

Peer Review File

Manuscript Title: Shock cooling of a red-supergiant supernova at redshift 3 in lensed images

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Congratulations on this exciting discovery! I find your result to be very plausible, and I only have minor comments / suggestions:

I agree your photometric redshift result $z \sim 3$ is confident. Still it is important to scrutinize:

BPZ and EAZY yield unambiguous results for all 3 lensed images. That's great. You might mention the next most likely result is $z \sim 0.2$, which is not only far less likely but also wouldn't be lensed. You might also note, if this is true, that you obtain similar redshift results even before GALFIT modeling and subtraction of the lensing galaxy, at least for G3, which is furthest away.

You show possible features in the spectrum that might correspond to [OII] at $z = 2.983$. What equivalent width would that correspond to? Would that be reasonable? Should we be surprised there aren't stronger lines detected by your spectroscopy? Or would that be easy to explain by low metallicity, for example?

Does the lensing geometry help constrain the redshift $z \sim 3$? What redshift would best fit the existing GLAFIC v4 lens model? Did the lensing galaxy mass change significantly after fixing the lensed source redshift to $z=2.95$? Presumably it still gives a reasonable mass-to-light ratio? Could you get a decent result for say $z \sim 1$, or are such low redshifts clearly ruled out by lensing alone? You say the positional offsets were $\sim 0.8''$ with the GLAFIC v4 lens model. What were they after you revised the model? Close to or better than the $0.1''$ uncertainty you assumed? (It certainly appears that way from Figure 1e.)

Regarding the lens model, can you say briefly why you are using GLAFIC. For example, it performed well in previous tests predicting simulated magnifications and observed time delays. I think I understand you are not using the broad range of other available lens models because they do not accurately reproduce this lensed galaxy, and you included that galaxy as a new constraint in your modeling. You might state that explicitly. Perhaps in future work it would be useful to try at least one other method to see if it gives consistent results and to quantify the systematic uncertainties due to lens modeling.

I understand your derived CCSN rates (based on two lensed CCSN in the HFF) are higher than those measured by Strolger+15 for CLASH + CANDELS. Can you combine the results from both studies? Would they meet in the middle near the prediction based on Madau & Dickinson 2014?

Did you make use of the "official" HFF image reductions, or did you re-reduce the data yourself?
Please clarify.

I just noticed a few small typos while reading:

L259 residue --> residual

Figure S-18 caption: Blue dashed lines is --> The blue dashed line is

In summary,

A. Summary of the key results

The authors have discovered one of the most distant supernovae known at $z \sim 3$ in archival Hubble imaging prior to the Frontier Fields program. Remarkably, three lensed images enable them to track its light curve rise and fall in 3 filters and in 3 epochs from the first day to 8 days after explosion. This enables them to constrain its pre-explosion radius and conclude it was likely a red supergiant. Finally, they derive a revised rate of core collapse supernovae, somewhat higher than the previous estimate, which was lower than predicted.

B. Originality and significance: if not novel, please include reference

This novel, significant discovery merits publication in Nature.

Everything else below is excellent, with some suggested improvements above:

C. Data & methodology: validity of approach, quality of data, quality of presentation

D. Appropriate use of statistics and treatment of uncertainties

E. Conclusions: robustness, validity, reliability

F. Suggested improvements: experiments, data for possible revision

G. References: appropriate credit to previous work?

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Referee #2 (Remarks to the Author):

This study presents novel work using multiple images from a lensed supernova to measure early emission from this event. Not only is this an exciting and creative technique, it also provides crucial information that teaches us about supernovae in a way that is usually not available.

Reading through the study, I feel the authors are mostly very careful with their analysis of the data and interpretation of the results. There are roughly three areas that I feel could use more investigation or clarification.

1. I have a concern with the data at the first epoch. Here the flux in F814W and F160W are comparable and more than F110W (although it is hard to read the F160W error bars in Figure 3 the way it is plotted). Such an ordering of fluxes is inconsistent with what would be expected for the

thermal emission that characterizes shock cooling emission. This suggests to me that that these early points should be more strongly scrutinized. Perhaps there is an issue with the template subtraction? If, for example, the true F160W point is much dimmer than what is found here, then the early emission is hotter and suggestive of a larger radius progenitor.

For these reasons, the authors need to better describe how robust this first epoch data is and investigate how strongly it impacts their conclusions. They should double check if there are any issues with their template subtraction and also explore how their radius measurements change if the errors bars are increased and/or the F160W point is incorrect.

2. Another potential problem I see is with the CC rate derived with a heterogeneous sample of events. Refsdal was 1987A-like and presumably had a blue supergiant progenitor while the radius inferred for this newer event suggests a Type IIP or IIL SN. But we know about the relative rates of these events which could be included into the rate analysis.

I suggest the authors explore how the relative rates of 1987A-like vs Type IIP/L-like SNe impacts their rate inferences at different redshifts. There should also be a discussion of whether the discovery of the Refsdal event suggests that 1987A-like SNe are more common at larger redshifts or not.

3. During the introduction there is discussion on how shock cooling emission can be used to measure SN progenitor radii. While this is true in principle, in practice (for most hydrogen-rich events) the shock cooling emission over the first ~20-30 days suggests dense, extended CSM surrounds many of the progenitors (see for example Morozova et al. 2016, ApJ, 838, 28). In fact, even the event the authors reference as showing a red supergiant-like radius (SN 2017eaw) actually shows evidence for early emission that can be explained by an earlier pre-supernova outburst (Morozova et al. 2020, ApJ, 891, 32).

For these reasons, I think the authors need to update and reframe their introduction better. If the ~500 R_{sun} radius they are inferring stands (after taking a more careful look at the first epoch data as described in point 1 above), the event in this study is actually pretty different than many other Type IIP/L supernova in that it does NOT have dense CSM around it. And it is interesting to think about and discuss why that may be.

Another way the authors can address the uniqueness of this event is to (rather than just comparing with semi-analytic theoretical models) compare with other early Type IIP/L light curves. Even if the same bands are not available at the same times, early emission is close to black body, and thus other Type IIP/L with early data can be extrapolated to see what they would look like if observed with these same HST filters. Such a comparison would provide better context on how unique this event really is.

Final note: Line 42, after "shocked material" there should be a reference to a classic paper on shocked cooling emission. I would suggest Nakar & Sari (2010) ApJ 725, 904

Referee #3 (Remarks to the Author):

This paper presents the first lensed (and hence highest redshift) supernova likely observed during its shock cooling phase. The shock cooling phase, lasting hours to days, can provide crucial information about the exploding star, which is otherwise inaccessible observationally. This is especially important at high redshifts, which allows testing of various stellar evolution theories.

The three images of a single supernova observed at three different intrinsic times, showing very different colors, is strong evidence that indeed the shock cooling phase has been captured here. This demonstrates the potential of the method, and opens the door to more such discoveries.

In addition, the paper uses this discovery to infer the rate of high-redshift supernovae that can be found via lensing and suggests that this rate is higher than expected. That is both interesting in its own right, and promising for the prospect of finding more such important events.

However, I have three main issues with the methods of the paper, that currently, in my view, preclude deducing robust conclusions from it:

1. The analysis fits a 10-parameter model to 9 data points. Thus any inference of physical parameters or even physical nature (such as RSG vs. BSG or CSM vs. no CSM) is weak. Furthermore this number of parameters is not clearly stated in the main text (and Figure 2, showing only 3 of the 10 parameters, could be misleading in this respect).

2. The analysis assumes that the source is a standard, maybe even a IIP, supernova (e.g. the first paragraph stating that the RSG progenitor is "consistent with the luminosity and colour of the supernova", the main text referring and comparing to IIP's, and the last sentence of the main text stating the source is "consistent with a normal supernova") but this is never clearly established. Neither the available data for the source, the inferred best fit parameters, nor the temperature evolution, are compared to that of any other observed supernova. In fact, the inferred temperature of 10^6 K for the first epoch seems very high compared to what has been seen in other supernovae, even within hours of explosion. The paper never considers the options of the source being some other type of supernova, or even other type of transient. For example, many shock cooling events observed in nearby cases are of Type IIb supernovae, which have different progenitors (typically YSGs with low mass extended envelopes). It is not clear if this or other options (maybe even companion shocking by a Type Ia SN) can be ruled out here, and hence what the physical implications of this discovery are.

3. The surprisingly high rate estimate is said to be reduced when considering also the non-discovery of lensed supernova in the CLASH program. However it is not stated then how significant the inferred high rate is in this case (it is only stated that it is "still higher" than expectations - but not whether the difference is still statistically significant or not). In addition, it might be interesting to discuss the implications of the possible higher rate, both for discovery and for the physics leading to the expected rates.

Additional comments:

* It is not stated in the main text how the blackbody temperatures in lines 94-95 were inferred - whether this is from the shock cooling models, blackbody fits, or other method.

* Line 114 refers to "model 1" and "model 2" but these names were never defined (in the rest of the paper, the models are given different names).

* Lines 116-117 state that the M20 RSG model gives a best fit radius of 577 R_{sun} , but later, in line 141 the M20-RSG model is said to give radii of 647 and 543 R_{sun} (one of these values is after altering the best fit lensing parameters by 1-sigma, but it is not clear which of them, and then what the other value refers to, and why both are different from the initial value mentioned in line 117).

* The paper favors the RSG model over the CSM models because "analyses of low-redshift hydrogen-rich SNe do not find evidence for such highly confined and dense CSM". But if the purpose of this discovery is to "study the massive stellar populations at high redshift" then projecting what we know at low redshift on to them defeats this goal completely.

* The paper favors the RSG model over other models because those are "not significantly better". Another way to phrase that is that the other models are equally good, in which case it is not clear why they are disfavored.

* It would be interesting to know how the inferred mass and SFR of the supernova host compare to those of known supernova hosts of various types.

* The color difference between the images is indeed striking, and is shown not likely to arise from noise. But it would be important to address (especially for non-expert readers) whether it could arise from different gravitational redshifts imparted on each image.

* Three R_V values are chosen but only one is shown in the figures. It might be useful to show the fits for all models tested, for all R_V values. Also, it is not clear why only one of the R_V values (4.05) is used in the rate simulations (line 393).

* It is not explicitly stated how the redshift of 0.375 was derived for the lens - whether it was through line identifications in the spectrum, a fit to the continuum only, or something else.

Minor issues:

* In the first (bolded) paragraph, it is stated that "very early observations have been possible only for approximately a dozen nearby supernovae" - This is not exactly true. Early observations are *possible* for many more supernovae, and are routinely obtained by ZTF, for example. Perhaps it is more accurate to say that shock cooling analysis has so far been performed only approximately *several* dozen nearby supernovae.

* In line 191 there is a reference to "all peaks", but it is not clear what these peaks are of.

* It is not stated what the horizontal dashed lines are in Figure S-7.

Author Rebuttals to Initial Comments:

Responses

Referees' comments:

Referee #1 (Remarks to the Author):

Congratulations on this exciting discovery! I find your result to be very plausible, and I only have minor comments / suggestions:

I agree your photometric redshift result $z \sim 3$ is confident. Still it is important to scrutinize:

BPZ and EAZY yield unambiguous results for all 3 lensed images. That's great. You might mention the next most likely result is $z \sim 0.2$, which is not only far less likely but also wouldn't be lensed. You might also note, if this is true, that you obtain similar redshift results even before GALFIT modeling and subtraction of the lensing galaxy, at least for G3, which is furthest away.

Thank you for pointing this out. We now add in the Methods section: "We note that the next most-probable values of the photometric redshift are $z \sim 0.2$, but we can exclude this redshift because it is in the foreground of the Abell 370 lensing cluster which has a redshift of 0.375."

You show possible features in the spectrum that might correspond to [OII] at $z = 2.983$. What equivalent width would that correspond to? Would that be reasonable? Should we be surprised there aren't stronger lines detected by your spectroscopy? Or would that be easy to explain by low metallicity, for example?

We have added a new calculation of the 3-sigma upper bound on the equivalent width, using the continuum of the host galaxy that best fits the HST photometry. We find that we would not expect to detect [O II] based on the samples from the MOSDEF survey. Now, in "Spectroscopic observations of the SN host galaxy" in the Methods section, we add "This analysis yields a 3-sigma upper bound of ~ 350 Angstroms, and we compare the bound to the distribution measured by the MOSDEF survey, as shown in the top-left panel in Fig. 3 from Reddy et al. (2018, ApJ). The host galaxy of the multiply imaged supernova we have discovered has a stellar mass of $\sim 2 \times 10^8$ solar masses, which is less massive than a majority of the MOSDEF galaxies that have [O II] lines detected by the survey. Therefore, we would not expect to detect [O II], based on the MOSDEF sample."

Does the lensing geometry help constrain the redshift $z \sim 3$? What redshift would best fit the existing GLAFIC v4 lens model? Did the lensing galaxy mass change significantly after fixing the lensed source redshift to $z=2.95$? Presumably it still gives a reasonable mass-to-light ratio? Could you get a decent result for say $z \sim 1$, or are such low redshifts clearly ruled out by lensing alone? You say the positional offsets were $\sim 0.8''$ with the

GLAFIC v4 lens model. What were they after you revised the model? Close to or better than the 0.1" uncertainty you assumed? (It certainly appears that way from Figure 1e.)

Yes, we thank the referee for raising how the lensing configuration places an independent constraint on the redshift. For our revised submission, we have reconstructed the multiple images using the published HFF lens models from MAST. Considering the small angular separations of the images, we are only able to use cluster lens models whose resolution is smaller than 0.06" per pixel. These are GLAFIC V4, Keeton V4, and Sharon V4. After fixing the host galaxy's redshift at $z = 3$, we vary the galaxy's source-plane coordinates to minimize the RMS residuals of the predicted and observed images. We can see that the Keeton V4 model yields the best result for $z = 3$. This new analysis is now described in the Methods section in "Modelling the gravitational lens."

We can instead allow the redshift of the SN to be a free parameter. We plot the RMS of the residuals of the predicted from the observed image positions as a function of the galaxy's redshift in Extended Data Fig. 3b. For these three high-resolution published models, we can see that, even for a very large redshift, we can still obtain a reconstruction with small RMS of the positions of the multiple images. This is because the Einstein radius of the lens for the published models is slightly smaller than the true value, but owing to the weak dependence of the distance ratio (D_{ls}/D_{os}) on the Einstein radius beyond $z \sim 3$ (for this cluster redshift) it is difficult to accommodate the mismatch by the change of the source redshift. But we are able to rule out low redshifts at $z \sim 1$, where the RMS of the angular separation of the predicted images is much larger than that at higher redshifts. We now add this analysis in the Methods section in "Modelling the gravitational lens."

The original GLAFIC V4 model predicts the velocity dispersion of the lens of 152 km/s, while our new GLAFIC model in which the host-galaxy image positions are included as constraints and the mass-to-light ratio becomes a free parameter predicts the velocity dispersion of the lens of 172 km/s. The difference between these velocity-dispersion values is small and within the scatter of the Faber-Jackson relation. When the SN redshift is fixed to $z = 1$, the velocity dispersion of the lens becomes much larger, ~ 268 km/s, and therefore its mass-to-light ratio significantly deviates from those of the other cluster member galaxies. This provides additional evidence to rule out such low redshifts by the lensing alone. In the revised Methods section in "Modelling the gravitational lens," we now mention, "As shown in Extended Data Fig. 3c, our revised GLAFIC model can reproduce the image positions with $\text{RMS} < 0.02$ ".

Regarding the lens model, can you say briefly why you are using GLAFIC. For example, it performed well in previous tests predicting simulated magnifications and observed time delays. I think I understand you are not using the broad range of other available lens models because they do not accurately reproduce this lensed galaxy, and you included that galaxy as a new constraint in your modeling. You might state that explicitly. Perhaps in future work it would be useful to try at least one other method to see if it gives consistent results and to quantify the systematic uncertainties due to lens modeling.

Yes, we felt that we ought to use a model that incorporates the host-galaxy images as constraints, and allows the light-to-mass ratio of the early-type lens to vary. Since the system was not identified before we found the SN, previous models were made with knowledge of it. We are able to fully explore degeneracy among the model parameters, which is impossible using the published HFF models.

As for why we used GLAFIC specifically, it has performed very well in previous simulation efforts including Meneghetti et al. (2017), and M. Oguri is a member of our team. We hope that, following the publication of the discovery, other cluster-lens modeling teams may use it as an additional constraint. In the revised main text, we point out that “To better reproduce the image positions and fully explore the degeneracy of the model parameters, we modelled the gravitational lens in the Abell 370 cluster using the GLAFIC modelling package”.

I understand your derived CCSN rates (based on two lensed CCSN in the HFF) are higher than those measured by Strolger+15 for CLASH + CANDELS. Can you combine the results from both studies? Would they meet in the middle near the prediction based on Madau & Dickinson 2014?

As we had described in the Methods section, we used the Strolger et al. (2015, ApJ) host-galaxy extinction prior distribution for the analysis in our initial submission. Strolger et al. (2015) adopted an exponential prior $\lambda_V \exp(-\lambda_V A_V)$ where $\lambda_V = 0.187$. Using this prescription, we obtained a best-fitting value of $k_{CC} = 0.162$ / solar mass, which is much greater than the $k_{CC} = 0.068$ / solar mass expected for a Salpeter initial-mass function (IMF) as derived by Madau & Dickinson (2014).

However, we believe the Strolger et al. (2015) dust prior is not an accurate representation of observations of core-collapse SNe, or even the samples of SNe used by Strolger et al. (2015) to determine the parameters of their exponential prior. As a general note, given a scale parameter of an exponential distribution, $\tau = 1 / \lambda_V$, the mean of the distribution is τ . Since Strolger et al. used $\lambda_V = 0.187$, the mean implied extinction A_V is 5.35 mag (i.e., $1/\lambda_V = 5.35$).

From private communication with L. Strolger, we learned that this value of λ_V was derived by fitting the A_V values for CCSN in Schmidt et al. (1994, ApJ) and Hamuy & Pinto (2002, ApJ).

However, the mean value of A_V for these same SNe in Schmidt et al. (1994) and Hamuy & Pinto (2002) is ~ 0.4 mag, *which is a factor of >10 smaller than the 5.35 mag used by Strolger et al. (2015)*. Therefore, the scale factor used by Strolger et al. is not accurate.

Indeed, 5.35 mag is also far larger than the A_V measured for host galaxies of core-collapse SNe (e.g., CCSN host galaxies from Kelly & Kirshner 2012, ApJ, yield a mean $A_V \approx 1.0$) and indicates a much larger color excess $E(B-V)$ than the measured values for CCSN host galaxies (e.g., Drout et al. 2011, ApJ; Prentice et al. 2016, MNRAS) for typical values of R_V .

In the revised manuscript, we instead use the distribution of CCSN host-galaxy extinction from 200+ CCSNe in the local universe as provided by Kelly & Kirshner (2012). As described in the Methods section in “Core-collapse SN rate from multiply imaged SNe,” and as shown in Extended Data Fig. 8a, we find that the distribution can be roughly described by an exponential with $\lambda_V \sim 1$. We use the exponential prior instead with $\lambda_V = 1$ to update our constraints on the volumetric rate of CCSNe.

Using an exponential prior that is more realistic than the Strolger et al. (2015) function yields a new constraint on the relationship between the SFR and the SN rate. As now described in the last paragraph of the main text, “given the measured star-formation density from ref. 47 and the two multiply imaged SNe, we find $k_{CC} = 0.0087^{+0.0071}_{-0.0046} M_{\odot}^{-1}$ and the maximum-likelihood-estimation (MLE) $k_{CC} = 0.0060 M_{\odot}^{-1}$, consistent with the expectation of $k_{CC} = 0.0068 M_{\odot}^{-1}$ for a Salpeter IMF and a range of progenitors’ initial masses between 8 and 40 $M_{\odot}$.”

We also adopt, to explore the dependence of our constraints on the exponential prior, the value of $\lambda_V = 2$ from Taylor et al. (2014, ApJ). We find a best-fit value of $k_{CC} = 0.0072^{+0.0059}_{-0.0038}$ and the MLE $k_{CC} = 0.0049$, and Madau & Dickinson (2014)’s value is within one sigma. Our results using these λ_V values, including $\lambda = 0.187$, are shown in Extended Data Fig. 8e. We also describe the results in the revised Methods section.

To provide an apples-to-apples comparison with the volumetric rate computed by Strolger et al. (2015), we compare our inferred rate computed with their $\lambda_V = 0.187$, in addition to computing our rate using more accurate values of λ_V . We note, as described in the Methods section (“Core-collapse SN rate from multiply imaged SN”; see Eq. 9) that the Strolger et al. calculation included an extraneous factor of h^2 . After correction, their value for k_{CC} is $0.0045 \pm 0.0008 M_{\odot}^{-1}$. When using $\lambda_V = 0.187$, we obtain $k_{CC} = 0.0216^{+0.0111}_{-0.0146} M_{\odot}^{-1}$, which is in tension ($p = 0.026$) with the value from Strolger et al., after correction for h^2 . In short, we find a greater frequency of core-collapse SNe than did Strolger et al. (2015), and our rate is in much better agreement with the expectation from the measured cosmic star-formation history and a Salpeter IMF, as described by Madau & Dickinson (2014).

Did you make use of the "official" HFF image reductions, or did you re-reduce the data yourself? Please clarify.

We used the “official” HFF images to measure the photometry of the SN host-galaxy images. In the revised Methods section, we state “We measured photometry of the multiple galaxy images G1, G2, and G3 from the published coadditions of HST imaging collected as part of the HFF survey.”

We created coadditions of HST imaging to measure photometry of the transient itself.. We have now created an online repository for our data as suggested by the editor. All of our own image coadditions are included in the repository.

I just noticed a few small typos while reading:

L259 residue --> residual

Figure S-18 caption: Blue dashed lines is --> The blue dashed line is

Thank you. These typos are now corrected.

In summary,

A. Summary of the key results

The authors have discovered one of the most distant supernovae known at $z \sim 3$ in archival Hubble imaging prior to the Frontier Fields program. Remarkably, three lensed images enable them to track its light curve rise and fall in 3 filters and in 3 epochs from the first day to 8 days after explosion. This enables them to constrain its pre-explosion radius and conclude it was likely a red supergiant. Finally, they derive a revised rate of core collapse supernovae, somewhat higher than the previous estimate, which was lower than predicted.

B. Originality and significance: if not novel, please include reference

This novel, significant discovery merits publication in Nature.

Everything else below is excellent, with some suggested improvements above:

C. Data & methodology: validity of approach, quality of data, quality of presentation

D. Appropriate use of statistics and treatment of uncertainties

E. Conclusions: robustness, validity, reliability

F. Suggested improvements: experiments, data for possible revision

G. References: appropriate credit to previous work?

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Referee #2 (Remarks to the Author):

This study presents novel work using multiple images from a lensed supernova to measure early emission from this event. Not only is this an exciting and creative technique, it also provides crucial information that teaches us about supernovae in a way that is usually not available.

Reading through the study, I feel the authors are mostly very careful with their analysis of the data and interpretation of the results. There are roughly three areas that I feel could use more investigation or clarification.

1. I have a concern with the data at the first epoch. Here the flux in F814W and F160W are comparable and more than F110W (although it is hard to read the F160W error bars in Figure 3 the way it is plotted). Such an ordering of fluxes is inconsistent with what would be expected for the thermal emission that characterizes shock cooling emission. This suggests to me that that these early points should be more strongly scrutinized. Perhaps there is an issue with the template subtraction? If, for example, the true F160W point is much dimmer than what is found here, then the early emission is hotter and suggestive of a larger radius progenitor.

For these reasons, the authors need to better describe how robust this first epoch data is and investigate how strongly it impacts their conclusions. They should double check if there are any issues with their template subtraction and also explore how their radius measurements change if the errors bars are increased and/or the F160W point is incorrect.

We agree that the error bars shown in Figure 3 were difficult to disentangle.

We have listed the fluxes and uncertainties in the Extended Data Table 1b, which are as follows:

F814W = 25.60 +/- 2.25 nJy, F110W = 20.20 +/- 7.10 nJy, F160W = 25.19 +/- 16.45 nJy .

These show that the F160W could (<0.3 sigma) be smaller than that of F110W, and it is not possible to rank the flux densities in a meaningful way. The F160W flux detection has a significance of only 1.5 sigma.

We have updated Fig. 3 to show the uncertainties more clearly.

We also now include constraints on the temperature at each epoch in Extended Data Fig. 7a, which shows that the first epoch alone does not provide a strong constraint on the progenitor's radius.

While the differences are not statistically significant, we have also revisited the photometry. The SN image closest to the galaxy lens is most delayed and therefore corresponds to the earliest epoch in the rest frame. Given its proximity to the early-type lens, it has, before subtraction of the lens' light distribution, greater contamination, especially in the two near-IR filters.

We have undertaken a very careful approach to subtraction of the early-type galaxy that goes beyond that typically used to create difference images. Hubble imaging taken at multiple epochs is, in general, acquired at multiple roll angles, which causes the diffraction spikes of the point-spread function (PSF) to rotate. These rotated diffraction spikes can leave large residuals, as is the case for the early-type galaxy.

We have taken the following steps to remove these diffraction spikes. We measure the PSF for each coadded set of images. We next develop a Sersic model for the unchanging early-type galaxy using GALFIT, and use that program to subtract the model, convolved with the PSF, from the imaging taken at both epochs.

We describe our technique in the revised Methods section as follows: "HST can have different roll angles taking images of the same field, resulting in different orientations of the diffraction spikes. Simply subtracting an HST image from a template may leave substantial residuals around bright objects, such as the lens galaxy in the centre of the multiply-imaged SN, as shown in Extended Data Fig. 1d. A proper way to remove the residuals from the misaligned diffraction spikes is to subtract a source model convolved with the PSF of each HST image before the image differencing. We used GALFIT to model sources in the coadded template images taken from the HFF survey. Extended Data Fig. 1a–1c show an example of our GALFIT modelling. We next modelled individual HST images taken during the SN event with source profiles fixed to their templates' GALFIT models."

In the Extended Data Figs. 1d and 1e, we compare the template-subtraction results with or without this process. We have also uploaded our GALFIT models, template and event images, and resulting difference images to an online repository, so that readers should be able to cross-check our analysis and reproduce our results.

We note that, for our initial submission, we used the F125W template for the F110W image (with the template's zeropoint rescaled to F110W), since there was no available F110W imaging of the galaxy acquired after the 2010 observations where we detect the SN. For the revised version of the paper, we have constructed an improved template, by combining available F105W and F125W imaging. The F105W has greater depth than that of the F125W imaging. The new template subtraction, however, produced an insignificant change in the flux constraints.

Finally, we have repeated the light-curve fitting using the M20-RSG and the P21 (with $M_{\text{csm}} < M_{\text{sun}}$) models where we inflate the error bars of the first F160W measurements by a factor of two. We find that the resulting radii of the progenitor given by the two models (524 R_{sun} and 474 R_{sun} , respectively) do not change significantly. We describe this analysis in the last paragraph in the Methods section in "Light-curve fitting."

2. Another potential problem I see is with the CC rate derived with a heterogeneous sample of events. Refsdal was 1987A-like and presumably had a blue supergiant progenitor while the radius inferred for this newer event suggests a Type IIP or IIL SN. But we know about the relative rates of these events which could be included into the rate analysis.

I suggest the authors explore how the relative rates of 1987A-like vs Type IIP/L-like SNe impacts their rate inferences at different redshifts. There should also be a discussion of whether the discovery of the Refsdal event suggests that 1987A-like SNe are more common at larger redshifts or not.

For our estimates of the volumetric CCSN rate in our submitted paper, we simulated populations of CCSNe adopting percentages and absolute magnitudes of different subtypes as given by Li et al. (2011, MNRAS). As a consequence, our constraints take into account our knowledge of the relative rates of events in the low-redshift galaxy population.

We appreciate the referee's suggestion, and have explored, in our updated simulations, how a much greater fraction of SN 1987A-like events (30%) among core-collapse SNe would affect our conclusions about the volumetric core-collapse rate. We find that our inferences about the volumetric rate have very little dependence on the fraction of events that are SN 1987A-like.

To describe this new analysis, we have added the following two paragraphs to the subsection, "Core-collapse SN rate from multiply imaged SNe" of the revised Methods section:

"SN Refsdal, the first-known example of a strongly lensed SN, and one of the most distant examples of a SN, was classified as SN 1987A-like (Kelly et al. 2016). The classification was surprising, because SN 1987A-like SNe account for 1-3% of CCSNe reported by low-redshift surveys (Pastorello et al. 2012, A&A; Taddia et al. 2016, A&A). One possible exception is the Caltech Core-Collapse Project which reported a fraction of 5% (Taddia et al. 2016, A&A).

We used our simulations to estimate the probability that, of the two first discoveries of strongly lensed CCSNe (by a cluster lens), at least one was SN 1987A-like. In our simulations, the absolute magnitudes of the SN 1987A-like SNe have a mean of -16 and a standard deviation of unity, as an approximation of the distribution shown in Fig. 5 of Kelly et al. (2016). As this figure shows, SN Refsdal would be a relatively luminous SN 1987A-like SN, with an absolute B-band magnitude of approximately -16.5 given a magnification of 15 for images S1–S3. Assuming that SN 1987A-like SNe account for 5% of CCSNe to $z \sim 4$, we find that we should expect 1.7 strongly lensed CCSNe, but only 0.09 of these should be SN 1987A-like SNe in the existing observations of the HFF cluster fields. Consequently, the discovery of SN Refsdal, a SN 1987A-like SN (Kelly et al. 2016) at $z = 1.49$, is not highly probable. However, the p value exceeds $p = 0.05$, given the two known examples."

We have rerun our simulation with an extreme assumption that SN 1987A-like SNe account for 30% of CCSNe in the lensed volume, and find that we should expect 1.76 strongly lensed SNe

in the existing observations of the HFF cluster fields. Consequently, inferences about the core-collapse rates are almost entirely unchanged (only $\sim 3.5\%$).

3. During the introduction there is discussion on how shock cooling emission can be used to measure SN progenitor radii. While this is true in principle, in practice (for most hydrogen-rich events) the shock cooling emission over the first ~ 20 -30 days suggests dense, extended CSM surrounds many of the progenitors (see for example Morozova et al. 2016, ApJ, 838, 28). In fact, even the event the authors reference as showing a red supergiant-like radius (SN 2017eaw) actually shows evidence for early emission that can be explained by an earlier pre-supernova outburst (Morozova et al. 2020, ApJ, 891, 32).

The referee raises an important comment, and we appreciate now that multiple analyses have yielded different conclusions about the presence of confined CSM around two of the better observed examples with early observations, SN 2013ej and SN 2017eaw. In both these instances, application of an analytical model from Rabinak & Waxman (2011) which incorporates no CSM provides reasonable fits to the observations and radii of 400 - $600 R_{\text{sun}}$ (Valenti et al, 2014) and $\sim 600 R_{\text{sun}}$ (Rui et al 2019), respectively. Morozova et al. develop numerical models that instead favor the presence of CSM. For SN 2013ej, they found the progenitor radius (star + dense CSM) of $2100 R_{\text{sun}}$ (Morozova et al. 2017). For SN 2017eaw, they added a $0.6 M_{\text{sun}}$ CSM onto a $841 R_{\text{sun}}$ RSG and improved the chi-square statistic fitting into the optical light curve (Morozova et al. 2020). For $K = 10^{18} \text{ g/cm}$ and a wind profile with $\rho \propto r^{-2}$ (Morozova et al. 2017), a $0.6 M_{\text{sun}}$ CSM extends up to a radius of $\sim 2200 R_{\text{sun}}$.

A key, important distinction, which we did not make clearly in our initial submission, was that our observations are unique in that they include rest-frame ultraviolet observations within the first day, which is unique among existing early-time SN observations (see Extended Data Figs. 7b–7f). As described by Sapir & Waxman (2017), these are very discerning and afford powerful constraints on the progenitor and CSM.

For SN 2013ej, the first UV observations are ~ 7 days after explosion (Huang et al. 2015). For SN 2017eaw, the first UV data are ~ 1.6 days after explosion (Szalai et al., 2019).

Our observations of the strongly lensed SN are unique in that we have UV (observer-frame $\sim 8000 \text{ \AA}$) observations within hours of explosion, and we now show that these provide constraints. We have rerun the fitting without the first rest-frame UV observation, and we obtain a CSM-shell radius of $952_{-398}^{+529} R_{\text{sun}}$ from the P21-CSM model, approximately twice as large as the original value ($\sim 480 R_{\text{sun}}$) we obtained from the P21-CSM model with the first rest-frame UV observation.

For these reasons, I think the authors need to update and reframe their introduction better. If the $\sim 500 R_{\text{sun}}$ radius they are inferring stands (after taking a more careful look at the first epoch data as described in point 1 above), the event in this study is actually pretty different than many other Type IIp/L supernova in that it does NOT have dense CSM around it. And it is interesting to think about and discuss why that may be.

Another way the authors can address the uniqueness of this event is to (rather than just comparing with semi-analytic theoretical models) compare with other early Type IIP/L light curves. Even if the same bands are not available at the same times, early emission is close to black body, and thus other Type IIP/L with early data can be extrapolated to see what they would look like if observed with these same HST filters. Such a comparison would provide better context on how unique this event really is.

We very much appreciate the referee's suggestion to compare the observations of our SN to those in the literature to assess agreement independent of models. We have undertaken this comparison, but we believe our high-redshift analysis may be more constraining than the nearby examples, since these lack very early UV observations. They are, of course, discovered at optical wavelengths from the ground, and Swift UV follow-up observations require a trigger and scheduling which push observations in almost all cases to more than one day later.

We have compared our observations to well-observed examples at low redshift that include those inferred to have CSM and those without CSM. We also have overplotted model fits to examine the differences among the CSM and non-CSM models at the very early times that our observations sample.

In Extended Data Fig. 7, we show these comparisons. We find that predictions of the models that contain CSM extend to over $1000 R_{\text{sun}}$ and those that do not have CSM or have $R < 1000 R_{\text{sun}}$ diverge in their predicted UV luminosities within the first day. Moreover, several of the best-studied examples in the literature (SN 2013fs and ASASSN-14gm) that have been inferred from their light curves to have substantial CSM at radii beyond $1000 R_{\text{sun}}$ similar to the examples from Morozova lack these critical early-time UV observations, as we show in Panels c and d.

Having a photosphere (optical depth $c/v > 1$) smaller than $\sim 1000 R_{\text{sun}}$ causes the UV luminosity to fall more rapidly following shock breakout, with subsequent rebrightening. As Panels c and d show, the UV light curves for objects with more extended CSM exhibit a close to monotonic decline.

Final note: Line 42, after "shocked material" there should be a reference to a classic paper on shocked cooling emission. I would suggest Nakar & Sari (2010) ApJ 725, 904

Thank you. We added this reference in the revised manuscript.

Referee #3 (Remarks to the Author):

This paper presents the first lensed (and hence highest redshift) supernova likely observed during its shock cooling phase. The shock cooling phase, lasting hours to days, can provide crucial information about the exploding star, which is otherwise inaccessible observationally. This is especially important at high redshifts, which allows testing of various stellar evolution theories.

The three images of a single supernova observed at three different intrinsic times, showing very different colors, is strong evidence that indeed the shock cooling phase has been captured here. This demonstrates the potential of the method, and opens the door to more such discoveries.

In addition, the paper uses this discovery to infer the rate of high-redshift supernovae that can be found via lensing and suggests that this rate is higher than expected. That is both interesting in its own right, and promising for the prospect of finding more such important events.

However, I have three main issues with the methods of the paper, that currently, in my view, preclude deducing robust conclusions from it:

1. The analysis fits a 10-parameter model to 9 data points. Thus any inference of physical parameters or even physical nature (such as RSG vs. BSG or CSM vs. no CSM) is weak. Furthermore this number of parameters is not clearly stated in the main text (and Figure 2, showing only 3 of the 10 parameters, could be misleading in this respect).

There are certainly fewer observations than have been used for these analyses in the low-redshift universe. As we have identified as part of our response to the referees, however, the very early UV observations are uniquely constraining when compared to existing observations of SNe in the nearby universe. The very early UV evolution places informative constraints.

We also have the four lens-model related parameters which nearby SN analyses lack, but we use informative priors on these.

While we agree with the referee that additional data would be better, from a quantitative perspective we have carried out a proper inference using the Bayes factor, which yields statistically significant evidence. While there are ten parameters, five of these are constrained by priors from the lens model, and we marginalize over these.

2. The analysis assumes that the source is a standard, maybe even a IIP, supernova (e.g. the first paragraph stating that the RSG progenitor is "consistent with the luminosity and colour of the supernova", the main text referring and comparing to IIP's, and the last sentence of the main text stating the source is "consistent with a normal supernova") but this is never clearly established. Neither the available data for the source, the inferred best fit parameters, nor the temperature evolution, are compared to that of any other observed supernova. In fact, the inferred temperature of 10^6 K for the first epoch seems very high compared to what has been seen in other supernovae, even within hours of explosion. The paper never considers the options of the source being some other type of supernova, or even other type of transient. For example, many shock cooling events observed in nearby cases are of Type IIb supernovae, which have different progenitors (typically YSGs with low mass extended envelopes). It is not clear if this or other options (maybe even companion shocking by a Type Ia SN) can be ruled out here, and hence what the physical implications of this discovery are.

We now include, in Extended Data Fig. 7, a detailed comparison with light curves of a sample of low-redshift SNe IIP, IIb, Ib/c, and Ia, as well as FBOTs that have available early-time UV observations required for comparison. These plots show that the light curves (and in some cases luminosities) of these examples are not good matches to our very early observations.

As the revised Extended Data Fig. 7 also shows, there are considerable uncertainties on the effective temperature at the first epoch, and the confidence interval overlaps with expected early-time temperatures of SNe. Since the object is in the Rayleigh-Jeans long-wavelength tail during the first epoch, its color will not be highly dependent on its effective temperature, which accounts for the large uncertainty. We provide the posterior distribution of temperature and luminosity parameters in Extended Data Figs. 1f, 1g, and 1h.

3. The surprisingly high rate estimate is said to be reduced when considering also the non-discovery of lensed supernova in the CLASH program. However it is not stated then how significant the inferred high rate is in this case (it is only stated that it is "still higher" than expectations - but not whether the difference is still statistically significant or not). In addition, it might be interesting to discuss the implications of the possible higher rate, both for discovery and for the physics leading to the expected rates.

Our discovery is of a SN that was missed by searches of imaging data of the HFF clusters. If we assume that no strongly lensed SNe were missed in searches of imaging of the CLASH survey, then our volumetric rate would fall by a factor of ~ 2 (still higher than blank-field results) and we obtain $p = 0.088$ in comparison with the Strolger et al. (2015) analysis, as we describe in the "Core-collapse SN rate from multiply imaged SNe" discussion in the Methods section.

While the Strolger et al. (2015) rates indicated a tension (too few SNe) with the SFR inferred from galaxy surveys at high redshift, given a Salpeter IMF, ours are in agreement. The improved agreement arises from both the higher rate of SN detections, and our identification of issues with the Strolger et al. (2015) dust prior and calculation of comoving volumes and distances (little h).

Additional comments:

*** It is not stated in the main text how the blackbody temperatures in lines 94-95 were inferred - whether this is from the shock cooling models, blackbody fits, or other method.**

The temperatures are inferred by blackbody fits, and we now make this clear in the revised manuscript. We thank the referee for the comment!

*** Line 114 refers to "model 1" and "model 2" but these names were never defined (in the rest of the paper, the models are given different names).**

We had intended "Model 1" and "Model 2" as hypothetical examples to explain the Bayes factor B_{12} . To make the presentation more clear, we now describe "two hypothetical models A and B."

*** Lines 116-117 state that the M20 RSG model gives a best fit radius of 577 R_{sun} , but later, in line 141 the M20-RSG model is said to give radii of 647 and 543 R_{sun} (one of these values is after altering the best fit lensing parameters by 1-sigma, but it is not clear which of them, and then what the other value refers to, and why both are different from the initial value mentioned in line 117).**

In the paper, we quote p16, p50, and p84 confidence intervals from the posterior. After an improved template subtraction, our principal constraint on the radius of $533^{+154}_{-119} R_{\text{sun}}$. The values shown in Fig. 3, for illustration, are $615^{+86}_{-76} R_{\text{sun}}$ and $509^{+74}_{-67} R_{\text{sun}}$ (lens parameters shifted by 1-sigma).

While all of these constraints are consistent, both the p50 value and uncertainties for the $615^{+86}_{-76} R_{\text{sun}}$ interval differ from those of our primary constraint, since we marginalize over the full uncertainties associated with the lens model (including covariance) for our primary estimate. The greatest likelihood (and probability, given uninformative priors) for the lens model parameters is not *exactly* the same favored by the lens model.

We have added the following sentence to the main text:

"The constraints in Fig. 3 are consistent with but shifted by a small amount from our primary estimate, since the SN model and photometry together favour larger magnifications (by $<\sim 1.2$ sigma) and smaller time delays (by $<\sim 0.5$ sigma)."

*** The paper favors the RSG model over the CSM models because "analyses of low-redshift hydrogen-rich SNe do not find evidence for such highly confined and dense CSM". But if the purpose of this discovery is to "study the massive stellar populations at high redshift" then projecting what we know at low redshift on to them defeats this goal completely.**

*** The paper favors the RSG model over other models because those are "not significantly better". Another way to phrase that is that the other models are equally good, in which case it is not clear why they are disfavored.**

We appreciate this point and agree, and we now have added the following sentence to the abstract,

"Highly confined and massive circumstellar material (CSM) at the same radius can also reproduce the light curve, but is unlikely since no similar low-redshift examples are known."

In the revision, we have also added to the discussion: "The CSM radius is smaller than the $\sim 1000 R_{\odot}$ inferred for a sample of SN-IIP explosions of RSGs (Morozova et al 2018), and the dense, compact P21-CSM model approximates the outer atmosphere of the M20 RSG model."

Both the RSG model and a model with a very massive ($\sim 1 M_{\odot}$) and confined ($\sim 500 R_{\odot}$) CSM can explain the observations.

We note the following: (1) The best-fit CSM mass is larger than the mass range shown in papers (e.g., Piro et al. 2021, Margalit 2021) describing the CSM models. (2) The combination of small radius and large mass for the CSM is uncommon in the local universe, as shown by Morozova et al. (2018; see, e.g., their Table 2). (3) The P21-CSM model does not have a well-constrained upper bound on the CSM mass, as shown by the distribution of MCMC samples from the P21-CSM model (Extended Data Fig. 5).

A reasonable view is that the CSM model is approximating the atmosphere of a red supergiant star of similar radius.

*** It would be interesting to know how the inferred mass and SFR of the supernova host compare to those of known supernova hosts of various types.**

Extended Data Fig. 2i shows the distribution of CCSN host-galaxy stellar mass and star-formation rate obtained from the Palomar Transient Factory (PTF) survey (Schulze et al. 2021), where the red star marks the stellar mass and the star-formation rate of the host galaxy of the newly discovered SN.

We find that the host falls on the "star-forming main sequence" of PTF host galaxies, and has a comparatively small stellar mass.

*** The color difference between the images is indeed striking, and is shown not likely to arise from noise. But it would be important to address (especially for non-expert readers) whether it could arise from different gravitational redshifts imparted on each image.**

In the revised Methods section, we now note "The earliest emission from an SN originates from the surface of the exploding progenitor. The differences between the photosphere's gravitational potential at the observed epochs are not sufficient to produce a detectable colour change

through HST filters. For example, based on Birkhoff's theorem, for a BSG with $R = 10 R_{\text{sun}}$ and $M = 20 R_{\text{sun}}$, the gravitational redshift of light from the BSG surface is 4.2×10^{-6} for a distant observer. Even for a compact white-dwarf progenitor with $R = 0.01 R_{\text{sun}}$ and $M = 1 M_{\text{sun}}$, the gravitational redshift is only 2.1×10^{-4} for light from the object's surface."

One might also wonder whether differences in the cluster's potential could produce a gravitational redshift. However, since the light from the background SN travels into and then out of the potential, the SN's light should not experience any net gravitational redshift.

*** Three R_V values are chosen but only one is shown in the figures. It might be useful to show the fits for all models tested, for all R_V values. Also, it is not clear why only one of the R_V values (4.05) is used in the rate simulations (line 393).**

The limit on the number of items in the Extended Data section of ten means that we are not able to plot for all the values of R_V . We have, however, uploaded the entire set of plots to the online repository for readers who would like to reproduce our results. For our estimate of the volumetric CCSN rate in our initial submission, we used a value of 4.05 following Strolger et al. (2015). In this revised manuscript, we also provide maximum-likelihood estimates for $R_V = 3.1$. As shown in Extended Data Fig. 8e, our inferences about the rate are not sensitive to the R_V choices.

*** It is not explicitly stated how the redshift of 0.375 was derived for the lens - whether it was through line identifications in the spectrum, a fit to the continuum only, or something else.**

It was measured spectroscopically, and we have now added a reference for the redshift of 0.375 (Struble et al. 1991), in the revised manuscript.

Minor issues:

*** In the first (bolded) paragraph, it is stated that "very early observations have been possible only for approximately a dozen nearby supernovae" - This is not exactly true. Early observations are *possible* for many more supernovae, and are routinely obtained by ZTF, for example. Perhaps it is more accurate to say that shock cooling analysis has so far been performed only approximately *several* dozen nearby supernovae.**

We now state: "Measurements of the progenitor's radius from early-time light curves exist for only a small sample of nearby SNe, and almost all lack constraining ultraviolet observations within a day of explosion.."

*** In line 191 there is a reference to "all peaks", but it is not clear what these peaks are of.**

We clarify this in the revised manuscript: "We next identified all peaks in the resulting difference images with flux exceeding a 3sigma signal-to-noise ratio."

*** It is not stated what the horizontal dashed lines are in Figure S-7.**

The horizontal green dashed line is at signal-to-noise ratio equal to 1. The horizontal black dashed line marks the zero flux. Owing to the length limit of the Extended Data, we omit the plots of the spectroscopic observations from the Extended Data, because none of the spectroscopic observations yields a significant detection of an emission line. We provide our reduced spectroscopic data in the online repository (<https://doi.org/10.5281/zenodo.6725770>) for readers to check the spectrum of the host galaxy we observed.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Thank you for thoroughly addressing my comments! I have no further comments and apologize sincerely for the delay replying. Congrats again on this exciting discovery!

Referee #2 (Remarks to the Author):

The authors thoroughly addressed the questions and comments of the referees. I am happy with accepting this paper for publication.

Referee #3 (Remarks to the Author):

I thank the authors for diligently addressing my comments. I find the results much more clearly presented now and appropriate for publication.

One remaining issue is that I could not find where the authors disfavor the BSG model. Indeed it has a lower Bayes factor vs. the null hypothesis compared to the RSG model, but the factor is still > 2 meaning it should be favored. Is this model indeed disfavored compared to the RSG or CSM models somewhere? If so, where (sorry if I missed it)? If not, why is the RSG model favored over it?

Author Rebuttals to First Revision:

Referees' Comments

Responses

Referees' comments:

Referee #1 (Remarks to the Author):

Thank you for thoroughly addressing my comments! I have no further comments and apologize sincerely for the delay replying. Congrats again on this exciting discovery!

Referee #2 (Remarks to the Author):

The authors thoroughly addressed the questions and comments of the referees. I am happy with accepting this paper for publication.

Referee #3 (Remarks to the Author):

I thank the authors for diligently addressing my comments. I find the results much more clearly presented now and appropriate for publication.

One remaining issue is that I could not find where the authors disfavor the BSG model. Indeed it has a lower Bayes factor vs. the null hypothesis compared to the RSG model, but the factor is still > 2 meaning it should be favored. Is this model indeed disfavored compared to the RSG or CSM models somewhere? If so, where (sorry if I missed it)? If not, why is the RSG model favored over it?

In the main text, we have rewritten the sentence

“When we assume that the CSM has a mass within $0.001\text{--}0.1\text{ }M_{\odot}$, representative of the values used by modelling efforts, there is very strong evidence for the M20 RSG model in comparison with the other models”

as

“When we assume that the CSM has a mass within a range of $0.001\text{--}0.1M_{\odot}$ that is consistent with those used by modelling efforts, the RSG model is significantly favoured in comparison to both the BSG and CSM models (with $\Delta \log B_{AB} > 2$).”

In addition, in the Methods section under “Bayes Factors”, we have added the following sentence to the end: “Since the Bayes factor is the ratio of the marginal likelihoods of two competing statistical models, the difference between two logarithmic Bayes factors corresponds to the models' logarithmic Bayes factor.”