift because galaxy number densities approximately halve per unit redshift, at least out to $z \sim 8$ (e.g. Bouwens et al. 2015; Finkelstein et al. 2015a). Conversely, the noise caused by chance projections decreases with redshift because galaxies are more spare and more strongly clustered at higher redshifts. Noise from chance projections dominates at $z \lesssim 6$ while Poisson noise begins to take over at higher redshifts.

*JWST*'s NIRSpec will greatly reduce the impact of chance projections by efficiently delivering precise redshifts out to $z \sim 9$ via the detection of strong rest-optical lines such as H α and [O III]

[8] The peak halo mass is nearly equal to the current halo mass for isolated haloes but may be significantly larger for satellites or close pairs where tidal stripping can lead to mass-loss (Reddick et al. 2013).

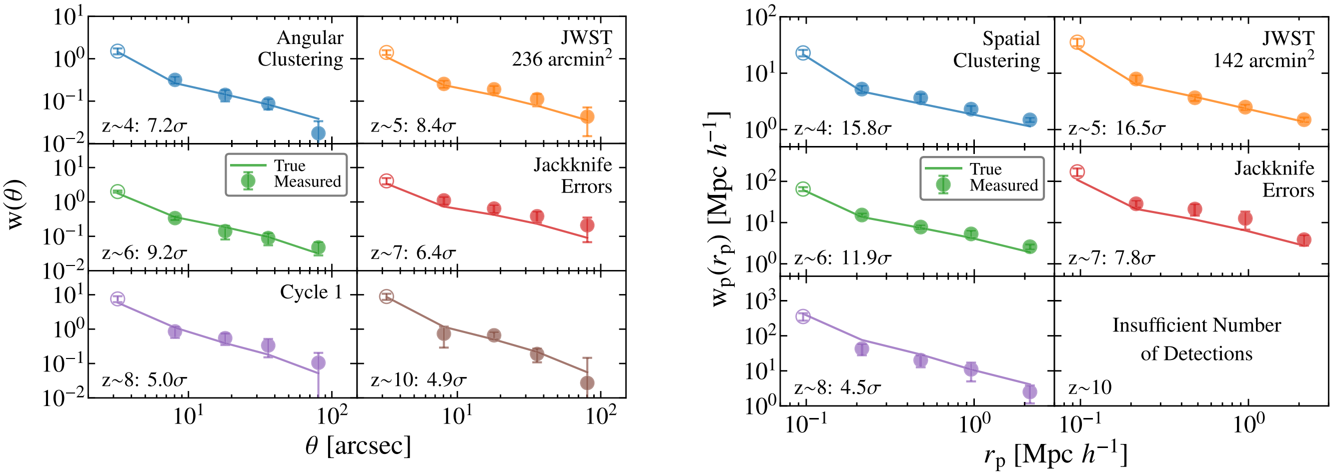

**Figure 3.** Typical simulated angular (left-hand panel) and projected spatial (right-hand panel) galaxy two-point correlation function measurements at $z \sim 4$–10 using the footprints and depths of the planned *JWST* Cycle 1 GTO program JADES. Measurements (markers) are the result of autocorrelating the entire mock galaxy sample selected to lie within the redshift interval of interest. The typical significance (using jackknife errors) from 100 simulated survey realizations is listed in the lower left of each panel (one per redshift bin). The true correlation functions (see Section 2.3) are shown with a solid line to illustrate accuracy. We do not consider spatial clustering measurements at $z \sim 10$ due to the low spectroscopic completeness expected in this regime (Section 2.2). The quality of spatial clustering measurements shown may require spectroscopic follow-up in addition to Cycle 1 observations due to possible target placement restrictions (see Appendix E).

(e.g. Chevallard et al. 2019; De Barros et al. 2019). This is of particular importance for the crowded $z \sim 4$–5 samples where chance projections over hundreds of Mpc dilute strong clustering signals on sub-Mpc scales. Fig. 3 illustrates that NIRSpec is capable of delivering ∼15$\sigma$ spatial clustering measurements at $z \sim 4$–5 over the 142 arcmin$^2$ JADES spectroscopic area, double the significance expected from angular measurements over the 236 arcmin$^2$ JADES NIRCam area. The spatial clustering measurements and significances shown in Fig. 3 assume spectroscopic follow-up of every $z \gtrsim 4$ candidate brighter than the limiting magnitudes listed in Table 1. As discussed in Appendix E, it is likely that only ∼50 per cent of such candidates can be targeted in Cycle 1 to avoid overlapping spectra. However, we choose to show results assuming 100 per cent follow-up completeness to illustrate results possible with dedicated spectroscopy in future cycles. These results are also possible with 50 per cent follow-up completeness given ≈4 × the coverage.[9]

We expect that angular clustering measurements will be more precise than spatial at $z \gtrsim 8$ for the following three reasons. First, chance projections occur less often at higher redshifts because number densities drop and clustering strengths increase. Secondly, angular clustering measurements will suffer less Poisson and cosmic variance because of wider coverage. Finally, spectroscopic completeness declines at $z \gtrsim 8$ as NIRSpec becomes less sensitive to strong [O III] emission. We do not consider spatial clustering measurements at $z \sim 10$ because NIRSpec's sensitivity to strong lines stops at $z \sim 9$, suggesting low spectroscopic completeness in this regime as discussed in Section 2.2. It is possible that dedicated deep spectroscopic follow-up of $z \sim 10$ photometric candidates would enable spatial clustering measurements at this redshift.

Notably, we find that angular clustering measurements will reach ∼5$\sigma$ significance at $z \sim 10$ within Cycle 1, rivalling current clustering measurements at $z \sim 7$ (Barone-Nugent et al. 2014). This conclusion persists even when adopting the most pessimistic $z \sim 10$ luminosity functions yet published so long as we take into account other planned Cycle 1 surveys. As detailed in Appendix B, adopting the ∼0.3 dex lower $z \sim 9$–10 luminosity functions reported by Bouwens et al. (2019) and Oesch et al. (2018) (relative to Bouwens et al. 2016a, used for the UM model) decreases the typical $z \sim 10$ angular clustering measurement significance from 4.9$\sigma$ to 3.0$\sigma$. However, this ignores the ∼0.8 mag increased depths within the 46 arcmin$^2$ deep JADES region (Williams et al. 2018). Cycle 1 observations will also include the 100 arcmin$^2$ Cosmic Evolution Early Release Science (CEERS; P.I. S. Finkelstein) program with NIRCam depths ∼0.4 mag shallower than those adopted in this work. Assuming $z \sim 10$ Schechter parameters of $M^*_{\rm UV} = -20.60$ and $\alpha = -2.3$ (Oesch et al. 2018), we estimate that combining all of these Cycle 1 observations will decrease Poisson noise by ∼60 per cent relative to the medium JADES program alone. Additional coverage provided by the CEERS program will also reduce cosmic variance. We therefore estimate that $z \sim 10$ angular galaxy clustering will be measured at ∼5$\sigma$ significance with Cycle 1 surveys, even assuming low galaxy number densities at this redshift.

When calculating the significances quoted above, we have omitted intrahalo clustering strength measurements due to model uncertainties in satellite clustering strengths (see Appendix F). Including the intrahalo term generally boosts the predicted significances by ∼2–3$\sigma$ (∼30–40 per cent). On the other hand, adopting a covariance matrix to remove correlations between different separation distances lowers our reported jackknife significances down to ∼6–7$\sigma$ at $z \sim 4$–6 and ∼3–5$\sigma$ at $z \sim 7$–10. This, however, does not impact our conclusions below on halo mass inferences because we adopted jackknife errors for this procedure.

## 3.2 Inferred halo mass precision with *JWST*

Here, we predict the halo mass precisions that will result from *JWST* galaxy clustering measurements at $z \sim 4$–10 using the procedure outlined in Section 2.4. To briefly review, we define halo mass

[9] Poisson noise is proportional to $1/\sqrt{D1D2}$ (Landy & Szalay 1993). Doubling galaxy number densities quadruples $D1D2$ and therefore reduces Poisson noise by a factor of 2. However, doubling the coverage with fixed number density only reduces Poisson noise by $1/\sqrt{2}$.

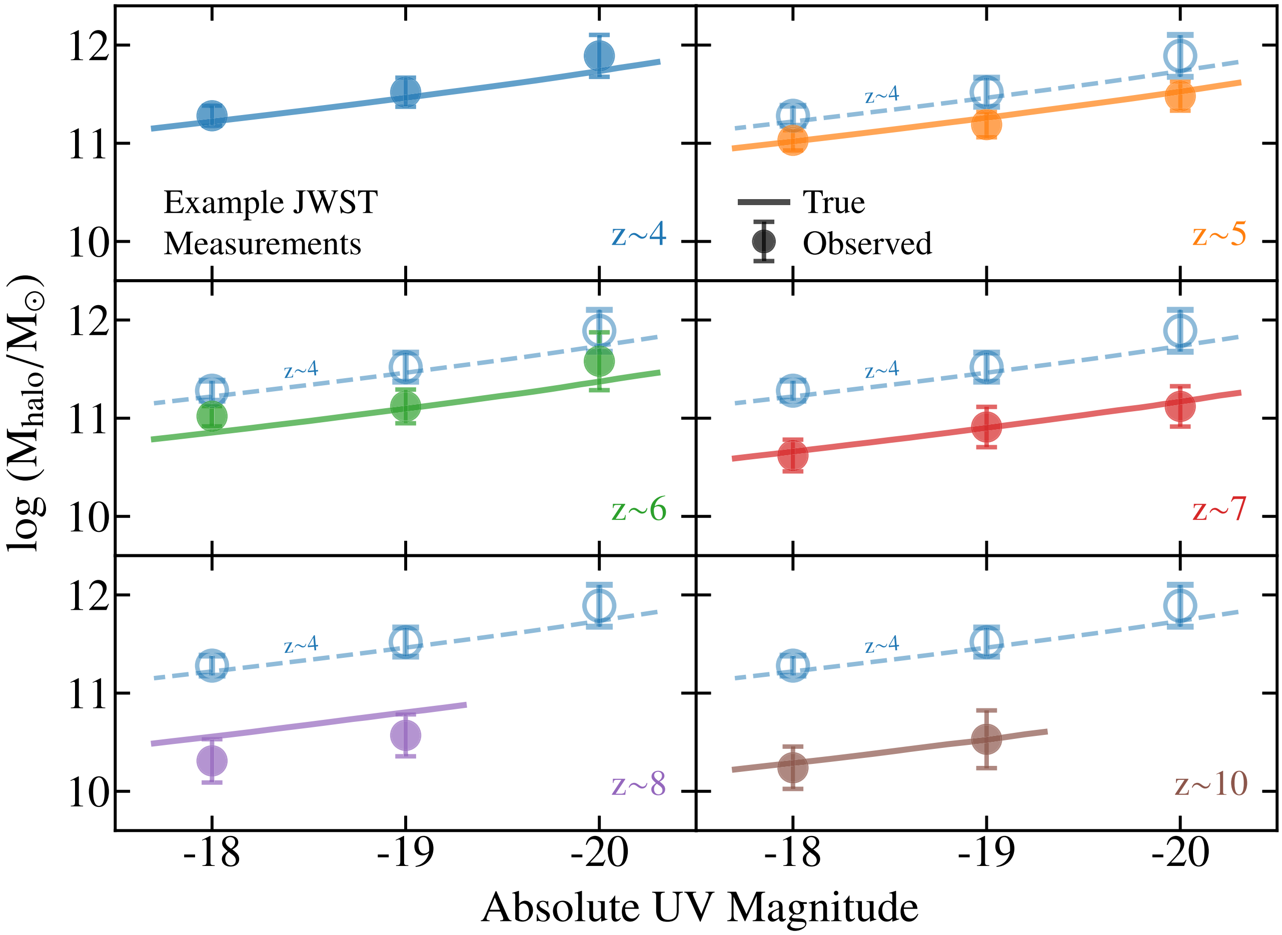

**Figure 4.** An example simulated measurement of the $M_{\rm UV}-{\rm M}_{\rm halo}$ relationship inferred from Cycle 1 *JWST* observations. Halo mass errorbars are taken from Table 2 using spatial clustering precisions at $z \sim 4$–6 and angular clustering precisions at $z \sim 7$–10. On the x-axis, we use the integer absolute UV magnitude most closely corresponding to the threshold apparent magnitude of each galaxy sample. The true $M_{\rm UV}-{\rm M}_{\rm halo}$ relation from the UM-VSMDPL mock catalogue is shown with a solid line at each redshift. The relation at $z \sim 4$ is shown in each panel to illustrate the evolution in the $M_{\rm UV}-{\rm M}_{\rm halo}$ relation seen from $z \sim 4$–10 in the simulated measurements.

precisions, $\sigma_{\rm halo}$, such that the inferred halo mass is within $\sigma_{\rm halo}$ dex of the true typical halo mass of the mock galaxy sample of interest in 68 of the 100 mock survey realizations. These values therefore represent the true errors in recovering halo masses, as opposed to the error estimates that would be obtained by marginalizing over observationally measured correlation function uncertainties obtained using, e.g. a jackknife approach.

In general, we find that $z \sim 4$–10 halo mass precisions will be $\lesssim$0.25 dex. The implications of this precision are illustrated in Fig. 4 where we compare the inferred $M_{\rm UV}-{\rm M}_{\rm halo}$ relations (markers) with the true relation (line) taken from the entire UM-VSMDPL mock catalogue.[10] Halo masses within the UM-VSMDPL catalogue decline by ~1 dex from $z \sim 4$ to $z \sim 10$ at fixed UV luminosity. If such a strong evolution does indeed exist in the real Universe, our results suggest that it will be easily possible to measure this evolution from future *JWST* observations.

As noted above, our reported precisions include both statistical and systematic components. The statistical component results from Poisson noise as well as field-to-field density fluctuations which alter the intrinsic clustering strengths at fixed halo mass. Angular clustering measurements also possess increased statistical uncertainties due to redshift contamination and chance projections. Combined, these statistical components contribute ~0.05–0.15 dex to the halo mass uncertainties. Systematic uncertainty arises from the required assumption of the true halo mass distribution of the observed galaxy sample. An incorrect assumption of the distribution may bias the observationally inferred halo masses. To estimate the resulting systematic halo mass uncertainties, we compare the true correlation functions (Section 2.3) of each mock galaxy sample with the correlation functions of haloes with a mass distribution following our adopted prescription (lognormal with 0.25 dex scatter) and median mass equal to that of the mock galaxy sample. These two correlation functions tend to differ by $\lesssim$0.1 dex for both our simulated angular and spatial clustering measurements, translating to ~0.05–0.15 dex systematic uncertainties in the inferred halo masses of our galaxy samples.

Table 2 lists our predicted true halo mass precisions for each galaxy sample at $z \sim 4$–10. In summary, we find that Cycle 1 angular clustering measurements will yield ~0.2 dex halo mass precisions for faint ($-18 \gtrsim M_{\rm UV} \gtrsim -19$) galaxies at $z \sim 4$–10 as well as ~0.3 dex precisions for bright ($M_{\rm UV} \sim -20$) galaxies at $z \sim 4$–7. Dedicated spectroscopic follow-up of $z \sim 4$–6 photometric

[10] The volume of the UM-VSMDPL mock catalogue is ~20 × that to be observed with the JADES program at each redshift.

candidates over the JADES/NIRSpec footprint would improve halo mass precisions by ∼0.1 dex at these redshifts. Specifically, spatial clustering measurements can yield high halo mass precision (∼0.10–0.15 dex) for faint galaxies as well as moderate precision (∼0.2 dex) for bright galaxies at $z \sim 4$–6.

Precisely inferring halo masses at high redshifts is of particular interest for settling the debate surrounding the evolution of the stellar–halo mass relation. This relation constrains the relative importance of various baryonic effects impacting galaxy evolution throughout cosmic time. Quantifying the evolution of this relation at high redshifts not only provides vital insight on the processes governing the early stages of galaxy evolution, but also the processes driving the ionizing output from galaxies during reionization. Many recent models have used observational data to infer the $z > 4$ evolution of the stellar-to-halo mass relation with differing conclusions. Significant (Behroozi, Wechsler & Conroy 2013c; Behroozi & Silk 2015; Finkelstein et al. 2015b; Harikane et al. 2016, 2018; Sun & Furlanetto 2016), moderate (Moster, Naab & White 2018), and little evidence of evolution (Mason et al. 2015; Rodríguez-Puebla et al. 2017; Stefanon et al. 2017; Tacchella et al. 2018) have all been claimed. This discrepancy is largely due to currently poor empirical constraints on both halo and stellar masses at high redshifts.

Current halo mass uncertainties from clustering measurements (and independent of galaxy number density measurements) are ≳0.5 dex at $z \sim 6$–7 (Barone-Nugent et al. 2014) while no halo mass constraints yet exist at $z \geq 8$. Our results therefore imply that future *JWST* observations will provide markedly improved halo mass constraints at $z \sim 6$–7 as well as the first constraints at $z \sim 8$–10. We also expect $z > 4$ stellar mass measurements to improve with *JWST*. Such measurements currently suffer from poor sensitivity ($m \sim 25$–26; Bouwens et al. 2015), significant confusion (e.g. González et al. 2011; Song et al. 2016; Stefanon et al. 2017), and systematic uncertainties due to nebular contamination (e.g. Stark et al. 2013; Song et al. 2016) in mid-infrared *Spitzer*/IRAC photometry. NIRCam's ∼100-fold increase in mid-infrared sensitivity and greatly improved point-source resolution (∼0.1 arcsec versus ∼2 arcsec) will eliminate the first two challenges while NIRSpec spectroscopy and NIRCam medium band photometry will mitigate the last. We thus conclude that future *JWST* observations will provide the first picture of the stellar–halo mass relation in the reionization era and substantially clarify how this relation evolves with redshift at $z > 4$.

## 4 ADDITIONAL APPLICATIONS WITH *JWST*

High-redshift clustering measurements will enable many other studies relevant to understanding early galaxy evolution. In Section 4.1, we discuss how well *JWST* will enable inferences on satellite fractions and the identification of satellites and satellite–host pairs at $z \gtrsim 4$. Such studies can be applied to inform models of sub-halo occupation and the environmental dependence on star formation in the early Universe. In Section 4.2, we simulate Cycle 1 measurements of $z \gtrsim 4$ galaxy pair fractions to assess how well *JWST* will provide insight on galaxy merger rates and the relative role of mergers to stellar mass build-up throughout cosmic history.

### 4.1 Constraining galaxy satellite fractions and quenching efficiencies at $z \gtrsim 4$

Small-scale (≲5 arcsec) galaxy clustering measurements can be used to infer galaxy satellite fractions. Inferences from current clustering measurements suggest that galaxy satellite fractions at $z \sim 4$–6 increase from ≲1 per cent for very bright ($M_{\rm UV} \lesssim -21$) galaxies to ∼5 per cent for bright ($M_{\rm UV} \sim -20$) galaxies (Ishikawa et al. 2017; Harikane et al. 2018; Hatfield et al. 2018). They also suggest that $z \sim 4$–6 galaxy satellite fractions decrease with redshift at fixed UV luminosity (e.g. Harikane et al. 2018). As illustrated in Fig. 3, our results suggest that *JWST* will deliver high-precision (∼5–10$\sigma$) small-scale galaxy clustering measurements for faint ($-18 \gtrsim M_{\rm UV} \gtrsim -19$) galaxies at $z \sim 4$–10. The satellite fractions of these samples are assumed to be 10–15 per cent from the UM model (including orphan satellites; Appendix F) with lower values at higher redshifts. We therefore predict that upcoming *JWST* surveys will soon make it possible to verify that satellite fractions decrease at earlier times and increase in fainter samples at $z \sim 4$–10.

In addition to quantifying the overall fraction of satellites, *JWST*'s spectroscopic capabilities will make it possible to determine, with high confidence, which $z \gtrsim 4$ galaxies are satellites and which are hosts. This capability is of particular interest for testing the influence of environment on star formation in the early Universe. Studies at $z \lesssim 2$ have found that satellites are systematically more quenched than field galaxies at fixed stellar mass where the quenching efficiency decreases with redshift (e.g. Kawinwanichakij et al. 2016). Empirically constraining satellite quenching efficiencies at $z \gtrsim 4$ would test model predictions of when environment begins to strongly influence galaxy evolution and how rapidly this dependence set in.

To assist future studies in these endeavours, we provide optimal parameters for selecting $z \gtrsim 4$ satellites and their hosts from future *JWST* surveys. We assume that satellites and centrals will be observationally selected using an isolation criteria commonly adopted in lower redshift studies (e.g. Tal et al. 2013; Kawinwanichakij et al. 2014). Specifically, we assume that a galaxy will be observationally identified as a central if it is the brightest galaxy (in the rest-UV) out to a specified maximum angular distance, $\theta_{\rm max}$, and inferred line-of-sight distance $\pi_{\rm max}$. We further assume that galaxies will be identified as satellites if they lie within $\theta_{\rm max}$ and $\pi_{\rm max}$ of an inferred central galaxy brighter than some limiting magnitude, $m_{\rm UV}^{\rm th,cen}$ where that central is inferred to be the satellite's host.

We simulate this selection process using mock galaxies from the UM-VSMDPL catalogue (Section 2.1) and compare the resulting satellite and host-galaxy populations to the true populations within that catalogue. Because the most significant impact of the host circumgalactic medium occurs within the virial radius (Zhang et al. 2019a), UM-VSMDPL haloes are defined as true satellites if they reside within the virial radius of a more massive halo which is defined as the true host. We account for spectroscopic completeness as described in Section 2.2 only considering mock galaxies brighter than the limiting spectroscopic magnitudes listed in Table 1. We fix $\pi_{\rm max}$ values to 10, 7, 5, 5, and 3 Mpc $h^{-1}$ at $z \sim 4$, 5, 6, 7, and 8, respectively (Section 2.3), corresponding to a redshift uncertainty of $\sigma_z \approx 0.02$. Because it will likely be difficult to observationally distinguish faint satellites from their hosts at separations of ≲2 × the host half-light radius, we conservatively ignore all potential satellites within 0.5 arcsec of a brighter neighbour. This threshold value was chosen assuming a typical half-light radii of 0.15 arcsec for very bright ($M_{\rm UV} = -21$) galaxies at $z \sim 4$ (Shibuya, Ouchi & Harikane 2015).

For each redshift interval, we test various maximum angular separations, $\theta_{\rm max}$, and central galaxy limiting magnitudes, $m_{\rm UV}^{\rm th,cen}$, seeking to optimize the number of correctly identified satellite and satellite–host pairs while keeping the fraction of contaminants to <20 per cent. We find that the parameter values listed in Table 3 are optimal. We choose a threshold 20 per cent contamination fraction

**Table 3.** Optimal parameters ($\theta_{\rm max}$ and $m_{\rm UV}^{\rm th,cen}$; defined in Section 4.1) to use when identifying satellites and satellite–host pairs (both the satellite and its specific host galaxy) at $z \sim$ 4–8 with >80 per cent confidence from spectroscopic *JWST* surveys. We also show the expected surface number of true identifiable satellites and satellite–host pairs using these parameters. Here, we have modelled spectroscopic completeness using the methods described in Section 2.2.

| | Satellites | | | Satellite–host pairs | | |
|---|---|---|---|---|---|---|
| Redshift | $\theta_{\rm max}$ (arcsec) | $m_{\rm UV}^{\rm th,cen}$ | # per arcmin$^2$ | $\theta_{\rm max}$ (arcsec) | $m_{\rm UV}^{\rm th,cen}$ | # per arcmin$^2$ |
| $z \sim 4$ | 4.25 | 27.5 | 0.65 | 3.50 | 26.0 | 0.30 |
| $z \sim 5$ | 3.75 | 28.0 | 1.05 | 3.25 | 27.0 | 0.75 |
| $z \sim 6$ | 3.25 | 28.3 | 0.45 | 3.00 | 27.3 | 0.35 |
| $z \sim 7$ | 2.75 | 28.5 | 0.15 | 2.25 | 28.0 | 0.15 |
| $z \sim 8$ | 2.50 | 28.2 | 0.05 | 2.50 | 27.7 | 0.05 |

because it is close to the lowest value that our mock procedure suggests will be possible with *JWST* (particularly for satellite–host identification) while still providing large samples of true satellites and satellite–host pairs. Statistical background subtraction procedures utilized in lower redshift studies (e.g. Tal, Wake & van Dokkum 2012; Kawinwanichakij et al. 2014) can be used to correct for contaminants.

Using the parameters listed in Table 3, we find that, on average, ≈0.65, 1.05, 0.45, 0.15, and 0.05 true satellites will be identifiable per arcmin$^2$ at $z \sim$ 4, 5, 6, 7, and 8, respectively. We also find that ≈0.30, 0.75, 0.35, 0.15, and 0.05 true satellite–host pairs will be identifiable per arcmin$^2$ at $z \sim$ 4, 5, 6, 7, and 8, respectively, again on average. The number of identifiable true satellite–host pairs is lower than the number of identifiable true satellites because a galaxy can be the satellite of another, more massive halo that may not necessarily be the brightest due to scatter in the observed $M_{\rm UV}$–M$_{\rm halo}$ relation.

These numbers suggest that Cycle 1 observations from the JADES and CEERS programs (∼200 arcmin$^2$ of photometric and spectroscopic coverage) will enable the identification of ∼200 satellites[11] at $z \sim$ 4–5. Studies of satellite quenching efficiencies at $z \sim 2$ have used ∼450 satellites (Kawinwanichakij et al. 2014, 2016), suggesting that Cycle 1 observations will begin pushing constraints on the environmental impact of star formation[12] to $z \sim 5$. Furthermore, given that current extragalactic *HST* legacy surveys cover ∼750 arcmin$^2$ (Grogin et al. 2011; Koekemoer et al. 2011), our results suggest that comprehensive follow-up of these regions with *JWST* would provide $z \sim$ 6–8 satellite quenching constraints comparable to what is currently available at $z \sim 2$. We conclude that *JWST* is capable of testing how satellite quenching efficiencies continue to evolve with both redshift and host star formation efficiency at $z \gtrsim 4$ and into the epoch of reionization.

It is worth discussing model dependencies of these conclusions. The UM model introduces orphan satellites (galaxies which no longer have an identifiable host halo in a dark matter simulation; Wang et al. 2006) to correct for the artificial disruption of low-mass satellites in simulations (Appendix F). Because orphan satellites constitute approximately 50 per cent of satellites at $z \sim$ 4–8 within the UM-VSMDPL catalogue, the accuracy of this satellite correction method impacts the numbers quoted above. In the worst-case scenario that no orphans should be introduced, we would be overestimating the number of identifiable $z \gtrsim 4$ satellites and satellite–host pairs by a factor of ≈2. Fortunately, independent empirical constraints of $z \sim$ 4–5 galaxy major merger pair fractions suggest that the high-redshift orphan populations introduced by the UM model are reasonable (Appendix F). It is also possible that $z \sim$ 4–8 galaxy number densities are underestimated in the UM-VSMDPL catalogue due to the empirical luminosity functions adopted by the UM model (Appendix B). If this is the case, we would expect the number of identifiable $z \gtrsim 4$ satellites and satellite–host pairs quoted above to rise.

### 4.2 Measuring major merger galaxy pair fractions at $z \gtrsim 4$

We now assess how well close galaxy pair fractions at $z \gtrsim 4$ can be measured from future *JWST* surveys. Pair fraction measurements can be applied to infer galaxy merger rates (e.g. Patton et al. 1997; Kartaltepe et al. 2007) and the relative contribution of mergers to the build-up of stellar mass throughout cosmic history (e.g. Mundy et al. 2017). Identifying close pairs also enables tests on the connection between mergers and AGN activity (e.g. Kocevski et al. 2012).

The Multi-Unit Spectroscopic Explorer (MUSE; Bacon et al. 2010) recently enabled the first spectroscopic galaxy pair fraction measurements at $z \sim$ 4–6 (Ventou et al. 2017). Because of *JWST*'s much wider field of view and ability to detect strong rest-optical lines to $z \simeq 9$, we expect that upcoming *JWST* surveys will deliver improved galaxy pair fractions measurements at $z \gtrsim 4$. We test this hypothesis directly by simulating $z \gtrsim 4$ galaxy pair fraction measurements using the same mock volumes and survey footprints described in Section 2.1 which reflect the planned Cycle 1 program, JADES (Williams et al. 2018). We also account for spectroscopic completeness using the methods outlined in Section 2.2.

We assume that galaxy pair fractions will be measured with *JWST* in a similar manner as described in Ventou et al. (2017), hereafter V17, which follows the methodology of de Ravel et al. (2009) and López-Sanjuan et al. (2013). Specifically, we identify a close pair as a system of two mock galaxies within a projected distance range of $r_{\rm p}^{\rm min} \leq r_{\rm p} \leq r_{\rm p}^{\rm max}$ and a rest-frame relative velocity of $\Delta{\rm v} \leq \Delta{\rm v}_{\rm max}$. As in V17, we adopt $r_{\rm p}^{\rm max} = 25\,{\rm h}^{-1}$ kpc. The $r_{\rm p}^{\rm min}$ value is set equal to the projected distance corresponding to 0.5 arcsec at the redshift of interest to account for the fact that it will be difficult to distinguish faint galaxies within ∼2 half-light radii of a very bright neighbour (cf. the 0.15 arcsec half-light radii of extremely luminous $z \sim 4$ galaxies; Shibuya et al. 2015). Rather than use the $\Delta{\rm v}_{\rm max} = 500\,{\rm km\,s}^{-1}$ value from V17, we adopt $\Delta{\rm v}_{\rm max} = 1000\,{\rm km\,s}^{-1}$ guided by the expected resolving power of JADES NIRSpec observations.[13] We do not expect that this increased choice of $\Delta{\rm v}_{\rm max}$ will significantly impact the fraction of identified close pairs that will eventually merge (see e.g. fig. 5 of Patton et al. 2000). For comparison, $\Delta{\rm v}_{\rm max} = 1000\,{\rm km\,s}^{-1}$ corresponds to $\pi_{\rm max} \approx 8$ and 6 Mpc $h^{-1}$ at $z = 4$ and 8, respectively.

We further consider only major mergers, which we define as galaxy pairs with a stellar mass ratio ≤1/4 to keep in convention with many other pair fraction studies (e.g. Man, Zirm & Toft 2016; Snyder et al. 2017; Mantha et al. 2018; Duncan et al. 2019).

[11] Here, we are assuming a NIRSpec MSA target placement efficiency of 50 per cent for Cycle 1 observations (Appendix E).

[12] We assume that it will be possible to separate star-forming and quenched galaxies using spectral energy distribution fits of *JWST* photometry and spectra to estimate rest-frame $U$–$V$ and $V$–$J$ colours as done in lower redshift studies (Kawinwanichakij et al. 2016).

[13] As noted in Section 2.2, we assume $R$∼100 at $z \sim$ 4–6 and $R$∼1000 at $z \sim$ 7–8. The MUSE resolving power is $R$∼3000 for Ly $\alpha$ at $z \sim$ 4–6.

Observed stellar masses are calculated by perturbing the true stellar mass of each mock galaxy (computed from the full star formation histories of the halo from the UM model) with a fixed scatter reflecting observational uncertainties. We assume that stellar mass uncertainties with *JWST* will be 0.2 dex at $z \sim 4$–8 though our conclusions are not significantly altered if we instead adopt an 0.3 dex uncertainty.

Similar to V17, we compute the major merger pair fraction, $f_{\rm MM}$, as

$$f_{\rm MM}(z) = \frac{N_{\rm p}^{\rm Corr}}{N_{\rm g}^{\rm Corr}} = \frac{\sum_{K=1}^{N_{\rm p}} C_1{}^{-1} C_2{}^{-1} \omega_K}{\sum_{i=1}^{N_{\rm g}} C_{\rm i}{}^{-1}}. \quad (6)$$

Here, $N_{\rm g}$ is the number of galaxies in the parent sample, $N_{\rm p}$ is the number of close pairs, $C_{\rm i}$ is the spectroscopic completeness of each mock galaxy as a function of magnitude and redshift where $i = 1,2$ corresponds to each galaxy in an observed pair $K$, and $\omega_{\rm K}$ is the area correction factor accounting for both survey boundaries and the minimum projected distance. That is

$$\omega_K = \frac{\left(r_{\rm p}^{\rm max}\right)^2}{\left(r_{\rm p}^{\rm max}\right)^2 - \left(r_{\rm p}^{\rm min}\right)^2} \times \frac{\pi \left(r_{\rm p}^{\rm max}\right)^2}{A_{\rm JWST}}, \quad (7)$$

where $A_{\rm JWST}$ is the survey area enclosing the circle of radius $r_{\rm p}^{\rm max}$ centred on the more massive source in pair $K$. $N_{\rm g}^{\rm Corr}$ and $N_{\rm p}^{\rm Corr}$ are then the completeness corrected number of parent galaxies and major merger galaxy pairs, respectively. We compute pair fraction uncertainties via a jackknife approach to account for sample variance. As with the simulated clustering measurements, ten jackknife samples are obtained by splitting the footprints of each JADES fields into five roughly equal-area regions split by constant right ascension.

We show the simulated JADES/NIRSpec major merger pair fraction measurements[14] at $z \sim 4$–8 in Fig. 5. We also plot for comparison the MUSE measurements from V17 at $z \sim 0$–6. We find that future *JWST* surveys will substantially improve current galaxy pair fraction measurements at $z \sim 4$–6 and establish the first measurements at $z \sim 7$–8. Specifically, we find that errors on measured major merger pair fractions will be $\lesssim$0.1 dex from $z \sim 4$–6 and $\sim$0.1–0.2 dex from $z \sim 7$–8, made possible by the much wider coverage enabled by *JWST*/NIRSpec ($>$100 arcmin$^2$) relative to MUSE ($\sim$10 arcmin$^2$). Given that we have not considered contributions from the JADES deep field nor pair fraction measurements made via photometry alone (cf. Duncan et al. 2019), it is plausible that $z \sim 10$ measurements will be possible within the first few cycles. We conclude that *JWST* will clarify the importance of major mergers at $z \gtrsim 4$ and into the epoch of reionization.

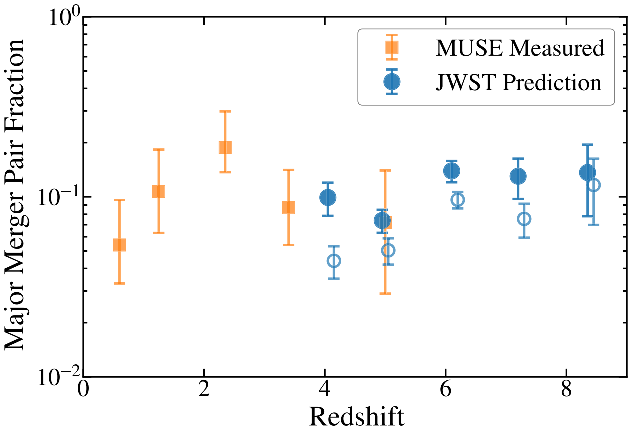

**Figure 5.** Simulated *JWST* measurements of the major merger galaxy pair fractions at $z \sim 4$–8 (blue) using the footprints and typical depths of the planned Cycle 1 GTO program, JADES. We also plot for comparison the MUSE measurements from Ventou et al. (2017) from $z \sim 0$–6 (orange). We find that future *JWST* surveys will substantially improve current galaxy pair fraction measurements at $z \sim 4$–6 and establish the first measurements at $z \sim 7$–8. We also show simulated pair fraction measurements ignoring all orphan satellites (Appendix F) with empty markers. We find that including orphans leads to better agreement with current observations when we match our simulated fields to existing surveys.

## 5 CONCLUSIONS

We have simulated *JWST* galaxy clustering measurements at $z \sim 4$–10 by adopting footprints and typical depths of the planned Cycle 1 GTO program, JADES (Williams et al. 2018), and utilizing an empirical model, the UNIVERSEMACHINE (Behroozi et al. 2019), to assign galaxy properties to haloes from a dark matter simulation. We have also assessed the ability of future *JWST* surveys to quantify galaxy satellite fractions, identify individual satellites, and measure galaxy major merger pair fractions at $z \gtrsim 4$. Conclusions from this study include:

(i) Planned Cycle 1 *JWST* surveys will measure galaxy angular clustering with $\gtrsim 5\sigma$ significance at $z \sim 4$–10. Dedicated spectroscopic follow-up over $\sim$150 arcmin$^2$ will enable $\sim$10–15$\sigma$ spatial clustering measurements at $z \sim 4$–6 and $\sim 8\sigma$ measurements at $z \sim 7$.

(ii) Halo mass uncertainties resulting from Cycle 1 angular clustering measurements will be $\sim$0.2 dex for faint ($-18 \gtrsim M_{\rm UV} \gtrsim -20$) galaxies at $z \sim 4$–10 as well as $\sim$0.3 dex for bright ($M_{\rm UV} \sim -20$) galaxies at $z \sim 4$–7. Dedicated spectroscopic follow-up over $\sim$150 arcmin$^2$ would yield $\sim$0.10–0.15 dex halo mass uncertainties for faint galaxies as well as $\sim$0.2 dex uncertainties for bright galaxies at $z \sim 4$–6. Future *JWST* observations will therefore provide the first constraints on the stellar–halo mass relation in the epoch of reionization and substantially clarify how this relation evolves with redshift at $z > 4$.

(iii) Cycle 1 observations will allow precise inferences on galaxy satellite fractions at $z \sim 4$–10 by enabling high-precision ($\sim$5–10$\sigma$) small-scale galaxy clustering measurements at these redshifts. *JWST* observations will therefore soon test sub-halo occupation models in the early Universe.

(iv) It will be possible to identify $\sim$200 individual satellites at $z \sim 4$–5 from Cycle 1 surveys. Furthermore, comprehensive follow-up of *HST* legacy surveys with NIRSpec would enable the identification of $\sim$500 satellites at $z \sim 6$–8. *JWST* will therefore be able to test the environmental dependence of star formation in the early Universe.

(v) Future *JWST* surveys can substantially improve current galaxy major merger pair fraction measurements at $z \sim 4$–6 and establish the first measurements at $z \gtrsim 7$. Specifically, we find that dedicated NIRSpec follow-up over $\sim$150 arcmin$^2$ would yield

[14] We note that the decrease in our simulated major merger pair fractions at $z \sim 5$ is due to sample variance. For these simulated measurements, we used the mock survey volumes best matching the two GOODS fields to be observed by JADES. Adopting other reasonably well matching mock survey volumes tends to yield slightly higher ($\approx$0.10–0.11) major merger pair fractions at $z \sim 5$, more consistent with a smooth evolution from $z \sim 4$–6.

$\lesssim$0.1 dex errors at $z \sim 4$–6 and $\sim$0.1–0.2 dex at $z \sim 7$–8. Such measurements can be used to quantify galaxy major merger rates and determine the relative role of mergers to the build-up of stellar mass into the epoch of reionization.

## ACKNOWLEDGEMENTS

The authors thank the referee for their thorough reading and very helpful, constructive comments. We also thank Gurtina Besla and Rachel Somerville for stimulating discussions. RE, DPS, and MR acknowledge funding from *JWST*/NIRCam contract to the University of Arizona, NAS5-02015. CCW acknowledges support from the National Science Foundation Astronomy and Astrophysics Fellowship grant AST-1701546. BER acknowledges a Maureen and John Hendricks Visiting Professorship at the Institute for Advanced Study, NASA contract NNG16PJ25C, and NSF award 1828315. GY acknowledges partial financial support under research grants AYA2015-63819-P (MINECO/FEDER) and PGC2018-094975-C21 (MICINN/FEDER).

The authors wish to thank V. Springel for allowing us to use the L-GADGET2 code to run the different Multidark simulation boxes, including the VSMDPL used in this work. The VSMDPL simulation has been performed at LRZ Munich within the project pr87yi. The CosmoSim data base (www.cosmosim.org) provides access to the simulation and the Rockstar data are a service by the Leibniz Institute for Astrophysics Potsdam (AIP).

This research made use of ASTROPY, a community-developed core PYTHON package for Astronomy (Astropy Collaboration 2013; Price-Whelan et al. 2018); MATPLOTLIB (Hunter 2007); NUMPY (Van Der Walt, Colbert & Varoquaux 2011); and SCIPY (Jones et al. 2001).

## REFERENCES

Allen M., Behroozi P., Ma C.-P., 2019, MNRAS, 488, 4916
Astropy Collaboration et al., 2013, A&A, 558, A33
Atek H. et al., 2015, ApJ, 814, 69
Bacon R. et al., 2010, in McLean I. S., Ramsay S. K., Takami H., eds Proc. SPIE Conf. Ser. Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III. SPIE, Bellingham, p. 773508
Barone-Nugent R. L. et al., 2014, ApJ, 793, 17
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 134
Behroozi P. S., Silk J., 2015, ApJ, 799, 32
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, ApJ, 762, 109
Behroozi P. S., Wechsler R. H., Wu H.-Y., Busha M. T., Klypin A. A., Primack J. R., 2013b, ApJ, 763, 18
Behroozi P. S., Wechsler R. H., Conroy C., 2013c, ApJ, 770, 57
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bhowmick A. K., Di Matteo T., Feng Y., Lanusse F., 2018, MNRAS, 474, 5393
Bouwens R. J. et al., 2014, ApJ, 793, 115
Bouwens R. J. et al., 2015, ApJ, 803, 34
Bouwens R. J. et al., 2016a, ApJ, 830, 67
Bouwens R. J. et al., 2016b, ApJ, 833, 72
Bouwens R. J., Stefanon M., Oesch P. A., Illingworth G. D., Nanayakkara T., Roberts-Borsani G., Labbé I., Smit R., 2019, ApJ, 880, 25
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Chabrier G., 2003, PASP, 115, 763
Chevallard J. et al., 2019, MNRAS, 483, 2621
Coil A. L., Mendez A. J., Eisenstein D. J., Moustakas J., 2017, ApJ, 838, 87
Conroy C., Gunn J. E., 2010, ApJ, 712, 833
Conroy C., Gunn J. E., White M., 2009, ApJ, 699, 486
Croton D. J., Gao L., White S. D. M., 2007, MNRAS, 374, 1303
De Barros S., Oesch P. A., Labbé I., Stefanon M., González V., Smit R., Bouwens R. J., Illingworth G. D., 2019, MNRAS, 489, 2355
de Ravel L. et al., 2009, A&A, 498, 379
Duncan K. et al., 2019, ApJ, 876, 110
Finkelstein S. L. et al., 2015a, ApJ, 810, 71
Finkelstein S. L. et al., 2015b, ApJ, 814, 95
Gao L., Springel V., White S. D. M., 2005, MNRAS, 363, L66
Giavalisco M. et al., 2004, ApJ, 600, L93
González V., Labbé I., Bouwens R. J., Illingworth G., Franx M., Kriek M., 2011, ApJ, 735, L34
González V., Bouwens R. J., Labbé I., Illingworth G., Oesch P., Franx M., Magee D., 2012, ApJ, 755, 148
Grazian A. et al., 2015, A&A, 575, A96
Grogin N. A. et al., 2011, ApJS, 197, 35
Guo H. et al., 2016, MNRAS, 459, 3040
Harikane Y. et al., 2016, ApJ, 821, 123
Harikane Y. et al., 2018, PASJ, 70, S11
Hatfield P. W., Bowler R. A. A., Jarvis M. J., Hale C. L., 2018, MNRAS, 477, 3760
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Illingworth G. D. et al., 2013, ApJS, 209, 6
Ishikawa S., Kashikawa N., Toshikawa J., Tanaka M., Hamana T., Niino Y., Ichikawa K., Uchiyama H., 2017, ApJ, 841, 8
Jiang F., van den Bosch F. C., 2016, MNRAS, 458, 2848
Jones E. et al., 2001, SciPy: Open source scientific tools for Python. Available at http://www.scipy.org/
Kartaltepe J. S. et al., 2007, ApJS, 172, 320
Kawinwanichakij L. et al., 2014, ApJ, 792, 103
Kawinwanichakij L. et al., 2016, ApJ, 817, 9
Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, MNRAS, 457, 4340
Kocevski D. D. et al., 2012, ApJ, 744, 148
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Labbé I. et al., 2013, ApJ, 777, L19
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Li C., Jing Y. P., Kauffmann G., Börner G., Kang X., Wang L., 2007, MNRAS, 376, 984
Livermore R. C., Finkelstein S. L., Lotz J. M., 2017, ApJ, 835, 113
López-Sanjuan C. et al., 2013, A&A, 553, A78
Lotz J. M., Jonsson P., Cox T. J., Croton D., Primack J. R., Somerville R. S., Stewart K., 2011, ApJ, 742, 103
Man A. W. S., Zirm A. W., Toft S., 2016, ApJ, 830, 89
Mantha K. B. et al., 2018, MNRAS, 475, 1549
Mason C. A., Trenti M., Treu T., 2015, ApJ, 813, 21
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Moster B. P., Somerville R. S., Newman J. A., Rix H.-W., 2011, ApJ, 731, 113
Moster B. P., Naab T., White S. D. M., 2018, MNRAS, 477, 1822
Mundy C. J., Conselice C. J., Duncan K. J., Almaini O., Häußler B., Hartley W. G., 2017, MNRAS, 470, 3507
Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009, MNRAS, 396, 19
Oesch P. A. et al., 2014, ApJ, 786, 108
Oesch P. A., Bouwens R. J., Illingworth G. D., Labbé I., Stefanon M., 2018, ApJ, 855, 105
Oke J. B., Gunn J. E., 1983, ApJ, 266, 713
Ono Y. et al., 2018, PASJ, 70, S10
Patton D. R., Pritchet C. J., Yee H. K. C., Ellingson E., Carlberg R. G., 1997, ApJ, 475, 29
Patton D. R., Carlberg R. G., Marzke R. O., Pritchet C. J., da Costa L. N., Pellegrini P. S., 2000, ApJ, 536, 153
Patton D. R., Torrey P., Ellison S. L., Mendel J. T., Scudder J. M., 2013, MNRAS, 433, L59
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton
Planck Collaboration et al., 2014, A&A, 571, A16
Planck Collaboration et al., 2016, A&A, 594, A13

Price-Whelan A. M. et al., 2018, AJ, 156, 123
Pujol A., Gaztañaga E., 2014, MNRAS, 442, 1930
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Ren K., Trenti M., Mason C. A., 2019, ApJ, 878, 114
Rodríguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016, MNRAS, 462, 893
Rodríguez-Puebla A., Primack J. R., Avila-Reese V., Faber S. M., 2017, MNRAS, 470, 651
Sheth R. K., Tormen G., 2004, MNRAS, 350, 1385
Shibuya T., Ouchi M., Harikane Y., 2015, ApJS, 219, 15
Snyder G. F., Lotz J. M., Rodriguez-Gomez V., Guimarães R. da S., Torrey P., Hernquist L., 2017, MNRAS, 468, 207
Somerville R. S., Lee K., Ferguson H. C., Gardner J. P., Moustakas L. A., Giavalisco M., 2004, ApJ, 600, L171
Song M. et al., 2016, ApJ, 825, 5
Springel V., 2005, MNRAS, 364, 1105
Stark D. P., 2016, ARA&A, 54, 761
Stark D. P., Ellis R. S., Bunker A., Bundy K., Targett T., Benson A., Lacy M., 2009, ApJ, 697, 1493
Stark D. P., Ellis R. S., Ouchi M., 2011, ApJ, 728, L2
Stark D. P., Schenker M. A., Ellis R., Robertson B., McLure R., Dunlop J., 2013, ApJ, 763, 129
Stefanon M., Bouwens R. J., Labbé I., Muzzin A., Marchesini D., Oesch P., Gonzalez V., 2017, ApJ, 843, 36
Sun G., Furlanetto S. R., 2016, MNRAS, 460, 417
Tacchella S., Bose S., Conroy C., Eisenstein D. J., Johnson B. D., 2018, ApJ, 868, 92
Tal T., Wake D. A., van Dokkum P. G., 2012, ApJ, 751, L5
Tal T., van Dokkum P. G., Franx M., Leja J., Wake D. A., Whitaker K. E., 2013, ApJ, 769, 31
Tang M., Stark D., Chevallard J., Charlot S., 2019, MNRAS, 489, 2572
Tinker J. L., Robertson B. E., Kravtsov A. V., Klypin A., Warren M. S., Yepes G., Gottlöber S., 2010, ApJ, 724, 878
Trenti M., Stiavelli M., 2008, ApJ, 676, 767
van den Bosch F. C., Ogiya G., 2018, MNRAS, 475, 4066
van den Bosch F. C., Aquino D., Yang X., Mo H. J., Pasquali A., McIntosh D. H., Weinmann S. M., Kang X., 2008, MNRAS, 387, 79
van den Bosch F. C., Ogiya G., Hahn O., Burkert A., 2018, MNRAS, 474, 3043
Van Der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Ventou E. et al., 2017, A&A, 608, A9
Wang L., Li C., Kauffmann G., De Lucia G., 2006, MNRAS, 371, 537
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
Williams C. C. et al., 2011, ApJ, 733, 92
Williams C. C. et al., 2018, ApJS, 236, 33
Zehavi I. et al., 2011, ApJ, 736, 59
Zehavi I., Contreras S., Padilla N., Smith N. J., Baugh C. M., Norberg P., 2018, ApJ, 853, 84
Zhang H., Zaritsky D., Behroozi P., Werk J., 2019a, ApJ, 880, 1
Zhang H., Eisenstein D. J., Garrison L. H., Ferrer D. W., 2019b, ApJ, 875, 132

## APPENDIX A: CLUSTERING RESULTS WHEN USING ABUNDANCE MATCHING TO ASSIGN UV LUMINOSITIES TO HALOES

Here, we investigate how the clustering results shown in Fig. 3 would change if we instead used abundance matching to assign UV luminosities to haloes within the UM-VSMDPL mock catalogue. For the sake of consistency, we abundance match to the same empirical high-redshift UV luminosity functions as those adopted by the UM model, namely the Finkelstein et al. (2015a) luminosity functions at $z \sim 4$–8 and the Bouwens et al. (2016a) luminosity functions at $z \sim 9$–10. We use the best-fitting analytic Schechter form at each redshift to generate the galaxy counts as a function of $M_{\rm UV}$ and follow the abundance matching with scatter algorithm described in Allen, Behroozi & Ma (2019). Here, we order haloes in the UM-VSMDPL mock catalogue by their peak mass and adopt a fixed scatter in $M_{\rm UV}$ of $\sigma = 0.5$ mag for simplicity. This value was chosen to be consistent with the scatter inferred by the UM model at high redshifts. We discuss how our results change if we instead adopt larger or smaller scatter values at the end of this section. All other methods for simulating clustering measurements remain as described in Section 2 with the exception that the best-fitting mock survey volumes (see Section 2.1) are re-chosen using the UV luminosities assigned via abundance matching. While these best-fitting volumes were originally selected using the same luminosity functions, the UV magnitudes assigned to each mock galaxy are different with abundance matching so we re-select the volumes for consistency.

As shown in Fig. A1, the typical clustering measurements and significances (again determined using simulated jackknife errors and ignoring the 1-halo term) obtained using abundance matching are consistent with those obtained using the UM model. The only particularly notable changes are at $z \sim 8$ and $z \sim 10$ where the measurement significances are $\sim 5\sigma$. At $z \sim 8$, the abundance matching approach yields slightly more optimistic measurement significances, moving from $5.0\sigma$ to $5.6\sigma$ for angular clustering and from $4.5\sigma$ to $4.9\sigma$ for spatial clustering. At $z \sim 10$, the UM model approach yields a slightly higher typical measurement significance ($4.9\sigma$) than abundance matching ($4.2\sigma$). This is because the $z \sim 10$ luminosity function inferred by the UM model is slightly higher than the best-fitting reported by Bouwens et al. (2016a), thereby lowering Poisson noise in pair counts.

We choose to emphasize results using the UM model for two reasons. First, our abundance matching approach does not account for possible evolution in the scatter of the $M_{\rm UV}$–$M_{\rm halo}$ relation with redshift or halo mass. Secondly, abundance matching does not account for the expectation that satellites have systematically lower star formation rates than centrals at fixed halo mass due to environmental quenching (e.g. van den Bosch et al. 2008; Kawinwanichakij et al. 2016). This means that satellites are less likely to be detected when using the UM model and explains why clustering strengths obtained via the UM model tend to be lower. The exception to this trend is at $z \sim 5$. At this redshift, the UV luminosity function inferred by the UM model is systematically lower than that reported by Finkelstein et al. (2015a) (see Fig. 1) meaning that the UM model is assigning fixed $M_{\rm UV}$ values to higher mass haloes which are more strongly clustered.

As noted in Ren, Trenti & Mason (2019), the clustering strength of a UV-selected galaxy population depends on the scatter of the $M_{\rm UV}$–$M_{\rm halo}$ relation. This is because bright galaxies are allowed to be hosted by lower mass haloes with larger scatter. To quantify the impact of the choice of scatter, we also perform this abundance matching procedure with $\sigma = 0.3$ and 0.7 mag. The typical angular clustering measurement significance changes negligibly at $z \sim 4$–7 for any $\sigma$ between 0.3 and 0.7 mag scatter, largely because the decrease in clustering strength (with larger scatter) is balanced out by the decrease in cosmic variance. However, at $z \sim 8$ and 10, the typical measurement significance is larger for smaller values of $\sigma$. This is because, at these redshifts, the samples are relatively small (given current observed luminosity functions) and therefore Poisson variance dominates the overall uncertainty in the simulated clustering measurements. Hence, the clustering

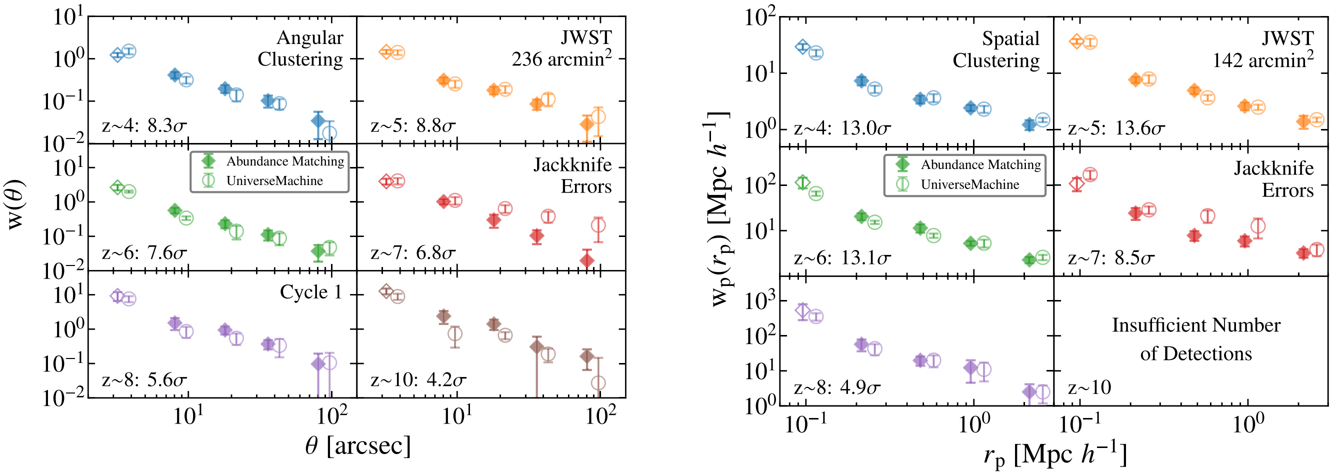

**Figure A1.** Same as Fig. 3 except that we show typical simulated *JWST* clustering measurements using $M_{\rm UV}$ values assigned via abundance matching with scatter with diamond markers. Original data points using $M_{\rm UV}$ values from the UNIVERSEMACHINE model are shown with empty markers and are slightly offset to the right for clarity. In the lower left of each panel, we display typical measurement significances using $M_{\rm UV}$ values from abundance matching where it can be seen (by comparing to Fig. 3) that these significances remain approximately the same between the two methods. Simulated clustering measurements are often lower when adopting the UNIVERSEMACHINE model as satellites are assigned systematically fainter $M_{\rm UV}$ values relative to abundance matching.

strength increases with smaller $\sigma$ while the uncertainties remain at similar values. Specifically, we find typical simulated measurement significances of 7.1, 5.6, and 5.2$\sigma$ for 0.3, 0.5, and 0.7 mag scatter, respectively, at $z \sim 8$. At $z \sim 10$, the typical simulated measurement significance is 4.9, 4.2, and 4.0$\sigma$ for 0.3, 0.5, and 0.7 mag scatter, respectively.

## APPENDIX B: THE IMPACT OF $Z \geq 4$ UV LUMINOSITY FUNCTION UNCERTAINTIES

We also test how our inferred clustering measurement significances depend on the $z \geq 4$ luminosity functions adopted by the UM model. Rather than re-run the UM model with different input luminosity functions, we use the abundance matching with scatter approach described in Appendix A. This approach is justified by the similar results obtained between the two methods (Appendix A).

At $z \sim$ 4–8, the Bouwens et al. (2015) luminosity functions are also commonly adopted in the literature. Galaxy number densities implied by this work are typically higher than those reported by Finkelstein et al. (2015a) which was adopted by the UM model. This suggests that using the Bouwens et al. (2015) luminosity functions will result in lower clustering strengths because galaxies at fixed UV magnitude are being assigned to lower mass haloes. However, we also expect cosmic variance to decrease for the same reason. Finally, Poisson noise should decrease due to higher overall pair counts. In the end, we find that $z \sim$ 4–8 clustering measurement significances either remain at very similar values or are boosted by $\sim 1\sigma$.

We re-simulate the $z \sim 10$ clustering strength measurements using the Bouwens et al. (2019) and Oesch et al. (2018) luminosity functions at $z \sim 9$ and $z \sim 10$, respectively. These luminosity functions are ~0.3 dex lower than those reported by Bouwens et al. (2016a) so we expect higher clustering strength measurements with larger cosmic and Poisson variance. We find that the typical $z \sim 10$ clustering measurement significance drops to 3.0$\sigma$ as a result. While these luminosity functions give more pessimistic results, we still find that Cycle 1 surveys are likely to enable ~5$\sigma$ clustering measurements at $z \sim 10$ (Section 3.1).

## APPENDIX C: THE IMPACT OF DARK MATTER SIMULATION SIZE

The VSMDPL dark matter simulation (Section 2.1) used here was chosen to due to its high mass resolution ($6.2 \times 10^6$ M$_\odot$ $h^{-1}$) capable of delivering realistic populations of low-mass haloes hosting faint $z \sim$ 4–10 galaxies detectable with Cycle 1 *JWST* surveys. However, we acknowledge that the size of the VSMDPL simulation (160 Mpc $h^{-1}$ on a side) limits the independence of the 500 mock survey volumes used to generate realizations of the GOODS fields. For example, the VSDMPL box is ~40 × the volume of the JADES/NIRCam GOODS-S field (146 arcmin$^2$) at $z \sim 7$ ($z$ = 6.7–7.7). While we only use the 10 best-fitting mock survey volumes for each GOODS field at each redshift (Section 2.1), it is possible that the UM-VSMDPL mock catalogue does not adequately capture the extent of cosmic variance possible with *JWST* surveys at $z \sim$ 4–10, even once constraining to the observed luminosity function in each surveyed field.

We test for strong effects due to dark matter simulation size by re-performing our clustering measurement simulations using the *Bolshoi–Planck* dark matter simulation (Klypin et al. 2016; Rodríguez-Puebla et al. 2016). *Bolshoi–Planck* adopts the same cosmology as VSMDPL, yet has a box length of 250 Mpc $h^{-1}$ yielding a 3.8 × larger volume. Because of the larger box size, *Bolshoi–Planck* has a poorer mass resolution ($1.6 \times 10^8$ M$_\odot$ $h^{-1}$) and becomes mass incomplete[15] at $\log(M_{\rm halo}/{\rm M}_\odot) \sim 9.75$ at $z \sim$ 4–10. Because this mass is higher than the lowest[16] median halo mass ($\log(M_{\rm halo}/{\rm M}_\odot) = 9.5$) used to generate grids of halo correlation functions (Section 2.4), it is likely that we cannot reliably model observational *JWST* halo mass uncertainties with *Bolshoi–Planck*. However, given that the typical halo mass of $M_{\rm UV} = -18$ $z \sim 10$

[15] The threshold of incompleteness was determined as the mass where the peak halo mass function from the *Bolshoi–Planck* simulation is 50 per cent of that from the VSMDPL simulation (Section H).

[16] This lowest grid point value was chosen to be 3$\sigma$ away from the typical halo mass of $M_{\rm UV} = -18$ galaxies at $z \sim 10$ in the UM-VSMDPL mock catalogue where $\sigma$ = 0.25 dex (Section 2.4).

galaxies from the UM-VSMDPL mock catalogue is $\sim$0.5 dex larger than the *Bolshoi–Planck* mass incompleteness threshold, it is likely that *Bolshoi–Planck* can adequately model the clustering of haloes hosting $M_{\rm UV} \lesssim -18$ galaxies at $z \sim 4$–10. We therefore re-simulate $z \sim 4$–10 JADES angular clustering measurements using *Bolshoi–Planck*.

We assign galaxy properties to haloes within *Bolshoi–Planck* using the UM model as described in Section 2.1 and use the best-fitting catalogue, hereafter referred to as the UM-BP mock catalogue. Adopting the same methods described in Sections 2.1−2.3, we simulate JADES/NIRCam angular clustering measurements at $z \sim 4$–10 and compare to the results presented in Fig. 3. We find that the typical simulated angular clustering measurement significances remain $\approx$7–9$\sigma$ at $z \sim 4$–7 with no systematic differences with redshift. At $z \sim 8$, the typical measurement significance lowers slightly from 5.0$\sigma$ to 4.7$\sigma$ and, at $z \sim 10$, it lowers from 4.9$\sigma$ to 4.1$\sigma$. Therefore, the only significant difference is at $z \sim 10$. This measurement significance decrease is largely due to the $\sim$0.15 dex lower $z \sim 10$ luminosity function in the UM-BP mock catalogue versus the UM-VSMDPL mock catalogue.[17] As noted in Appendix B, this decrease is well within the uncertainties of current $z \sim 9$–10 luminosity functions. Furthermore, as described in Section 3.1, we still find that Cycle 1 surveys are likely to enable $\sim$5$\sigma$ clustering measurements at $z \sim 10$ once we account for the increased depth across the JADES deep region as well as the 100 arcmin$^2$ CEERS program. Therefore, we find that our conclusions are robust to the size of the dark matter simulation used to generate our mock catalogues.

## APPENDIX D: MODELLING PHOTOMETRIC SELECTION OF HIGH-REDSHIFT GALAXIES

### D1 Determining optimal colour cuts

To ensure that our simulated high-redshift galaxy clustering measurements are realistic, we account for selection completeness. We assume that high-redshift galaxies will be photometrically selected with *JWST* using colour cuts as commonly done with *HST* imaging (e.g. Stark et al. 2011; González et al. 2012; Oesch et al. 2014; Bouwens et al. 2015, 2019). To determine the most optimal colour cuts for $z \sim 4$–10 galaxy selection, we utilize the JAGUAR package which provides *HST* and *JWST* broad-band photometry for mock galaxies from $z = 0.2$ to $z = 15$. The properties of the high-redshift JAGUAR mock galaxies are empirically constrained by stellar mass functions, stellar mass to UV luminosity relations, and UV luminosity to UV slope relations. See Williams et al. (2018) for further details.

We first add noise to JAGUAR mock galaxy photometry 50 times for each source using the medium *JWST* depths listed in table 5 of Williams et al. (2018) and the deep GOODS-S *HST* depths listed in table 1 of Bouwens et al. (2015). Specifically, these (5$\sigma$) depths are $m = 28.8$, 29.4, 29.5, 29.7, 29.8, 29.4, 28.8, 29.4, 28.9, and 29.1 for the F070W, F090W, F115W, F150W, F200W, F277W, F335M, F356W, F410M, and F444W *JWST* bands, respectively, and $m = 28.7$, 28.0, and 27.5 for the F435W, F606W, and F775W *HST* bands, respectively. We then iterate through potential photometric redshift selection criteria in various redshift intervals at $z \sim 4$–10 using a combination of Lyman-break, rest-UV continuum, rest-optical emission line, and Balmer-break colour cuts, seeking to maximize selection completeness and minimize the fraction of low-redshift contaminants.

We find that the colour cuts listed in Table D1 are optimal. These colour cuts separate galaxies into approximate photometric redshift bins of [3.7,4.4], [4.4,5.5], [5.5,6.7], [6.7,7.7], [7.7,9.0], and [9.0,11.0] which we refer to as the $z \sim 4$, 5, 6, 7, 8, and 10 intervals, respectively. Within the context of JAGUAR, they successfully select >50 per cent of galaxies at apparent rest-UV magnitudes of $m_{\rm UV} = 28.0$, 28.5, 28.8, 29.2, 29.2, and 29.5 at $z \sim 4$, 5, 6, 7, 8, and 10, respectively, while simultaneously introducing only 2–15 per cent fractions of low-redshift contaminants. We use these colour cuts to simulate both photometric and spectroscopic completeness as described in Section 2.2.

### D2 Accounting for contaminants in angular clustering measurements

After applying colour cuts to the JAGUAR catalogue, we find that low-redshift ($z < 3$) contaminants will comprise $\sim$2–15 per cent of the photometrically selected $z \sim 4$–10 galaxy samples used for simulated angular clustering measurements. We assume that these low-redshift contaminants will not be clustered with one another because none of the low-redshift contaminant distributions are sharply peaked in redshift space within the context of JAGUAR. We therefore insert a sample of randomly positioned objects into the angular clustering measurements so that they populate a fixed fraction of the total sample equal to our estimated low-redshift contamination fractions (Williams et al. 2011). Specifically, these fractions are 2.4 per cent, 9.1 per cent, 2.1 per cent, 5.3 per cent, 1.8 per cent, and 16.3 per cent at $z_{\rm phot} \sim 4$, 5, 6, 7, 8, and 10, respectively. Because the vast majority of these contaminants are within one magnitude of our adopted limiting magnitudes (listed in Table 1), we do not introduce these random points in the brighter threshold samples.

There will also be high-redshift contaminants which lie slightly outside the targeted redshift interval in photometrically selected samples. These sources may be clustered with each other and galaxies which lie inside the targeted redshift interval. We therefore select high-redshift contaminants according to modelled photometric selection completeness as a function of redshift and magnitude using the colour cuts in Table D1.

Unclustered contaminants decrease measured angular clustering strengths. For the case of low-redshift contaminants where each galaxy is completely unclustered relative to all other galaxies in the sample, autocorrelation clustering strengths are reduced by a factor of $(1 - f_{\rm low})^2$ where $f_{\rm low}$ is the low-redshift contamination fraction. However, for the case of high-redshift contaminants, some of these sources may be clustered. We therefore treat the effective fraction of high-redshift contaminants as $f_{\rm high,\,eff} = X_{\rm eff} \times f_{\rm high}$ where $f_{\rm high}$ is the inferred fraction of high-redshift contaminants and $X_{\rm eff}$ is some value between 0 and 1. Adopting $X_{\rm eff} = 0$ assumes that the high-redshift contaminants are as equally clustered as the intended galaxy sample. Conversely, adopting $X_{\rm eff} = 1$ assumes that every high-redshift contaminant is completely unclustered with all other galaxies in the sample. We find that fixing $X_{\rm eff} = 0.75$ leads to angular clustering strength measurements that best match the 'true' clustering strengths (Section 2.3) which only include galaxies inside the targeted redshift interval. Therefore, we take the total

[17] The luminosity function decrease arises because the $z \sim 10$ peak halo mass function from the *Bolshoi–Planck* simulation is $\sim$0.2 dex lower than that from VSMDPL while the UM model maintains a similar $z \sim 10$ $M_{\rm UV}$−$M_{\rm halo}$ relation when assigning galaxy properties to fit the range of input observational data.

**Table D1.** Photometric redshift colour cuts and non-detection criteria designed to be used with medium depth *JWST* photometry along with deep optical *HST* photometry. The $\mathrm{SN_{IVW}}(435,606,775)$ values are the inverse-variance weighted signal-to-noise values of the F435W, F606W, and F775W optical *HST* filters. All galaxies are also required to have a $\chi^2 > 25$ using all filters redder than and including F200W.

| Redshift | Colour cuts | Non-detection criteria |
|---|---|---|
| $z \sim 10$ | $\mathrm{F115W}-\mathrm{F150W}>1.6 \wedge -2.15 <\mathrm{F150W}-\mathrm{F200W}<0.55 \wedge$<br>$\mathrm{F115W}-\mathrm{F150W}>0.1(\mathrm{F150W}-\mathrm{F200W})+1.6$ | $\mathrm{SN(F090W)}<2 \wedge \mathrm{SN(F070W)}<2 \wedge$<br>$\mathrm{SN_{IVW}}(435, 606, 775) <2.5$ |
| $z \sim 8$ | $\mathrm{F090W}-\mathrm{F115W}>2 \wedge \mathrm{F150W}-\mathrm{F200W}<0.4 \wedge$<br>$\mathrm{F090W}-\mathrm{F115W}>2.6(\mathrm{F090W}-\mathrm{F115W})+2 \wedge \mathrm{F410M}-\mathrm{F444W}>0$ | $\mathrm{SN(F070W)}<2 \wedge \mathrm{SN_{IVW}}(435, 606, 775) <2.5$<br>$\wedge$ Not in $z \sim 10$ selection |
| $z \sim 7$ | $\mathrm{F090W}-\mathrm{F115W}>1.3 \wedge \mathrm{F150W}-\mathrm{F200W}<0.4 \wedge$<br>$\mathrm{F090W}-\mathrm{F115W}>2.6(\mathrm{F090W}-\mathrm{F115W})+1.3$ | $\mathrm{SN(F070W)}<2 \wedge \mathrm{SN_{IVW}}(435, 606, 775) <2.5$<br>$\wedge$ Not in $z \sim 8$–10 selection |
| $z \sim 6$ | $\mathrm{F070W}-\mathrm{F090W}>1 \wedge \mathrm{F115W}-\mathrm{F150W}<0.35 \wedge$<br>$\mathrm{F070W}-\mathrm{F090W}>2.4(\mathrm{F070W}-\mathrm{F090W})+1 \wedge \mathrm{F410M}-\mathrm{F444W}>-0.1 \wedge$<br>$\mathrm{F410M}-\mathrm{F444W}>-1.16(\mathrm{F200W}-\mathrm{F277W}) + 0.28$ | $\mathrm{SN(F435W)}<2 \wedge \mathrm{SN(F606W)}<2 \wedge$<br>Not in $z \sim 7$–10 selection<br>– |
| $z \sim 5$ | $\mathrm{F090W}-\mathrm{F115W}<0.24 \wedge \mathrm{F200W}-\mathrm{F277W}>0.28 \wedge$<br>$\mathrm{F410M}-\mathrm{F444W}>-1.06(\mathrm{F200W}-\mathrm{F277W})+0.34 \wedge$<br>$\mathrm{F410M}-\mathrm{F444W}>-0.1$ | $\mathrm{SN(F435W)}<2 \wedge$<br>Not in $z \sim 6$–10 selection |
| $z \sim 4$ | $\mathrm{F115W}-\mathrm{F150W}<0.25 \wedge$<br>$\mathrm{F335M}-\mathrm{F356W}<1.92(\mathrm{F150W}-\mathrm{F200W})-0.52$ | $\mathrm{SN(F435W)}<2 \wedge$<br>Not in $z \sim 5$-10 selection |

contamination fraction of each galaxy sample as $f_{\mathrm{low}} + 0.75 \times f_{\mathrm{high}}$. It will, in principle, be possible to more accurately forward model these effects once *JWST* begins to supply a large sample of spectroscopic redshifts at $z > 4$.

### D3 Estimating the impact of foreground objects

Bright foreground objects (e.g. low-redshift galaxies and stars) will inhibit the identification of some high-redshift galaxies within the JADES geometric footprint. Assuming that these bright sources are not strongly clustered and that they will not mask out significant patches of the sky,[18] the impact of these foreground sources will be to increase the Poisson noise in high-redshift clustering measurements proportional to the on-sky fraction masked out. We estimate the fraction of masked out area for each redshift selection interval using current *HST* images. Specifically, we calculate the signal-to-noise (using SEXTRACTOR; Bertin & Arnouts 1996) in 10 000 randomly placed circular apertures within the public GOODS F435W, F606W, and F775W mosaics (Giavalisco et al. 2004). Each of these apertures has a radius of 0.3 arcsec, similar to that typically used for $m{\sim}28.5$ galaxies at $z \sim 7$ (Bouwens et al. 2014).

We find that ≈8 per cent of apertures have SN(F435W)>2 and ≈12 per cent have SN(F435W)>2 or SN(F606W)>2. This suggests that only ∼10 per cent of $z \sim 4$–6 galaxies will be masked out by bright foreground sources. To estimate the fraction of masked out area for $z \sim 7$–10 selection, we use the extremely deep F850LP image from the XDF (Illingworth et al. 2013) as this image has depth similar to the planned medium JADES survey in F070W and F090W (Bouwens et al. 2015). We find that ≈15 per cent of apertures[19] within this F850LP mosaic have S/N>2. However, because the PSF of *JWST* will be ≈2 × narrower than that of *HST*, this likely represents an upper limit on the fraction of $z \sim 7$–10 galaxies not selected due to bright foreground sources.

## APPENDIX E: ACCOUNTING FOR SPECTROSCOPIC FOLLOW-UP EFFICIENCY

In our simulated clustering measurements, we adopt the survey design of the planned Cycle 1 JADES program which includes 26 NIRSpec MSA pointings. Every MSA pointing will be comprised of $R{\sim}100$ prism exposures as well as exposures with the $R{\sim}1000$ G235M and G395M grisms. Each $R{\sim}100$ MSA pointing can hold ∼200 sources (NIRSpec team, private communication). Using the JAGUAR mock catalogue, we find that the typical JADES $R{\sim}100$ prism depths will provide ≳50 per cent spectroscopic redshift completeness for $z \sim 4$–6 galaxies brighter than the limiting magnitudes set by our modelled photometric selection (Table 1). Applying this photometric selection to the UM-VSMDPL catalogue (Section 2.1), we expect that ∼5300 galaxies will be photometrically selected to be at $z \sim 4$–6 over the 142 arcmin$^2$ of JADES/NIRSpec coverage. Therefore, the 26 MSA pointings are sufficient to follow-up nearly every $z \sim 4$–6 candidate with $R{\sim}100$ spectroscopy assuming that each MSA pointing is completely filled with high-redshift galaxies.

Due to the smaller dispersion, each $R{\sim}1000$ pointing will likely only be able to hold ∼75 sources (NIRSpec team, private communication) suggesting a maximum of ∼2000 sources over 26 pointings. The JADES $R{\sim}1000$ depths will provide ≳50 per cent spectroscopic redshift completeness for $z \sim 7$ and 8 galaxies brighter than $m_{\mathrm{UV}} = 29.0$ and 28.7, respectively, within the context of JAGUAR. Applying modelled photometric selection to the UM-VSMDPL mock catalogue, we expect ∼800 such that $z \sim 7$–8 candidates to be photometrically selected over the 142 arcmin$^2$ of JADES/NIRSpec coverage, well below the maximum ∼2000. We therefore assume that every $z \sim 7$ and 8 photometric candidate brighter than our assumed spectroscopic limiting magnitudes (see Table 1) will also be spectroscopically followed up with NIRSpec.

While the above calculations show that the 26 JADES MSA pointings can, in principle, hold targets equal to the number of $z \gtrsim 4$ photometric candidates, we have not accounted for target placement restrictions to avoid overlapping spectra. Based on our experience with ground-based spectroscopy, a maximum of ∼50 per cent of available targets can plausibly be placed on a multiobject mask. We therefore discuss how the spatial clustering results presented in Section 3.1 would change if we assume a 50 per cent spectroscopic follow-up efficiency for the JADES Cycle 1 program. Assuming that targets of different magnitudes are assigned the same follow-up priority, cosmic variance will not be altered. However, Poisson error will approximately double as a result because $\sigma_{\mathrm{Poisson}} \propto 1/\sqrt{D1D2} \propto 1/\Sigma$ (Landy & Szalay 1993) where Σ is galaxy surface density. This suggests that spatial clustering will be measured with similar ($z \sim 4$–5) or weaker ($z \gtrsim 6$) precision relative to angular clustering with the Cycle 1 JADES

[18] This is a particularly well motivated assumption across the extragalactic GOODS fields as these fields were selected to be devoid of extremely bright foreground galaxies and stars.

[19] Only 2.5 per cent of apertures have $\mathrm{SN_{INV}}(435,606,775) < 2.5$.

program alone. Additional dedicated spectroscopic follow-up will be necessary to achieve the $\sim15\sigma$ precisions quoted in Fig. 3. Nonetheless, spatial clustering measurements are more desirable (if measured with similar significance) as they remove the systematic uncertainty associated with redshift contamination.

We acknowledge that this approach ignores effects that may lead to a slit placement efficiency that is dependent on the local target surface density. In particular, in instances where multiple satellites are hosted by one central galaxy, we would expect that the central would have a higher probability of being targeted versus the satellites if the spectroscopic survey is aimed at building a uniform sample across luminosity and/or mass. Fortunately, there are several low-redshift techniques (e.g. Li et al. 2007; Zehavi et al. 2011; Patton et al. 2013) that can be applied to correct for any systematic impact this may have on the clustering measurements. However, because satellite populations remain highly uncertain at high redshifts (Appendix F), it is difficult to predict which technique(s) will be most appropriate until *JWST* begins delivering data.

## APPENDIX F: SATELLITE EVOLUTION IN THE UNIVERSEMACHINE MODEL

The UM model accounts for both satellite and central halo evolution, allowing us to investigate not only central galaxy clustering, but satellite–satellite and satellite–central clustering as well. During the halo identification process, haloes are identified as true satellites if they reside within the virial radius of a more massive halo. It is well known that multiple systematic effects lead to the premature disruption of low-mass satellites within dark matter simulations (e.g. van den Bosch et al. 2018; van den Bosch & Ogiya 2018). The UM model compensates for these effects by allowing disrupted satellites to continue orbiting their host as orphan galaxies (Wang et al. 2006) until their expected mass loss (computed via Jiang & van den Bosch 2016) reaches a preset threshold. This threshold is tuned to match $z = 0$–1 clustering constraints for star-forming and quiescent galaxies and it is assumed not to evolve with redshift. The net effect is to extend satellite lifetimes by $\sim$25 per cent at all redshifts; however, the total satellite fraction is always $\leq$15 per cent at $z > 4$.

We note that our simulated galaxy major merger pair fraction measurements (Section 4.2) are sensitive to the orphan populations introduced by the UM model. Ignoring all orphans reduces the measured pair fractions by 0.3 dex at $z\sim4$ and 0.2 dex at $z\sim5$, which would make our simulated measurements less consistent with observations (Fig. 5). This suggests that the satellite correction methodology adopted by the UM model is reasonable at high redshifts and we therefore adopt this model as a plausible representation of satellite behaviour at $z > 4$ even as we acknowledge that clustering constraints from *JWST* will inform better models in the future.

## APPENDIX G: JACKKNIFE ERRORS ON THE TWO-POINT CORRELATION FUNCTION WITH *JWST*

Jackknife and bootstrap methods are commonly used to determine the error on observationally measured two-point correlation functions. However, it is known that these errors should be interpreted with caution as they do not necessarily follow a Gaussian error distribution nor are they necessarily accurate (Norberg et al. 2009). On one hand, this is because clustering depends not only on halo mass but also environment, halo formation time, concentration, and subhalo occupation (e.g. Sheth & Tormen 2004; Gao, Springel & White 2005; Wechsler et al. 2006; Pujol & Gaztañaga 2014). This phenomenon is broadly referred to as 'assembly bias' and can be particularly strong in the narrow pencil-beam fields *JWST* is designed to survey. Thus, if *JWST* survey volumes are not sufficiently large to capture the full extent of clustering variations for a given galaxy sample, clustering strength errors will be underestimated. Furthermore, the number of subsamples into which the observational data are split must be sufficiently large to accurately derive the true errors.

We test whether these concerns apply to future *JWST* surveys by asking how often the 'true' two-point correlation functions (Section 2.3) are enclosed by jackknife errors for the 100 simulated clustering measurement realizations for each galaxy sample (Section 2.1). Jackknife errors are calculated by dividing each of the two GOODS fields into five roughly equal area segments split by constant right ascension for a total of 10 jackknife samples. We find that these jackknife errors have typical $1\sigma$ and $2\sigma$ confidence intervals of $\sim$70 per cent (i.e. jackknife errors enclose the true correlation functions in 70 of the 100 realizations) and 94 per cent, respectively, with JADES angular clustering measurements and $\sim$62 per cent and 88 per cent, respectively, with JADES spatial clustering measurements. Because these confidence intervals are not highly different from those expected of a Gaussian distribution and because we wish to predict observational clustering measurements, we adopt jackknife errors throughout this work.

## APPENDIX H: EVOLUTION OF PEAK HALO MASS FUNCTION AT Z ∼ 4–10

Here, we derive a smooth analytic fit to the redshift evolution of the peak halo mass function from $z\sim4$–10. We calculate the intrinsic peak halo mass function at snapshot redshifts spanning $z = 3.5$–10 from the VSMDPL dark matter simulation (Section 2.1) using only haloes with masses of $\log_{10}(M_{\rm peak}/M_\odot) > 9.5$. After experimentation with various functional forms, we find that the following form best matches the intrinsic evolution over these

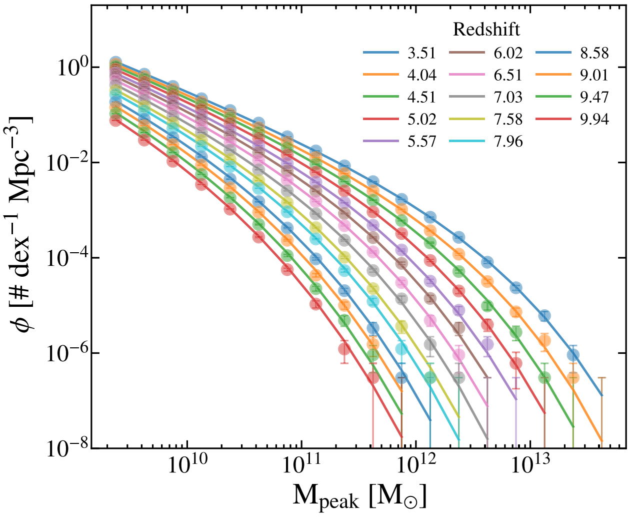

**Figure G1.** Evolution in the peak halo mass ($M_{\rm peak}$) function at $z = 3.5$–10. The markers show data points from the UM-VSMDPL mock catalogue (Section 2.1) with Poisson errors. The lines show the fit from equation (H1) at each redshift.

redshifts:

$$\log_{10}\phi\left(M_{\rm peak}, z\right) = \log_{10}\phi^{*} + \alpha\left(x - M^{*}\right) - \log_{10}(e)\,10^{\,\beta\left(x-M^{*}\right)}. \qquad \text{(H1)}$$

Here $x \equiv \log_{10}(M_{\rm peak}/{\rm M}_{\odot})$. We find that it is not necessary to leave $\alpha$ (which represents the faint-end slope) as a free parameter and fix it at $\alpha\ = -0.95$. The evolutionary forms of the other parameters are found to be:

$$\beta\,(z) = -0.0185\,z + 0.488 \qquad \text{(H2)}$$

$$M^{*}\,(z) = 10.944\ -\ 0.652\,(z - 4.558) + 0.204\ \log_{10}\left(1 + 10^{\,0.820(z-4.558)}\right) \qquad \text{(H3)}$$

$$\log_{10}\phi^{*}\,(z) = -1.361\ +\ 0.533\,(z - 4.690) - 0.0714\ \log_{10}\left(1 + 10^{\,1.722(z-4.690)}\right). \qquad \text{(H4)}$$

This analytic evolutionary form agrees with the VSMDPL data at $z = 3.5$–10 to within either $2\sigma$ (Poisson) or 0.10 dex at each data point (binned by 0.25 dex in $M_{\rm peak}$). These data points and the fit from equation (H1) are shown in Fig. G1.

We note that these halo mass functions include the contribution from satellites identified by our adopted halo finder (ROCKSTAR; Behroozi et al. 2013a) as well as orphan satellites that have been introduced to account for the premature disruption of low-mass satellites within dark matter simulations (Appendix F). Excluding orphan satellites would have a very small impact on the intrinsic peak halo mass function at $z\ = 3.5$–10, decreasing it by at most 0.04 dex.

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