Peer Review File

Manuscript Title: A population of red candidate massive galaxies ~600 Myr after the Big Bang

Redactions – unpublished data

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This is a well-motivated study, searching for relatively mature high-redshift galaxies at early times (redshifts 7 < z < 11), from early-release JWST multi-band NIRCam imaging, by looking for objects which appear to display two distinct breaks in their spectral energy distributions (the Lyman break at rest-frame 1216 angstroms, and the Balmer break at rest-frame 3600 angstroms, with the latter being indicative of early star-formation activity). Although this is one of many studies which have aimed to select high-redshift galaxies from the early public JWST imaging, this is the only one which explicitly aims to select galaxies via this double break technique – in this sense it is well-focused and worthy of consideration for publication in Nature. The key science goals here are to get an early insight into the number density of "older" galaxies at these early times, and to measure their stellar masses – the stellar mass density already in place at z ~ 10 has clear implications for the prevalence and timing of even earlier star-formation activity.

The authors report 14 galaxies selected using their "double break" method from their own reduction of the public JWST NIRCam imaging taken as part of the CEERS Early Release Science programme. One advantage of the use of public data is that other research groups can check the validity of studies such as this, but unfortunately the authors have not provided the coordinates of their final selected massive high-redshift galaxy candidates. This certainly needs to be rectified in any final published version of this paper, but my understanding is that this may just be an oversight due to time pressure, and indeed the authors have been happy to share the coordinates with some of my collaborators (without knowledge that I am refereeing this paper!). As a result I am able to confirm that this galaxy sample is pretty solid, with generally robust photometric redshifts – in particular, for the 7 proposed high-mass galaxies highlighted in Fig. 1 and presented in Fig. 2, an independent analysis, based on a completely independent reduction of the data, yields similarly high redshifts.

However, despite this apparently successful sample selection, there are some important scientific issues which need to be properly addressed before this paper can be considered appropriate for publication in Nature. First, in addition to the lack of object coordinates, there is a lack of (vital) detail on the modelling, and hence of the accuracy with which physical parameters have been extracted. In particular, clearly tabulated (and explained) uncertainties on redshift and stellar mass are required in the main paper. This matters in part because it is clear that some of the redshifts have been (modestly) over-estimated – for example object 21834 is detected in the F115W imaging and so cannot lie at z = 10.83 as stated in Fig 3. A modest reduction in adopted redshift might seem unimportant, but it has already become clear (e.g. Carnall et al. 2022) that a relatively small change in redshift (towards lower z) can favour emission-line contributions to the second (redder) photometric step, weakening the evidence for a Balmer break, in turn resulting in a younger inferred age, which can then result in a rather dramatic reduction in inferred stellar mass.

This is just one example (although in our reduction we find that all 7 galaxies are detected in F115W, and hence none of 14924, 16624 or 21834 can be at z > 9.5 as reported), but it helps to illustrate the broader point. Before asserting so boldly that these objects are all extremely massive, to an extent which potentially threatens the LambdaCDM cosmological model, it is incumbent on the authors to properly expose the full range of redshift/age/mass parameter space allowed by the data. I am not saying the results are wrong, but such a potentially important (and arguably surprising) result deserves/requires a more careful analysis than is currently presented.

In essence, then, I am requesting a rather more thorough and comprehensive analysis of the data, especially given the strongly *correlated* uncertainties in derived physics quantities, with special focus on the potential contribution of emission lines to what are proposed to be continuum breaks – the authors are clearly aware of this issue but it is really not obvious that the full parameter space of the redshift probability function in tandem with the full range of strong emission-line galaxy templates has been properly explored (despite the discussion of alternative templates in the methods section). Indeed, it might be worth playing devil's advocate, assuming that *all* of the supposed Balmer break is due to emission lines, and then checking whether the inferred redshift for each object is reasonable or not (and whether the quality of the resulting fit is acceptable). But really this should just be a straightforward job of exploring the acceptable fit space while marginalising over all parameters, including redshift.

Also not discussed at all (as far as I could spot) is the extent to which *selection* on the basis of spectral breaks will inevitably have biased (i.e. enhanced) the strengths of the inferred breaks in the final sample (and hence biased ages and stellar masses towards high values). A little bit of source injection and re-extraction simulation work would help here.

The detail of this will of course need to appear in an expanded version of the Methods section, but nonetheless it should be possible to propagate the results back to the main text in a way which better enables a proper discussion of the uncertainties, and any caveats which might apply to the headline conclusion. Provided the results hold up, this would significantly strengthen the paper, and ultimately enhance its impact. This is exciting work, potentially appropriate for publication in Nature, but it first deserves to put on a more solid footing.

Finally, I would note that in addition to the requested key table summarizing physical quantities with full uncertainties, the figures (especially Figs 3 and 4) and the associated captions could also be much improved/polished.

Referee #2 (Remarks to the Author):

The paper, "A very early onset of massive galaxy formation" by Labbe et al. paper presents a sample of 14 massive galaxy candidates at z~7-11, identified within one of the first publicly-released infrared imaging datasets from the James Webb Space Telescope (JWST), i.e., CEERS. These massive galaxy candidates have a median stellar mass of ~10^10 Msun, and, if robust, their cosmic abundance at these masses far exceeds the predictions from previous, rest-UV-selected stellar mass functions. The subject matter of this paper is compelling, and showcases the potential of JWST for constraining the abundance of a wide range of galaxy stellar populations in the early universe. At the same time, I have significant concerns about the lack of information provided in this paper regarding basic properties of the individual targets in the 14-galaxy sample, the methodology, and the surprising implications of such a high stellar mass density in massive systems at z~7-11. I will describe these concerns below, along with a bunch of minor comments. Indeed, the claims in the paper regarding the stellar mass density in massive systems are fairly extraordinary, and therefore require extraordinary (or even, merely, sufficient) evidence, which is not currently provided. The authors must address the concerns listed below before the paper can be considered for publication.

MAJOR COMMENTS

1. Missing galaxy information: There is a surprisingly small amount of detail provided about the 14 sample objects. There is no information about their coordinates (which is non-standard compared to other recent JWST papers, e.g., Donnan et al. 2022, Naidu et al. 2022). There is no table listing the z_phot and z_phot uncertainty for each galaxy, its photometric measurements and all derived stellar population parameters using the preferred methodology. Such information is crucial to include, for readers and others working in the field, in the interests of testing and comparing results for the same public dataset. I also note that stellar masses are listed (and plotted) without error bars.

2. Details lacking regarding methodology: On a similar theme, I found the descriptions of methodology throughout the paper to be too brief, or entirely lacking. Here are some specific examples, where the paper should be expanded so that the results can be fully appreciated:

i. The paper adopts a two-color selection technique, plus non-detection at optical wavelengths. The two-color selection criteria should have been motivated by model tracks in color-color diagrams, showing how galaxies covering a range of different stellar population parameters (including a range of nebular emission strengths) and redshifts populate the space of F150W-F277W and F277-F444W.

ii. I found that the description of the derivation of stellar masses lacked clarity. Given the closer connection of the Prospector-alpha derived masses to a physical picture of associated star-formation histories, it was unclear to me why the EAZY stellar masses were adopted as "fiducial" (if in fact that was done). It was also unclear to me whether models with EAZY templates modified to include enhanced rest-optical emission-line fluxes were adopted as "fiducial" or were simply mentioned as an alternative. In particular, I wonder this for the 5 galaxies for which the enhanced emission-line template set provided a better fit (as described in the "Fit to the Photometry" section of "Methods"). A more complete discussion of the systematic uncertainties associated with individual stellar mass estimates (i.e., when allowing for significant nebular emission) is required. I mention this point in the context of model degeneracy between a strong Balmer Break and strong nebular emission.

iii. Another key result in the paper is the stellar mass density contained in the massive systems presented here. Yet, there is no discussion included regarding the calculation of cosmic volume for these sources, and the manner in which cosmic variance is reflected in the number density error bars. When presenting cosmic densities, it is crucial to provide at least a minimum description of such calculations, at the very least in the "Methods" section.

iv. It is stated that no galaxies with log(M*/Msun)>10 have been found at z>8 in rest-UV-selected samples. I would have liked to see more discussion of this point. Specifically, the authors should clarify if this absence results because the rest-UV (i.e., observed near-IR photometry points) of the higher M/L galaxies presented here would be fainter than the pre-existing near-IR photometric limits. This is a case where it would be very useful to include the F150W magnitudes of the 7 sample galaxies with M>10^10 Msun, to demonstrate if they are fainter than current HST/F160W limits. For example, in a paper by one of the co-authors (Stefanon et al. 2021, ApJ, 922, 29, Table 1), data over ~10 arcmin^2 down to H_160~29 AB are analyzed. It seems that such limits should be sufficient for detecting galaxies like those presented here (and at least ~1-3 of them, given the 7 galaxies with log(M*/Msun)>10 in the 40 arcmin^2 of CEERS). There is a comment at the end of the caption of 3 that the two most massive galaxies were previously mis-identified as being at z~1, but some further clarification about why log(M*/Msun)>10 galaxies were not recovered in previous high-redshift stellar mass function work would strengthen the paper.

3. Lack of consideration of implications in the context of the dark matter halo mass function: One of the key results of the paper is that the observed stellar mass in galaxies with log(M*/Msun)>10 is significantly higher at z~8 and z~10 than predicted by existing rest-UV-selected stellar mass functions. In fact, I would argue that the stellar mass density associated with the galaxies presented here is *uncomfortably* (perhaps unphysically) large. Specifically, (again referring to previous work by one of the co-authors), the stellar mass density shown in Figure 4 for z~10 galaxies with log(M*/Msun)>10 (~10^6 Msun/Mpc^3) is roughly equivalent to the *total* mass in dark matter halos down to 10^10.72 Msun at z~10 (see Stefanon et al. 2021, Table 6). Given typical assumptions regarding the stellar-to-halo mass ratio, the galaxies presented here should only populate dark matter halos with M>10^12 Msun, with a significantly lower associated total mass density. The strong tension between the inferred stellar mass density and the total mass density in dark matter halos that can host massive galaxies must be considered here.

A related concern is whether a galaxy with M*~9x10^10 Msun (14924, z=9.9) can be produced in a standard framework. According to Behroozi and Silk (2018, Figure 2), such massive galaxies should have space density of 10^-9 /Mpc^3 at z=10. This tension needs to be discussed as well.

One additional angle that should be considered is that the constant inferred stellar mass density in massive systems between z~8 and z~10 implies rapid evolution in the galaxy stellar-mass-to-halo-mass relation, given the rapid evolution in the dark-matter-halo mass functon(e.g., see Oesch et al. 2018, ApJ, 855, 105).

MINOR COMMENTS

1. Figure 1: I have a few comments on this figure. First, it should be stated which set of modeling assumptions (EAZY vs. Prospector-alpha, enhanced emission templates or not, etc.) is adopted.

There should be error bars on stellar masses, and it should be clarified in the caption why two of the red stars are enlarged relative to the rest (is it simply the two most massive galaxies?). Also, it appears that, above z~8.5, the majority (7/10) of F444W-detected galaxies satisfy the two-color selection criteria. Is that expected?

2. p. 2, end of first paragraph on page. It is asserted that the most massive galaxy in the sample 38094 has two nearby companions with approximately the same redshift, "indicating that is in a group or proto-cluster." Given the likely uncertainties in photometric redshifts (which are not listed in the paper), nearby companions on the sky may not be physically associated. This statement should be softened.

3. Figure 3: I recommend showing panels for all 14 galaxies and making them all the same size. It is difficult to read the p(z) panels, given that they are immersed without borders within the rest of each SED panel. I recommend including the p(z) information as separate, adjacent panels. I recommend listing errors for both z_phot and M*.

4. Figure 4 and associated text in the paper. It is stated in the paper text that "the dashed lines assume errors of a factor of 3" -- errors in what? There are curves in this figure labeled "S21", which I am assuming is for "Stefanon et al. 2021", but these labels are nowhere defined. In the caption, it is stated, "Dashed lines are derived from Monte Carlo simulations that perturb the masses of individual galaxies. The perturbations are drawn from a log-normal distribution with a 1-sigma width of 0.48 dex." Again, it is unclear which galaxies are being perturbed, and why the 1-sigma width of 0.48 dex was adopted. Some referencing and motivation is required here.

5. I recommend one more round of proofreading for small typos.

Author Rebuttals to Initial Comments:

Response to Referee #1

We are happy to read that the referee considers our finding important and worthy of consideration by Nature.

Referee:
One advantage of the use of public data is that other research groups can check the validity of studies such as this, but unfortunately the authors have not provided the coordinates of their final selected massive high-redshift galaxy candidates. This certainly needs to be rectified in any final published version of this paper, but my understanding is that this may just be an oversight due to time pressure, and indeed the authors have been happy to share the coordinates with some of my collaborators (without knowledge that I am refereeing this paper!). As a result I am able to confirm that this galaxy sample is pretty solid, with generally robust photometric redshifts – in particular, for the 7 proposed high-mass galaxies highlighted in Fig. 1 and presented in Fig. 2, an independent analysis, based on a completely independent reduction of the data, yields similarly high redshifts.

Response:
We are encouraged that the referee can straightforwardly repeat the selection. This was our intention by selecting on basic observables.

The referee is correct that the omission of coordinates was an oversight due to time pressure. It was certainly not our intention to obfuscate! We made the coordinates, photometry, and fitting outcomes available right away on a website linked to the arxiv posting, and responded immediately to email requests from several researchers. The source catalog was downloaded a dozen times, so we hope that the omission in the original submission was not too much of an impediment.

Two tables with source information, one with all photometric measurements (Extended Data Table 1) and one with coordinates, and fiducial stellar masses, and redshifts (Extended Data Table 2) are included in the revision. A fits file with all source information and modeling results will also be made available on the github site[1].

Referee:
However, despite this apparently successful sample selection, there are some important scientific issues which need to be properly addressed before this paper can be considered appropriate for publication in Nature. First, in addition to the lack of object coordinates, there is a lack of (vital) detail on the modelling, and hence of the accuracy with which physical parameters have been extracted. In particular, clearly tabulated (and explained) uncertainties on redshift and stellar mass are required in the main paper. This matters in part because it is clear that some of the redshifts have been (modestly) over-estimated – for example object 21834 is detected in the F115W imaging and so cannot lie at z = 10.83 as stated in Fig 3. A modest reduction in adopted redshift might seem unimportant, but it has already become clear (e.g. Carnall et al. 2022) that a relatively small change in redshift (towards lower z) can favour emission-line contributions to the second (redder)photometric

[1] https://github.com/ivolabbe/massive

step, weakening the evidence for a Balmer break, in turn resulting in a younger inferred age, which can then result in a rather dramatic reduction in inferred stellar mass.

This is just one example (although in our reduction we find that all 7 galaxies are detected in F115W, and hence none of 14924, 16624 or 21834 can be at z > 9.5 as reported), but it helps to illustrate the broader point.

Response
The referee is correct on all accounts, and we are grateful for them pointing this out.

On the last, clearly important point, we agree that our initial redshifts for the sources at z~10 were somewhat overestimated. We had been aware of this effect and had already corrected the redshift of one z~10 source due to a clear F115W detection, but we agree that essentially all sources are visible in F115W and therefore should be at z<9.5.

Our initial redshift overestimates appear to have been driven by too conservative (large) uncertainties in the JWST short wavelength photometry, leading to formal non-detections in F115W and higher redshift z>9.5 solutions. In our improved reductions (mitigating 1/f noise in particular) the photometric uncertainties in F115W are lower and all but one are now clearly detected in F115W. The revised photometric redshifts for these sources are now z=8.7-9.1.

Three further changes were made. The first is that the pre-flight photometric zeropoints that we used were incorrect, making the SEDs on average redder than they should be. STScI released a zeropoint update on July 29, which improved the flux calibration for the Long Wavelength bands, but further testing showed that the Short Wavelength bands were made worse. Two authors of our paper (Gabriel Brammer and Ivo Labbe) independently derived new zeropoints, using complementary techniques (stars in the LMC and fits to galaxies in the CEERS field). These zeropoints turned out to be consistent with one another within 3%, and consistent with two additional independent analyses of stars in the globular cluster M92 (Boyer et al. 2022, Nardiello et al. 2022). Now, finally, the NIRCAM calibrations are under control.

Second, in the initial submission we used redshifts and masses from the EAZY photometric redshift code, with Prospector only used as a check. We now expanded the analysis and explicitly use and compare four different mass and redshift measurements: EAZY, Prospector, and two setups of the Bagpipes code. (We also test a fifth method, templates that were proposed in a research note by Steinhardt et al. (2022) specifically for extremely high redshift galaxies, but these templates do not fit the photometry well and assume a vastly different IMF.) Our fiducial masses and redshifts are the medians of these four methods, with well-motivated uncertainties. This is all described in the substantially revised and expanded Methods section.

Third, for consistency with previous mass functions at the highest redshift (e.g., Song+2016, Stefanon+2021,2022) and with IMFs that were derived for the cores of massive early-type galaxies (Treu+2010, Cappellari+2012, van Dokkum+2017), we adopt a Salpeter IMF.

In the end, these modifications produced changes in the masses and redshifts for individual sources and a general lowering of the mass density, but the overall conclusions do not change. The largest change is for 16624 (the bluest source with a clear excess in F444W), whose mass decreased by ~1 dex, in equal part due to the combination of NIRCam recalibration and the contribution of [OIII]5007 to F444W due to somewhat lower redshift, as the referee suspected. However, other sources change remarkably little, e.g., ~<0.1 dex for 35300 (log(M)=10.4 at *z*=9.1, with a strong F115W-F150 > 2 mag break and no indication of strong emission lines from F410M-F444W). Similarly, the most massive source 38094 (log(M)~11 at z=7.5) did not change at all. Interestingly, many of the most massive sources exhibit only a weak or no correlation between stellar mass and redshift and produce high masses irrespective of the presence of strong emission lines.

This is likely because the F410M filter helps to constrain the contribution of emission lines, improving redshifts and stellar mass estimates (see e.g., Roberts-Borsani+2021) and because the spectral energy distributions (SEDs) of the sources in this sample are very red throughout the NIRCam filters (rest-frame UV and optical), as opposed to a blue (flat in f_nu) SED with only one or two bands boosted by emission lines. Such red SEDs require substantial reddening (and high M/L) even for young stellar populations with strong emission lines.

In fact, the overall SEDs of the sources in this sample are much redder than any galaxy from the literature at similar redshifts. To illustrate this, we add a new panel to Figure 3, which shows the average rest-frame SED compared to galaxies previously found in studies with *HST* and *Spitzer*. The JWST sample clearly shows a Lyman break, anchoring their redshifts, but the entire SED is qualitatively different and much redder at all wavelengths than previously-studied objects at similar redshifts.

Referee:
Before asserting so boldly that these objects are all extremely massive, to an extent which potentially threatens the LambdaCDM cosmological model, it is incumbent on the authors to properly expose the full range of redshift/age/mass parameter space allowed by the data. I am not saying the results are wrong, but such a potentially important (and arguably surprising) result deserves/requires a more careful analysis than is currently presented.

*In essence, then, I am requesting a rather more thorough and comprehensive analysis of the data, especially given the strongly *correlated* uncertainties in derived physics quantities, with special focus on the potential contribution of emission lines to what are proposed to be continuum breaks – the authors are clearly aware of this issue but it is really not obvious that the full parameter space of the redshift probability function in tandem with the full range of strong emission-line galaxy templates has been properly explored (despite the discussion of alternative templates in the methods section). Indeed, it might be worth playing devil's advocate, assuming that *all* of the supposed Balmer break is due to emission lines, and then checking whether the inferred redshift for each object is reasonable or not (and whether the quality of the resulting fit is acceptable). But really this should just be a straightforward job of exploring the acceptable fit space while marginalising over all parameters, including redshift.*

Response
The referee is right, of course. We should note that when writing the paper, we did not appreciate the extent of the tension that these early massive galaxies presented to theoretical models. **Redacted**. We therefore agree that when our results are viewed in this light, the estimates should be robust.

As also discussed above, we expanded and modified our analysis of the stellar masses as follows: 1) we update the EAZY templates in a manner that is less ad-hoc, 2) we changed the SFH prior of Prospector to favor younger ages, and thus emission lines over Balmer Breaks, 3) added an additional analysis with Bagpipes to evaluate the dependence on stellar population fitting codes, 4) adopt a “devil’s advocate” strong emission line model. Finally, we adopt the median across the 4 methods as our fiducial redshift and stellar mass, rather than favoring one method in particular. A comparison of all the stellar mass and redshift solutions for each object is presented in Extended Data Figure 4.

1. EAZY Rather than arbitrarily boosting emission lines of one existing red template, we generate two templates with strong emission lines (650-1100A)using FSPS (one red $f\lambda$=constant, one blue $f\nu$=constant), so that the emission line strengths and mass-to-light ratios are self-consistent with more flexibility.

In addition, we refit the sample with an upper limit of $z < 9.5$, taking the F115W detection as a strong prior on redshift and removing the solution space with strong Balmer Breaks at z~10 to avoid erroneously high stellar masses. We note though that with the improved reductions, zeropoints, and updated templates, the best-fit EAZY redshifts are already at $z < 9.5$, so the prior only removes a high-mass tail of the probability distribution.

2. Prospector. Prospector allows for the parameter exploration the referee requests. While we initially used Prospector masses only as a sanity check on the EAZY masses, in the new version we use them explicitly. The Prospector results favor relatively old ages (~0.3 Gyr) and Balmer Breaks for most galaxies, after marginalizing over the entire parameter space (which includes strong emission lines). This is partly the result of Prospector’s default SFH prior, which is flat in time. We therefore also adopt a strongly rising SFH as a prior (note that Prospector in all cases fits a flexible non-parametric SFH). We additionally adopt a stellar mass function prior; since the z~6-10 mass function is unknown, we adopt the observed mass function at z=3 as a conservative choice. This effectively downweighs high-mass solutions. Finally, we run Prospector in normal – rather than neural-net accelerated – mode to avoid a situation where the neural-net would be used in a parameter space where it was not sufficiently trained. This takes some time: a full Prospector run for a sample of 13 galaxies takes about a week.

In the end, as described in the Methods section, the Prospector results assuming two different SFH priors do not show a large variation. In particular, although some galaxies require contributions of Balmer breaks independent of priors and assumptions, with mass-weighted ages 100 Myr (14924, 35300) to 200 Myr (38094), the runs with a rising star formation rate prior show a much increased contribution from emission lines, with a median mass-weighted stellar age of the entire sample of only 30 Myr, an order of magnitude smaller than the age of the universe at z=10, or the cosmic time elapsed between $7 < z < 10$. This shows that despite the much lower overall ages and stronger emission lines, the masses are quite robust, as the new fits with younger ages require more reddening (median Av~1.5) resulting in high M/L.

3. Bagpipes. While Prospector allows a sufficient exploration of available parameter space, we expanded the analysis to include results with a different popular fitting code, Bagpipes, as well, to explore the dependence of the result on the stellar population fitting method. Details of the Bagpipes setup are provided in the Methods section. Having already explored SFH priors and flexible SFHs, we adopt here simply a constant SFH and restrict redshifts to z <

9.5, finding that the stellar masses and redshift agree well with the EAZY and Prospector outcomes (mean offsets in stellar masses < 0.1 dex), with the exception of 2 (out of 13) galaxies (7274, 25666) where Bagpipes favors z~6.5 solutions rather than z~8, due to confusion between Hb+[OIII] and Ha+[NII]+[SII] emission lines. These galaxies do not impact the results in this paper.

4. Extreme emission lines. We also use Bagpipes to explore "forced" models that maximize the emission line contribution to the reddest filter, F444W, to assess the robustness of the results when we do not search for the best fit in a wide parameter space but only in a restricted space where the intrinsic rest-frame UV - optical continuum colors are as blue as possible. We do this by selecting a delayed τ SFH with a minimum τ>0.5 Gyr. This forces strongly rising SFRs and very young ages at high redshift. In addition, we limit the redshifts to z<9.0, which forces the strong Hb+[OIII] complex to fall in the F444W filter.

The resulting fits produce extremely (and unrealistically) young median ages (~10 Myr) with extremely strong emission lines, except for 11184 (170 Myr) and 35300 (80 Myr). The fits are generally acceptable (albeit at somewhat worse chi2, by a median factor 1.3, compared to the EAZY fits). Interestingly, while for some galaxies the stellar masses are significantly lower, the emission line favorable setup still recovers 5/6 massive $\log(M_*/M_\odot) > 10$ galaxies, including some of the most massive 38094 (z=7.51, log(M)=11.2) and 35300 (z=8.8, log(M)=10.3). The mean difference in stellar mass between EAZY and this model is $\log(M_{bagpipes\_rising}/M_{EAZY}/) = -0.1$ dex for the massive galaxies.

In other words, most stellar masses change within the uncertainties and the average systematic differences are small. If we adopt the redshift and stellar masses of each 4 models above by themselves, we would find the same key result within the uncertainties (the very high cumulative mass density at $M_* > 10^{10}$ $M_\odot$)).

We also test a fifth method, EAZY templates that were proposed in a research note by Steinhardt et al. (2022) specifically for extremely high redshift galaxies, but these templates do not fit the photometry well, produce inconsistent redshifts, and assume a vastly different IMF, which can not be tested with the photometric data. More details are provided in the Methods section.

Fiducial masses and redshift. Given the good correspondence between the 4 methods for all but a few sources, we adopt the median across these 4 setups as our fiducial redshift and stellar mass: for each object we remove the lowest and highest value and average the remaining two. The same is done for the corresponding uncertainties. While a combination of methods is a bit more involved to reproduce, we think there is merit to not favor any particular outcome in this case, because the results agree well.

Referee
*Also not discussed at all (as far as I could spot) is the extent to which *selection* on the basis of spectral breaks will inevitably have biased (i.e., enhanced) the strengths of the inferred breaks in the final sample (and hence biased ages and stellar masses towards high values). A little bit of source injection and re-extraction simulation work would help here.*

Response
That is an excellent point. This effect can be particularly pronounced when the selection is made at low SNR. We estimate this effect by removing the lowest S/N band of each color

selection from the analysis, refitting, and verifying that the best-fit model SED is consistent with the flux in the selection band, which would not be the case if they were scattered high/low because of selection effects.

For F150W-F277W, the F150W has lowest S/N, so we remove this band and re-fit with EAZY keeping the redshift fixed. Then we compare the predicted F150W model fluxes to the measured ones. We repeat this for F277W-F444W, dropping the lower S/N F277W, refitting, and comparing the model F277W to the observed. In neither case do we detect an offset, probably because the SEDs are very high SNR on average, especially in the NIRCam Long Wavelength. Qualitatively, this can also be seen in the new panel of Fig 3, which shows the average SED of the sample galaxies. The SED is uniformly red, not just in the selection bands.

Referee
The detail of this will of course need to appear in an expanded version of the Methods section, but nonetheless it should be possible to propagate the results back to the main text in a way which better enables a proper discussion of the uncertainties, and any caveats which might apply to the headline conclusion. Provided the results hold up, this would significantly strengthen the paper, and ultimately enhance its impact. This is exciting work, potentially appropriate for publication in Nature, but it first deserves to put on a more solid footing.

Response
We have taken the advice of the referee to heart and as described above significantly expanded the analysis and description in the Methods section. We hope that the analysis is now sufficient.

Referee:
Finally, I would note that in addition to the requested key table summarizing physical quantities with full uncertainties, the figures (especially Figs 3 and 4) and the associated captions could also be much improved/polished.

Response:
We added Tables with full photometry (Extended Data Table 1) and location, redshift, stellar mass, with full uncertainties (Extended Data Table 2). We also significantly updated Figure 3 and 4 and their captions.

Referee #2 (Remarks to the Author):
The paper, "A very early onset of massive galaxy formation" by Labbe et al. paper presents a sample of 14 massive galaxy candidates at z~7-11, identified within one of the first publicly-released infrared imaging datasets from the James Webb Space Telescope (JWST), i.e., CEERS. These massive galaxy candidates have a median stellar mass of ~10^10 Msun, and, if robust, their cosmic abundance at these masses far exceeds the predictions from previous, rest-UV-selected stellar mass functions. The subject matter of this paper is compelling, and showcases the potential of JWST for constraining the abundance of a wide range of galaxy stellar populations in the early universe. At the same time, I have significant concerns about the lack of information provided in this paper regarding basic properties of the individual targets in the 14-galaxy sample, the methodology, and the surprising implications of such a high stellar mass density in massive systems at z~7-11. I will describe these concerns below, along with a bunch of minor comments. Indeed, the claims in the paper

regarding the stellar mass density in massive systems are fairly extraordinary, and therefore require extraordinary (or even, merely, sufficient) evidence, which is not currently provided. The authors must address the concerns listed below before the paper can be considered for publication.

MAJOR COMMENTS

Referee:
1. Missing galaxy information: There is a surprisingly small amount of detail provided about the 14 sample objects. There is no information about their coordinates (which is non-standard compared to other recent JWST papers, e.g., Donnan et al. 2022, Naidu et al. 2022). There is no table listing the z_phot and z_phot uncertainty for each galaxy, its photometric measurements and all derived stellar population parameters using the preferred methodology. Such information is crucial to include, for readers and others working in the field, in the interests of testing and comparing results for the same public dataset. I also note that stellar masses are listed (and plotted) without error bars.

Response:
The referee is correct, of course. The lack of detail was a significant omission due to time pressure. A table with key measurements (coordinates, fluxes, masses, redshifts) had been prepared but did not make it into the submitted paper due to last minute compiling issues. We made essential information (including the coordinates) available right away on a website and added a link to the arxiv posting, but this obviously belongs in the main paper! In addition, a better description of the methodology, results, and uncertainties is necessary.
The full photometric measurements are now provided in Extended Data Table 1. Key information, including location, redshift, stellar mass, and corresponding uncertainties are now in Extended Data Table 2.

We intend to publish the template set on zenodo, with a link in the paper. We are happy to consider other ways to make supporting material public if the referees or Nature editors wish.

In the current revision of the paper we adopt the median of 4 different methods as our fiducial redshift and stellar mass (see response to #referee 1) and show a comparison of the masses, redshifts, and their uncertainties for all methods in Extended Data Figure 4. Unfortunately, there is no easy way to include the full parameter output for all models (or in the case of Prospector even for one!) in a human readable table. However, we have made available the full model output in fits catalog format on the website[2].

Redshift and stellar mass uncertainties for the sample are now also included in Figure 1 and 3.

Referee:
2. Details lacking regarding methodology: On a similar theme, I found the descriptions of methodology throughout the paper to be too brief, or entirely lacking. Here are some specific examples, where the paper should be expanded so that the results can be fully appreciated:
i. The paper adopts a two-color selection technique, plus non-detection at optical wavelengths. The two-color selection criteria should have been motivated by model tracks in

[2] https://github.com/ivolabbe/massive

color-color diagrams, showing how galaxies covering a range of different stellar population parameters (including a range of nebular emission strengths) and redshifts populate the space of F150W-F277W and F277-F444W.

Response

Perhaps surprisingly, but model tracks in color-color diagrams were never used to motivate the selection. As is common for first explorations of new observations, one of the sources was found serendipitously by inspecting a batch of sources with red observed F115W-F444W color. Closer examination of the SED showed a ‘double break’, suggesting that the source is high redshift at high confidence and potentially massive, but also that a simple color criterion could select galaxies with similar spectral shapes. With some trial and error we arrived at the selection and sample presented here. We note that we adjusted the color selection slightly to F150W – F277W< 0.7 to reselect the same galaxies after the NIRCam flux calibration update. While it is traditional to project selections of models onto color space, there is risk of hindsight bias if we now try to motivate it after the fact. Ultimately, the most relevant support is the result from full and flexible model fitting.

The real motivation is that in previous work, we employed a similar ad-hoc selection using wide-field HST and Spitzer imaging, by selecting bright Lyman Break galaxy candidates only on HST-optical non-detections and red Spitzer/IRAC 3.6-4.5 colors (Roberts-Borsani+2016) without ever resorting to traditional two-color diagrams. This approach turned out to be highly selective and successful (e.g., a sample of 4 bright z=7-9 galaxies out of 100k sources, with 100% spectroscopic confirmation).

The critical difference here is that the SEDs in this sample are very red, much redder than previously found with HST + Spitzer. This is illustrated by the rest-frame average SED of the galaxies, which is in a new panel that was added to Figure 3. Furthermore, we have now explored the effects of different priors and different fitting codes on the redshifts and masses, as explained in detail in the response to referee 1. An important aspect is the well-known relation between M/L ratio and color: red galaxies require high M/L ratios, irrespective of whether they are old or young+dusty. The main uncertainty is the contribution of emission lines to the broad band photometry, and we have now assessed that explicitly (see response to referee 1 and expanded Methods section).

Referee

ii. I found that the description of the derivation of stellar masses lacked clarity. Given the closer connection of the Prospector-alpha derived masses to a physical picture of associated star-formation histories, it was unclear to me why the EAZY stellar masses were adopted as "fiducial" (if in fact that was done). It was also unclear to me whether models with EAZY templates modified to include enhanced rest-optical emission-line fluxes were adopted as "fiducial" or were simply mentioned as an alternative. In particular, I wonder this for the 5 galaxies for which the enhanced emission-line template set provided a better fit (as described in the "Fit to the Photometry" section of "Methods"). A more complete discussion of the systematic uncertainties associated with individual stellar mass estimates (i.e., when allowing for significant nebular emission) is required. I mention this point in the context of model degeneracy between a strong Balmer Break and strong nebular emission.

Response:

We completely agree with the referee that the description of the derivation of the stellar masses was confusing. Initially, the main motivation to adopt the EAZY masses were ease of use, speed, and reproducibility. Running the full 42,000 sources of the entire catalogs takes only a few minutes, allowing to place the stellar masses in a broader context. We noticed that the default EAZY template set produced unreasonably high masses, therefore at high redshift (z>7) we only consider EAZY masses with the updated strong emission line template set.

In contrast, running Prospector is more involved and less well-defined, precisely because it has much more flexibility (non-parametric SFHs, age-dependent complex dust models, mass and SFH priors, etc). In addition, it is many orders of magnitude slower, preventing quick turnaround experimentation or updates when the photometry changes.

As described in detail in the response to referee #1 and in the revised Methods section, we significantly updated and expanded the model fitting, by constructing the new EAZY template set in a more principled way, describing (and using) the Prospector fitting results explicitly, and exploring model fits with the Bagpipes software (including a setup that strongly favors emission lines over Balmer breaks). Given the good correspondence between the methods, we now adopt as our "fiducial" masses and redshifts the median of the fitting results across the 4 methods. The methods section has been updated to describe the new way in which the masses were derived.

Referee
iii. Another key result in the paper is the stellar mass density contained in the massive systems presented here. Yet, there is no discussion included regarding the calculation of cosmic volume for these sources, and the manner in which cosmic variance is reflected in the number density error bars. When presenting cosmic densities, it is crucial to provide at least a minimum description of such calculations, at the very least in the "Methods" section.

Response:
This is indeed an obvious omission. We added the following description to the methods section.

"**Volume.** Stellar mass densities for galaxies with $M_* > 10^{10}$ $M_\odot$ are calculated by grouping the galaxies in two broad redshift bins ($7 < z < 8.5$ and $8.5 < z < 10$). At z~8.5 the Lyman Break moves through the F115W filter, allowing galaxies to be reliably separated into the two bins. The cosmic volume is estimated by integrating between the redshift limits of each bin over 38 sq arcmin, making no corrections for contamination or incompleteness. Any incompleteness would increase the derived stellar mass densities. The key result is driven by the 3 most massive galaxies (14294,35300,38094) which have well-determined redshifts with no P(z) solution at low redshift and a mean uncertainty σ(z)~0.25, much smaller than the width of each redshift bin. Therefore contamination should not be an issue. Cosmic variance is about 30%, calculated using a web calculator[63]. The error bars on the densities are the quadratic sum of the Poisson uncertainty and cosmic variance, with the Poisson error dominating."

Referee:
iv. It is stated that no galaxies with log(M/Msun)>10 have been found at z>8 in rest-UV-selected samples. I would have liked to see more discussion of this point. Specifically, the authors should clarify if this absence results because the rest-UV (i.e., observed near-IR photometry points) of the higher M/L galaxies presented here would be fainter than the pre-existing near-IR photometric limits. This is a case where it would be very useful to include the F150W magnitudes of the 7 sample galaxies with M>10^10 Msun, to demonstrate if they*

are fainter than current HST/F160W limits. For example, in a paper by one of the co-authors (Stefanon et al. 2021, ApJ, 922, 29, Table 1), data over ~10 arcmin^2 down to H_160~29 AB are analyzed. It seems that such limits should be sufficient for detecting galaxies like those presented here (and at least ~1-3 of them, given the 7 galaxies with log(M/Msun)>10 in the 40 arcmin^2 of CEERS). There is a comment at the end of the caption of 3 that the two most massive galaxies were previously mis-identified as being at z~1, but some further clarification about why log(M*/Msun)>10 galaxies were not recovered in previous high-redshift stellar mass function work would strengthen the paper.*

Response:
We agree with the referee that it is important to understand why these galaxies were not found with HST+Spitzer, and we now include in Table 1 the full photometry for the massive sources to facilitate a better comparison to literature. Roberts-Borsani et al. (2016) successfully selected bright *z*=7-9 galaxies using HST + red Spitzer colors in the same HST field (but outside the current JWST coverage), including a galaxy which was spectroscopically confirmed to be at *z*=8.68 (Zitrin et al. 2015). However, these sources were substantially brighter in the rest-frame UV (H160~25) and bluer (H-4.5~1.0). Two sources were present in published 3DHST catalogs, but placed at z~1. The reason for this difference is that the existing wide field F160W limits were comparatively shallow (H~27) and the large gap between F125W and F814W makes it difficult even for the two brightest sources (38094, with F150W=26.6 and 14924, with F150=27.6) to be reliably selected as *z*=7-9 dropouts.

Indeed, as the referee suspects, the massive sample is very faint with a median of F150W ~ 28 and F444W between 26-27 AB. Some sources similar to the ones in our sample could perhaps have been detected in the deeper parts of Stefanon et al. (2021), but that study was limited by the 2 magnitude shallower Spitzer/IRAC bands, and for faint galaxies relied on stacking analysis of Spitzer/IRAC in bins of (H160-based) rest-frame UV luminosity due to the overall low SNR in the Spitzer/IRAC bands.

We note in that regard, that, similar to what has been found at *z*=2-6 (e.g. Davidzon+2015) the most massive galaxies are not the most UV-luminous ones, and that the typical galaxy of the same UV luminosity is much bluer with lower M/L. In Stefanon+2021, galaxies of our sample would appear in their MUV=-19.5 bin (2^{nd} faintest bin) and their 4.5 micron fluxes would have been diluted with large numbers of blue low-mass galaxies.

We added to the main text:
"We note that these galaxies are relatively faint in the rest-frame UV (median F150W~28 AB), and previous wide-field studies with HST and Spitzer[36] of individual galaxies did not reach the required depths to find this population."

Referee
3. Lack of consideration of implications in the context of the dark matter halo mass function: One of the key results of the paper is that the observed stellar mass in galaxies with log(M/Msun)>10 is significantly higher at z~8 and z~10 than predicted by existing rest-UV-selected stellar mass functions. In fact, I would argue that the stellar mass density associated with the galaxies presented here is *uncomfortably* (perhaps unphysically) large. Specifically, (again referring to previous work by one of the co-authors), the stellar mass density shown in Figure 4 for z~10 galaxies with log(M*/Msun)>10 (~10^6 Msun/Mpc^3) is roughly equivalent to the *total* mass in dark matter halos down to 10^10.72 Msun at z~10*

(see Stefanon et al. 2021, Table 6). Given typical assumptions regarding the stellar-to-halo mass ratio, the galaxies presented here should only populate dark matter halos with M>10^12 Msun, with a significantly lower associated total mass density. The strong tension between the inferred stellar mass density and the total mass density in dark matter halos that can host massive galaxies must be considered here.

A related concern is whether a galaxy with M~9x10^10 Msun (14924, z=9.9) can be produced in a standard framework. According to Behroozi and Silk (2018, Figure 2), such massive galaxies should have space density of 10^-9 /Mpc^3 at z=10. This tension needs to be discussed as well.*

Response:
These are excellent point, which we upon submission had admittedly not fully considered. **Redacted.**

We would also note that in this revision, the mass and redshift of 14924 is now lower in both redshift and stellar mass (z=9.1 M*=3x10^10 Msun, due to improved photometry which showed a detection in F115W, lowering the redshift and the updated NIRCam zeropoints, which caused the SEDs to become bluer and the stellar mass to go down (see response to referee #1).

Referee:
One additional angle that should be considered is that the constant inferred stellar mass density in massive systems between z~8 and z~10 implies rapid evolution in the galaxy stellar-mass-to-halo-mass relation, given the rapid evolution in the dark-matter-halo mass functon(e.g., see Oesch et al. 2018, ApJ, 855, 105).

Response
We thank the referee for this important scientific angle to consider. With the updated stellar masses and redshifts the densities at z~9 are now lower than at z~8. Given the large uncertainties, however, it is difficult to draw firm conclusions about the evolution between z~8 and z~9. We intend to revisit these questions with future larger JWST samples.

MINOR COMMENTS

Referee:
1. Figure 1: I have a few comments on this figure. First, it should be stated which set of modeling assumptions (EAZY vs. Prospector-alpha, enhanced emission templates or not, etc.) is adopted. There should be error bars on stellar masses, and it should be clarified in the caption why two of the red stars are enlarged relative to the rest (is it simply the two most massive galaxies?). Also, it appears that, above z~8.5, the majority (7/10) of F444W-detected galaxies satisfy the two-color selection criteria. Is that expected?

Response:
We thank the referee for the useful suggestions! We updated the caption of Figure 1 to read:

"Shown in gray circles are EAZY-determined redshifts and stellar masses using emission-line enhanced templates (Salpeter IMF) for objects with S/N> 8 in the F444W band. Fiducial redshifts and masses of the bright galaxies (F444W < 27 AB) that satisfy our double-break selection are shown by the large red symbols."

Error bars are now included on the stellar masses and redshifts for the sample.

We noticed that in the original version of Figure 1 we did not show all S/N>8 galaxies in the catalog, but only those with F444W < 27AB (this information was missing in the caption). The revised Figure 1 now includes all S/N(F444W)>8 galaxies, including those with F444W > 27 AB).

Even so, it is certainly suggestive that amongst the galaxies with F444W < 27AB, most seem to have been picked up with the double break criteria. It could indicate a rest-frame U-V versus V color-magnitude similar to found at lower redshift, where optically brighter galaxies are redder, but at the same time have some residual UV flux due to unobscured sightlines. Complex dust geometries are common for star forming galaxies at low redshift and high redshift. But in all honesty, I think it is fair to admit that we were as surprised as anyone by the discovery of this sample.

Compared to a mass-selected sample, however, it should be noted that our requirements of two breaks (favoring a somewhat bluer UV) selects sources with slightly *lower* M/L than possible and redder, massive galaxies with fainter F444W > 27 AB could have been missed.

Referee:
2. p. 2, end of first paragraph on page. It is asserted that the most massive galaxy in the sample 38094 has two nearby companions with approximately the same redshift, "indicating that is in a group or proto-cluster." Given the likely uncertainties in photometric redshifts (which are not listed in the paper), nearby companions on the sky may not be physically associated. This statement should be softened.

Response:
The referee is correct that without a spectroscopic confirmation the nearby galaxies might just be projections, and that we have not measured redshifts.

We softened the language to:
"It has two nearby companions with a similar break in their optical to near-IR SEDs, suggesting that the galaxy may be in a group."

Referee:
3. Figure 3: I recommend showing panels for all 14 galaxies and making them all the same size. It is difficult to read the p(z) panels, given that they are immersed without borders within the rest of each SED panel. I recommend including the p(z) information as separate, adjacent panels. I recommend listing errors for both z_phot and M.*

Response:

We adopted the suggestions of the referee. It was difficult to accommodate all 14 galaxies with separate P(z) panels in a legible manner, so we opted to show separate SED panels for the 6 massive sources in the main text Figure 3 and have an additional Figure 3 in the Extended Data section for the remaining 7. We added a grid of lines at every dz=1 to the P(z) panels to improve legibility. Finally, we added uncertainties for redshift and stellar mass to the figure as well.

Referee:
4. Figure 4 and associated text in the paper. It is stated in the paper text that "the dashed lines assume errors of a factor of 3" – errors in what? There are curves in this figure labelled "S21", which I am assuming is for "Stefanon et al. 2021", but these labels are nowhere defined. In the caption, it is stated, "Dashed lines are derived from Monte Carlo simulations that perturb the masses of individual galaxies. The perturbations are drawn from a log-normal distribution with a 1-sigma width of 0.48 dex." Again, it is unclear which galaxies are being perturbed, and why the 1-sigma width of 0.48 dex was adopted. Some referencing and motivation is required here.

Response:
The label S21 is now defined in the caption, and we removed the perturbation as it is confusing and not well motivated. The curves are now simply the Schechter fits as derived in Stefanon et al. (2021). We note that S21 does not correct for the scatter in M/L, so a direct comparison is appropriate without further corrections.

Referee
5. I recommend one more round of proofreading for small typos.

Response

We performed more careful reading of the manuscript for typos and language.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I have read the revised manuscript "A very early onset of massive galaxy formation" and the authors' response to my comments/questions, and have to say I am very impressed by the effort they have invested to address virtually all the key points and to revise the manuscript accordingly. The level of detail on the analysis, and the overall quality of the text and figures is much improved. And it makes sense that the authors have taken the opportunity to reanalyse their data with the corrected ZPs (which, independently, I know they have done carefully and correctly).

In the spirit of openness it is good to see Extended Data Tables 1 and 2, and I appreciated the authors being open to revisiting the redshifts once the detections in F115W became more clear.

The expanded analysis comparing 4 different mass and redshift measurements is also very welcome, including the authors' adoption of my suggested "devil's advocate" strong emission line model.

So lots of positives, but I have one criticism and then a query which, unfortunately is a fundamental one.

My criticism is that I think the move to the Salpeter IMF is poorly justified, and, could be seen as an attempt to minimise the reduction in the masses for several of the objects between version 1 and version 2 of the manuscript. I strongly recommend the authors go back to the same IMF as before, and hence be more transparent about the stellar mass changes that have been produced by the more extended analysis.

My query is that I remain mystified as to why, given the range of galaxy models now explored, the masses change so little, or rather the range of masses is not greater. It is, for example, now known (e.g. Endsley et al. 2022) that some other authors find up to two orders of magnitude lower masses for some of these objects, usually as a result of fitting models which are very young, with strong emission lines (which can also influence the derived photometric redshift), and yet such solutions really don't seem to be present in Extended Data Figure 4. Now one can argue that very young models are unrealistic, and the authors make this point in their response, but they still say they haven't rejected them and that the masses stay high. This seems quite mysterious, as is the comment on page 5 of the authors' response that "new fits with younger ages require more reddening resulting in high M/L" - I really don't understand this comment.

I got a member of my research group to do their own fits to these objects, with corrected zero points, and while we agree that 38094 is the most massive galaxy in the sample, we find log(M*/Msun) = 10.5, and find all the other galaxies in the sample to have stellar masses below 10^10 Msun. We also like Endsley et al. find objects in this sample for which our derived stellar masses are up to two orders of magnitude lower than presented here. So I am left bemused, given the expanded nature of the analysis presented here, why the authors have apparently not found such solutions at all, even if they then might reasonably argue against their plausibility.

There is no escaping the importance of this, given the most celebrated result of this paper is the threat to the standard LambdaCDM cosmological model - with the masses derived by the researchers in my group there is no such threat, or at least it is much diminished.

So I urge the authors to explore why they have not found these much lower mass solutions. Two points might be worth checking.

1) has the mass and redshift degeneracy been allowed for properly in the results shown in Extended Data Figure 4 - i.e. could it be that the redshift has been locked at the median and then the range of masses derived for that single redshift? I don't think this is what has been done, but I thought it worth asking

2) In the use of Bagpipes has a linear or a logarithmic prior on age been uses - if linear, then I strongly suggest redoing with a logarithmic prior, because the linear prior the fraction of the prior at ages less than 10 Myr is only ~0.1% and young models are therefore disfavoured.

Bottom line - it is worth exploring the analysis a bit further to uncover these lower mass solutions (which clearly exist) and then the authors would be in a stronger position - they could still argue that some of these solutions are physically unrealistic, and could probably still defend the main results of the paper.

Oh, and please go back to the previous IMF.

Referee #2 (Remarks to the Author):

I have read the responses and revised manuscript for "A very early onset of massive galaxy formation" by Labbe et al. I appreciate the fact that the authors have gone to significant lengths to provide a more thorough analysis than the one contained in their initial submission. In particular, I commend the authors for describing their efforts to obtain robust photometric zeropoints and explaining their methodology in a level of detail that more closely meets the standard for publication. Overall, I am satisfied with the comprehensive responses from the authors in their reply to the referee reports, and the updates to the text, figures, and methodology.

However, I do have one significant remaining concern, which must be addressed before the paper can be accepted for publication:

The authors maintain that the stellar mass density in the sample of massive sources they have discovered is on order 10^6 Msun/Mpc^3 at z~8-9 -- even based on their more thorough methodology. In particular, the most surprising results pertain to the discovery of exceptionally massive galaxies, e.g., one at z=9.1 with log(M*/Msun)=10.5 (ID=14924) and one at z=7.5 with log(M*/Msun)=11.1 (ID=38094). Given that the result about the high stellar mass density is, according to the authors, "driven by the 3 most massive galaxies (14924, 35300, 38094)" the values of the stellar masses of these objects are of particular importance.

In this context, I would like to draw the authors' attention to a recent paper by Endsley et al. 2022 (https://ui.adsabs.harvard.edu/abs/2022arXiv220814999E/abstract), in which it is shown that the mass of ID 38094 (called 47624 in Endsley+2022) may be as low as log(M*/Msun)=8.64 (or 8.91, assuming a Salpeter IMF) -- i.e., *2 orders of magnitude* lower than claimed in this paper. Alternatively, Endsley+2022 arrive at best-fit model for 38094 that includes an AGN (supported by a point-source morphology) and is described by log(M*/Msun)~10. Labbe+2022 claim that the mass of ID 38094 is robust at log(M*/Msun)~11, and yet I feel that such (frankly extraordinary) claims regarding the high mass for this galaxy must address these alternate and seemingly just as robust claims for a significantly lower mass found by Endsley+2022. The alternate solutions demonstrated by Endsley+2022 for 38094 cause me to wonder about the robustness of the masses of, in particular, the two most massive z>9 galaxies (14924 and 35300), in addition to that of 38094.

Extending to the rest of the sample, I also note that Endsley+2022 find log(M*/M)=7.5 (or 7.8, assuming a Salpeter IMF) for ID 21834 (called 41858 in Endsley+2022), which is reported as having log(M*/Msun)=9.7 by Labbe+2022 -- a factor of 85 difference in mass. None of the 4 inferred stellar masses for 21834 in Extended Data Figure 4 come close to such a low value, which suggests at best that the systematic uncertainty in stellar mass has been severely underestimated by Labbe+2022 (at worst, that the stellar mass has been overestimated by ~two orders of magnitude).

The question of the masses of these galaxies will be resolved much more conclusively with spectroscopic observations. Such measurements will constrain both the strengths of the nebular emission lines that perturb photometric datapoints as well as hopefully the dynamical masses of the galaxies. Until then, I remain skeptical of the key result in the current paper regarding stellar masses whose very existence challenges a standard picture for galaxy formation within the LCDM framework.

Redacted.
Indeed, if the stellar masses of the galaxies presented here can be significantly lower (by 1-2 orders of magnitude), the tension with respect to LCDM is significantly abated. As I wrote in my previous report, I feel that the connection with the LCDM context belongs as a short discussion in the Labbe+2022 paper itself.

Minor comments:

1. The sample of galaxies changed from 14 to 13. While this does not need to be addressed in the revised paper itself, I am curious what happened to the sample. Did the updated photometric zeropoints result in one of the sample galaxies no longer satisfying the "double-break" criteria? Was this sample change only with respect to a single galaxy (i.e., one dropping out), or was it more complex?

2. The cosmic volume calculation (now elaborated) that factors into the inferred stellar mass densities seems extremely rudimentary. Is it possible to obtain at least a rough estimate of how

much completeness corrections would affect the inferred stellar mass densities plotted in Figure 4?

3. The statement "Galaxy 14924 is the most extreme object in terms of its mass and redshift" puzzled me, given the greater mass (but I guess lower redshift) of 38094. Is the point here that M*=3e+10 Msun at z=9.1 is more extreme than M*=1.5e+11 at z=7.5, in terms of the characteristic mass at each redshift? Perhaps a clarification here would help.

Author Rebuttals to First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I have read the revised manuscript "A very early onset of massive galaxy formation" and the authors' response to my comments/questions, and have to say I am very impressed by the effort they have invested to address virtually all the key points and to revise the manuscript accordingly. The level of detail on the analysis, and the overall quality of the text and figures is much improved. And it makes sense that the authors have taken the opportunity to reanalyse their data with the corrected ZPs (which, independently, I know they have done carefully and correctly).

In the spirit of openness it is good to see Extended Data Tables 1 and 2, and I appreciated the authors being open to revisiting the redshifts once the detections in F115W became more clear.

The expanded analysis comparing 4 different mass and redshift measurements is also very welcome, including the authors' adoption of my suggested "devil's advocate" strong emission line model.

We thank the referee for these remarks!

So lots of positives, but I have one criticism and then a query which, unfortunately is a fundamental one.

My criticism is that I think the move to the Salpeter IMF is poorly justified, and, could be seen as an attempt to minimise the reduction in the masses for several of the objects between version 1 and version 2 of the manuscript. I strongly recommend the authors go back to the same IMF as before, and hence be more transparent about the stellar mass changes that have been produced by the more extended analysis.

We realize that the move to a Salpeter IMF may come across as an attempt to minimize the differences. However, it was a genuine mistake; in fact, in the original manuscript the comparison of our data with the previous mass functions was flawed as two different IMFs were used. The previous studies in this area used a Salpeter IMF, and we used a Chabrier IMF. We are now using the default IMF of previous mass function studies, so that our mass densities can be directly compared to the literature, and would prefer to continue doing that rather than adjust the literature data.

More to the point, in response to the new referee reports we changed the tone and language in the paper completely, acknowledging that there are major unresolved uncertainties associated with the masses, as explained below. These uncertainties far exceed the factor of ~1.6 uncertainty introduced by the IMF.

My query is that I remain mystified as to why, given the range of galaxy models now explored, the masses change so little, or rather the range of masses is not greater. It is, for example, now known (e.g. Endsley et al. 2022) that some other authors find up to two orders of magnitude lower masses for some of these objects, usually as a result of fitting models which are very young, with strong emission lines (which can also influence the derived photometric redshift), and yet such solutions really don't seem to be present in Extended Data Figure 4. Now one can argue that very young models are unrealistic, and the authors make this point in their response, but they still say they haven't rejected them and that the masses stay high. This seems quite mysterious, as is the comment on page 5 of the authors' response that "new fits with younger ages require more reddening resulting in high M/L" - I really don't understand this comment.

I got a member of my research group to do their own fits to these objects, with corrected zero points, and while we agree that 38094 is the most massive galaxy in the sample, we find log(M*/Msun) = 10.5, and find all the other galaxies in the sample to have stellar masses below 10^10 Msun. We also like Endsley et al. find objects in this sample for which our derived stellar masses are up to two orders of magnitude lower than presented here. So I am left bemused, given the expanded nature of the analysis presented here, why the authors have apparently not found such solutions at all, even if they then might reasonably argue against their plausibility.

There is no escaping the importance of this, given the most celebrated result of this paper is the threat to the standard LambdaCDM cosmological model - with the masses derived by the researchers in my group there is no such threat, or at least it is much diminished.

We investigated the differences with Endsley and find that we are able to reproduce the low mass solutions for 38094 (their 47624) (log M*/Msun = 8.9 (Salp) +/- 0 at z=6.93 +/- 0.01) and also 21834 (their 41858).

The origin of the difference is a combination of 1) assuming an SMC extinction law, 2) an age prior weighted towards very young ages, and 3) high S/N ratio in NIRCam long wavelength filters (which puts all the weight in the fits there). Together these aspects can produce lower redshifts and lower masses by up to factors of 100, albeit at specific, almost discrete redshifts. We explain these aspects below, and also in the revised paper.

We also modified and added figures: we added to Extended Data Figure 4 additional Bagpipes fits that highlight the impact of varying these three modeling choices. In Extended Data Figure 3, we also contrast the high-mass continuum fit with the low-mass line-dominated solutions to illustrate the quality of the fits and highlight differences. In the new Extended Data Figure 5 we show that at specific redshifts line-dominated and continuum-dominated SEDs become nearly indistinguishable due to aliasing. In the new Extended Data Figure 6 we show that the line-dominated fits are preferentially placed in these narrow aliased redshift windows.

Dust: For the reddest, most massive sources in our sample, the adoption of an SMC-type extinction law seems to be critical for finding low-mass solutions. The reason is that SMC laws have a strong wavelength dependence (steep in the FUV, flatter in optical). However, these fits do not reproduce the overall shape of the SEDs in our sample well. This can be clearly seen in Extended Data Figure 3, where the SMC based fits have strongly “curved” continuum, which are generally too steep in the rest-UV and too flat in the rest-optical (F356W,F410W,F444W bands). When combined with an age sampling that emphasizes very low ages (and disfavors e.g. Balmer breaks) this leads to solutions with ages of a few Myr at specific redshifts.

SMC-type attenuation is generally not thought to be appropriate for these rest-frame optically red galaxies. Empirical studies (e.g., Salmon et al. 2016 and Salim et al. 2018), have shown that larger optical depths and optical reddening correspond to grayer (shallower) attenuation curves. This effect is expected theoretically (e.g. Gordon+2000, Chevallard+2013) and seen in simulations (Narayanan+2018). The low optical depth regime tends to produce steep (SMC-like) attenuation curves due to wavelength dependent scattering, whereas optically thicker geometries produce flatter (Calzetti-like) dust laws.

Note that for our Bagpipes fits we previously used Calzetti, but in this revision we switch to a Salim+2018 attenuation law, which accounts for the above correlation. A Salim+2018 law is steep (SMC-like) at low Av and flat (Calzetti-like) at high Av. In our Prospector fits steepness of the attenuation law is a free parameter and marginalized over. All galaxies are fit with high Av=1-2.5 using either the Salim et al (2018) law or a free dust slope, indicating that a flatter slope is more appropriate than SMC for this sample.

Age priors: For the bluer galaxies of the sample, lower mass solutions can be found with Bagpipes regardless of the assumed dust law if logarithmic age priors are assumed (see Extended Data Figure Table 4). Switching to linear age priors produces high mass solutions, indicating that for these galaxies the results depend on assumptions about priors.

Prospector uses age priors that track the evolution of the global CSFRD, favoring young ages, which should theoretically be the best-informed prior (certainly better than either logarithmic or linear, which are both extremes). Prospector finds exclusively high mass solutions, and ages which appear reasonable for the redshift. The Prospector (mass weighted) ages range from 16 Myr (39575) to 190 Myr (for 38094) with a median of 34 Myr for the whole sample.

Effect of a high SNR in some bands: Since JWST is still in early days of calibration, it is useful to see what happens if an error floor of 10% is imposed on the photometry. In that case we find high mass solutions for all but 1 galaxy with Bagpipes and a Salim+18 law, regardless of a logarithmic vs linear age prior. Comparing to the best-fits models without a noise floor indicates that very high SNR in the NIRCam red filters can drive solutions that effectively ignore the blue part of the SED (and e.g., the presence and location of a Lyman break), leading to inconsistent redshifts between the red and the blue part of the SED. Calibration uncertainties is what motivates us to adopt a 5% error floor in all model fits.

While low mass ultra young models cannot be firmly excluded with the currently available data, there are four aspects of these solutions that are worth noting.

1. Ages. While ages <5 Myr are formally allowed, they are not plausible in any scenario, as the galaxy would not be causally-connected - 10^9 Msun of star formation would have started spontaneously on timescales less than a dynamical time. In addition, the probability of catching an object at that precise moment is low - given the ~300 Myr search window at z=7-10. It would suggest there are >60 older and more massive galaxies for every galaxy in our sample!

2. Redshifts. The P(z) of the young, line-dominated solutions are extremely narrow and clustered at nearly discrete redshifts. As can be seen in the Endsley fit for 38094 the formal uncertainty on the redshift P(z) is vanishingly small (z=6.93 +/- 0.01 !), as is the reported formal error on age and mass (0).

The reason is that line-dominated models can only reproduce the smooth red SEDs if a strong line is placed *exactly* in the narrow overlap at the edges of F356W, F410M, F444W filters at 3.9-4.0 micron (e.g., [OIII] at z=6.9). At this redshift one line boosts all three filters, similar to a continuum dominated SED, with great flexibility in colors due to the rapidly varying transmission at the filter edge. The only redshifts where this occurs are z=5.6, 6.9, and 7.7, corresponding to [OIII] or Hα entering F410M and F444W and moving out of F356W and F410M (see Extended Data Figure 5).

While finding one 5 Myr galaxy exactly in this dz=0.01 window could be luck, using the Endsley+22 assumptions, we find 10/13 galaxies in our sample are fit with ages <5 Myr and only at these discrete redshifts z=5.6, 6.9, 7.7 (see Methods and extended Data Figure 4).

Such an age and P(z) distribution for the sample, at precisely the redshifts where this fortuitous overlap between filters occurs, is obviously highly implausible. For our sample the 10 ultra young galaxies are placed in ~< 8% of the redshift range between z=5-9 (see Fig 3). Such a distribution would not be expected for an F277W-F444W selected sample, *even if the entire sample were emission line dominated,* because [OIII] and/or Ha contribute to F444W over 95% of the redshift range z=5-9 (see Extended Data Figure 6). So for every galaxy at z=5.6, 6.9, 7.7 we would expect ~13 galaxies at redshifts that are in between.

A more likely explanation is that this extreme distribution of age and redshift is the result of a fitting artefact, due to undersampling of the SED by the filters and resulting aliasing. The aliasing is more pronounced if extreme emission lines are favored, which appears to be the case when assuming SMC dust and age sampling favoring extremely young ages.

3. Treatment of uncertainties. In 9/13 cases we find considerably lower chi-2 for continuum dominated solutions than line-dominated solutions. For example, for 38094 the best-fit SMC / young model is chi2=110 versus chi2=23 for the continuum model. This is not seen by Endsley+22 who find chi~60 for both line and continuum dominated fits.

One possible reason for the difference with Endsley is that we adopt a 5% floor to account for calibration errors, which penalizes a large discrepancy in a single filter more than small discrepancies in many filters. This is especially important because the SNR in NIRCam LW can reach SNR>100 (300 for 38094!). Without the inclusion of an error floor, the fits can effectively ignore the lower SNR blue part of the SED. The Endsley paper does not specify if they assumed an error floor, but this could make a significant difference in the relative chi-2 between continuum and line-drive fits. It is remarkable that despite the highly constrained redshift, the line-dominated solution does not reproduce the F410M-F444W color, both here and in Endsley, even though the F410M-F444W color is an excellent tracer of emission lines.

Another possible reason why Endsley did not find a difference in chi2 for continuum and line fits could be that the Endsley+22 pre-print results are affected by calibration errors. Inspection of the NIRCamPhotoCal_MRieke_Aug22.txt zeropoint file used for their paper shows offsets of up to 0.2 mag across the NIRCam SW+LW bands compared to the latest STScI calibrations. We note that our calibrations are consistent with the latest STScI calibrations to within 0.03 mag.

4. Turning to AGN, it seems generically implausible that AGN would be prevalent in low mass galaxies as the ratio of the black hole mass to the galaxy mass would likely exceed standard values by a huge factor.

More to the point, while some sources may appear like point sources on account of the visible PSF diffraction pattern, a recent paper by Finkelstein+22 and our own analysis (Baggen et al. in prep) of this sample both indicate resolved sizes of ~0.5 kpc (~2 pixels) in F444W, which effectively rules out an AGN being dominant at that wavelength.

However, given the stakes (challenging the well-established LCDM model) we agree that we cannot simply discount such models. We therefore changed the tone and language of the paper to emphasize that the galaxies we find are candidates that need to be confirmed with spectroscopy. We now state that we are finding a significant population of red galaxies at very high redshift z=7-9, but that the interpretation is uncertain. We hope that this change in emphasis, the investigation of the differences with Endsley, and explicit acknowledgement of possibly large systematic errors in the masses of individual objects addresses the concerns of the referee.

So I urge the authors to explore why they have not found these much lower mass solutions. Two points might be worth checking.

1) has the mass and redshift degeneracy been allowed for properly in the results shown in Extended Data Figure 4 - i.e. could it be that the redshift has been locked at the median and then the range of masses derived for that single redshift? I don't think this is what has been done, but I thought it worth asking

All uncertainties of Bagpipes and Prospector are derived from Bayesian sampling of the posterior and therefore reflect the mass redshift degeneracy for a given model.

However, we show above and in Methods that the difference with Endsley is related to model choices themselves (e.g., attenuation law), choice of age priors, and treatment of uncertainties, which are not necessarily reflected in the mass-redshift sampling of the posterior.

The discrepancies are therefore systematic and driven by modeling assumptions.

2) In the use of Bagpipes has a linear or a logarithmic prior on age been uses - if linear, then I strongly suggest redoing with a logarithmic prior, because the linear prior the fraction of the prior at ages less than 10 Myr is only ~0.1% and young models are therefore disfavoured.

Yes, that is an excellent point. The Bagpipes fits indeed used a linear prior in age. We now also include bagpipes fits with logarithmic sampling (and both SMC and Salim+18 dust) to highlight which galaxies are sensitive to these assumptions.

The Prospector fits use a piecewise-constant ‘nonparametric’ SFH, and a central expectation following the cosmic SFRD (with modest adjustments depending on galaxy mass, Wang et al. submitted). This amounts to rapidly rising SFR at early times, so weighted strongly towards young ages but in a way informed by previous observations of CSFRD. Prospector adopts a continuity prior (Leja+19a) disfavoring rapid changes in the SFH.

Bottom line - it is worth exploring the analysis a bit further to uncover these lower mass solutions (which clearly exist) and then the authors would be in a stronger position - they could still argue that some of these solutions are physically unrealistic, and could probably still defend the main results of the paper.

We thank the referee for the suggestions. As explained above, we investigated various model assumptions, finding that a combination of SMC dust law, age sampling favoring low ages, and unrestricted uncertainties can produce very young very low mass solutions.

We think there are good arguments based on the age, P(z) distributions, uncertainties (chi2, error floor) and choice of dust law that disfavor these solutions, and that continuum dominated, high-mass solutions are far more plausible for the sample presented here. Nevertheless, we changed the tone of the paper to emphasize that the masses are highly uncertain.

Oh, and please go back to the previous IMF.

See above; the Salpeter IMF is the assumption in the literature we are comparing to so we prefer to keep it.

Referee #2 (Remarks to the Author):

I have read the responses and revised manuscript for "A very early onset of massive galaxy formation" by Labbe et al. I appreciate the fact that the authors have gone to significant lengths to provide a more thorough analysis than the one contained in their initial submission. In particular, I commend the authors for describing their efforts to obtain robust photometric zeropoints and explaining their methodology in a level of detail that more closely meets the standard for publication. Overall, I am satisfied with the comprehensive responses from the authors in their reply to the referee reports, and the updates to the text, figures, and methodology.

We thank the referee for these remarks.

However, I do have one significant remaining concern, which must be addressed before the paper can be accepted for publication:

The authors maintain that the stellar mass density in the sample of massive sources they have discovered is on order 10^6 Msun/Mpc^3 at z~8-9 -- even based on their more thorough methodology. In particular, the most surprising results pertain to the discovery of exceptionally massive galaxies, e.g., one at z=9.1 with log(M*/Msun)=10.5 (ID=14924) and one at z=7.5 with log(M*/Msun)=11.1 (ID=38094). Given that the result about the high stellar mass density is, according to the authors, "driven by the 3 most massive galaxies (14924, 35300, 38094)" the values of the stellar masses of these objects are of particular importance.

In this context, I would like to draw the authors' attention to a recent paper by Endsley et al. 2022 (https://ui.adsabs.harvard.edu/abs/2022arXiv220814999E/abstract), in which it is shown that the mass of ID 38094 (called 47624 in Endsley+2022) may be as low as log(M*/Msun)=8.64 (or 8.91, assuming a Salpeter IMF) -- i.e., *2 orders of magnitude* lower than claimed in this paper. Alternatively, Endsley+2022 arrive at best-fit model for 38094 that includes an AGN (supported by a point-source morphology) and is described by log(M*/Msun)~10. Labbe+2022 claim that the mass of ID 38094 is robust at log(M*/Msun)~11, and yet I feel that such (frankly extraordinary) claims regarding the high mass for this galaxy must address these alternate and seemingly just as robust claims for a significantly lower mass found by Endsley+2022. The alternate solutions demonstrated by Endsley+2022 for 38094 cause me to wonder about the robustness of the masses of, in particular, the two most massive z>9 galaxies (14924 and 35300), in addition to that of 38094.

Extending to the rest of the sample, I also note that Endsley+2022 find log(M*/M)=7.5 (or 7.8, assuming a Salpeter IMF) for ID 21834 (called 41858 in Endsley+2022), which is reported as having log(M*/Msun)=9.7 by Labbe+2022 -- a factor of 85 difference in mass. None of the 4 inferred stellar masses for 21834 in Extended Data Figure 4 come close to such a low value, which suggests at best that the systematic uncertainty in stellar mass has been severely underestimated by Labbe+2022 (at worst, that the stellar mass has been overestimated by ~two orders of magnitude).

The question of the masses of these galaxies will be resolved much more conclusively with spectroscopic observations. Such measurements will constrain both the strengths of the nebular emission lines that perturb photometric datapoints as well as hopefully the dynamical masses of the galaxies. Until then, I remain skeptical of the key result in the current paper regarding stellar masses whose very existence challenges a standard picture for galaxy formation within the LCDM framework.

We agree, and we have made significant changes to the paper in response. First, we expanded the analysis, are able to reproduce and explain the Endsley result, and can trace back the low mass solutions to specific assumptions for dust attenuation law, age sampling, and treatment of uncertainties, as described in detail to the response to referee #1. We agree that these assumptions represent a source of systematic uncertainty that is not reflected in the random uncertainties of the models

We do find that Endsley+22 fit to 38094 and fits to our broader sample using their modeling assumptions (SMC dust), have some unusual qualities, and the distributions on the whole seem implausible: extremely young ages at a few narrow, near discrete, redshifts spikes, often with higher chi2 than continuum dominated solutions, and/or disagreeing with the location of the Lyman-break. In the case of AGN, the Endsley preprint did not include a quantitive analysis of sizes, only that the source looks like a point source. However, JWST is very deep and diffraction patterns are seen for sources that are small but resolved. In our own analysis we find that the sample is resolved with median (~0.5 kpc ~2pix), as does work by Finkelstein+22 for an overlapping subsample. Finally, the Endsley modeling as posted on the arxiv is based on preliminary NIRCam zeropoints with known offsets of up to 0.2 mag. It is unclear if they assumed an error floor, without which calibration uncertainties can have significant impact on the fits due to the very high SNR of NIRCam long wavelength photometry.

Nevertheless, we do acknowledge that many of these arguments are circumstantial and ultimately this question can only be fully resolved with spectroscopy.

We now state that the masses are uncertain, as they may suffer from known (exotic lines, AGN) or unknown errors. Following the referee's suggestion we now say in the paper that spectroscopy is needed to verify the redshifts and the masses.

As explained in our response to referee 1, we think we understand the differences with Endsley, but we are not arguing that their masses are conclusively wrong in all cases and ours are right. The key result of our study is that a population of galaxies with red rest-frame optical colors can be found at high redshift with JWST. And using standard tools and reasonable assumptions, high M/L are found. We now say that IF the red colors are caused by high M/L ratios then the implications are profound for galaxy formation.. However, this interpretation can only be tested with follow-up spectroscopy.

Redacted. Indeed, if the stellar masses of the galaxies presented here can

be significantly lower (by 1-2 orders of magnitude), the tension with respect to LCDM is significantly abated. As I wrote in my previous report, I feel that the connection with the LCDM context belongs as a short discussion in the Labbe+2022 paper itself.

We follow the referee's suggestion and in the revised version we refer to the two papers that have appeared on the arXiv about this.

Minor comments:

1. The sample of galaxies changed from 14 to 13. While this does not need to be addressed in the revised paper itself, I am curious what happened to the sample. Did the updated photometric zeropoints result in one of the sample galaxies no longer satisfying the "double-break" criteria? Was this sample change only with respect to a single galaxy (i.e., one dropping out), or was it more complex?

The photometry underwent significant changes due to (location-dependent) zeropoints changes, and to a lesser extent improvements in processing (1/f noise), which changed the noise properties (causing galaxies to scatter within the uncertainties).

As a result the colors of all galaxies changed modestly and the original color-selection produced a (smaller) sub-sample. To retain as many galaxies as possible to the original sample and analysis, we modified the F150W-F277W cut to 0.7 (from 0.5). This recovered 13 out of the 14 original galaxies. To reproduce the sample perfectly is not possible, because the zeropoint changes were detector dependent, therefore different for each source.

2. The cosmic volume calculation (now elaborated) that factors into the inferred stellar mass densities seems extremely rudimentary. Is it possible to obtain at least a rough estimate of how much completeness corrections would affect the inferred stellar mass densities plotted in Figure 4?

We agree that the volume calculation is simplistic. However, traditional completeness corrections require strong assumptions of the underlying mass and SED distributions of the galaxies. We did however performed a simple simulation to estimate a selection function (Methods and Extended Data Figure 6) assuming all galaxies look like our sample. The typical selection efficiency is ~50% between z=7-9, but depends on the assumed SEDs. However, such a refined treatment does not seem warranted given that the main (orders of magnitude) uncertainty in our study is the interpretation of the red colours of the galaxies (redshifts, M/L).

3. The statement "Galaxy 14924 is the most extreme object in terms of its mass and redshift" puzzled me, given the greater mass (but I guess lower redshift) of 38094. Is the point here that M*=3e+10 Msun at z=9.1 is more extreme than M*=1.5e+11 at z=7.5, in terms of the characteristic mass at each redshift? Perhaps a clarification here would help.

We thank the referee for highlighting this. The high mass end of the dark matter halo mass function evolves very rapidly between z = 7 - 9, so the cumulative comoving number density corresponding to an observed *M** is lower for M*=3e10 Msun at z=9.1 than for M=1e11 at 7.5.

The differences are small though, and both could be considered equally extreme within the (likely underestimated) stellar mass uncertainties. After including more methods for deriving mass, the median masses changed slightly so 35300 and 14924 traded places.

We therefore changed the text to read “Galaxy 35300 is the most extreme object in terms of its redshift”

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have made a major effort to revise this manuscript, and have properly investigated all the issues raised in my referee report. I also appreciate the conscientious and thoughtfully detailed response letter.

On balance I think this represents a proper and careful revision of the manuscript extending (sensibly and appropriately) to a revised title and changes to the abstract, which is actually quite refreshing to see.

In my view this is a much better, more impressive, and still very interesting paper which can still lay claim to a result worthy of publication in Nature but now carries all the appropriate caveats (many of which are interesting, including the very narrow redshift intervals required for the low-mass solutions, which can reasonably be used to push back against their plausibility).

I would now be happy to see this published in Nature.

Referee #2 (Remarks to the Author):

I have read the latest round of responses and revised manuscript for the paper now entitled, "A population of red, candidate massive galaxies 500-700 Myr after the Big Bang." The authors have gone to even more lengths to address the second round of referee reports, and have significantly toned down some of the claims in the paper. Specifically, they articulate which results are more robust (the colors of their galaxies) and which require further, spectroscopic confirmation (the masses and cosmological implications of their galaxies).

The authors have now considered low-mass solutions in detail, which I appreciate. However, one argument that I don't fully agree with is the authors' dismissal of the young (~5 Myr) ages associated with low-mass solutions. Indeed, they argue that such ages are implausible as they are shorter than the dynamical timescales of the sample galaxies. However, the authors haven't actually measured the dynamical timescales of the galaxies. There are size estimates, but of course no velocity (e.g., dispersion) measurements, which are required for an estimate of t_dyn~size/v. Given typical sizes of <1 kpc, plausible velocity dispersions could yield dynamical times as short as 5-10 Myr.

However, I don't see the need to debate things further and feel that the paper is ready for publication. I leave it to the editor to decide whether the paper, in its revised, more empirically focused, state, belongs in Nature or Nature Astronomy.

Author Rebuttals to Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have made a major effort to revise this manuscript, and have properly investigated all the issues raised in my referee report. I also appreciate the conscientious and thoughtfully detailed response letter.

On balance I think this represents a proper and careful revision of the manuscript extending (sensibly and appropriately) to a revised title and changes to the abstract, which is actually quite refreshing to see.

In my view this is a much better, more impressive, and still very interesting paper which can still lay claim to a result worthy of publication in Nature but now carries all the appropriate caveats (many of which are interesting, including the very narrow redshift intervals required for the low-mass solutions, which can reasonably be used to push back against their plausibility).

I would now be happy to see this published in Nature.

We thank the referee for the remarks!

Referee #2 (Remarks to the Author):

I have read the latest round of responses and revised manuscript for the paper now entitled, "A population of red, candidate massive galaxies 500-700 Myr after the Big Bang." The authors have gone to even more lengths to address the second round of referee reports, and have significantly toned down some of the claims in the paper. Specifically, they articulate which results are more robust (the colors of their galaxies) and which require further, spectroscopic confirmation (the masses and cosmological implications of their galaxies).

The authors have now considered low-mass solutions in detail, which I appreciate. However, one argument that I don't fully agree with is the authors' dismissal of the young (~5 Myr) ages associated with low-mass solutions. Indeed, they argue that such ages are implausible as they are shorter than the dynamical timescales of the sample galaxies. However, the authors haven't actually measured the dynamical timescales of the galaxies. There are size estimates, but of course no velocity (e.g., dispersion) measurements, which are required for an estimate of t_dyn~size/v. Given typical sizes of <1 kpc, plausible velocity dispersions could yield dynamical times as short as 5-10 Myr.

However, I don't see the need to debate things further and feel that the paper is ready for publication. I leave it to the editor to decide whether the paper, in its revised, more empirically focused, state, belongs in Nature or Nature Astronomy.

Referee 2 is correct that without measured velocity dispersions, and accurate corresponding sizes, dynamical time are uncertain. We therefore reflect the caveat of the referee in the text in Methods:

Added:
$10^{8.5} M_{\odot}$ of star formation would have started spontaneously on timescales less than a dynamical time (although dynamical times are uncertain until velocity dispersions and corresponding sizes are measured).