Peer Review File

Manuscript Title: At least one in a dozen stars exhibits evidence of planetary ingestion

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors present a detailed analysis of 125 co-moving stellar pairs (a combination of co-natal and just co-moving) searching for signatures of planet engulfment. In addition to the usual detection signature in the form of trends in abundance differences as a function of condensation temperatures, the authors present a new religious selection based on Bayesian statistics, which can differentiate between abundance differences resulting from atomic diffusion and planet engulfment. With this substantial dataset and new analysis method, they determine the fraction of stars to have engulfed a planet to be 8%

This is a comprehensive and detailed study, presenting for the first time a statistically significant dataset to investigate the question of planet engulfment. The work is well presented and is a significant and important step towards understanding the formation of planetary systems. I have some comments (see below) for the authors to consider, and following satisfactory completion, I recommend the paper for publication.

Major points:

1) I don't completely understand some aspects of the control sample and its use.

- The control sample is comprised of 34 co-moving but not co-native stars. What is the purpose of them being co-moving? I assume this was just stars in the initial sample, which the authors found that had separations larger than the co-native range. I think it is more important to emphasize the similarities of stellar parameters between pairs in this sample rather than the fact that they are co-moving. Also, it should be noted whether this sample covers the same parameter space as the co-native sample.

- Some stars in the control sample may also have engulfed planets, actually 2.7(8%) of them according to the results of this paper, but that doesn't seem to be taken into account in the analysis. I realize that a planet engulfment signature is weak and hard to find for non-co-native stars, but it seems to be assumed that no stars in the control sample carry this signature.

- It is mentioned that the abundance differences seen for the control sample stars are due to atomic diffusion and inhomogeneous interstellar medium (ISM). How will an inhomogeneous ISM result in a trend in abundance difference as a function of condensation temperature? I suppose it can create single false positives. Still, the distribution for this diagnostic (Tcondensation) is the same in the two samples, which seems unlikely if inhomogeneous ISM is the main driver of abundance differences for the control sample. Hence, does this suggest that atomic diffusion is the main driver?

- In extended Fig. 2 (page 22), the control sample doesn't overlap with the mock noise sample, suggesting that the abundances of the control sample carry a signature of one or more processes, which could be atomic diffusion. It would be useful to see a mock sample with this signature, see where those stars lie, and how distinguishable they are from the planet engulfment sample.

- Extended Fig. 3 In this figure, it looks like the abundance differences of the co-moving pairs are more affected by atomic diffusion than the co-native pairs. What is the reason for this? As I understand, the stars are selected in a similar manner, so even if the control sample pairs were not born from the same cloud, they can be assumed to be in the same evolutionary state, the difference mainly being different chemical birth compositions.

2) I would like to see demonstrated that the difference in stellar parameters of the co-native pairs doesn't induce differences in absorption line strengths on a level important for this study.

- L.592 Say the majority of the pairs have Teff_diff < 300 and logg_diff < 0.3. How much does the equivalency width of a given absorption feature change with a 300K and/or 0.3dex change in Teff and logg? Is this smaller than the planet engulfment signature or of the same size?

3) It is unclear to me how much this result depends on the earth composition being used for the dynasty models.

How much can the planet composition be varied in the dynasty model? And how were different options tested? L.677 briefly mentions a combined composition, including carbonaceous chondrites, but it is not clear to me how large the change was in composition and if other compositions could have been tested. My main question is, how much does this result depend on the planet competition being used? I would assume that previous results using only the condensation temperature diagnostic do not depend that much on the composition of the engulfed planet.

4) Is there an overlap between the sample presented here and the literature sample of stars with planet engulfment signatures? And would the literature candidates be selected as candidates with this method?

Minor points:

Fig. 1: I suggest making a panel where the two spectra are plotted on top of each other and provide the parameters of the two stars in the caption.

L.113: What is the likelihood of both stars engulfing a planet, and would this be detected? Are there any statistics on binary systems with planets - do both stars have planets, or do the planet usually orbit just one of the stars, or both?

Fig. 2: Please comment on the fact that Al and Na seem to be outliers here. Is that always the case, or just in this star? Why, and what influence does it have on the fit?

Fig. 3a: I suggest changing the color scheme so white is co-native fraction = 100%.

L.228: What is the distribution of |T_cond slope| for the seven pairs that fulfill all criteria? Is it the 7 with the largest slope, or are they evenly distributed? i.e., would a more stringent cut on |T_cond slope| help to identify these systems?

L.604: is isotopic split included for Ba, and with what isotopic ratios?

Extended Fig. 1: Sulfur seems to exhibit large variations between telescopes/instruments. What is the cause of these, and should this warrant an exclusion of this element? Also, for pair 38 and pair 108, many of the points lie outside the one sigma range and all to either plus (pair 38) or minus (pair 108). Which telescope pairs are these, and does this influence the results?

L.681: It is unclear if an uncertainty on Fcz is computed and included in the Bayesian statistics. If not, please comment on this.

L.735: The atomic diffusion models from Dotter et al., 2017 are used. Are there other models in the literature that could have been used, and how would that influence the results?

Does the fact that the sample consists of co-native binary stars affect the ratio of planet engulfments? i.e., is it possible to say something about whether planet engulfment is more or less likely to happen in a binary system?

Thermohaline mixing is mentioned very briefly at the end of the text. Please expand a bit on how this might influence the result.

Referee #2 (Remarks to the Author):

This interesting work describes a detailed differential study of the chemical compositions of stars formed from the same molecular cloud. Since these stars are still on the main sequence, one would expect the surface composition to be the same for the two stars in each pair, inherited from the common molecular cloud they were born in. However, in some cases they find subtle differences which they compare to models based on three scenarios: the accretion of Earth-like material, the abundance variations associated to atomic diffusion, and the situation in which the abundances are identical (the null hypothesis).

Based on Bayesian statistics the study finds there are changes between the compositions of the stars in the same pair consistent with the accretion of planetary material in 7 cases out of 91, which leads the authors to conclude that 8% of the stars suffered such accretion.

The paper is well written, and the methodology well explained in general. The figures are appropriate and easy to understand.
There have been similar studies before, referred to in this paper, but this new study improves the consistency in the analysis methodology, the methodology itself, and enlarges significantly the sample. This, in addition to the interest of the topic for the public, makes me recommend the paper for publication in Nature.

However, there are a number of issues I cannot understand well and I think need to be addressed before giving the green light.

1- The authors point out in the very first paragraph that the 'ingestion of planetary material' is equivalent to 'planet engulfment'. I do not agree. The first term describes, in my view, much more accurate what they are measuring in the sample stars. All the material consistent with the addition of Earth material to the star could come a debris disk around the star, and not from an actual planet.

2. Not only one cannot distinguish accretion of actual planets from debris from a (proto)planetary disk, but one cannot distinguish 'adding' planetary material from 'not subtracting' it to form other planets. In fact, in the study the authors compare their results to reference [24], where the differences between the Sun and solar twins are attributed to planet formation rather than planet accretion.

I think these issues should be acknowledged upfront, and the title modified accordingly.

3. I'm not sure I understand the quoted average slope of abundance changes with condensation temperature. If the changes are for star A relative to star B (and the designations are interchangeable) I would naively expect the average for the sample to be 0.0 (or any number smaller than the uncertainty). What am I missing?

4. The criteria ii and iii described in page 5 appear to be arbitrary, chosen based on the distribution of the actual sample under analysis. Any changes in the parameters chosen will directly affect the inferred statistics on the rate of stars with accreted planetary material. Can you offer some idea on how much your inferred rates will change with reasonable changes in the parameters adopted for these criteria?

In addition, I have a number of issues/suggestions for the authors to consider:

-page 3, in would be useful to include here the ranges in Teff,logg and Fe/H for the stars in the sample. I know the information is in the additional material, but I consider such information critical (at least the range of spectral types).

-figure 1 is presented as evidence of abundance variations, but subtle changes in Teff can mimic these variations due to different line excitation energies. Can you mention/show the deltaTeff for these two stars and the theoretically predited change it implies?

-not all planets have the same composition of Earth. One could argue, however, that Earth-mass planets are more frequent than giant gas planets and the latter will probably leave too small a signature to be detected.

-page 6, paragraph 2: a strictly differential analysis works well for solar twins (or twins in general) but your stars are not, so you have to worry about parameter determination, not just differential abundances. I'm missing a description (in the supplementary material) of how those parameters are derived so that systematic errors are minimized.

-page 6: you mention reference [26] finding a correlation betrween the chemical anomalies and the stellar mass. Do you find such correlation?

-methods, sample: a few of your stars must have been included in detailed studies in the literature. How well your parameters agree with those?

-methods: the line-list adopted ... is avaiable upon request ...
also later a similar statement is made about the data. This is not acceptable
by Today's standards for reproducibility. The data and tables should be made
available in a digital repository, such as Zenodo.

-methods: presumably the model atmospheres, synthesis code, etc. apply as well to the parameter determination, not just the abundance inference. Please, state clearly or explain.

-methods: One of the authors has written papers in which a dependence of the strenght of the O triplet with stellar activity was reported. Does that deserve a comment in this context?

-methods, page 14: it is true that the agreement suggests the differential analysis has effectively suppresed systematics, but other systematics due to imperfect modeling must be present, given the deltaTeff can be as large as 300 K. This should be acknowledged, as such errors are not accounted for in the description/analysis (3D NLTE effects are only accounted for in a particular case).

-methods, line 661: Beyesian -> Bayesian

-methods: not everyone agrees on the physical description of diffusion. How different are the recipes adopted in MESA from alternative models? Are they similar or span a wide range that could affect the conclusions of this paper?

Referee #3 (Remarks to the Author):

The manuscript "At least one in a dozen stars has engulfed planets" is an interesting contribution to the literature which aims to constrain the occurrence rate of planet engulfment events using chemical differences in binary star pairs. Seven newly identified candidate systems for planet engulfment are found from a sample of 91 close binary pairs, substantially increasing the number of such systems known and providing an occurrence rate estimate of 8% for these events.

This paper advances the field by employing a larger sample of binary pairs than previously used for such studies, and by developing a new statistical methodology for distinguishing planet engulfment signatures from atomic diffusion. The goals and general methods of the paper are similar to previous works (Spina et al 2021, published in Nature Astronomy; and Behmard et al 2023, published in MNRAS), but comparison to these studies is appropriately addressed in the text. In particular, this work employs a much larger set of elemental abundances than the Spina paper, which allows for more confident detection of planetary signatures; and it uses a much larger sample size and more precise measurements than the Behmard paper. For these reasons, I would regard this work as a significant advance to the field.

The data set used in this work is excellent: by selecting co-moving stellar pairs from Gaia with a reproducible selection function, the authors have achieved an unprecedentedly large and complete sample of twin-star binaries. The analysis techniques employed to derive abundance measurements are appropriate and the precision achieved is well-justified with validation tests.

The statistical methodologies are generally sound. The Bayesian analysis developed in this work is an improvement over traditional methods (e.g. fitting a line to the abundance vs condensation temperature relation, which can only provide a rough metric for planetary material). A few specific suggestions for improvement are made below.

I have some reservations about the robustness of the conclusions as presented in this work; in particular, I think the title overstates the certainty of the results. As discussed in the final paragraphs of the main text, the "planet signatures" found are still ambiguous in their cause: alternative models to engulfment exist, including dust trapping as a natural byproduct of the planet formation process (Booth & Owen 2020). Furthermore, the analysis tests only three models to explain the observed abundance differences (planets, atomic diffusion, and the null hypothesis of the stars having identical compositions), which is not an exhaustive list of potential causes for differing abundances. Thus, while the evidence is promising, it seems premature to strongly conclude that these systems must have ingested planets. The engulfment hypothesis could be tested further by examining lithium signatures (as done in Spina et al 2021). Knowledge of the planetary systems (or lack thereof) around these stars would also be a valuable constraint for future work.

Specific suggested changes:
1. Line 62: "conservative occurrence rate": I would just call this the estimated occurrence rate. In this context "conservative" implies that the number is calculated as a lower limit.
2. Lines 68-81 (second paragraph): This introductory section seems like it would not read very clearly to a reader who is not an expert in the concepts being discussed. Specifically: "Ingestion of planetary material… will increase the elemental abundances of the host star" does not clearly convey that the refractory elements will increase more than volatile, which in turn obscures why it

matters that “elemental abundance differences as a function of the dust condensation temperature… were reported.” In the next sentence, it’s not obvious what “the signal” refers to. Some edits to clarify these points would make the intro more effective.

3. Lines 110-113: Cite a justification for the cutoff at 10^6 AU?

4. Lines 214-217: I don’t think any conclusions can be drawn about the prevalence of atomic diffusion among the far co-moving pairs. The distant pairs are not co-natal and therefore the null hypothesis (that they have the same abundances) is not applicable.

5. Lines 222-226 (& methods generally): Another available metric is the confidence with which the ingested planet mass is non-zero. This should already be available because the Bayesian method results in a posterior predictive distribution on the parameter M_E. You might impose a cut that requires that M_E be equivalent to 3 sigma away from zero. This cut, which frames the problem of detecting planet ingestion in terms of parameter estimation rather than model selection, would be more robust to inaccuracy in the Bayesian evidence calculations (which are notoriously difficult – see e.g. Nelson et al (2020), AJ 159, 73 for one discussion of uncertainty in evidences and their impact on Bayesian model selection).

6. Lines 265-268: The assertion that a different analysis technique applied to the Spina+2021 sample would result in “a much smaller number of engulfment instances” is speculative, based on the assumption that the occurrence rate found in this work is correct and therefore other work should be revised to match it. I don’t think this is an appropriate speculation to make, and would recommend cutting this sentence.

7. Figure 4: Is panel A really needed? The evidence for a given model without comparison to any alternatives does not seem to contain useful information.

8. Lines 596-598: How were the equivalent widths measured? (e.g. fitting Gaussians with IRAF, an automated code like ARES, etc)

9. Lines 668-678: It is confusing that M_E, a key parameter of the model, does not appear in these equations. I recommend adding another equation to show how M_{X,E} is calculated from M_E. Relatedly, the description of M_{X,E} as “the total mass of element X” seems incomplete– perhaps it should be “the total mass of element X introduced by planet ingestion.”

Author Rebuttals to Initial Comments:

Response to the referees

We thank the three referees for their careful, constructive and generally positive comments.

Referee #1 (Remarks to the Author):
The authors present a detailed analysis of 125 co-moving stellar pairs (a combination of co-natal and just co-moving) searching for signatures of planet engulfment. In addition to the usual detection signature in the form of trends in abundance differences as a function of condensation temperatures, the authors present a new religious selection based on Bayesian statistics, which can differentiate between abundance differences resulting from atomic diffusion and planet engulfment. With this substantial dataset and new analysis method, they determine the fraction of stars to have engulfed a planet to be 8%

This is a comprehensive and detailed study, presenting for the first time a statistically significant dataset to investigate the question of planet engulfment. The work is well presented and is a significant and important step towards understanding the formation of planetary systems. I have some comments (see below) for the authors to consider, and following satisfactory completion, I recommend the paper for publication.

Major points:
1) I don't completely understand some aspects of the control sample and its use.
- The control sample is comprised of 34 co-moving but not co-native stars. What is the purpose of them being co-moving? I assume this was just stars in the initial sample, which the authors found that had separations larger than the co-native range. I think it is more important to emphasize the similarities of stellar parameters between pairs in this sample rather than the fact that they are co-moving. Also, it should be noted whether this sample covers the same parameter space as the co-native sample.

Response: Our original purpose includes confirming the co-natal boundary, which is found to be ~10^6 AU (Nelson et al. 2021; Yong et al. 2023). In addition, this will ensure that we select the whole co-moving sample homogeneously without systematic bias. We agree with the referee that it's more important to clarify the similarity and difference between the co-natal and the control sample. The control sample does cover a similar parameter space as the co-natal sample as stated below:
Parameter range for far pairs: Teff: 5300 - 6400 K; logg: 3.9 - 4.6 [cgs]; [Fe/H]: -0.3 - 0.3
Parameter range for close pairs: Teff: 5000 - 6500 K; logg: 3.9 - 4.6 [cgs]; [Fe/H]: -0.6 - 0.4
However, the control sample generally has relatively larger Teff differences within each pair (e.g., <deltaTeff>_far = 236 K with a standard deviation of 123 K; <deltaTeff>_close = 138 K with a standard deviation of 117 K) as well as larger abundance differences (e.g., <delta[Fe/H]>_far = 0.08 dex with a

standard deviation of 0.11 dex; <delta[Fe/H]>_close = 0.02 dex with a standard deviation of 0.04 dex) due to the fact that they are not co-natal. They are selected with the same manner and criteria. However, the stars in a given far-apart pair (unlike those in a ‘true’ co-natal pair) just naturally seem to be much less similar to each other. This also agrees with the wide binary sample in El-Bady et al. (2021). We have commented on this briefly in the Methods at L.504-508.

- Some stars in the control sample may also have engulfed planets, actually 2.7(8%) of them according to the results of this paper, but that doesn't seem to be taken into account in the analysis. I realize that a planet engulfment signature is weak and hard to find for non-co-native stars, but it seems to be assumed that no stars in the control sample carry this signature.

Response: We agree that there could be potential candidates with planet signatures in the control sample. But it’s hard to disentangle them from other sources (e.g., atomic diffusion and ISM noise) from our indicators (Tcond trends or deltalnZ). In addition, these pairs are so far apart that the basic assumption that they had initially the same composition might not work anymore. Therefore we focus on the co-natal sample in this study, which enables us to draw the conclusion with confidence. The control sample mainly helps us to assess the selection criteria and implies that relying solely on the commonly used indicator T_cond trends may lead to false positives. We have emphasized this before discussing the selection criteria in the main text at L.159-161.

- It is mentioned that the abundance differences seen for the control sample stars are due to atomic diffusion and inhomogeneous interstellar medium (ISM). How will an inhomogeneous ISM result in a trend in abundance difference as a function of condensation temperature? I suppose it can create single false positives. Still, the distribution for this diagnostic (Tcond) is the same in the two samples, which seems unlikely if inhomogeneous ISM is the main driver of abundance differences for the control sample. Hence, does this suggest that atomic diffusion is the main driver?

Response: There is no easily accessible model for ISM inhomogeneity for us to test and compare quantitatively. Thus we can only assess the impact qualitatively. We agree with the referee that it is less likely for ISM noise to mimic abundance trends, and atomic diffusion is the favoring driver for the control sample (due to larger Teff differences). It is still possible that ISM diffusion can correlate with atomic number because lighter elements (CNO) are generated more homogeneously through stellar wind, but heavier elements are from clustered SNe. Indeed, both ISM diffusion and especially atomic diffusion correlate with atomic number, which can affect the Tcond trend. We can only state this briefly at L.152-158 without further discussion due to the word limit.

- In extended Fig. 2 (page 22), the control sample doesn't overlap with the mock noise sample, suggesting that the abundances of the control sample carry a signature of one or more processes, which could be atomic diffusion. It would be useful to see a mock sample with this signature, see where those stars lie, and how distinguishable they are from the planet engulfment sample.

Response: We appreciate the referee’s suggestion. However, we find that in the quoted figure (now Extended Fig. 3), the control sample still has some overlap with the mock noise sample (unlike the co-natal sample that doesn’t overlap). We would like to note that the mock noise sample should not necessarily have overlap with either the control nor the co-natal sample. In addition, we cannot directly plot the mock atomic diffusion sample to the current figure because the original sample is based on 21 elements, while the atomic diffusion models are only available for 8 elements. The Bayesian evidence will be scaled with the number of elements adopted. That’s also the reason we didn’t plot the distributions of relative Bayesian evidence in the same panel (e.g., current Fig. 2). Therefore we consider that it may be beyond the scope of the current work because our main conclusion is based on the co-natal sample (L.159-161), while we tend not to draw a conclusion from the control sample because the key assumption for this study (i.e, co-natal stars born with the same composition) does not apply anymore. It might be worth pursuing in further detail in a future work, but it will not affect our main results and conclusion.

- Extended Fig. 3 In this figure, it looks like the abundance differences of the co-moving pairs are more affected by atomic diffusion than the co-native pairs. What is the reason for this? As I understand, the stars are selected in a similar manner, so even if the control sample pairs were not born from the same cloud, they can be assumed to be in the same evolutionary state, the difference mainly being different chemical birth compositions.

Response: The reason is possibly due to larger differences in relative stellar parameters (e.g., deltaTeff) in the control sample. The control sample may consist of many pairs born with different chemical compositions, and can have different evolutionary states. Although we selected these stars in the same manner (in terms of colours and colour differences), the stars in a given far-apart pair (unlike those in a 'true' co-natal pair) just naturally seem to be much less similar to each other. This also agrees with the binary sample in El-Bady et al. (2021). We have clarified this with the revision stated above and added: "(possibly due to larger differences in the relative stellar parameters such as Δ\teff)" in the caption (L.1093-1095) of the quoted figure (now Extended Fig. 4).

2) I would like to see demonstrated that the difference in stellar parameters of the co-native pairs doesn't induce differences in absorption line strengths on a level important for this study.
- L.592 Say the majority of the pairs have Teff_diff < 300 and logg_diff < 0.3. How much does the equivalency width of a given absorption feature change with a 300K and/or 0.3dex change in Teff and logg? Is this smaller than the planet engulfment signature or of the same size?

Response: We would like to emphasize that the co-natal pairs are not assumed to share the same stellar parameters in terms of Teff and logg. During the parameter determination, the EW differences will naturally lead to parameter differences, which have been properly accounted for when transferring the EWs to abundances of individual lines (via curve of growth) to establish the excitation/ionization balance. The details of differential parameter determination have been published in Yong et al. (2023). We limited the details due to the word limit.
For differential analysis, the uncertainties in Teff_diff and logg_diff are: ~16 K and ~0.03, same for the pairs of stellar twins (Teff_diff < 200K & logg_diff < 0.2) and the pairs with 200K < Teff_diff <= 300K and 0.2 < logg_diff <= 0.3. The uncertainties in stellar parameters have also been properly accounted for uncertainties in abundance differences.
In addition, in our Bayesian analysis we also included a parameter 'sigma_scatter' to account for extra systematic errors from e.g., (3D) NLTE effects due to relatively larger parameter differences. The intrinsic scatter is basically smaller than 0.01 – 0.015 dex, much weaker than the planet signals detected in e.g., the 7 reported systems (~ 0.05 – 0.15 dex). In addition, it can only mimic the signal if the systematics also correlate with the Tcond trend, which is not the case because the uncertainties only largely correlate with the total usable lines. We have added a sentence to emphasize the magnitude of planet signals detected in our seven new systems at L.216-218 in the main text, and refined the sentence at L510-512 and L.616 in the Methods to address the above points.

3) It is unclear to me how much this result depends on the earth composition being used for the dynasty models. How much can the planet composition be varied in the dynasty model? And how were different options tested? L.677 briefly mentions a combined composition, including carbonaceous chondrites, but it is not clear to me how large the change was in composition and if other compositions could have been

tested. My main question is, how much does this result depend on the planet composition being used? I would assume that previous results using only the condensation temperature diagnostic do not depend that much on the composition of the engulfed planet.

Response: We mainly tested the planetary ingestion model with bulk Earth composition. Previously we also tested it with the combination of bulk Earth and CI chondrites but found no clear difference (< 5%) in terms of the relative Bayesian evidence (ln(Z) & deltaln(Z)). We could certainly test more cases for different type of planets with different compositions but there is no certainty of how realistic these are because we know very little and don't have accurate models for other types of planets (e.g., Jupiter, Saturn, Neptune etc.), even in the solar system. The tests made in this work are probably the best we can achieve at this stage. We agree that we can not exclude that a small number of pairs may carry a signature different to the abundance pattern using bulk earth composition but they can't be tested quantitatively until we have realistic planet compositional models in future. That's part of the reason we conservatively present our occurrence rate and use the words "At least" in the title. In addition, although Tcond trend is independent of "models", it is also subjected to more false positives as we discussed above (can be driven by atomic diffusion, and is very sensitive to a few elements, such as C and O). We have briefly commented about the test with the combination of bulk Earth and CI chondrites at L.633-634 in the Methods. We could not expand the discussion due to the word limit.

4) Is there an overlap between the sample presented here and the literature sample of stars with planet engulfment signatures? And would the literature candidates be selected as candidates with this method?

Response: Our sample contains 14 common pairs with previous literature (Nagar et al. 2020; Spina et al. 2021) that have been analyzed with high-precision spectroscopy. Most of them (9 out of 14 pairs) were found to have no clear abundance difference, and 5 pairs were claimed to be chemical anomalies (but not necessarily carrying planet signatures). Our results agree with the literature, if applying the same criteria for chemical anomalies (see also Yong et al. 2023). Two common pairs (69 and 38) show potential planet signatures and one is confirmed in this study. Another common pair (HIP 34407/34426) was reported to show planet signature based on the Tcond trend in literature (Ramirez et al. 2019; Nagar et al. 2020). This pair was not selected as a candidate in this study. Although our study confirms that this pair shows large abundance differences (~ 0.16 dex) and significant Tcond trend (9 * 10^{-5} dex/K), it can be explained by the null hypothesis with deltaln(Z) near zero. We emphasize again that relying solely on the Tcond trend may lead to false positives, because it can be driven by random abundance variation or a few elements (especially C and O). Indeed, the Tcond trend for this pair only exists if we fit for the whole Tcond range, and it will be much weaker if we set the fitting range to be > 500K: (2.2+/-1.3) * 10^{-5} dex/K, or > 1000K: (3.5+/-3.0) * 10^{-5} dex/K, further supporting our claim. In addition, one common pair (HD 133131A/B) was compared with previous studies and showed no systematic differences (Extended Fig. 2b). A future study to re-visit all the previous binaries (and even solar twins) using the new Bayesian approach would benefit this field. However, due to the word and page limits, we cannot add the above discussion to the manuscript, but only commented briefly about comparison of stellar parameters of these 14 pairs at L.512-515 in the Methods (suggested by Referee #2).

Minor points:
Fig. 1: I suggest making a panel where the two spectra are plotted on top of each other and provide the parameters of the two stars in the caption.

Response: We have made the suggested change for Fig. 1 (now Extended Fig. 1) and provided the stellar parameters of the two stars in the caption (L.949-950).

L.113: What is the likelihood of both stars engulfing a planet, and would this be detected? Are there any statistics on binary systems with planets - do both stars have planets, or do the planet usually orbit just one of the stars, or both?

Response: Given the low ingestion rate we find here, the chance of both stars engulfing a planet would be very low (i.e. ~8% * 8% -> ~0.6%), and most likely the signatures will be canceled out or weakened significantly. The occurrence rate in (wide) binary systems can be similar to that in the single stars (Hirsch et al. 2021). It would be less common for planets orbiting around both stars to our knowledge. In addition, we do not see any trend with pair separation (delta_s), which may indicate that the rate is not affected by pair separation (for wide binaries with delta_s >~ 1000 AU). We added a relevant sentence at L.148-149 in describing the Tcond trends. But we think it might be too early to make any claim regarding occurrence rate in single stars vs. binaries.

Fig. 2: Please comment on the fact that Al and Na seem to be outliers here. Is that always the case, or just in this star? Why, and what influence does it have on the fit?

Response: This is only the case for a few pairs and we believe it's just coincidence (especially for Na). The possible reason could be: (1) there are only two Al lines available, making the measurement more uncertain; (2) occasionally the spectra lines may be affected by bad pixels or asymmetry in the wings that they are not well measured. The influence on the results is minor. For example, we find the numbers for Pair 124 in Fig. 2 to be 12.4 (instead of 13.8) * 10^{-5} dex/K for the slope of the Tcond trend, and 11.7 (instead of 12.3) for the Bayesian indicator deltaln(Z), if both Al and Na are excluded. We note although the outlier elements can exist for some pairs, this will not affect our results qualitatively.

Fig. 3a: I suggest changing the color scheme so white is co-native fraction = 100%.

Response: We thank the referee's suggestion. The white part on the left was beyond the resolution of the simulation (Kamdar et al. 2019), but it is assumed that at these small separations, a vast majority of them are co-natal (but not necessarily 100% at large deltaV). We have refined the sentence in the caption (L.361-362) of Fig. 3a to clarify this. We are happy to make the suggested change, if the referee feels strongly about this.

L.228: What is the distribution of |T_cond slope| for the seven pairs that fulfill all criteria? Is it the 7 with the largest slope, or are they evenly distributed? i.e., would a more stringent cut on |T_cond slope| help to identify these systems?

Response: The |T_cond slope| for the 7 systems are more significant in average than the pairs without detection. But they are not necessarily with the largest slopes. As shown in Fig. 3, large Tcond trends (e.g., in the control sample) may still lead to false positives due to random/larger abundance variation. That's also the reason we developed the new Bayesian indicator for more stringent detection. A more stringent cut on |T_cond slope| would not affect our conclusion. We have refined the relevant sentences at L.702-707 (now in the Methods) to emphasize our point.

L.604: is isotopic split included for Ba, and with what isotopic ratios?

Response: For Ba we only took into account the corrections for Hyper-fine structure splitting (HFS) for mainly three lines (5853A, 6141A, and 6496A) and the atomic data of components were taken from the Kurucz line-list [38]. Implementing HFS splitting already ensures that the Ba lines used in this analysis are not saturated, and adding isotopic splitting does not change our results.

Extended Fig. 1: Sulfur seems to exhibit large variations between telescopes/instruments. What is the cause of these, and should this warrant an exclusion of this element? Also, for pair 38 and pair 108, many of the points lie outside the one sigma range and all to either plus (pair 38) or minus (pair 108). Which telescope pairs are these, and does this influence the results?

Response: Surfur is relatively harder to measure given there are only one or two weak lines available. But it's an important indicator in terms of planet signatures/compositions. Therefore we prefer to include it. As stated above, excluding one outlier element would not affect our results, especially the Bayesian indicator (deltaln(Z)).
Regarding the internal comparison, Pair 38 was observed with Magellan and Keck telescopes. The slightly larger systematics for this pair are expected due to different instruments (i.e., different resolution and spectral coverage) and the lines available for measurement, which was also found in Bedell et al. (2014). Pair 108 was observed with the Magellan telescope in two nights. But this pair is actually a triple system, and we switched the reference and object for differential analysis, where the fixed stellar parameters for star 'A' are slightly different (see Yong et al. 2023). More importantly, we didn't achieve enough S/N for this pair (S/N ~ 100 – 150 per pixel) for the first night due to worse seeing, while we did achieve a better S/N of ~200 per pixel for the second night. This can bring in some extra systematics. The differences are slightly larger but still under 0.03-0.04 dex for most elements (except for S). It will not influence our results because we treat them as different pairs if observed with different instruments, and most pairs in our sample were observed using the same instrument in the same night. We also emphasize that for pairs observed in consecutive nights, we utilized the combined spectrum (better S/N). The individual spectra observed in different nights were only used for the purpose of internal comparison. We have commented on this in the caption of this figure (now Extended Fig. 2) at L.1002-1005 for further clarification.

L.681: It is unclear if an uncertainty on Fcz is computed and included in the Bayesian statistics. If not, please comment on this.

Response: The uncertainty on Fcz is not computed and included. Fcz is not set as a free parameter to fit in our Bayesian analysis because it directly correlates/degenerates with M_E. We note the Fcz from MESA calculations agree with Pinnasualt et al. (2001) and we believe it's reasonably estimated for a star with given Teff and mass range. More importantly, changing Fcz or adding uncertainties will mainly affect the inferred mass of planetary material (M_E) but not the relative Bayesian evidence that we adopted to identify the candidates. In addition, we also tested using another set of Fcz with fixed values of: 2.5% for stars with 5500-6100K; 1% for hotter stars > 6100K; and 7.5% for cooler stars < 5500K. There is almost no difference in terms of the Bayesian indicator deltaln(Z). We didn't include this due to the word limit and have added a sentence at L.610-612 in the Methods to address that M_E can be scaled with Fcz and it will not affect the relative Bayesian evidence.

L.735: The atomic diffusion models from Dotter et al., 2017 are used. Are there other models in the literature that could have been used, and how would that influence the results?

Response: Other models are not publicly available to the best of our knowledge. They also don't usually make isochrones. The results aren't likely to change substantially from using a different set of models. Indeed most diffusion models with radiative acceleration included are using data from the Opacity Project (OP). The only exception is Michaud's group who gained access to the atomic data from OPAL, while no other group can access or compare the difference between the OP and OPAL. In addition, the impact from different sets of models should be similar in terms of differential abundance pattern since the zero-point doesn't matter much in this case. We have added a comment on this at L.688-692 in the Methods.

Does the fact that the sample consists of co-native binary stars affect the ratio of planet engulfments? i.e., is it possible to say something about whether planet engulfment is more or less likely to happen in a binary system?

Response: Our goal of using wide, co-natal binaries is mainly for differential analysis. In our opinion, the rate might be similar to that of single stars (Hirsch et al. 2021). We noticed that our reported seven detections are evenly distributed in terms of their pair separations, which may imply that the rate of planet engulfment/ingestion isn't necessarily affected by whether they are binary systems or not. The fact that we do not see any trend as a function of delta_s is also suggestive. However, it might be too premature to make any conclusion, because our sample is much more widely separated than typical binary systems.

Thermohaline mixing is mentioned very briefly at the end of the text. Please expand a bit on how this might influence the result.

Response: Thermohaline mixing can weaken/obscure the detection signals in general (Deal et al. 2020; but the timescale and exact amount is unclear), which would naturally lead to a relatively lower detection rate. That's also the reason we conservatively report the occurrence rate, because some signals of planet engulfment may have weakened below our detection limit (~0.05 dex). In addition, while thermohaline mixing is poorly constrained in 1D stellar models (Tayar & Joyce 2022), hydrodynamical simulations suggest low thermohaling efficiency in stellar environments (Traxler et al. 2011; Brown et al. 2013). We have refined the relevant sentences at L.192-195 in the main text for further clarification.

Referee #2 (Remarks to the Author):
This interesting work describes a detailed differential study of the chemical compositions of stars formed from the same molecular cloud. Since these stars are still on the main sequence, one would expect the surface composition to be the same for the two stars in each pair, inherited from the common molecular cloud they were born in. However, in some cases they find subtle differences which they compare to models based on three scenarios: the accretion of Earth-like material, the abundance variations associated to atomic diffusion, and the situation in which the abundances are identical (the null hypothesis). Based on Bayesian statistics the study finds there are changes between the compositions of the stars in the same pair consistent with the accretion of planetary material in 7 cases out of 91, which leads the authors to conclude that 8% of the stars suffered such accretion.

The paper is well written, and the methodology well explained in general. The figures are appropriate and easy to understand. There have been similar studies before, referred to in this paper, but this new study improves the consistency in the analysis methodology, the methodology itself, and enlarges significantly the sample. This, in addition to the interest of the topic for the public, makes me recommend the paper for publication in Nature.

However, there are a number of issues I cannot understand well and I think need to be addressed before giving the green light.

1- The authors point out in the very first paragraph that the 'ingestion of planetary material' is equivalent to 'planet engulfment'. I do not agree. The first term describes, in my view, much more accurate what they are measuring in the sample stars. All the material consistent with the addition of Earth material to the star could come from a debris disk around the star, and not from an actual planet.

Response: We agree that using the term “ingestion of planetary material” would be more accurate and general regarding the sources/scenarios of the detected chemical signatures (e.g., from planet engulfment and/or accretion disk). Although we note that, if the signatures were imprinted from disk accretion, the timescale would be short (<~ 10 Myr for proto-planetary disc and <~ 100 Myr for debris disc) when the stellar convection zone is still thick (especially 30-40% at ~ 10 Myr). This will make it hard to imprint detectable signatures, especially from proto-planetary discs. For example, as shown in Huhn & Bitsch (2023), the predicted chemical signatures are generally smaller than 0.05 dex (~0.02 – 0.04 dex). Therefore for our reported seven systems (with abundance enhancement ~ 0.05 – 0.15 dex), planet engulfment (or later accretion event due to instability) is still favored over other scenarios. This would have important implications in terms of (in)stability and long-term evolution of planetary systems. To be more general, we have changed the title and the relevant text accordingly. We have refined the paragraph about ‘Implications’ in the main text (L.196-221) to emphasize this and expand the discussion a bit further.

2. Not only one cannot distinguish accretion of actual planets from debris from a (proto)planetary disk, but one cannot distinguish 'adding' planetary material from 'not subtracting' it to form other planets. In fact, in the study the authors compare their results to reference [24], where the differences between the Sun and solar twins are attributed to planet formation rather than planet accretion. I think these issues should be acknowledged upfront, and the title modified accordingly.

Response: We agree with the referee and have made suggested changes as stated above. We have adopted a more general term for the title and the main text to cast a wider net of explanation as suggested. The magnitude of the signal favors “planet engulfment” in explaining the detected signatures in our seven systems instead of “planet formation”, especially if considering the timescale issue.

3. I'm not sure I understand the quoted average slope of abundance changes with condensation temperature. If the changes are for star A relative to star B (and the designations are interchangeable) I would naively expect the average for the sample to be 0.0 (or any number smaller than the uncertainty). What am I missing?

Response: The average here is for absolute value of |T_cond slopes|, which is therefore positive. The value is indeed around zero for T_cond slopes (-0.26 +/- 0.5 with standard deviation of 4.5, in the unit of * 10^{-5} dex/K). We note this part has been moved to the Methods due to the word limit.

4. The criteria ii and iii described in page 5 appear to be arbitrary, chosen based on the distribution of the actual sample under analysis. Any changes in the parameters chosen will directly affect the inferred statistics on the rate of stars with accreted planetary material. Can you offer some idea on how much your inferred rates will change with reasonable changes in the parameters adopted for these criteria?

Response: Our assessment is made based on the distributions of mock noise and signal sample, as well as the control sample (far co-moving pairs). Indeed our criteria are very stringent that even cut off potential signatures carried in the control sample (as stated by Referee #1). Considering the fact that some weak planet signals may exist but below our detection limit, our inferred rate (8%) represents a conservative limit. Following the referee’s suggestion, we also tested the inferred rate with reasonable changes and summarized the results below. For criteria II (deltaln(Z) > 3.5), changing the number of 3.5 to: (a) 4 or 4.5 will lead to no change; (b) 5 will lead to six detections (7%); (c) 2.5 or 3 will lead to 10–11 detections (11–12%). For criteria III (deltaln(Z)_atom > 2.5), changing the number of 2.5 to 1.5 or 2 will lead to 8–9 detections (9–10%). Indeed they are still within uncertainties of the quoted rate of 8% if assuming a Poisson noise (~ 3%), and will not change our results and conclusion in principle. We have refined the sentence in the main text (L.170) by adding "($\pm$3\%, if assuming a Poisson noise)". We couldn't expand the discussion further due to the word limit.

In addition, I have a number of issues/suggestions for the authors to consider:
-page 3, it would be useful to include here the ranges in Teff, logg and Fe/H for the stars in the sample. I know the information is in the additional material, but I consider such information critical (at least the range of spectral types).

Response: We have added the range of spectral types in the main text (L.90) as suggested.

-figure 1 is presented as evidence of abundance variations, but subtle changes in Teff can mimic these variations due to different line excitation energies. Can you mention/show the deltaTeff for these two stars and the theoretically predicted change it implies?

Response: We emphasize that co-natal systems are not assumed to have the same stellar parameters (e.g., Teff), and the differences in our observed EWs and measured abundances due to parameter differences (e.g., deltaTeff) have been properly accounted for (see also Yong et al. 2023). We omitted the details due to the word limit.
For Pair 124, the differential stellar parameter for deltaTeff is: -139 +/- 15 K (HD 185689 w.r.t. HD 185726). We tested by changing the Teff of HD 185689 to be the same as HD 185726 with fixed EWs. This will lead to a change in (differential) abundances of O_NLTE, Si, Fe, and Ni from [0.052, 0.129, 0.158, 0.173] to [-0.158, 0.158, 0.222, 0.236], respectively. The change (assuming two stars have the same Teff) would make the planet signals (in terms of refractories vs. volatiles) even larger, which implies that it cannot mimic the abundance differences (so as planet signals) detected in this pair.

-not all planets have the same composition of Earth. One could argue, however, that Earth-mass planets are more frequent than giant gas planets and the latter will probably leave too small a signature to be detected.

Response: We agree and appreciate the suggestion from the referee regarding the point that super-Earths are more frequent than giant gas planets, and thus leaving more predictable/detectable signatures. For further clarification, we mainly tested the planetary ingestion model with bulk Earth composition. Previously we also tested it with the combination of bulk Earth and CI chondrites but found no clear difference (< 5%) in terms of the relative Bayesian evidence (ln(Z) & deltaln(Z)). We could certainly test more cases for different type of planets with different compositions but there is no certainty of how realistic these are because we know very little and don't have accurate models for other types of planets (e.g., Jupiter, Saturn, Neptune etc.), even in the solar system. The tests made in this work are probably the best we can achieve at this stage. We agree that we can not exclude that a small number of pairs may carry a signature different to the abundance pattern using bulk earth composition but they can't be tested quantitatively until we have realistic planet compositional models in future. That's part of the reason we conservatively present our occurrence rate. We have briefly commented about the test with the combination of bulk Earth and CI chondrites in the Methods (L.633-634) and added some relevant discussion in the main text (L.205-210).

-page 6, paragraph 2: a strictly differential analysis works well for solar twins (or twins in general) but your stars are not, so you have to worry about parameter determination, not just differential abundances. I'm missing a description (in the supplementary material) of how those parameters are derived so that systematic errors are minimized.

Response: The details of determination of stellar parameters and their uncertainties are presented in Yong et al. (2023). We summarized the determination process briefly at L.491-508 and omitted the details in this paper due to the word limit. For stars with deltaTeff +/- 200K, they can still be considered as stellar twins and the systematic errors are minor. We note the typical deltaTeff is 112 K (with a standard deviation of 117 K) for our co-natal sample. We only have 10 close pairs with deltaTeff ~ 300 K, but based on Yong et al. (2023) their uncertainties (especially line-to-line scatter) do not increase much. For example, for co-natal pairs with deltaTeff = 100K, the average error in [Fe/H] = 0.012 dex; for co-natal pairs with deltaTeff = 300k, the average error in [Fe/H] = 0.014 dex. In addition, this should not create a planet-related trend because the uncertainties do not correlate with Tcond or atomic number. Previous studies (Liu et al. 2016a,b; 2020; 2021) indicate that deltaTeff can be as large as 200-300K but they do not affect the accurate determination of differential stellar parameters so as their relative abundances. To our knowledge, this approach remains the state-of-the-art in differential analysis.

-page 6: you mention reference [26] finding a correlation between the chemical anomalies and the stellar mass. Do you find such correlation?

Response: Yes, we find a similar trend against Teff, which was discussed and published in Yong et al. (2023; Fig. 13-14). We didn't include this due to the word limit. In addition, we would like to emphasize that chemical anomalies do not necessarily carry planet signatures.

-methods, sample: a few of your stars must have been included in detailed studies in the literature. How well your parameters agree with those?

Response: We checked this and found 14 common pairs which were included in previous literature with high-precision spectroscopy (Nagar et al. 2020; Spina et al. 2021). Our stellar parameters agree well with the literature values for those 14 common pairs, with an average Teff difference of -16 K (standard deviation of 34 K) and an average [Fe/H] difference of 0.002 dex (standard deviation of 0.009 dex). Most of these pairs (9 out of 14 pairs) were found to have no clear abundance difference, and 5 pairs were claimed to be chemical anomalies (but not necessarily carrying planet signatures). Our results agree with the literature, if applying the same criteria for chemical anomalies (see also Yong et al. 2023). In addition, one common pair (HD 133131A/B) was compared with previous studies and showed no systematic differences. We have added a few sentences about the comparison with literature values for these 14 common pairs in the Methods (L.512-515).

-methods: the line-list adopted ... is available upon request ...
also later a similar statement is made about the data. This is not acceptable by Today's standards for reproducibility. The data and tables should be made available in a digital repository, such as Zenodo.

Response: We agree and have included the final line-list as well as the differential abundances as two ASCII tables in the Supplementary information. If the paper gets accepted, we will be delighted to also include all the data and tables in a data repository at Zenodo.

-methods: presumably the model atmospheres, synthesis code, etc. apply as well to the parameter determination, not just the abundance inference. Please, state clearly or explain.

Response: Yes, indeed both parameter determination and abundance analysis were conducted carefully in the same/consistent approach (in terms of EWs measurement, model atmospheres, analysis code, etc.) and each step of the data analysis was carried by myself. The stellar parameters and the details of determination have been presented and published in Yong et al. (2023). We can only summarize the process of parameter determination (L.491-508) and focus more on the abundance analysis (L.525-529) due to the word limit. We also added “(consistent with parameter determination)” to the sentence in the main text (L.91) to address this.

-methods: One of the authors has written papers in which a dependence of the strength of the O triplet with stellar activity was reported. Does that deserve a comment in this context?

Response: We tend to discard young and active stars (also fast rotators) during sample selection and observations. Since each pair of stars are selected to be similar (in terms of stellar type and parameters), the potential systematics from different stellar activities within a given pair should be minor. Unfortunately we couldn’t find a space to add a comment about this due to the word and page limits.

-methods, page 14: it is true that the agreement suggests the differential analysis has effectively suppresed systematics, but other systematics due to imperfect modeling must be present, given the deltaTeff can be as large as 300 K. This should be acknowledged, as such errors are not accounted for in the description/analysis (3D NLTE effects are only accounted for in a particular case).

Response: We agree that the systematics from e.g., 3D NLTE can be more severe for pairs with relatively larger deltaTeff (~ 300K), while for stellar twins with deltaTeff of <~ 200K, the impact should be minor except for O based on our previous experience (e.g., Liu et al. 2018,2020). We have applied and adopted 3D NLTE corrections for the oxygen triplet. Here we have tested and estimated the impact of 3D NLTE for a few elements (Fe, O, Na, Mg, Al, and Mn) for two pairs (Pair 22 and 17) with deltaTeff of 200K and 300K, respectively. The effects/corrections are summarized below:

Pair 22 (deltaTeff ~ 200K): deltaFe = 0.012, deltaMg = -0.004, deltaMn = -0.010, deltaNa = 0.005, deltaAl = -0.005;

Pair 17 (deltaTeff ~ 300K): deltaFe = 0.020, deltaMg = -0.007, deltaMn = -0.014, deltaNa = 0.006, deltaAl = -0.009;

where deltaX represents the average differential corrections (NLTE - LTE)_{B-A} (line-by-line) for a given element.

The differential corrections (NLTE - LTE) are generally between 0.01 - 0.015 dex for the majority of our co-natal sample, thus will not affect our results and conclusion. We have added a few sentences to address this point in the Methods (L.533-537).

In addition, we find that the seven systems showing planet signatures do not have large parameter differences. Indeed most of them have deltaTeff ~ 100 – 150 K, indicating that the signatures are not likely due to additional systematics induced by parameter differences.

-methods, line 661: Beyesian -> Bayesian

Response: Done

-methods: not everyone agrees on the physical description of diffusion. How different are the recipes adopted in MESA from alternative models? Are they similar or span a wide range that could affect the conclusions of this paper?

Response: Our atomic diffusion models follow the Burgers equation to describe the physical description of diffusion, which is also well agreed between different models. There are two mathematical formalisms that have been implemented: by Michaud and by Thoul et Al. (1994), which are equivalent by Greg Feiden. MESA uses the Thoul version, while Michaud's version/model is not publicly available. People tend to differ in the atomic data and in the extra mixing that may be applied near the surface. As stated in our response to Referee #1, most diffusion models with radiative acceleration included are using data from the Opacity Project (OP). The only exception is Michaud's group who gained access to the atomic data from OPAL, while no other group can access or compare the difference between the OP and OPAL. We used the one by Vandenberg et al. (2012), which is based on Michaud's "turbulent diffusion" model. In addition, other atomic diffusion models are not publicly available to the best of our knowledge. They also don't usually make isochrones. The results aren't likely to change substantially from using a different set of models, especially for the co-natal sample with smaller parameter differences. We have added a comment on this point in the Methods (L.688-692).

Referee #3 (Remarks to the Author):

The manuscript "At least one in a dozen stars has engulfed planets" is an interesting contribution to the literature which aims to constrain the occurrence rate of planet engulfment events using chemical differences in binary star pairs. Seven newly identified candidate systems for planet engulfment are found from a sample of 91 close binary pairs, substantially increasing the number of such systems known and providing an occurrence rate estimate of 8% for these events.

This paper advances the field by employing a larger sample of binary pairs than previously used for such studies, and by developing a new statistical methodology for distinguishing planet engulfment signatures from atomic diffusion. The goals and general methods of the paper are similar to previous works (Spina et al 2021, published in Nature Astronomy; and Behmard et al 2023, published in MNRAS), but comparison to these studies is appropriately addressed in the text. In particular, this work employs a much larger set of elemental abundances than the Spina paper, which allows for more confident detection of planetary signatures; and it uses a much larger sample size and more precise measurements than the Behmard paper. For these reasons, I would regard this work as a significant advance to the field.

The data set used in this work is excellent: by selecting co-moving stellar pairs from Gaia with a reproducible selection function, the authors have achieved an unprecedentedly large and complete sample of twin-star binaries. The analysis techniques employed to derive abundance measurements are appropriate and the precision achieved is well-justified with validation tests.

The statistical methodologies are generally sound. The Bayesian analysis developed in this work is an improvement over traditional methods (e.g. fitting a line to the abundance vs condensation temperature relation, which can only provide a rough metric for planetary material). A few specific suggestions for improvement are made below.

I have some reservations about the robustness of the conclusions as presented in this work; in particular, I think the title overstates the certainty of the results. As discussed in the final paragraphs of the main text, the "planet signatures" found are still ambiguous in their cause: alternative models to engulfment exist, including dust trapping as a natural byproduct of the planet formation process (Booth & Owen 2020). Furthermore, the analysis tests only three models to explain the observed abundance differences (planets, atomic diffusion, and the null hypothesis of the stars having identical compositions), which is not an exhaustive list of potential causes for differing abundances. Thus, while the evidence is promising, it seems premature to strongly conclude that these systems must have ingested planets. The engulfment hypothesis could be tested further by examining lithium signatures (as done in Spina et al 2021). Knowledge of the planetary systems (or lack thereof) around these stars would also be a valuable constraint for future work.

Response:
- We agree that the title can be more general regarding the inferred events/scenarios. We have changed the title ("planet engulfment" -> "ingestion of planetary material") and relevant text, which was also suggested by Referee #2. We note the planet engulfment scenario is still favoured over the other few scenarios for our reported candidates. The main reason is that, for the disc scenario mentioned by the referee (the Booth & Owen 2022 paper, as well as the Hühn & Bitsch 2023 paper), we consider these early signatures to have been mixed or cleaned up by the billions of years of evolution that the stars have undergone before we observed them now. Meanwhile for our seven candidates, the detected signatures are significant at a level of ~0.05 – 0.15 dex, which is stronger than the predictions (0.02 – 0.04 dex) from e.g., Huhn & Bitsch (2023) assuming the disc accretion scenario.
- Of course we cannot exclude disc accretion scenarios from this study (signatures may be too small to detect), therefore our inferred rate might serve as a conservative estimate. While there are many possible scenarios, this is why we build the model of planet engulfment and test the Bayesian evidence for the observation. This is not exhaustive, but certainly makes our detection a strong and possible candidate.
- We have refined the 'Implications' in the main text (L.196-221) to emphasize the above points.
- Regarding the search for potential planetary systems, it is certainly interesting and needed in the near future. We are collecting LCOGT spectra to look for close binaries but long-term follow-up (at least 6-12 months) is required to confirm any RV variations, if they exist. In addition, the resolution of LCOGT spectra are not sufficient to detect super-Earth so we plan to follow up our candidates with HARPS. We want to add that this could open a new window to understand late instabilities and long-term evolution of planetary systems.
- We appreciate the referee's suggestion about lithium signatures. However, this is more complicated due to a few factors, such as Pre-Main-Sequence Li depletion and dependence on stellar parameters (e.g., Teff and age). For example, even for pairs (with similar stellar parameters) without large relative Bayesian evidence and no Tcond trend, their lithium abundances are still very different. It thus warrants a careful treatment that is outside the scope of this study.

Specific suggested changes:
1. Line 62: "conservative occurrence rate": I would just call this the estimated occurrence rate. In this context "conservative" implies that the number is calculated as a lower limit.

The reason we stated the rate as 'conservative' limit is because: (1) some potential planet signatures may be too weak to detect; (2) we are strict with our selection criteria that may even cut-off potential signatures carried by e.g., the control sample (as suggested by Referee #1); (3) different signatures (due to different planet compositions) may exist in addition to the bulk Earth composition (although accurate models are not available to test quantitatively). Following the referee's suggestion, we have avoided using the word 'conservative' directly and only stated "At least" when describing our new detections.

2. Lines 68-81 (second paragraph): This introductory section seems like it would not read very clearly to a reader who is not an expert in the concepts being discussed. Specifically: "Ingestion of planetary material… will increase the elemental abundances of the host star" does not clearly convey that the refractory elements will increase more than volatile, which in turn obscures why it matters that "elemental abundance differences as a function of the dust condensation temperature… were reported." In the next sentence, it's not obvious what "the signal" refers to. Some edits to clarify these points would make the intro more effective.

Response: We thank the referee’s suggestion and have made some edits for the relevant sentences at L.76-79 in the main text to better explain/connect the chemical signatures of planets to the correlation between abundance differences and Tcond. Meanwhile, based on the Editor’s suggestions, we have to summarize the background in 2-3 sentences and place them into lines 47-56. The rest of the ’introductory section’ has been removed or moved to the Methods due to the word limit.

3. Lines 110-113: Cite a justification for the cutoff at 10^6 AU?

Response: We have cited Kamdar et al. (2019) and Nelson et al. (2021) as suggested at L.64 in the ‘introductory paragraph’, which demonstrated in both theory and observation that the boundary between co-natal and non-co-natal systems is likely at ~ 10^6 AU. This is also confirmed with our own study (Yong et al. 2023) based on precise relative abundance results and the associated chemical homogeneity.

4. Lines 214-217: I don’t think any conclusions can be drawn about the prevalence of atomic diffusion among the far co-moving pairs. The distant pairs are not co-natal and therefore the null hypothesis (that they have the same abundances) is not applicable.

Response: Some of our far pairs have strong Tcond trends, but we found that they are more likely due to atomic diffusion than pure/random abundance scatter. We therefore consider it as one of the possible explanations for these non co-natal pairs. ISM inhomogeneity might be another source but we can’t test it quantitatively. It also shows the limitation of Tcond trend. We have refined the relevant sentence (now at L.130-131) as below:
“... demonstrating that the observed abundance differences within some of these pairs are potentially affected by atomic diffusion.”

5. Lines 222-226 (& methods generally): Another available metric is the confidence with which the ingested planet mass is non-zero. This should already be available because the Bayesian method results in a posterior predictive distribution on the parameter M_E. You might impose a cut that requires that M_E be equivalent to 3 sigma away from zero. This cut, which frames the problem of detecting planet ingestion in terms of parameter estimation rather than model selection, would be more robust to inaccuracy in the Bayesian evidence calculations (which are notoriously difficult – see e.g. Nelson et al (2020), AJ 159, 73 for one discussion of uncertainty in evidences and their impact on Bayesian model selection).

Response: We acknowledge the referee's concern regarding potential numerical sampling noise within the Bayesian evidence calculation. However, we contend that this issue is not significant in our analysis for several reasons. Firstly, our computational model involves sampling only two parameters — M_E and sigma_scatter — which mitigates the complexity and potential for noise. Additionally, the posterior distribution in our case is predominantly unimodal, ensuring that the Bayesian evidence attains robust convergence. Empirical tests, such as random bootstrap sampling, reinforce this assertion, inducing a mere 5% variation in the evidence, which is within acceptable limits for our purposes.
Furthermore, we concur that the adoption of a threshold based on M_E is theoretically a viable approach. Nonetheless, the practical application of such a threshold is intrinsically linked to the effective temperature (T_eff), due to the variability in convective mass across different stellar classifications. Implementing a T_eff-dependent threshold introduces a layer of complexity that may not be justified, as it could be perceived as more arbitrary compared to the systematic application of Bayesian evidence. Consequently, we maintain that our methodology, while not without its assumptions, provides a more straightforward and less speculative avenue for interpreting the data.

6. Lines 265-268: The assertion that a different analysis technique applied to the Spina+2021 sample would result in “a much smaller number of engulfment instances” is speculative, based on the assumption that the occurrence rate found in this work is correct and therefore other work should be revised to match it. I don’t think this is an appropriate speculation to make, and would recommend cutting this sentence.

Response: We agree with the referee and have removed this sentence accordingly. We also refined the relevant sentences at L.180-189 to explain that: (1) our results (e.g., fraction of chemical anomalies) are generally consistent with Spina+2021; (2) the different rates mainly stem from the more homogeneous sample in this study and the analysis methods (relative Bayesian evidence & Tcond trend) being applied using 20+ elements.

7. Figure 4: Is panel A really needed? The evidence for a given model without comparison to any alternatives does not seem to contain useful information.

Response: We consider panel A in Figure 4 to be somehow important because it will enable easier and direct model comparison for future studies. It also sets a baseline at an absolute scale, even though we are mainly interested in the differences in Bayesian differences. We will keep it for now, but are happy to remove it if the referee feels strongly about it, since it is not pertinent to the manuscript.

8. Lines 596-598: How were the equivalent widths measured? (e.g. fitting Gaussians with IRAF, an automated code like ARES, etc)

Response: The EWs were measured using a semi-automatic code similar to {\sc REvIEW} (\url{https://github.com/madeleine-mckenzie/REvIEW}), with each line being visually inspected. The details are also presented in Yong et al. (2023). We have added a sentence to describe the EWs measurement at L.520 in the Methods.

9. Lines 668-678: It is confusing that M_E, a key parameter of the model, does not appear in these equations. I recommend adding another equation to show how M_{X,E} is calculated from M_E. Relatedly, the description of M_{X,E} as “the total mass of element X” seems incomplete– perhaps it should be “the total mass of element X introduced by planet ingestion.”

Response: We have refined the description of M_{X,E} and added another equation at L.597 in the Methods (now Equation 3) to clarify the calculation of M_{X,E} and M_E at L.603-604 as suggested:
“M_{X,E} = f_{X,E} * M_E
where f_{X,E} is the mass fraction for element X being accreted assuming the bulk Earth composition (Asplund 2009), M_{X,E} is the total mass of element X introduced by planet ingestion.”

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I thank the authors for the careful and thorough reply to the referee comments. The responses to the original review are appropriate, and the authors have made additional changes that strengthen the paper.
I have no further comments and recommend this paper for publication.

Referee #2 (Remarks to the Author):

The responses to the comments I provided are fully satisfactory. This is an excellent research report for Nature, and I'm sure there will be follow-up publications, from the same group and others.

Referee #3 (Remarks to the Author):

I am satisfied with the edits to the paper and recommend its acceptance. In particular, the rewritten introductory paragraph reads quite clearly.

My only remaining suggestions are as follows:
1. While the rephrasing of "planet engulfment" to "ingestion of planetary material" in the title and body text is an improvement, it still implies a greater certainty in the inferred scenario than I think is warranted. Specifically, it is possible that rather than an overabundance of refractory elements in one star due to planet ingestion, we are actually seeing an underabundance of refractories in the other star due to planet formation (as in the original explanation of Melendez et al 2009, or as in the case of dust trapping due to giant planet formation). This explanation would not fit under the umbrella of "ingestion of planetary material." Could the title say something like "exhibits chemical signatures of planets / planet formation / planet interactions"?

I recognize that at some point this debate becomes splitting hairs, although I do think the title is important. If the editor and other reviewers are happy with the title as is, I am willing to be overruled on this!

2. (Related to the authors' response to my suggestion #5 in the initial review:) I'm glad that empirical tests like bootstrap validate the Bayesian evidence, and I agree with the authors' point that the straightforward unimodal distribution of parameters makes the evidence calculation relatively more reliable here than in some other cases. It may be worth discussing this (at least mentioning the bootstrap test results) in the supplementary materials.

Author Rebuttals to First Revision:

Response to the referees

We thank the three referees again for their careful reading and constructive comments.

Referee #3 (Remarks to the Author):
I am satisfied with the edits to the paper and recommend its acceptance. In particular, the rewritten introductory paragraph reads quite clearly.

My only remaining suggestions are as follows:
1. While the rephrasing of "planet engulfment" to "ingestion of planetary material" in the title and body text is an improvement, it still implies a greater certainty in the inferred scenario than I think is warranted. Specifically, it is possible that rather than an overabundance of refractory elements in one star due to planet ingestion, we are actually seeing an underabundance of refractories in the other star due to planet formation (as in the original explanation of Melendez et al 2009, or as in the case of dust trapping due to giant planet formation). This explanation would not fit under the umbrella of "ingestion of planetary material." Could the title say something like "exhibits chemical signatures of planets / planet formation / planet interactions"?

I recognize that at some point this debate becomes splitting hairs, although I do think the title is important. If the editor and other reviewers are happy with the title as is, I am willing to be overruled on this!

Response:
We acknowledge the possibility that the formation of "terrestrial" planets may remove refractory material (or that giant planets may trap refractory material), potentially inducing a similar abundance trend, though on a reduced scale. To ensure a more balanced and inclusive narrative, we have revised the first two sentences and the last sentence of the summary paragraph to further introduce both scenarios as below:
1st and 2nd sentences: "*Stellar chemical compositions can be altered by ingestion of planetary material [1,2] and/or planet formation which removes refractory material from the proto-stellar disc [3,4]. These ``planet signatures" appear as correlations between elemental abundance differences and the dust condensation temperature (T_{cond}; 3,5,6).*"
Last sentence: "Our study doubles the number of known instances of planetary ingestion to date and …" -> "Our study provides new evidence of planet signatures and …"

In the last paragraph, we have added further/neutral discussion of these alternative interpretations, and explained why we favor planetary ingestion as the plausible interpretation, particularly for our reported candidates (refer to Methods for detailed discussion) as below:
"*(b) the formation of terrestrial planets that remove refractory material from the proto-planetary discs [2];*"
"*The above processes are likely to imprint smaller amounts of abundance differences (0.02 – 0.04 dex; 5–10%) when compared to that from the later accretion/ingestion events, due to different sizes of stellar convection zone (see Methods).*"

Although we prefer to keep our original title, we are open to adopting the title suggested by the editor: *"Co-natal binary stars with different abundances of refractory material"*, if deemed necessary.

2. (Related to the authors' response to my suggestion #5 in the initial review:) I'm glad that empirical tests like bootstrap validate the Bayesian evidence, and I agree with the authors' point that the straightforward unimodal distribution of parameters makes the evidence calculation relatively more reliable here than in some other cases. It may be worth discussing this (at least mentioning the bootstrap test results) in the supplementary materials.

Response: We have added a brief description of the bootstrap test results in the Methods as below: "*The posterior distribution in our case is predominantly unimodal, ensuring robust convergence of the Bayesian evidence. Random bootstrap sampling reinforces this assertion, inducing a mere 5\% variation in the evidence (within detection limits).*"