Peer Review File

Manuscript Title: The 13CO-rich atmosphere of a young accreting super-Jupiter

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

The paper entitled "The 13CO-rich atmosphere of a young accreting super-Jupiter" by Zhang et al. details observations of an exoplanet candidate spectrum that shows signatures of a carbon isotope in the form of 13CO. So far, isotopes in the atmospheres of planetary mass objects have not been detected, but could hold important and unique information regarding a planet's formation history. The manuscript reports a detection and enrichment of 13CO, which could potentially be explained by an ice fractionation process. As such, this would be an important and novel detection, which could also motivate future efforts as new isotope aspect of chemistry could be explored in exoplanets. The results would be of wide interest to the community so nature is an appropriate journal. The overall quality of the data look to be high, and their retrieval methods and checks with cross correlation provide good evidence of the detection. Given the importance of the detection and the subtleness of the signal, it is important to run additional tests to help verify the detection. I have several suggested efforts which can help demonstrate the signal has multiple spectral features identified and/or is repeatable in their dataset.

In examining Figure 1 and Figure 2, there is only one feature I can clearly pick out by eye that is potentially due to 13CO, the feature on the left in Fig 1 (a) which is labeled and near 2.345 microns. The second strong 13CO feature on the right labeled on Fig 1 (a) near 2.374 microns is not so obviously seen in the figures, despite is being only a bit weaker than the other feature. In fact Figure 2 (left) seems to show the opposite where the residual flux plotted in black is opposite of the expected 2.374 micron 13CO model. Can 13CO be still be detected without this strongest feature near 2.345 microns? In this case, I think it's important to demonstrate that there are multiple spectral features detected in the data, such that the correct identification of the 13CO species does not all hinge on one feature. The current analysis both with the retrieval and cross-correlation can be run excluding the 2.345 micron range, this can help quantify the detection significance of multiple 13CO features being detected.

The data was obtained from two observing nights. Is the signal repeatable in both nights? Given the methods indicates the spectrum from night 1 was more noisy than night 2, it would be reassuring to see if a detection could still be made on night 1 alone, albeit at a lower significance. If night 2 is dominating the SNR, then perhaps the sub-exposures for each night could be used to show repeatably through the night. In addition to the test above, I suggest the authors run additional tests on sub-sets of the data to show the signal is repeatable. Presumably with a 6-sigma detection, the signal should still show up at about the 4-sigma level when dividing the data in half and analyzing it separately, so should give high confidence statistically.

The full retrieval model does not include many opacity sources. Can the 13CO feature at 2.345 microns be caused by any other absorber? Perhaps another molecule in the EXOMOL database also has strong features at the same wavelength. Additional tests/retrievals ruling out other potential molecules/absorbers could help show the uniqueness of the 13CO signal, and marginalize the over the possibility of other absorbing species. Showing 13CO is robust against other possible signatures would help lend credence to isotope being the only credible source of absorption. Such an additional analysis would have additional importance if only the 2.345 micron absorption feature was detected at high confidence.

Extended Table 1 needs to show the fit values and 1-sigma ranges as well as the priors. Also, the fitting statistics (chi^2, degrees of freedom, ect.) need to be given such that the reader can fully asses the statistical quality of the result.

Referee #2 (Remarks to the Author):

Measuring the isotopic composition of an accreting young exoplanet would indeed be very important and worthy of publication in Nature. However, I am not convinced of the reality of the detection. Figure 1 shows virtually no effect of 13CO. In the "a" inset panel, the reduced model seems to agree with the observations within 2-sigma at all wavelengths. The maximum deviation occurs just to the right of the 13CO bandhead, but that difference is less than two sigma based on the error bar shown on the Figure, and the difference between the full model and the reduced model is less than 1-sigma. Figure 2 is more convincing, as is the retrieval in Extended Data Figure 2. Nevertheless, I think the authors need to strengthen the case for detection of 13CO.

With regard to the reality of the detection, I have the following specific questions about the analysis:

1. Have the authors extended the cross-correlation shown on Figure 2 to a much larger span of wavelength (i.e, 2.10 to 2.42 microns)? Do other peaks appear at wavelengths that do not coincide with CO bands?

2. In spite of the statement in the paper that stellar contamination is negligible, I am not convinced of that. The first night was said to have "unstable atmospheric conditions", and on the second night, the seeing was as poor as 1.07 arc-sec. I think the authors need to demonstrate quantitatively that these poor atmospheric conditions did not result in stellar contamination. The star might well have enhanced 13C, and possible contamination could plausibly account for the entire apparent result in the paper.

3. The telluric spectrum shows strong absorption between 2.316 and 2.358 microns, and that covers the wavelength of the strongest 13CO bandhead. Imperfect telluric correction could masquerade as planetary 13CO absorption. Especially since atmospheric conditions were "unstable" on the first night, imperfect telluric correction seems plausible. Can the authors conclusively prove that the apparent planetary 13CO absorption is not due to imperfect telluric correction?

As regards the interpretation of the planetary spectrum, I am slightly concerned that the relative opacity in 12CO versus 13CO (and hence the conclusions concerning the 12C/13C ratio) may be biased by undersampling in the forward models, since the paper states that the models used every 10th point in high-resolution opacity tables. I realize that opacity sampling is an established technique, but the 10th point choice seems arbitrary. Do the models change with different sampling?

Author Rebuttals to Initial Comments:

We thank the referees very much for their careful reading of our manuscript and insightful comments. Their main concerns were about the reliability of the ^{13}CO signal, which inspired us to devise an improved analysis (involving an extra iteration between the planet-atmospheric modelling and telluric removal), and to do the extra tests suggested by the referees.

We detect the signal in both nights independently, and also independently for the two bandheads - of course at relatively low signal-to-noise. We also show that improper telluric corrections cannot cause the ^{13}CO signal.

In addition, we now also present new analysis of archival data of an easily-accessible young brown dwarf, serving as a proxy for super-Jupiters like TYC 8998 b. The ^{13}CO signal is very clearly present, as expected from our TYC 8988 b observations (albeit with a more typical $^{12}CO/^{13}CO$ ratio). It shows that these isotopologue features are readily observable, and their abundances can be assessed - providing intriguing implications for planet formation.

Below we go step by step through the comments of the referees.

Yours sincerely,

Ignas Snellen & Yapeng Zhang, on behalf of all authors.

Addressing the comments of Referee #1:

1. In examining Figure 1 and Figure 2, there is only one feature I can clearly pick out by eye that is potentially due to ^{13}CO, the feature on the left in Fig 1 (a) which is labeled and near 2.345 microns. The second strong 13CO feature on the right labeled on Fig 1 (a) near 2.374 microns is not so obviously seen in the figures, despite being only a bit weaker than the other feature. In fact Figure 2 (left) seems to show the opposite where the residual flux plotted in black is opposite of the expected 2.374 micron ^{13}CO model. Can ^{13}CO still be detected without this strongest feature near 2.345 microns? In this case, I think it's important to demonstrate that there are multiple spectral features detected in the data, such that the correct identification of the ^{13}CO species does not all hinge on one feature. The current analysis both with the retrieval and cross-correlation can be run excluding the 2.345 micron range, this can help quantify the detection significance of multiple ^{13}CO features being detected.

We improved our analysis by performing a new telluric correction after removing the best-fit planet atmospheric model, to repeat the atmospheric retrieval again. This has improved the fit, in particular at the positions of some strong telluric lines. We subsequently performed the proposed test by separating the wavelength region into two parts, 2.34-2.37 and 2.37-2.40 µm, then calculating the cross-correlation in each part. The cross-correlation results are shown in panel d of the Extended Data Figure 4. The CCF for both bands show peaks at zero velocity. While the signal from the second band is less significant, we think this is not surprising since the SNR of the data at the second band is lower due to stronger telluric absorption features. The residuals show some correlated noise as a result of the imperfect telluric correction (Extended Data Figure 5). We simulated the cross-correlation function by adding random noise to the ^{13}CO model and then correlating it with the noise-free version of itself, showing that we do indeed expect a lower signal from the second bandhead.

2. The data was obtained from two observing nights. Is the signal repeatable in both nights? Given the methods indicate the spectrum from night 1 was more noisy than night 2, it would be reassuring to see if a detection could still be made on night 1 alone, albeit at a lower significance. If night 2 is dominating the SNR, then perhaps the sub-exposures for each night could be used to show repeatably through the night. In addition to the test above, I suggest the authors run additional tests on sub-sets of the data to show the signal is repeatable. Presumably with a 6-sigma detection, the signal should still show up at about the 4-sigma level when dividing the data in half and analyzing it separately, so should give high confidence statistically.

We performed this test by doing the same analysis (both the retrieval and cross-correlation as detailed above) on the data of each individual night to check the repeatability of the detection. The ^{13}CO signal is present both nights at a lower significance, as suggested by the cross-correlation peaks in panel b of the Extended Data Figure 4. The retrieval results are shown in the Extended Data Figure 3. The data taken in both nights result in consistent constraints on the ^{12}CO/^{13}CO ratio and other parameters.

3. The full retrieval model does not include many opacity sources. Can the ^{13}CO feature at 2.345 microns be caused by any other absorber? Perhaps another molecule in the EXOMOL database also has strong features at the same wavelength. Additional tests/retrievals ruling out other potential molecules/absorbers could help show the uniqueness of the ^{13}CO signal, and marginalize the over the possibility of other absorbing species. Showing ^{13}CO is robust against other possible signatures would help lend credence to isotope being the only credible source of absorption. Such an additional analysis would have additional importance if only the 2.345 micron absorption feature was detected at high confidence.

In the retrieval models, we include the major molecular opacity sources (CO, H_2O, CH_4, NH_3) that play a role at the K-band. We inspected other potential absorbers given the expected temperature, including CO_2, HCN, and C_2H_2. None of them show significant features at the same wavelength region as ^{13}CO. Moreover, we performed retrieval analysis to marginalize over these absorbers. Adding these molecules to the retrieval model does not change the inferred value of any parameter. Therefore, it is not plausible that the detected ^{13}CO feature is caused by these molecules.

4. Extended Table 1 needs to show the fit values and 1-sigma ranges as well as the priors. Also, the fitting statistics (chi^2, degrees of freedom, ect.) need to be given such that the reader can fully assess the statistical quality of the result.

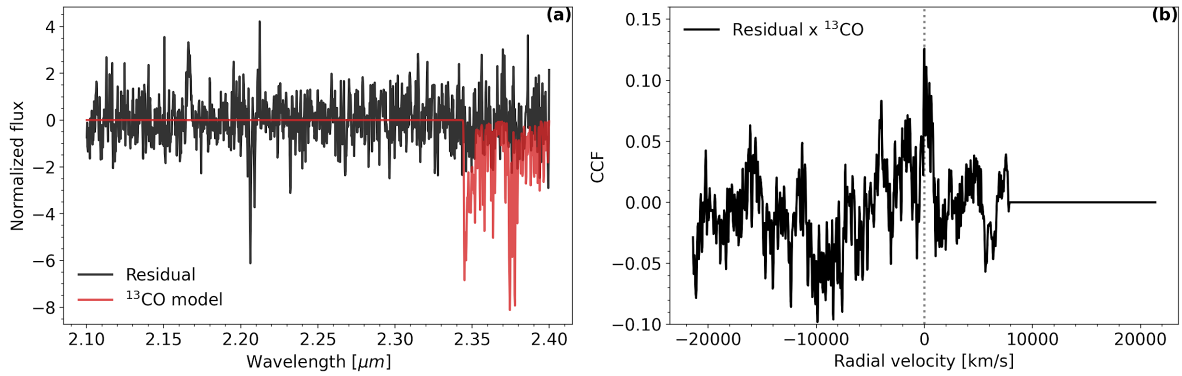
The statistics have been added to the Extended Data Tables.

Addressing the comments of Referee #2:

1. Figure 1 shows virtually no effect of ^{13}CO. In the "a" inset panel, the reduced model seems to agree with the observations within 2-sigma at all wavelengths. The maximum deviation occurs just to the right of the ^{13}CO bandhead, but that difference is less than two sigma based on the error bar shown on the Figure, and the difference between the full model and the reduced model is less than 1-sigma. Figure 2 is more convincing, as is the retrieval in Extended Data Figure 2. Nevertheless, I think the authors need to strengthen the case for detection of ^{13}CO.

The referee is correct in saying that Figure 1 provides very little evidence for ^{13}CO. Figure 1 is really meant as an overview-figure, showing the entire observed spectrum redward from the strong telluric CO_2 bands, not the actual detection of ^{13}CO (for which we have Figure 2). Differences per pixel are indeed less than 2-sigma. It is by adding up the signal over tens of pixels that we obtain the 6-sigma detection.

2. Have the authors extended the cross-correlation shown on Figure 2 to a much larger span of wavelength (i.e, 2.10 to 2.42 microns)? Do other peaks appear at wavelengths that do not coincide with CO bands?



The requested test is shown here. In the left panel the residuals of the observed spectrum (in black) and the ^{13}CO model (in red). The right panel shows the cross-correlation between the two over a very wide velocity range (20,000 km/s corresponds to ~0.15 μm; the 2.166 μm Brackett γ emission and 2.21 μm sodium lines are masked). The ^{13}CO signal gives by far the strongest peak.

3. In spite of the statement in the paper that stellar contamination is negligible, I am not convinced of that. The first night was said to have "unstable atmospheric conditions", and on the second night, the seeing was as poor as 1.07 arcsec. I think the authors need to demonstrate quantitatively that these poor atmospheric conditions did not result in stellar contamination. The star might well have enhanced ^{13}C, and possible contamination could plausibly account for the entire apparent result in the paper.

Sorry, we were not clear about this. What we meant here is that the stellar contamination is small, not that we ignored it. We did remove the background starlight when optimally extracting the planetary spectra. We fitted the flux of pixels outside the window of the target using a linear function and then subtracted the background at each wavelength. As for the poor seeing conditions, even though a few frames were taken under non optimal seeing, they barely contribute to the combined spectrum because we weighed the individual spectra according to the SNR when adding them up. The final spectrum is certainly dominated by the high-quality frames.

4. The telluric spectrum shows strong absorption between 2.316 and 2.358 microns, and that covers the wavelength of the strongest ^{13}CO bandhead. Imperfect telluric correction could masquerade as planetary ^{13}CO absorption. Especially since atmospheric conditions were "unstable" on the first night, imperfect telluric correction seems plausible. Can the authors conclusively prove that the apparent planetary ^{13}CO absorption is not due to imperfect telluric correction?

We checked this by cross-correlating the telluric model with the ^{13}CO model. There should be a cross-correlation peak (positive or negative) if an overcorrection or undercorrection had caused the ^{13}CO signal. This cross-correlation function is given in panel b of Extended Data Figure 5, showing no peak. To further check this, we artificially increase/decrease the strength of lines in our telluric model by 10%, 20% and 30%, and remove the tellurics from the observed spectrum to mock data with imperfect telluric correction (i.e. the tellurics either over- or under-subtracted from the spectrum). Then we perform the same cross-correlation analysis and see how the signal changes. If the signal is caused by imperfect telluric correction, we would expect the CCF signal to appear enhanced in these mock data. However, we saw the opposite that the signal is buried in stronger noise due to these imperfect telluric corrections. Based on this, we conclude that telluric absorption is not responsible for the detection.

5. As regards the interpretation of the planetary spectrum, I am slightly concerned that the relative opacity in ^{12}CO versus ^{13}CO (and hence the conclusions concerning the $^{12}C/^{13}C$ ratio) may be biased by undersampling in the forward models, since the paper states that the models used every 10th point in high-resolution opacity tables. I realize that opacity sampling is an established technique, but the 10th point choice seems arbitrary. Do the models change with different sampling?

We thank the referee very much for this point. The sampling of every 10th point turned out to be somewhat too aggressive. We benchmark the retrievals using different sampling (e.g. 10th, 5th, 2nd, and full). The 5th-point sampling provides the same results as the 2nd and the full table. Therefore, we changed it to 5th-point sampling.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I thank the authors for their revised manuscript, and especially for all the additional tests that have been included as requested. The brown dwarf analysis is also a welcome addition, as a detection in that object with higher SNR and resolution seems reliable. However, I still have significant doubts as to the reliability of the exoplanet 13CO detection.

In examining the spectra of night 1 and night 2, there are some potential 13CO features that do appear on both nights and are relatively free of telluric lines. Notably a trio of lines between 2.361 and 2.366 microns seem to match well night-to-night and with the 13CO model. A few other similar lines which match the model also appear in both datasets, making it fairly easy to see how a cross-correlation signal would generally match up. However, for both nights, the two very strongest expected lines at 2.345 and 2.374 microns do not show up in the data. If the spectral identification is correct, it is hard to believe the strongest lines on both bands would be hidden from view, while weaker lines would be more easily detectable. Note that the two strong 13CO lines at 2.345 and 2.374 microns do not seem to be directly overlaid by strong telluric lines. The strong 13CO feature at 2.378 microns does fall within a telluric line, and can perhaps be discounted, but it is still worrisome. If telluric lines are to blame for the decreased significance of the second banded, then it stands to reason telluric correction is still an issue at some level. Thus, with these tests I'm convinced there are some absorption features in the spectra, and they do look to be repeatable night to night. However, without showing the strongest expected 13CO lines are seen at the intensities expected compared to weaker lines, I still have doubts as to the correct spectral identification of the species.

For the test showing the extended CCF, it is true that zero-velocity is still the most significant, but it also seems worrying that there are other negative peaks that seem to regularly get to the -0.1 level which is not too much weaker than the 0.13 value.

Referee #2 (Remarks to the Author):

In response to my first point, the authors reply that Figure 1 is not meant to illustrate the detection, that is done is Figure 2. I agree that the cross-correlation in Figure 2 indicates a detection, albeit a weak one. But many readers will look eagerly at Figure 1 to see the detection, and the text encourages that ("...the difference between the two models is apparent at the 13CO bandheads around 2.345 µm..."). I suggest adding a remark to the caption, and clarifying in the text, that Figure 1 is meant as overview, and the actual detection is shown in Figure 2.

I thank the authors for their reply to my request to see the cross-correlation over a large swath of wavelength. Indeed, I am convinced that the positive peak at zero nails down the detection, not only because it's the largest peak (and positive), but also because it's at the expected location.

I am similarly satisfied with the reply to my points about possible stellar and telluric contamination, and I'm glad that my comment on undersampling in the forward model was useful.

Overall, this detection is weak but it is above the noise when averaged over wavelength, and it's physically reasonable. So I am now convinced that it is real. The detection of 13C in the brown dwarf adds to the plausibility and interest in this result. I recommend this paper for publication in Nature, but I repeat my suggestion above about clarifying Figure 1.

Author Rebuttals to First Revision:
We thank the referees for reviewing the second version of our manuscript, and for all the time and energy they have put into their reports.

About the main remaining concern of Referee-1: As we show in our manuscript, we detect 13CO at a significance of ~6 sigma. It is generally accepted that this is more than sufficient to claim a detection (often at 4-5 sigma). This statistical significance is derived from the atmospheric retrieval analysis, where the models with 13CO fit the data much better than models without 13CO.

The next thing is to somehow visualize this in the paper. This is not easy, because the signal does not originate from one or two features, but from two bandheads and a dozen or more lines – which combine to provide the 6-sigma signal. This automatically means that the individual 13CO lines should be barely visible in the spectrum. Some will show up, others will not, depending on how the random noise is distributed in the spectrum. This is indeed what we see. A cross-correlation then helps, because it is also a way to add signals from different parts of the spectrum. Indeed, a clear cross-correlation peak is seen with the expected auto-correlation structure. Hence, we have a quantitative assessment of the detection from the retrieval analysis, and a qualitative assessment from the cross-correlation signal.

Referee-1 is worried that "*the two very strongest expected lines at 2.345 and 2.374 microns do not show up in the data*". However, these lines individually are expected to be at about the level of the noise. Note that they are only barely stronger than the other features (one should look at the red curve of Fig. 2, not the blue curve in the previous version of the manuscript). They deviate from the best-fit model (i.e. the red curve) at about a ~1 sigma level. Indeed, renderings of the model + artificial noise often show spectra without the two lines visible. That the cross-correlation function over the very wide range (1300 pixels - see our first response letter) shows some other positive and negative peaks, another concern of Referee-1, does not worry us. It is not strange that on scale-lengths of ten-thousands of km/s the cross-correlation function shows low-frequency variations. Indeed, where the negative peaks are seen (-10000 km/s) the cross-correlation function is negative for thousands of km/s. Other bumps can be caused by poorly removed telluric lines – all very far away from the 13CO bandhead regions. The cross-correlation function we show in the manuscript (overall a smaller, more relevant range) is for qualitative purposes only. It has the shape we expect and peaks at the expected position. The quantitative, statistical assessment comes from the retrievals. Note that all high-resolution spectroscopic observations of exoplanet atmospheres (e.g. Ehrenreich et al. Nature 2020; Hoeijmakers et al. Nature 2018; Brogi et al. Nature 2012) rely on combining many lines without detecting them individually.

In this newest version of our manuscript we also have transferred some of the more technical bits of the main text to the Method Section to comply to the ~1500 word limit. In addition, we have incorporated the remaining comment by Referee-2 on the relation between Figs. 1 and 2. In this new version we have merged the two figures and simplified it a bit, so that (we think) this is no longer an issue.

Referee 2's comment on referee 1's report:
The argument in favor of the detection is that it's based on a cross-correlation, integrating over all of the 13CO features rather than detecting individual lines. Since the detection is weak even in the cross-correlation, some individual lines being hidden by noise doesn't bother me. Indeed, if all individual lines were clearly seen (or even just the strongest ones) then the S/N for the cross-correlation would be much higher than it is. So I don't expect to see individual lines, and it seems plausible to me that noise could mask the stronger lines while making weaker lines seem strong. ("Noise is noisy").

But having argued that, I think the authors should be able to prove that point, i.e. demonstrate quantitatively that not seeing some individual lines is consistent with the S/N indicated by the cross-correlation. "

Referee 1's comment on referee 2's report:
The cross correlation technique he made work for CO in HD209458b has been used widely since then, and has given the community overall solid results. But the method is not bullet proof, there have been cases in the literature of signals not being repeatable night-to-night, or signals of that are a bit hard to believe physically.

Big advantages for the cross-correlation technique in close-in transiting planets is that there are several consistency checks that tie the atmospheric signal to the planet very convincingly. For one, the signal occurs at the correct system velocity, and two you can often see the Doppler trail move with the known planet orbit over a given night. In these cases, the signal more or less has to be coming from the planet.

In this case with a directly imaged planet, there’s no possibility to see the Doppler signal move with the orbit. There is also the complication that the 13CO spectral features are very very similar to 12CO, especially at the relatively low resolution of 4500 of this data. So the 12CO has to be very accurately modeled and removed. If there are 12CO residuals, then cross correlating with the similar looking 13CO model could pick up those residual 12CO signals at the same zero velocity, and show up as a significant fit. Thus, while the signal is seen at zero velocity and is certainly a good check, I don’t find that as convincing as other IR cross-correlation measurements.

My thinking is that the atmospheric models from the spectral retrieval could be preferentially fitting the strongest highest SNR CO lines very well (constraining that part of the atmosphere well), but not doing as good a job on the weaker noisier CO lines which could be sensitive to different atmospheric regions which could change the intensities slightly in ways the model may have a hard time picking up. Thus the strongest lines could be well modeled/subtracted away completely with the 12CO model, but leaving traces of the weaker lines which the atmosphere model uses 13CO to further correct, even if the signal is mostly all still 12CO. Granted the data are noisy and noise is certainly noisy, but this concern is quite consistent with not seeing signals in the two strongest lines for both nights.

I think the authors should prove their point that not seeing individual lines (especially the stronger ones) is consistent with the S/N of the cross-correlation. In addition, to help elevate my concern above, in the observed residuals, the weaker lines could be also be shown to cross-correlate significantly better with 13CO than a further 12CO model (which is given some further flexibility to account for some modeling uncertainties).

Also, for the previous tests done (e.g. individual band heads), it would be good to state the confidence level associated with those tests. It’s otherwise quite hard to associate a CCF to a confidence level.