Peer Review File

Manuscript Title: Accelerated Formation of Ultra-Massive Galaxies in the First Billion Years

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

This is a very nice result.
The paper is is extremely well written.
This work is absolutely deserving of being published in Nature.

I have no comments on the text. I tried to find things, but couldn't.
Sincere apologies for taking so long with this review.

My only question is about Figure 1. It was not clear to me how the Halpha line map was constructed.
Could you please describe how these were constructed in the appendix ?

Referee #2 (Remarks to the Author):

1. The manuscript is aimed at "massive optical dark galaxies" - this is a topic that has been beaten to the dead horse in Nature, or many
other journals. There is a clear progression in increasing z of the galaxy samples or the topic discussed, but the overall themes,
results, conclusions remain the same. Back in mid 1990s, there is the paper by Guiderdoni et al. by Nature, then SCUBA, Herschel, other
ground mm-wave variations, including ALMA, in many papers in Nature, then Wang et al. around z of 3-5, then we are here now at z of 5-8.

2. The galaxy sample the authors discuss is important, but the present manuscript does not do justice to the data they have obtained.
While they mention 36 galaxies in the first paragraph, the paper is focussed on the three brightest galaxies. Unfortunately, 2 of these 3
galaxies have already been the subject of other papers, including in Nature. The third in this paper is sort of new.
The brightest is a single emission line detection. In this day and age with JWST and ALMA, it is a bit of an embarrassment to the galaxy formation community as a whole to read a Nature submission where valuable page space is spent on modeling arguments related to the redshift of a galaxy. If the team thinks the single line z with H-alpha is correct, then it wouldn't be that hard to confirm with a second line. The capability to do that naturally exists within JWST.

The sample also includes well-known galaxies like HDF850.1, a struggle going back to early days of SCUBA and Hughes et al. Nature paper in 1998, but finally resolved in Walter+ 2012, again in Nature.

Instead of concentrating on three brightest, whose modeling and other uncertainties can negatively impact the conclusions, the paper should really be
about the full sample. Fig 2 has some of this but details are not clear. Also as Fig 2 shows, those three galaxies are not really ruled out
by LCDM. And it is also not just those 3 galaxies, there are others from before. One could also naturally argue why select these three, when
at z=7-8 they seem to have two equally interesting galaxies? Looking at this figure, I am afraid the point of this whole paper is totally unclear to me.

3. It is a good useful result that with JWST Halpha one can get Halpha SF. The paper would have a stronger impact if the manuscript had
discussed all properties of the sample and not concentrate on three of the brightest. There is a mention of CIGALE SFs, including all of
long wavelength data, but no SEDs are shown and one cannot make any conclusions on those SFs.

4. Fig 3 left analysis seems a little bit disjoint with Fig 2 left. Fig 2 caption says efficiency 50% is ok; Fig 3 left would imply an issue with
even efficiency 20% that for S1 is a factor of x1000. This is not captured in Fig 2 left, but I believe that plot. Arguments similar to Fig 3 left
- this is not the first paper to try and make up, there are many related to more massive things like galaxy clusters - have not lived up to the
test of time.

5. Structural parameters are probably useful, but again I am unclear what is the information content here. Fig 3 right shows one of the galaxies with ALMA dust. Instead of some rest optical structures, there is probably a whole lot more to be gained by a combined JWST, ALMA study, apart from the fact
that dust is extended.

Author Rebuttals to Initial Comments:

We are very grateful to the referee for the very helpful report. We revised and improved the manuscript in red by considering all suggestions and comments from the referee's report. Detailed explanations are given just below each comment or suggestion.

Referee #2 (Remarks to the Author):
1. The manuscript is aimed at "massive optical dark galaxies" - this is a topic that has been beaten to the dead horse in Nature, or many other journals. There is a clear progression in increasing z of the galaxy samples or the topic discussed, but the overall themes, results, conclusions remain the same. Back in mid 1990s, there is the paper by Guiderdoni et al. by Nature, then SCUBA, Herschel, other ground mm-wave variations, including ALMA, in many papers in Nature, then Wang et al. around z of 3-5, then we are here now at z of 5-8.

Answer: Thanks to the Referee for pointing out the concerns. The report made us realize that we failed to present the result in a sufficiently focused and clear way that is necessary for a Nature paper. We realize that the wording of our introduction gave too much weight to the red and dusty property of distant massive galaxies, which is not the main point of this work.

In fact, the key point of our paper is the existence of spectroscopically confirmed extremely massive galaxies in the early Universe and its implications for existing models of galaxy formation, mainly through the efficiency of the conversion of baryons into stars. With their masses of ~10^11 Msun at zspec=5-6, these galaxies fall in the same "difficult to explain" region of the redshift - mass plane as the ~10^10 Msun, z=7-8 galaxies of Labbe et al (2023, Nature, 616, 266), but - and this is the key point of this paper - they are spectroscopically confirmed and have more reliable masses (owing to their slightly lower redshifts and the fact that we can rule out significant AGN contributions).

In the broader context, the discovery of unexpectedly massive galaxies in the early universe is potentially one of the most important results so far from JWST in observational cosmology, but there is a lot of understandable skepticism. With these redshift measurements, we have addressed the most important uncertainty in this result. We note that the galaxies presented in this article do not challenge the LCDM cosmological framework, contrary to what some studies suggested in the early months of JWST exploitation. Our study is concerned only with existing models of galaxy formation, which need to be revised to explain such a high efficiency in converting baryons into stars.

We have now revised the abstract and the following paragraphs to make the paper more focused on the science of ultra-massive galaxies.

2. The galaxy sample the authors discuss is important, but the present manuscript does not do justice to the data they have obtained.
While they mention 36 galaxies in the first paragraph, the paper is focussed on the three brightest galaxies. Unfortunately, 2 of these 3 galaxies have already been the subject of other papers, including in Nature. The third in this paper is sort of new.
The brightest is a single emission line detection. In this day and age with JWST and ALMA, it is a bit of an embarrassment to the galaxy formation community as a whole to read a Nature submission where valuable page space is spent on modeling arguments related to the redshift of a galaxy. If the team thinks the single line z with H-alpha is correct, then it wouldn't be that hard to confirm with a second line. The capability to do that naturally exists within JWST.

Answer: The key point of our paper is to discuss the implications of spectroscopically confirmed extremely massive galaxies that are located in a redshift range that guarantees robust stellar determination. Therefore, we focus on the three ultra-massive galaxies throughout the whole paper. The 36 galaxies are the parent sample based on our sample selection.

None of the 3 galaxies has been published in Nature. S1 was briefly mentioned in a few previous papers (e.g., Xiao+23), but its redshift and stellar mass were unknown. S2 is a submillimetre galaxy GN10, that was pointed out previously (e.g., Riechers et al. 2020, 895, 81, APJ). S3 is a completely new and previously unknown source.

S1 indeed has a single line detected, but its zphot = 5.61 agrees perfectly with zspec = 5.58, we dedicate a section to discuss the reliability of its redshift in the Supplementary Materials. The emission line of S1 is about 14 sigma and is not bright enough to ensure that we can detect a second line. The non-detection of a second line indicates the possible low metallicity in the Halpha emitting region of this galaxy, which needs to be confirmed with follow-up NIRSpec observations.

The sample also includes well-known galaxies like HDF850.1, a struggle going back to early days of SCUBA and Hughes et al. Nature paper in 1998, but finally resolved in Walter+ 2012, again in Nature.

Answer: HDF850.1 is an optically dark galaxy in the GOODS-N field, therefore entered our catalog through sample selection. But HDF850.1 is not part of the sample with our three main sources.

Instead of concentrating on three brightest, whose modeling and other uncertainties can negatively impact the conclusions, the paper should really be about the full sample. Fig 2 has some of this but details are not clear. Also as Fig 2 shows, those three galaxies are not really ruled out by LCDM. And it is also not just those 3 galaxies, there are others from before. One could also naturally argue why select these three, when at z=7-8 they seem to have two equally interesting galaxies? Looking at this figure, I am afraid the point of this whole paper is totally unclear to me.

Answer: Given that our key point is the existence of spectroscopically confirmed extremely massive galaxies in the early Universe, we focus on three ultra-massive galaxies rather than the whole sample.

We agree with the referee that the two galaxies at z=7-8 might mislead the reader, which was not sufficiently clarified in the paper. Our main reason for choosing these three massive galaxies at z~5 is related to the robustness of the very high Mstar/Mhalo ratio, which implies a high efficiency of star formation in the distant Universe far exceeding the maximum efficiency (~20%) found in previous statistical studies. If we had selected the range z=7-8, then we would have suffered from the caveat that even JWST cannot sufficiently probe the rest-frame visible light for galaxies at such distances, and therefore the stellar mass determination could be more inaccurate. In addition, among the two sources at z=7-8, one has a red point source morphology, indicating a possible AGN contribution, and the other has only a single emission line at limited S/N (~8 sigma), so both the redshift and stellar mass may be uncertain.

We have now added more explanation in the paper (mainly at lines 138-148; lines 105-107) and in the caption of Fig. 2 to explain why z=7-8 sources were not included in the main discussion of our results.

3. It is a good useful result that with JWST Halpha one can get Halpha SF. The paper would have a stronger impact if the manuscript had discussed all properties of the sample and not concentrate on three

of the brightest. There is a mention of CIGALE SFs, including all of long wavelength data, but no SEDs are shown and one cannot make any conclusions on those SFs.

Answer: Due to the limited coverage of NIRCam F444W of 4-5 μm, only for 26/36 galaxies in our sample can we probe the Halpha line, while Hbeta is not covered. In this case, we can only derive the Halpha SF for 26 galaxies without dust attenuation correction, which is only a lower limit of the total SFR. Since we are focused on ultra-massive galaxies in the paper, we have briefly discussed the properties of the entire sample in the main text and in the Supplementary section.

For the long-wavelength data, we used it for the FIR SED fitting only. To break the energy balance, we performed the UV-to-NIR SED fit and the FIR SED fit separately. We note that assuming an energy balance for such optically dark galaxies might lead to an underestimation of the infrared luminosity, hence, the SFR (e.g., see Xiao et al. 2023). With the UV-to-NIR SED fit, we obtained the stellar masses and SFRs corrected for dust attenuation. With the FIR SED fit, we obtained their total infrared luminosity and SFR_IR. Given the limited FIR to millimeter photometry, we fitted S1 and S3 using the IR template (Schreiber et al. 2018) and/or CIGAGLE. For S2, the so-called GN10, since no new data points were available, we directly used the IR photometry and SFR_IR from Riechers et al. 2020 instead of fitting the FIR SED ourselves. We do not show the FIR spectra because of the limited FIR and submillimeter data available and the fact that the SFR is not the focus of this paper.

4. Fig 3 left analysis seems a little bit disjoint with Fig 2 left. Fig 2 caption says efficiency 50% is ok; Fig 3 left would imply an issue with even efficiency 20% that for S1 is a factor of x1000. This is not captured in Fig 2 left, but I believe that plot. Arguments similar to Fig 3 left
- this is not the first paper to try and make up, there are many related to more massive things like galaxy clusters - have not lived up to the test of time.

Answer: In Fig. 2, we show that the three galaxies are so massive that the baryon-to-star conversion efficiency needs to be 50%, which is two to three times higher than the maximum observed efficiency (~20%; the efficiency of Milky Way-like galaxies), therefore challenging the empirical and numerical models of galaxy formation. In Fig. 3, we further verify their extreme properties in an opposite way — what would happen if the efficiency of these galaxies was relatively less intense, similar to "normal" galaxies? We therefore assume that these galaxies have the same efficiency as the Milky Way-like galaxies, i.e., 20%, and then derive the required halo mass based on our stellar mass measurements. We then compare the theoretical and observed number densities of dark matter halos, i.e., to study the probability of detection of these galaxies. The derived number density of S1 is ~1100 times lower than the observations. It means that the detection probability of S1 should be only 1/1100=0.0009 in the FRESCO field, however, we have now observed it. This shows, in another way, that the efficiency of these galaxies is much higher than we expected.

We have now updated the relative sentences (lines 149-150) to make our statement clearer.

5. Structural parameters are probably useful, but again I am unclear what is the information content here. Fig 3 right shows one of the galaxies with ALMA dust. Instead of some rest optical structures, there is probably a whole lot more to be gained by a combined JWST, ALMA study, apart from the fact that dust is extended.

Answer: The structural parameters and the right panel of Fig. 3 are meant to discuss the build-up of such ultra-massive galaxies, and demonstrate that they are not dominated by a point-source, i.e. a possible AGN. Currently, we only have ALMA data at 1.1mm for S1, which is limited for an in-depth study.

However, we have successfully obtained ALMA time in Cycle 10 this year with priority A (will get the data next year). We will perform the resolved [CII]158um emission line observations, which can allow us to do further studies such as dynamical mass, total star formation rate, and gas distribution.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

I thought the manuscript was good before, but I think it is now even clearer.

Referee #3 (Remarks to the Author):

I have read the manuscript "Ultra-Massive Galaxies in the First Gyr Unveiled by JWST Challenge Galaxy Formation Models" by M. Xiao et al., submitted for publication in Nature. I acknowledge that this manuscript appears to have been in the refereeing process for a while already, so I apologize to the authors for the delayed report. That being said, I have read the manuscript with a fresh eye and compiled my report before reading the history of the process to provide the authors with an unbiased assessment. After reading the previous communication, I do see some overlapping concerns.

Overall, I think this work likely has the potential to be suitable for publication in Nature, if the findings can be put on a more solid quantitative footing that reveals the true level of progress brought to the field. Only with these quantitative details, I would be able to make a clearer recommendation.

I refrain from making suggestions about language details (e.g., "hotly debated", "astonishingly", etc.), which I presume will be fixed in the language editing process with the editorial office.

This manuscript deals with the unexpectedly high number of massive galaxies uncovered by JWST in the very early universe. Some such systems were previously known, and commonly identified as dusty starburst galaxies or "submillimeter galaxies". Indeed, some of the galaxies identified by the authors, HDF850.1 and GN10 = S2, correspond to some of the first such galaxies seen in submillimeter surveys back to the late 90s, which typically lacked optical/near-infrared counterparts and delayed actual redshift measurements to the 2010s. As such, the main topic of this work has been debated for several decades (and such galaxies being "missed by optical observations" in itself is not new), and the hope is that JWST can deliver the missing information (especially reliable stellar masses) to solidify the claims that these are already-massive galaxies, and to better quantify their number densities, since previous studies may only have revealed the "tip of the iceberg". With this motivation in mind, the general subject area of the study is not "new", and the impact work thus has to be mainly judged by how much progress it brings to this high-interest subject, and to what degree these can be considered interesting new constraints on the early galaxy formation process.

First, do these observations uniquely challenge our understanding of galaxy formation? Such claims have been made ever since the initial discovery of z>4 submillimeter galaxies (which needs to be presented for context here), which seemed to be in excess of models of galaxy formation. To understand if the problem had been solved in the meantime, and/or how much worse it gets with the new JWST data, it would be desirable to quantify how bad the discrepancy to models is, how "wrong" stellar masses would need to be to alleviate the problem, and what the increase is compared to previous suggested number densities of massive dusty starbursts in the literature. Also, given that some of the examples, like HDF850.1 and GN10 = S2, are now known to apparently live in overdense regions (e.g., Sun et al. 2023; Herard-Demanche et al. 2023), it would be good to state early on how this could potentially bias the results (at least for the sources in that redshift range). There also needs to be some more discussion on sample completeness levels.

Then, there needs to be more information on how much the stellar masses with the new data improved, compared to previous estimates (e.g., based on GN10 = S2), to evaluae how much any "conflict" with models is sharpened by what is reported here.

Next, the assumed baryon-to-stellar mass conversion efficiency should be considered with realistic uncertainties, which makes it possible to better evaluate the uncertainties in volume densities, and therefore, the "real" level of disagreement with models. Also, once again, the likely overdensity of the field at z~5.3 needs to re-enter the comparison to expected field occurrence rates of such galaxies.

In more general terms, the presentation of the sample as "three galaxies and the rest" is somewhat confusing, and the authors should consider what they really want to do here: Either interpret the full sample using the better statistics, with perhaps worsening error bars as f(z), OR only consider the three sources for the full analysis if these are the only ones considered to be robust for the broader interpretation. The latter case has the potential disadvantage of providing small numbers, more similar to sample sizes previously known from submillimeter studies, and for providing less constraints on the overall shape of the mass function than the broader sample.

As to the finding of 6-sigma "clumps" in ALMA data, I would recommend that the authors consider their analysis in the context of, e.g., the work of Hodge et al. (2016), who have studied based on models the reliability of modest-significance clumps in ALMA data once the imaging restrictions are properly accounted for. As it stands, the identification of clumps does not appear solid.

Detailed comments:

- Fig. 1, right: I find the observed spectrum hard to make out in this display. I would suggest to add a histogram style outline to the data.

- Fig. 3: Given the large abundance of +/-4 sigma peaks, how was the noise level determined in the ALMA data, how deeply was the data cleaned, and what baseline weighting was adopted?

Author Rebuttals to First Revision:

We are very grateful to the referee for the very helpful report. We revised and improved the manuscript in red color by considering all suggestions and comments from the referee's report. Detailed explanations are given just below each comment or suggestion.

[Referee Report]
Referee #3 (Remarks to the Author):

I have read the manuscript "Ultra-Massive Galaxies in the First Gyr Unveiled by JWST Challenge Galaxy Formation Models" by M. Xiao et al., submitted for publication in Nature. I acknowledge that this manuscript appears to have been in the refereeing process for a while already, so I apologize to the authors for the delayed report. That being said, I have read the manuscript with a fresh eye and compiled my report before reading the history of the process to provide the authors with an unbiased assessment. After reading the previous communication, I do see some overlapping concerns.

Overall, I think this work likely has the potential to be suitable for publication in Nature, if the findings can be put on a more solid quantitative footing that reveals the true level of progress brought to the field. Only with these quantitative details, I would be able to make a clearer recommendation.

Answer: Thanks for the useful suggestions. We have now added more quantitative statements in the paper, in terms of a comparison of the pre-JWST studies and our results as well. Specifically, 1) in our sample, 95% of dusty star-forming galaxies (and 2/3 ultra-massive galaxies) did not have spec-z and stellar mass information before JWST. It is only with the JWST NIRCam/grism data that we can perform statistical analyses as we do in this paper; 2) we have now added 1sigma uncertainties to the values (e.g., SFRD, probabilities in Fig.3) reported in the paper; 3) We have also added the uncertainties in our main plots, Fig.1 and Fig.3, which take into account the cosmic variance from simulations. These make our conclusions more robust.

Meanwhile, we have reframed the paper to make it more quantitative. The key point of this paper is that previous findings of the high baryon-to-star conversion efficiency of massive galaxies (e.g. Labbe+23, Casey+23, Boylan-Kolchin2023) that were presented as a challenge to galaxy formation models (or even the LCDM cosmological model) are very uncertain, as these results were initially based only on photometric observations and missed spectroscopic confirmation of the sources. Some of these sources have already been found to be lower redshift AGNs with lower stellar masses (e.g., Kocevski+2023, Wang+24). So whether these massive galaxies indeed have high baryon-to-star conversion efficiency remains unknown. We systematically studied a flux-limited spectroscopically confirmed sample at z~5-9, with robust redshifts and stellar masses, and found that 1) none of these galaxies have an efficiency larger than unity that would "break" the LCDM model; 2) some sources (i.e. the three ultra-massive galaxies at z~5-6 presented in the paper) require high efficiency, that challenge models of star formation in galaxies, in terms of the conversion efficiency of baryons into stars.

I refrain from making suggestions about language details (e.g., "hotly debated", "astonishingly", etc.), which I presume will be fixed in the language editing process with the editorial office.

Answer: We have now removed these words.

This manuscript deals with the unexpectedly high number of massive galaxies uncovered by JWST in the very early universe. Some such systems were previously known, and commonly identified as dusty starburst galaxies or "submillimeter galaxies". Indeed, some of the galaxies identified by the authors, HDF850.1 and GN10 = S2, correspond to some of the first such galaxies seen in submillimeter surveys back to the late 90s, which typically lacked optical/near-infrared counterparts and delayed actual redshift measurements to the 2010s. As such, the main topic of this work has been debated for several decades (and such galaxies being "missed by optical observations" in itself is not new), and the hope is that JWST can deliver the missing information (especially reliable stellar masses) to solidify the claims that these are already-massive galaxies, and to better quantify their number densities, since previous studies may only have revealed the "tip of the iceberg". With this motivation in mind, the general subject area of the study is not "new", and the impact work thus has to be mainly judged by how much progress it brings to this high-interest subject, and to what degree these can be considered interesting new constraints on the early galaxy formation process.

First, do these observations uniquely challenge our understanding of galaxy formation? Such claims have been made ever since the initial discovery of z>4 submillimeter galaxies (which needs to be presented for context here), which seemed to be. To understand if the problem had been solved in the meantime, and/or how much worse it gets with the new JWST data, it would be desirable to quantify how bad the discrepancy to models is, how "wrong" stellar masses would need to be to alleviate the problem, and what the increase is compared to previous suggested number densities of massive dusty starbursts in the literature. Also, given that some of the examples, like HDF850.1 and GN10 = S2, are now known to apparently live in overdense regions (e.g., Sun et al. 2023; Herard-Demanche et al. 2023), it would be good to state early on how this could potentially bias the results (at least for the sources in that redshift range). There also needs to be some more discussion on sample completeness levels.

Answer: We thank the referee for pointing this out. We fully agree that the sources studied here are likely just a somewhat lower luminosity version of dusty star-forming galaxies compared to SMGs. The number densities of the lower luminosity DSFGs are significantly (>~100x; e.g., Wang+19, Gómez-Guijarro+22) higher, however, compared to traditional SMGs resulting in a much stronger tension with models. We have now highlighted the connection to SMGs in the summary paragraph and added some comparisons of our sample with the pre-JWST results in the main text.

We would like to highlight that the strength of our observations with JWST is that we can now routinely obtain redshifts for sources with SEDs similar to SMGs but with luminosities that extend into the regime of more normal, main-sequence galaxies at z>5. With the grism data, we obtain a complete sample of such sources. Additionally, we would like to clarify that our subsample of dusty star-forming galaxies (We now generally replace the name "Optically dark/faint galaxies" in this version with the more general name "DSFGs") is selected by optical and NIR wavelengths, not submillimeter wavelengths. In other words, most of our DSFGs are not submillimeter galaxies and/or dusty “starburst” galaxies. For example, in the FRESCO-South field, only 1/11 of our DSFGs are detected by SCUBA and ALMA, that is S1. This leads to great difficulty if we want to focus on and discuss the specific topic of "submillimeter galaxies" and quantify their respective number densities, as our sample is not a submillimeter-selected complete sample.

Following the referee’s suggestion, we have largely rephrased the logic of this paper. What is "new" about this paper is that it is the first systematic study of a Halpha and [OIII] flux-limited complete sample at z>5 (impossible to be done before JWST), focusing on two key issues that have recently been debated in the literature: whether the LCDM model has been challenged, and whether the efficiency of baryon conversion into stars represents a challenge for existing galaxy formation models [galaxy formation model has been challenged]. Compared to the previous studies, this is the first time we have systematically confirmed, through spectroscopy, that ultra-massive galaxies have unexpectedly high baryon-to-stellar mass conversion efficiency. Only the JWST grism observations (which provide spectroscopic information while not needing to target pre-selected galaxies) allow us to carry out this kind of systematic study.

As the referee pointed out, in fact, two of the three ultra-massive galaxies at z~5-6 might be located in an overdense region (e.g., Herard-Demanche et al. 2023). We have now considered the effect of the cosmic variance on our results. With the help of the FLAMINGO simulation suite, we have derived the typical 1sigma scatter (and 3sigma upper limit) in the most massive expected haloes in Fig.1. We have also added the cosmic variance effect in Fig. 3. Note that our FRESCO survey covers non-contiguous regions of GOODS-S and GOODS-N, resulting in a reduced cosmic variance impact as compared to a contiguous survey of the same size. In addition, we have also discussed that the high efficiency of the ultra-massive galaxies might be related to the overdense environment at the end of the main part of this paper.

Then, there needs to be more information on how much the stellar masses with the new data improved, compared to previous estimates (e.g., based on GN10 = S2), to evaluate how much any "conflict" with models is sharpened by what is reported here.

Answer: Among the DSFGs in this paper, only 2/36 (GN10 and HDF850.1) already had spec-z and stellar mass measurements prior to our JWST observations. However, only now have we finally detected the stellar continuum for the remaining 34 sources, for a 'direct' estimate of their stellar masses. For S2 (GN10), the pre-JWST stellar mass (logM* = $11.08_{-0.03}^{+0.02}$; Riechers et al. 2020) was derived from optical-to-radio SED fitting (incl. upper limits from HST and photometric data from IRAC) and is consistent with our measurement (logM* = $11.18_{-0.17}^{+0.23}$) within uncertainties. As for HDF850.1, the pre-JWST stellar mass (logM* = $10.89_{-0.22}^{+0.15}$; Serjeant & Marchetti 2014) estimated based on the NIR to the radio bands (J, H WFC3, IRAC 5.8 and 8.0 μm, ISO 15 μm, MIPS 24, 70 and 160 μm, SCUBA 450 μm, SCUBA 850 μm and IRAM 1.3mm; The *J*, *H*, 15 μm, 70 μm and 160 μm data are upper limits.) are also consistent with our measurement (logM* = $10.58_{-0.17}^{+0.14}$) within uncertainties.
However, out of the remaining 95% of the galaxies, none had sufficient information to be attributed a relatively robust photometric redshift, and even more a stellar mass. Specifically, S1 and S3 were detected in the submillimeter wavelength, but this didn't even allow for estimates of their photometric redshift (and stellar mass) because they were not detected in the UV-to-optical wavelengths before JWST. We have now added a detailed comparison to the paper.

Next, the assumed baryon-to-stellar mass conversion efficiency should be considered with realistic uncertainties, which makes it possible to better evaluate the uncertainties in volume densities, and therefore, the "real" level of disagreement with models. Also, once again, the likely overdensity of the field at z~5.3 needs to re-enter the comparison to expected field occurrence rates of such galaxies.

Answer: Thanks for the useful suggestions. We have now included the effect of cosmic variance in our paper, please see the answer above.

In more general terms, the presentation of the sample as "three galaxies and the rest" is somewhat confusing, and the authors should consider what they really want to do here: Either interpret the full sample using the better statistics, with perhaps worsening error bars as f(z), OR only consider the three sources for the full analysis if these are the only ones considered to be robust for the broader interpretation. The latter case has the potential disadvantage of providing small numbers, more similar to sample sizes previously known from submillimeter studies, and for providing less constraints on the overall shape of the mass function than the broader sample.

Answer: Thanks for the suggestion. Following this, we have rephrased the logic of this paper and now first discuss the full sample and then highlight in more detail the three most efficient sources. Please also see the detailed explanation above. Briefly, to answer the key questions in this paper, that is whether the present data challenge existing galaxy assembly models or even the LCDM cosmological framework. 1) We first study a sample of spectroscopic-confirmed galaxies at z~5-9 and find that none of them have an efficiency larger than unity that would otherwise challenge the LCDM cosmological framework. This is especially highlighted in the new panel c in Fig 1, where we now show the distribution of the inferred baryon-to-stellar mass conversion efficiency. Panel b further shows how the DSFGs are populating the high-mass end of the z>5 galaxy population; 2) We do find some sources (especially the three ultra-massive galaxies at z~5-6) exhibiting a high baryon-to-star conversion efficiency that challenges the currently favored models of galaxy formation and growth.

As to the finding of 6-sigma "clumps" in ALMA data, I would recommend that the authors consider their analysis in the context of, e.g., the work of Hodge et al. (2016), who have studied based on models the reliability of modest-significance clumps in ALMA data once the imaging restrictions are properly accounted for. As it stands, the identification of clumps does not appear solid.

Answer: Considering the current length of the paper has far exceeded the usual limit, and that the ALMA data is not the key and not strongly related to the main results of this paper, we have removed this part on the ALMA dust distribution. We are very grateful for the referee's suggestions, which we will incorporate in our future work by not highlighting too much the 6-sigma "clumps" based on the current limited data.

Detailed comments:

- Fig. 1, right: I find the observed spectrum hard to make out in this display. I would suggest to add a histogram style outline to the data.

Answer: Done.

- Fig. 3: Given the large abundance of +/-4 sigma peaks, how was the noise level determined in the ALMA data, how deeply was the data cleaned, and what baseline weighting was adopted?
Answer: We have now removed this figure and relevant parts of the paper, given the limited length of the paper, and also to make the whole text logically clearer, as noted earlier. We are very grateful for this comment. We will incorporate this comment in our future work.

Other changes:
We added a new co-author (Seunghwan Lim) who helped with the FLAMINGO simulations to estimate the cosmic variance effects.

Reviewer Reports on the Second Revision:

Referee #3 (Remarks to the Author):

I have read the revised version of the manuscript "Ultra-Massive Galaxies in the First Gyr Unveiled by JWST Challenge Galaxy Formation Models" (original title) by M. Xiao et al., submitted for publication in Nature.

The authors have done a thorough job in addressing my previous report. As such, I have only two follow-up points for clarifications. I recommend to accept the manuscript after the authors had a chance to make these changes.

1) The term "DSFG" is somewhat ill-defined in the literature. In some definitions, these are only the most extreme specimens in the authors' sample based on high submillimeter fluxes, while in others, it encompasses all galaxies with detectable dust. An ideal definition would perhaps lie somewhere in between, requiring some minimum amount of dust obscuration or dust mass fraction to be considered "dusty" (e.g., there are galaxies with quite low Av in ED Table 1). I am not advocating that the authors have to use a specific definition, but I would like to ask them to provide a precise definition somewhere in the manuscript to make sure readers have a correct impression of the sample.

2) I would recommend to also adjust the ED figures to the same 1D histogram style of the spectra as the new version of Fig. 2.

Author Rebuttals to Second Revision:

We are very grateful to the referee for the very helpful report. We revised and improved the manuscript in red color by considering all suggestions and comments from the referee's report. Detailed explanations are given just below each comment or suggestion.

[Referee Report]
Referee #3 (Remarks to the Author):

I have read the revised version of the manuscript "Ultra-Massive Galaxies in the First Gyr Unveiled by JWST

Challenge Galaxy Formation Models" (original title) by M. Xiao et al., submitted for publication in Nature.

The authors have done a thorough job in addressing my previous report. As such, I have only two follow-up points for clarifications. I recommend to accept the manuscript after the authors had a chance to make these changes.

1) The term "DSFG" is somewhat ill-defined in the literature. In some definitions, these are only the most extreme specimens in the authors' sample based on high submillimeter fluxes, while in others, it encompasses all galaxies with detectable dust. An ideal definition would perhaps lie somewhere in between, requiring some minimum amount of dust obscuration or dust mass fraction to be considered "dusty" (e.g., there are galaxies with quite low Av in ED Table 1). I am not advocating that the authors have to use a specific definition, but I would like to ask them to provide a precise definition somewhere in the manuscript to make sure readers have a correct impression of the sample.

Answer: Thanks for the very helpful suggestion. We fully agree that the definition of "DSFG" is a bit vague in general. We have now defined the DSFG that we called in our paper (shown in red in the first paragraph).

2) I would recommend to also adjust the ED figures to the same 1D histogram style of the spectra as the new version of Fig. 2.

Answer: Done. We have now adjusted the ED figures 3-5 to match the Fig. 2.